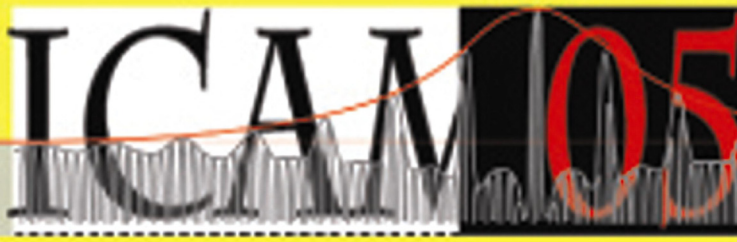


Proceedings of International Conference on Applied Mathematics 2005

22-26 AUGUST 2005
AULA TIMUR
ITB BANDUNG INDONESIA

ISBN: 90-365-2244-7

Proceedings of International Conference on Applied Mathematics 2005

ICAM05

Published by

Centre for Mathematical Modelling and Simulation (P2MS)
Institut Teknologi Bandung
Labtek III, Lt 1, Jl. Ganesha 10
Bandung 40132
INDONESIA

Phone/Fax: +62 +22 250 8126

E-mail: admin@labmath-itb.or.id, website: www.labmath-itb.or.id

ISBN: 90-365-2244-7

About ICAM05

International Conference on Applied Mathematics 2005 (ICAM05) aims to bring together mathematicians and scientist from other areas who apply mathematical modelling and techniques in one of the many application domains. Except invited lectures by renowned scientist, contributions from participants in parallel sessions constitute the main part of the programme. Participants are invited to submit proposals for mini-symposia or contributed papers. Special EPAM-symposia report about results of a 5 year collaboration project between Indonesia and The Netherlands. The conference also aims to support the initiation or extension of linkages between institutes in ASIAN countries and institutes from other parts of the world. Major sponsor of the conference is the Royal Netherlands Academy of Arts and Sciences.

The role of mathematics in science, technology and modern society is continuously expanding. New developments in mathematical modelling, methods and ideas are used and often triggered by classical and new application domains. Classical applications areas like the natural sciences, engineering and technology are joined by more modern domains for decision support and modelling, like life sciences, financial engineering, etc. This conference is aimed to present a broad scope of all such activities to cover both new advances in mathematics, as well as applications from various areas of applications. To specify the activities and contributions, two types of keywords are used to indicate the mathematical method and the application domain. Math key-words to describe the main mathematical method (non-exhaustive) are: Analysis, ODE & PDE, Statistics, Probability, Operations Research, Optimisation, Discrete Mathematics, Signal analysis, Systems & Control, Numerical Analysis, Scientific Computing, Mathematical Modelling, Mathematical Physics, etc.. Application key-words to describe main application domain (non-exhaustive) are: Surface and internal water waves, Coastal engineering, Optics, Seismology, Financial engineering, Telecommunication, etc. This proceedings contains invited lecturers and contributions of other participants. The organisation of this proceedings follows the conference programme.

Preface

We like to warm heartedly welcome all participants to the International Conference on Applied Mathematics. We hope that you enjoy the scientific ambiance of the conference where you meet and make friends while crafting future collaborations. We also like to welcome you to Bandung and we wish that the hospitality of the people will make this conference a memorable event.

The International Conference on Applied Mathematics (ICAM05) is held at Institut Teknologi Bandung from 22 till 26 August 2005 and is hosted by the Centre for Mathematical Modelling and Simulation (P2MS) ITB. The event is attended by approximately 230 participants from 13 countries. This conference marks the end and reports on the success of the Extended Programme in Applied Mathematics (EPAM) 2000/2005. The programme is part of the Scientific Programme Indonesia – Netherlands (SPIN) funded by the Royal Netherlands Academy of Arts and Sciences and by various local grants. EPAM consists of six projects jointly coordinated by Dutch and Indonesian project leaders, and the results will be presented in separate EPAM Symposia.

During the conference there will be, besides the EPAM Symposia, fourteen invited lectures by experts coming from 10 different countries from west to east. Moreover, over 160 papers will be presented by national and international scientists. The abstracts of all the contributions are collected in the conference booklet; a separate CD contains the full proceedings of the conference, including full papers of many contributions.

It is our honour that the Rector of Universitas Syiah Kuala (UNSYIAH) will be present in a special session on Tuesday. That day will have contributions that focus on Geo-Mathematics, with topics like tsunamis, flooding, warning, and mathematical modelling. For this occasion the rector will present a speech on the current situation of the badly affected region hit by the December 2004 tsunami. A declaration of support for staff of UNSYIAH will be read in public; it is our hope that scientists attending this conference will support the declaration.

We would like to thank all invited lecturers, EPAM project leaders, and all other participants to contribute to the success of the conference. To Prof. Djoko Santoso, Rector of ITB who will formally open the meeting on Sunday evening, we would like to express our gratitude.

To Dr. Rinovia Simanjutak, the manager of the conference, we are very thankful for beautiful organization of the conference. We greatly appreciate the efforts of Rini, Helena, Adri, Yudith, Noor, Sena, RK, as well as the secretaries of P2MS, Elis and Fiska, who provided very helpful and total support for the preparation as well as during the conference.

One more time to all participants, we wish that you enjoy the conference and we hope to see you again at some other event.

Bandung, 20 August 2005
Andonowati, E. van Groesen, R.K. Sembiring

Table of Contents

MS 5 Laboratory Wave Generation

MS 6 Indonesian PhD Students Minisymposium 1: Statistics and Neural Network

MS 7 Mathematics Education 3: University Practice of Modelling

MS 8 Geo-Mathematics 1

CP 5 Theoretical Fluid Dynamics

CP 6 Discrete Mathematics

CP 7 Dynamical Systems

CP 10 Statistics

CP 11 Financial Mathematics and Neural Networks

CP 12 Diffusion and Flow in Porous Media

CP 15 Mathematical Biology

CP 16 Mathematics Education

A Mathematical Model for Geomagnetic Reversals

J.J. Duistermaat, Peter Hoyng

University of Utrecht, The Netherlands

Abstract: The earth's magnetic fields has reversed its polarity many times in history, where the polarity remained the same during very long time intervals, between about 100,000 years and many million years, whereas the reversals took place in relatively short time intervals of about 1,000 years. The lengths of the time intervals between subsequent reversals form an irregular sequence with a large variation, which make the reversals look like a stochastic process.
Several years ago Peter Hoyng asked me some questions about the mathematical model which he had for the reversals. This consisted of a stochastic perturbation of a deterministic dynamical system with two symmetric asymptotically stable equilibria, of which the domains of attraction were separated by the stable manifold of a saddle point. The magnetic field is equal to zero at the saddle point and has opposite polarities at the two stable equilibria. Against the flow of the deterministic system, the stochastic perturbations can build up and push the system from close to one of the stable equilibrium points into the domain of attraction of the other one, after which it quickly moves to the other one.
Then the process repeats itself with reversed roles of the stable equilibria. It is one of the properties of such stochastic processes that with probability one a reversal will take place, but also that most attempts fail and the expectation time for the next reversal is very long. The asymptotic behaviour of the expectation time of the reversals and the most probable path of escape is given by the theory of Freidlin and Wentzell from the 1970's. In this theory, the leading coefficients in the asymptotic formulas are described by a variational equation, which in turn leads to a Hamiltonian system. An interesting feature of Hoyng's model is that the stochastic perturbation is proportional to the magnetic field, and therefore vanishes at the unstable saddle point.
This leads among others to the conclusion that the most likely escape route makes a detour and approaches the saddle point tangentially to the stable manifold of the saddle point, instead of going straight from the stable equilibrium to the unstable one.

Application of linear optical networks in quantum optics

Takayoshi Kobayashi, Hai-bo Wang, Yongmin Li, Satoru Odate
University of Tokyo, Tokyo, Japan
CREST, Japan Science and Technology Corporation (JST).

Abstract. We analyzed and constructed quantum networks which consists of all linear optical component. It is shown that the strong nonlinear effects caused by postselection strategies based on single photon technologies is enough to realize some nonlinear quantum operation.
Key-words: linear optics, quantum optics, quantum state preparation

1 Introduction

The squeezed state [1], which has smaller noise fluctuation in one of the quadrature components than that of the coherent state, has been widely used in quantum optics and in many other branches of quantum physics[2]. Recently, the interest of squeezed states has been rekindled in quantum information and communication fields, in which the quantum entanglement generated from squeezed states plays an essential role. By utilizing entanglement shared by sender and receiver together with local operations and classical communication, various feats of quantum communication, such as quantum teleportation[3], quantum dense coding, entanglement swapping and the construction of quantum network [4] have been experimentally demonstrated with continuous variable quantum systems. Till now, the proposed or realized squeezed state mainly builded on the nonlinear physical processes. A few years ago, significant progress was achieved by proposals for quantum gate using only passive linear optics and projection measurement[5]. These proposals show that strong nonlinear effects can be implemented by exploiting postselection strategies based on single photon technologies.

In this paper we will discuss a specific scheme for generating a photon-number squeezed state. Quite different from the nonlinear processes used before, the scheme proposed here consists of all linear optical component. The arrangement of the paper is as follows. We first introduce a quantum component, "*quantum high/low-pass filter (Quantum HPF/LPF)*", which consists of linear optics components. The output characteristic of the quantum filter is also analyzed. Then we show that it is possible to generate a photon-number squeezed state by using cascaded quantum low-pass filter and high-pass filter. Preliminary results in this paper have been submitted in Ref.[6].

2 Quantum high/low-pass filter with linear optics

Figure 1 shows the optical circuit of a quantum high-/low-pass filter. A similar setup has been used to demonstrate the QND measurement of Fock states [7, 8].

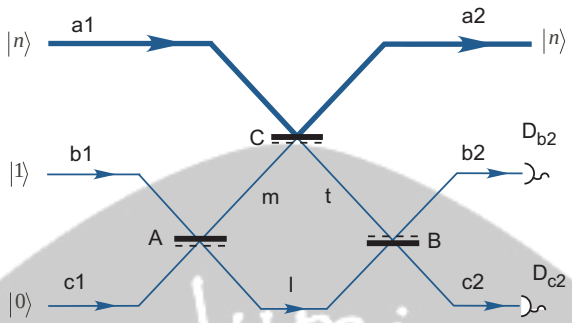


Figure 1: Optical circuit of a quantum filter. Beam splitters A, B, and C are assumed to be asymmetric in phase. Reflection off the "dash" surface of each beam splitter produces a sign change. D_{b2} and D_{c2} are single-photon detectors. Quantum filter is realized when the detectors D_{b2} and D_{c2} counts one photon and only one photon and another detector counts no photon. b_1 and c_1 are ancilla modes.

We analyze the circuit for arbitrary input state and show how it can work as a quantum filter. The operator input-output relations between the two input modes (a_{in}, b_{in}) and the corresponding output modes (a_{out}, b_{out}) have the general form $a_{out} = \sqrt{\eta_i}\ a_{in} + \sqrt{1-\eta_i}\ b_{in}, b_{out} = \sqrt{1-\eta_i}\ a_{in} - \sqrt{\eta_i}\ b_{in}$. Here η_i and $1-\eta_i$ are the intensity reflectivity and transmittivity, $i = a$, b, and c corresponding to the beam splitters A, B, and C, respectively.

The input and output operators in the Heisenberg picture are connected by

$$\hat{a}_1 = \sqrt{\eta_c}\,\hat{a}_2 - \sqrt{\eta_b(1-\eta_c)}\,\hat{b}_2 + \sqrt{(1-\eta_b)(1-\eta_c)}\,\hat{c}_2,$$

$$\hat{b}_1 = \sqrt{\eta_a(1-\eta_c)}\,\hat{a}_2 + [\sqrt{\eta_a\eta_b\eta_c} + \sqrt{(1-\eta_a)(1-\eta_b)}\,]\,\hat{b}_2 + [\sqrt{\eta_b(1-\eta_a)} - \sqrt{\eta_a(1-\eta_b)\eta_c}\,]\,\hat{c}_2,$$

$$\hat{c}_1 = \sqrt{(1-\eta_c)(1-\eta_a)}\ \hat{a}_2 + [\sqrt{(1-\eta_a)\eta_b\eta_c} - \sqrt{\eta_a(1-\eta_b)}\ ]\ \hat{b}_2 - [\sqrt{(1-\eta_a)(1-\eta_b)\eta_c} + \sqrt{\eta_a\eta_b}\ ]\hat{c}_2,$$

Let a Fock state $|n\rangle$ impinges on the input ports a_1. At the same time, a single-photon state $|1\rangle$ is injected into the ancilla mode $b1$ and the other ancilla mode $c1$ is set unoccupied. For a time sysmmetric linear network such as that in Fig.1, the output state can be directly obtained from the input state

$$|\psi\rangle_{in} = |n\rangle_{a1}|1\rangle_{b1}|0\rangle_{c1} = \frac{1}{\sqrt{n!}}(\hat{a}_1^+)^n\hat{b}_1^+|0,0,0\rangle.$$

In this paper, we consider only cases where one of the detectors D_{b2} and D_{c2} counts one photon and only one photon (OPOOP) and another detector counts

no photon. In this case, n photons will appear at the outport port a_2. We obtain the output state

$$|\psi\rangle_{out} = C_1(n)|n\rangle_{a2}|1\rangle_{b2}|0\rangle_{c2} + C_2(n)|n\rangle_{a2}|0\rangle_{b2}|1\rangle_{c2} \tag{1}$$

with

$$C_1(n) = (\sqrt{\eta_c})^{n-1}[\sqrt{(1-\eta_a)(1-\eta_b)\eta_c} - (n - n\eta_c - \eta_c)\sqrt{\eta_a\eta_b}]$$
$$C_2(n) = (\sqrt{\eta_c})^{n-1}[\sqrt{(1-\eta_a)\eta_b\eta_c} + (n - n\eta_c - \eta_c)\sqrt{\eta_a\eta_b}]$$

First, we discuss the application of the filter to an input state of $|n_w\rangle_{a1}|1\rangle_{b1}|0\rangle_{c1}$. Provided the photons are indistinguishable, the conditional interference of the filter can be maximized by setting η_a, η_b and η_c to satisfy the follow equation

$$\begin{aligned} 0 < \eta_c < n_w/(n_w+1); \\ C_1(n = n_w) = 0; \\ C_2(n = 0) = 0. \end{aligned} \tag{2}$$

The solutions of Eq.(2) are given as

$$\begin{aligned} \eta_a = 1/(n_w+1)(1-\eta_c), \\ \eta_b = \eta_c/n_w(1-\eta_c). \end{aligned} \tag{3}$$

So that, if n_w photons exist in path a_1, the probability of the state $|n\rangle_{a2}|1\rangle_{b2}|0\rangle_{c2}$ is given by $|C_1(n = n_w)|^2 = 0$. The final output state is the second term in Eq.(1). Therefore, there is one photon output at port c_2 and no photon at port b_2. On the contrary, if no photon in path a_1, the final output state is the fist term of Eq.(1). A single photon appears at output port of b_2 and no photon at port c_2. These operations will succeed with 100% probability. Therefore, the QND of $|n_w\rangle$ state will be achieved when one photon appears at output port c_2 [7]. For convenience, the point n_w is called the "*working point*" for a quantum filter.

When photons with different number from n_w is injected to this scheme, the output is not so clear and the QND cannot in general be achieved. Next, We consider the output characteristic of this scheme for more generalized input state and show that this scheme may be treated as a quantum HPF/LPF. The input state at the port a_1 may be expanded in terms of the number states as $|\Psi\rangle_{a1} = \sum C_n|n\rangle_{a1}$. The input state for the quantum filter is given by $|\psi_{in}\rangle = \sum_{n=0}^{\infty} C_n|n\rangle_{a1}|1\rangle_{b1}|0\rangle_{c1}$. The reflectivities of each beam splitters are set as Eq.(3). After passing through the setup, the output state becomes

$$\sum_{n=0}^{\infty} C_n[C_1(n)|n\rangle_{a2}|1\rangle_{b2}|0\rangle_{c2} + C_2(n)|n\rangle_{a2}|0\rangle_{b2}|1\rangle_{c2}] \tag{4}$$

One may use the technique of conditional state preparation[10], in which the state is extracted by triggering signal. In our case, the triggering signal is one of the detectors D_{b2} and D_{c2} counts OPOOP while another detector counts no photon.

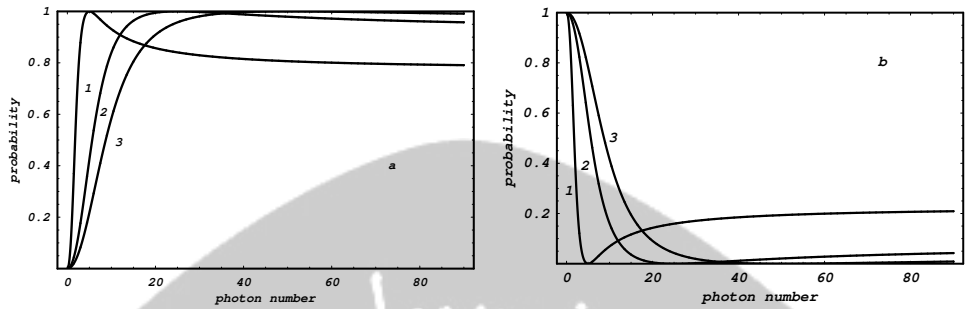


Figure 2: The response of quantum HPF and LPF working at several different working points. (a) The response of a quantum HPF. The lines 1, 2, and 3 have different work points of $n_{w1} = 5$, $n_{w2} = 25$ and $n_{w3} = 50$. η_c is set to be $\eta_c = n_w/(n_w + 1) - 0.3$. (b) The response of a quantum LPF. The lines 1, 2, and 3 have different work point $n_{w1} = 5$, $n_{w2} = 25$ and $n_{w3} = 50$. η_c is set to be $\eta_c = n_w/(n_w + 1) - 0.3$.

Depending on the triggering signal, the filter will work as a quantum HPF or LPF. For example, when the detector D_{c2} counts OPOOP and the detector D_{b2} counts no photon, the output state is reduced to the second part of Eq.(4) and the first term in Eq.(4) is discarded by post-selection. The curve in Figure 2(a) shows how the normalized output probability of the filter varies with photon number. It can be seen that the curve in Fig.2 (a) shows a typical response like a classical HPF. In this way, a quantum HPF is realized in a quantum domain. When the detector D_{b2} counts OPOOP and the detector D_{c2} counts no photon, the filter is worked as a quantum LPF with the response function shown in Fig. 2(b). Figure.2 also show the response functions of filters with several different working point. The working points of line 1, 2, and 3 are set to n_{w1}=5, n_{w2}=25, and n_{w3}=50, respectively. The other parameter is $\eta_c = n_w/(n_w + 1) - 0.3$. It can be seen that the slopes of the filters become steeper when the n_w decrease. However, the transmission probability for high photon number also decrease when n_w becomes small. It is well known that the slope and transmittance probability at higher photon-number side are two important parameters for a good high-pass filter. So that, the filter constructed with linear optics is not an ideal quantum filter.

The clearest physical description of the quantum filter properties is that of a QND measurement for the n-photon state. At the working point of $n = n_w$, the quantum filter can be perfectly demonstrated. The imperfect QND measurement provides the response function of a quantum HPF or LPF. It should be pointed out that the quantum filter, shown in our paper, is different from a classical filter. The response of the classical filter has a definite transmittance for each input physical quantity. However, the curves shown in Fig.2 is a probability distribution function for each input photon-number, which gives the conditional probability that the photons appear at the output port a_2 of the filter while the triggering event happens. This filter can thus be used to generate a sub-Poissonian state generally generated from nonlinear process before.

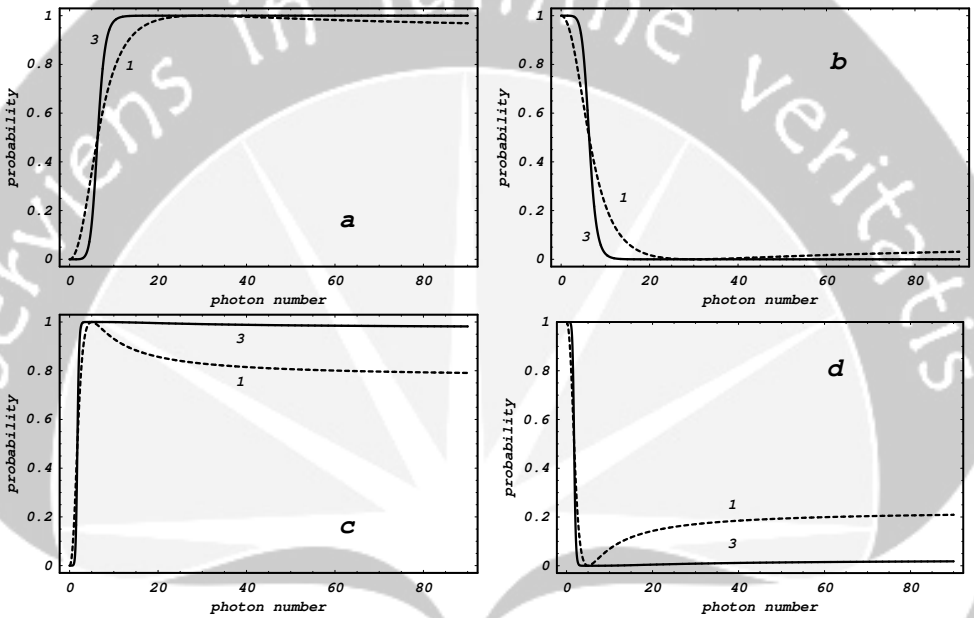


Figure 3: The response of high-order quantum HPF and LPF. (a,b) The lines 1, and 3 are corresponding to first- and third-order HPFs or LPFs. The parameters are set to be $n_w = 30$, $\eta_c = n_w/(n_w + 1) - 0.3$. (c,d) The lines 1, and 3 are corresponding to first- and third-order HPFs or LPFs. The parameters are set to be $n_w = 5$, $\eta_c = n_w/(n_w + 1) - 0.3$.

The output characteristic of the quantum filter can be improved by using second- or higher-order quantum HPF. The simplest way to make a second-order filter is just to cascade two quantum filters, that is, to connect one after the other, so the input state must go through the first one and then the second. The solid lines in Figure.3 show the response for third-order quantum HPF or third-order LPF. The working points of filters shown in Fig.3(a) and Fig.3(b) are $n_w = 30$. The working points of filters shown in Fig.3(c) and Fig.3(d) are $n_w = 5$. The responses of the first-order filters are also comparatively shown in Fig.3 (the dashed lines). It can be seen that the slopes of the third-order filters are steeper than the first-order one. And also, the normalized high-pass or low-pass characteristic of the third-order filter is also improved.

3 Photon-number squeezed state by cascaded quantum low- and high-pass filters

Consider the schematic setup for the generation of sub-Poissonian state illustrated in Fig.4. The input state $|\psi_{in}\rangle = |\alpha\rangle$ for the setup is a coherent state, which can be written as a superposition of the Fock states in the form

$$|\alpha\rangle = e^{-|\alpha|^2/2} \sum_{m=0}^{\infty} \frac{\alpha^m}{\sqrt{m!}} |m\rangle.$$

The coherent state passes through a quantum LPF firstly and then a quantum HPF. The quantum LPF and HPF may have different working points n_1 and n_2 respectively. The sub-Poissonian state preparation is successful when the operations of both quantum filters work properly. The output state can then be derived as

$$|\psi_{out}\rangle \quad = e^{-|\alpha|^2/2} \textstyle\sum_{m=0}^{\infty} \frac{\alpha^m}{\sqrt{m!}} C_{1-LPF} C_{2-HPF} |m\rangle.$$

Figure 5 shows an example that generates a sub-Poissonian state for a specific case of this scheme. The input light is a coherent state with $|\alpha| = \sqrt{2}$. The solid line in Fig.5 gives the probability distribution of the state $|\psi_{out}\rangle$ after the filter pair. The Poissonian distribution $P(m)$ of a coherent state is also shown in Fig.5 (dashed line). For convenience, the probability distribution of the sub-Poissonian state has been magnified to the same level as the coherent state. It can be seen that the width of photon-number distribution of $|\psi_{out}\rangle$ is much narrower than that of the coherent state. The parameters are set to be $n_1 = 20$, $n_2 = 10$, $\eta_{c1} = n_1/(n_1 + 1) - 0.4$, and $\eta_{c2} = n_2/(n_2 + 1) - 0.4$. The efficiency of successful preparation of the sub-Poissonian state in our scheme can be defined as the ratio between the postselected photons and the injected coherent photons. The probability of the sub-Poissonian state shown in Fig.5 is calculated to be 1.3%, which is a acceptable level. More squeezing of the photon-number distribution with reduced successful probability can be realized by using higher-order LPF and HPF.

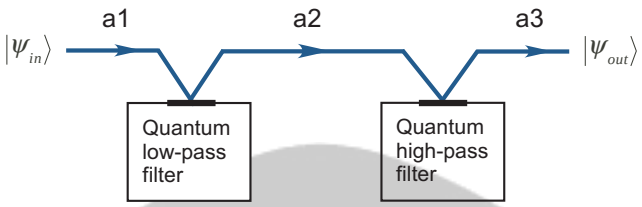


Figure 4: Schematic for the generation of sub-Poissonian state. The coherent state passes through a quantum LPF firstly and then a quantum HPF. The quantum LPF and HPF may have different working points n_1 and n_2.

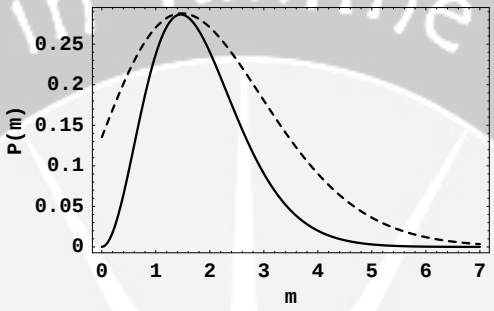


Figure 5: The distribution functions $P(m)$ of the generated sub-Poissonian state for specific cases of this scheme. The dash line shows the Poissonian distribution of a coherent state with $|\alpha| = \sqrt{2}$. The parameters are set to be $n_{w1} = 20$, $n_{w2} = 10$, $\eta_{c1} = n_{w1}/(n_{w1} + 1) - 0.4$, and $\eta_{c2} = n_{w2}/(n_{w2} + 1) - 0.4$.

It should be noted that the quantum filters, as depicted in our paper, is a nondeterministic quantum element the operation of which is conditioned on the detection of an auxiliary photon. The generation of the sub-Poissonian state can be seen as a collapse of the entanglement among photons at output ports of a_2, b_2 and c_2. However, the state preparation is made by postselection of the relevant events in the record of the measurements on the two detectors, which can be made after physical measurements. In fact, because of only one photon exists at the output ports of b_2 or c_2, all the n photon components which happen appear at output port a_2 will contribute to the non-postselected output state. So that no wave-function collapse actually occurs from non-postselected output state to sub-Poissonian state.

Conclusion

In conclusion, we have firstly introduced and analyzed the output characteristic of quantum high- and low-pass filter which consist of only passive linear optics. It was shown that a sub-Poissonian state can be generated from a coherent state by using a band-pass filter which consists of quantum LPF and HPF. The generated

sub-Poissonian state in a free propagating optical mode can then be used to other purposes. The difference between the proposed scheme and those proposed before is that the measurement-caused nonlinear effect, instead of traditional nonlinear process, is used to alter the photon distribution of a coherent light.

References

[1] J. N. Hollenhorst (1979), *Phys. Rev. D.*, **19**, 1669; C. M. Caves (1981), *Phys. Rev. D.*, **23**, 1693.

[2] W. M. Zhang, D. H. Feng and R. Gilmore (1990), *Rev. Mod. Phys.*, **62**, 867; V. V. Dodonov (2002), *J. Opt. B: Quantum Semiclass. Opt* (2002). **4**, R1; R. Loudon and P. L. Knight (1987), *J. Mod. Opt.*, **34**, 709.

[3] A. Furusawa et al. (1998), *Science*, **282**,706; Y.-H. Kim et al. (2001), *Phys. Rev. Lett.*, **86**, 1370; W. P. Bowen et al. (2003), *Phys. Rev. A.*, **67**, 032302.

[4] H. Yonezawa, T. Aoki, and A. Furusawa (2004), *Nature*, **431**, 430.

[5] E. Knill, R. Laflamme, and G. Milburn (2001), *Nature (London)*, **409**, 46.

[6] Hai-Bo Wang, Y.M. Li, S. Odate, and T. Kobayashi(2005), *Phys. Rev. D.*, **72**, 013822.

[7] K. Sanaka (2005), *Phys. Rev. A.* **71**, 021801 .

[8] G. J. Pryde, J. L. O'Brien, A. G. White, S. D. Bartlett, and T. C. Ralph (2004), *Phys. Rev. Lett.*, **92**, 190402.

[9] H.B. Wang and T. kobayashi, Phys. Rev. A 71, 021802 (2005).

[10] J. Laurat, T. Coudreau, N. Treps, A. Maitre, and C. Fabre (2003), *Phys. Rev. Lett.*, **91**, 213601; Y. Zhang, K. Kasai, M. Watanabe (2002), *Opt. Lett.*, **27**, 1244.

Takayoshi Kobayashi: Department of Physics, Graduate School of Science, University of Tokyo,7-3-1 Hongo, Bunkyo, Tokyo, 113-0033, Japan
Phone: +81-3-5841-4227. Fax: +81-3-5841-4165.
E-mail: kobayashi@phys.s.u-tokyo.ac.jp

Hai-bo Wang: Post-Dr. researcher at Department of Physics, Graduate School of Science, University of Tokyo,7-3-1 Hongo, Bunkyo, Tokyo, 113-0033, Japan
Core Research for Evolutional Science and Technology(CREST), Japan Science and Technology Corporation (JST).
E-mail: wang@femto.phys.s.u-tokyo.ac.jp or obiahw@yahoo.com.cn

SIMULATING THREE-DIMENSIONAL VORTEX INTERACTIONS WITH VORTEX METHODS

H. Ishikawa[a], S. Izawa[b], M. Kiya[c]

[a] Tokyo University of Science, Tokyo, Japan
[b] Tohoku University, Sendai, Japan
[c] Kushiro National College of Technology, Kushiro, Japan

Abstract. Several examples of vortex interaction are simulated by three-dimensional vortex methods. The first is the interaction between a straight vortex tube and a vortex ring. Modes of the interaction appears be classified into three categories in terms of circulation of the vortex ring relative to that of the vortex tube. The second is the collision of two impulsively started round jets issuing from circular nozzles of the same diameter. At the head-on collision, a periodic deformation is generated along the periphery of the colliding vortex rings, developing into smaller vortex rings (ringlets) which move in the radial direction.
Key-words: vortex method, vortex interaction, mixing layer, vortical structure, colliding jets, vortex ring, vortex tube

1 Introduction

It is essential to study vortex interactions in order to understand dynamics of turbulent flows in engineering applications in which flows are mostly turbulent. Enhancement of heat and mass transfer in turbulent shear flows is realized by manipulation of vortical structure in the flows. Enhanced turbulent mixing also delays or prevents flow separation from a solid surface by introducing high momentum flow. The turbulent mixing may be understood in terms of interaction of multiple-scale vortices.

Numerical simulation has the advantage of investigating detailed mechanism of vortex dynamics. This paper presents examples of vortex interaction simulated by three-dimensional vortex method [1][2]. Vortex method is one of the mesh-less type numerical simulations of fluid dynamics, calculating evolution of a vortical region by a collection of vortex blobs with overlapping cores. The vortex method has merits in simulations of vortical flows in that this method employs vortex elements which give us direct information on evolution of vortex structure in the flows [3].

The first example is the interaction of a vortex ring with a vortex tube. The vortex tube is a model of rolling-up vortices in a plane mixing layer, while the vortex ring is a simple model of external vortices introduced into the mixing layer to manipulate its growth or mixing [4]. The interaction of a vortex ring with mixing layer vortices, consisting of five vortex tubes in a linear arrangement was also actually investigated. The second example is the interaction of vortices in two impulsively started round jets impinging head-on or at right angles.

2 Vortex method

Vortex method calculates evolution of a vortical region by a collection of vortex blobs with overlapping cores. A vortex blob is defined by its position $\boldsymbol{x}^\alpha$, vorticity $\boldsymbol{\omega}^\alpha$, volume $d^3\boldsymbol{x}^\alpha$ and cut-off radius σ_α. Strength of a vortex blob is denoted by $\boldsymbol{\gamma}^3 = \boldsymbol{\omega}^\alpha d^3\boldsymbol{x}^\alpha$. Vorticity field $\boldsymbol{\omega}$ at a time t is given by

$$\boldsymbol{\omega}^\alpha(\boldsymbol{x},t) = \frac{1}{\sigma_\alpha^3} p\left(\frac{\left|\boldsymbol{x}-\boldsymbol{x}^\alpha(t)\right|}{\sigma_\alpha}\right)\boldsymbol{\gamma}^\alpha \tag{1}$$

where $p(\cdot)$ is the smoothing function of the cut-off radius σ_α. Evolution of the position $\boldsymbol{x}^\alpha$ is described by the Biot-Savart law in Lagrangian form, as follows

$$\frac{d\boldsymbol{x}^\alpha}{dt} = -\frac{1}{4\pi}\sum_\beta \frac{r_{\alpha\beta}^2 + (5/2)\sigma_\beta^2}{\left(r_{\alpha\beta}^2+\sigma_\beta^2\right)^{5/2}}\boldsymbol{r}^{\alpha\beta}\times\boldsymbol{\gamma}^\beta \tag{2}$$

Where $\boldsymbol{r}^{\alpha\beta} = \boldsymbol{x}^\alpha - \boldsymbol{x}^\beta$, $r^{\alpha\beta} = \left|\boldsymbol{r}^{\alpha\beta}\right|$. Evolution of the vorticity is described by vorticity equation without viscous diffusion of vorticity

$$\frac{d\boldsymbol{\gamma}^\alpha}{dt} = \left(\boldsymbol{\gamma}^\alpha\cdot\nabla\right)\boldsymbol{u}^\alpha \tag{3}$$

In this paper the viscous diffusion of vorticity is represented by a core-spreading model or a particle-exchange scheme.

A core-spreading model [1][5] is argued not to yield the exact solution of Navier-Stokes equations even in the limit of infinitely many vortex blobs (8). However, this model is expected to give a fairly good approximation to the exact solution within a finite time after the start of flow [9]. Moreover, the core-spreading model appears to be useful in the sense that it reproduces with tolerable accuracy the gross feature of evolution of large-scale vortices, as demonstrated by a number of applications especially in two-dimensional flows. This model is also attractive because less number of vortex blobs are required than the particle-exchange method.

A particle-exchange scheme, the essence of this method is in the point which the diffusion term of vorticity can be approximated by an integral operation. The details of this method are given in Degond and Mas-Gallic [6] and Winckelmans and Leonard [7].

3 Vortex ring – vortex tube interaction

The interaction of a vortex tube with a vortex ring is simulated by the core-spreading model. In this case the important issue is change of modes of interaction as a function of circulation of the vortex ring Γ_R relative to that of the vortex tube Γ_F. The circulation ratio Γ_R / Γ_F ($=\lambda$) is varied from 0.5 to 2.0. The thickness of the ring and the tube is chosen to be the same, the radius being $0.27R$, where R is the radius of the vortex ring. Reynolds number based on the velocity of translation of the ring and its diameter is 500.

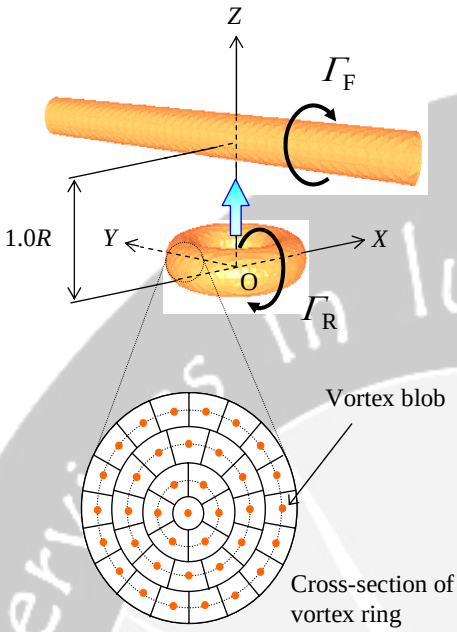


Figure 1. Initial arrangement of vortex ring and vortex tube. Cross section of vortex ring shows arrangement of vortex blobs. R is diameter of vortex ring.

The cross section of both vortices was initially divided into four layers of the same thickness, each layer consisting of vortex blobs of particular vorticity in such a way that the vorticity distribution in the cross section is approximately Gaussian (Fig. 1). The total number of vortex blobs is 37 in the cross section. The length of the vortex tube is chosen as $20R$, which was found by preliminary calculations to be long enough to eliminate the end effects on its interaction with the vortex ring.

The simulations revealed three modes of the interaction depending on the circulation ratio λ. When the ratio λ is of the order of 0.5, the vortex ring is deformed and wrapped around the vortex tube to be finally engulfed into the latter. The vortex tube experiences relatively small deformation during the interaction (Fig. 2). The region of interaction is localized in the sense that the deformation of the vortex tube does not propagate outwards in the axial direction.

On the other hand, when circulation of vortex ring is equal to that of vortex tube, the cut-and-reconnection occurs between the two vortices (Fig. 3). A part of the vortex ring is replaced by a part of the vortex tube to form a new vortex ring, which moves away from the tube almost in the same direction as that of the original vortex ring. The remaining part of the vortex ring now fills in the removed part of the tube.

When the circulation ratio is greater than approximately 1.5, the vortex ring passes through the vortex tube, keeping its initial shape. The vortex tube, on the other hand, is partially cut in the region of interaction. Details of interaction also depend on the initial relative position of the vortex ring and the tube. The effects of the initial position will be described for $\lambda = 1$ because this case is most representative. If a part of the ring meets the tube at a part of anti-parallel vorticity, these parts

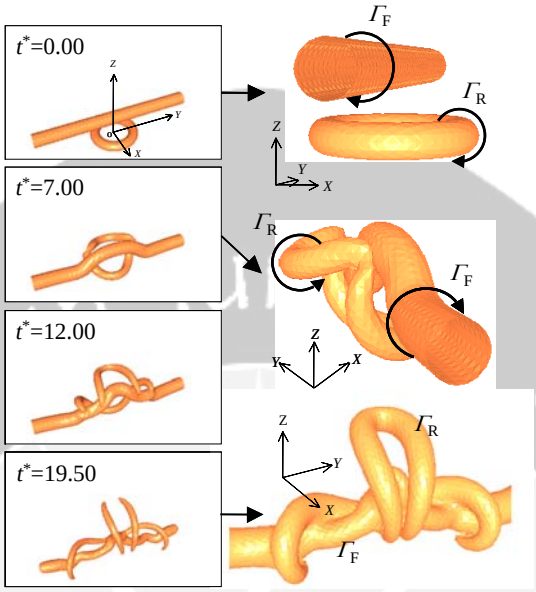


Figure 2. Vortex ring interacting with a vortex tube. Circulation ratio λ = 0.5. Time advances from top to bottom.

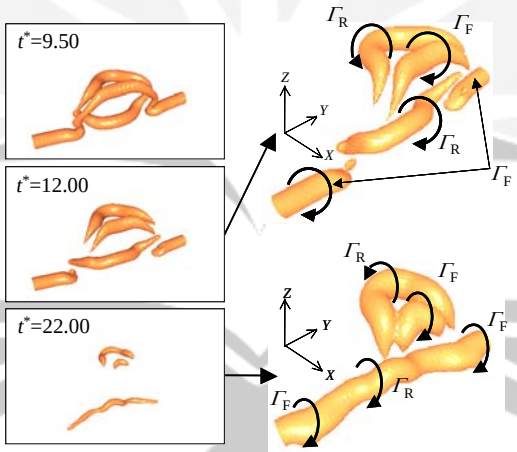


Figure 3. Vortex ring interacting with a vortex tube. Circulation ratio λ = 1.0. Time advances from top to bottom.

form a vortex-pair-like structure, which move away and will eventually be dissipated, while the other parts of both vortices will reconnect to form a single, deformed vortex tube. On the other hand, if both vortices meet at parts of parallel

vorticity, the local coalescence occurs. The remaining part of the vortex ring translates approximately in the original direction.

4 Vortex ring – mixing layer vortices interaction

In last section, the vortex tube is a model of a rolling-up vortex in a plane mixing layer while the vortex ring is a simple model of external vortices introduced into the mixing layer to manipulate its growth or mixing [4][10]. In this section, the whole mixing layer was represented by five vortex tubes in a linear arrangement for streamwise direction. These vortex tubes, having the same cross section as the vortex ring, consist of 37 vortex blobs. The initial vorticity within the cross section is also the same as that of the ring. However, whole mixing layer is tending to rotate by its self-induced velocity. In order to prevent this rotation, other five dummy vortex tubes are arranged in the both side of the mixing layer vortex tubes. The cross section of dummy vortex tubes consists of 7 vortex blobs in 2 layers. The initial vorticity distribution in the cross section is the third-order Gaussian as the same as the vortex ring. The streamwise distance of adjacent vortex tubes is chosen as 1.5R. The spanwise length of the vortex tube is chosen as 20R.

Fig.4 shows the interaction the mixing layer vortices with a vortex ring λ = 1.0. The pair vortex rotating in the same direction as the mixing layer vortices (VR_++VF_3) , generates large vortices in upstream side. On the other hand, another pair vortex rotating in the opposite direction (VR_-+VF_2) move in the downstream direction by self-induced velocity. The vortex ring experiences large stretching in the streamwise direction. The vortex ring directly interacts with almost five vortex tubes.

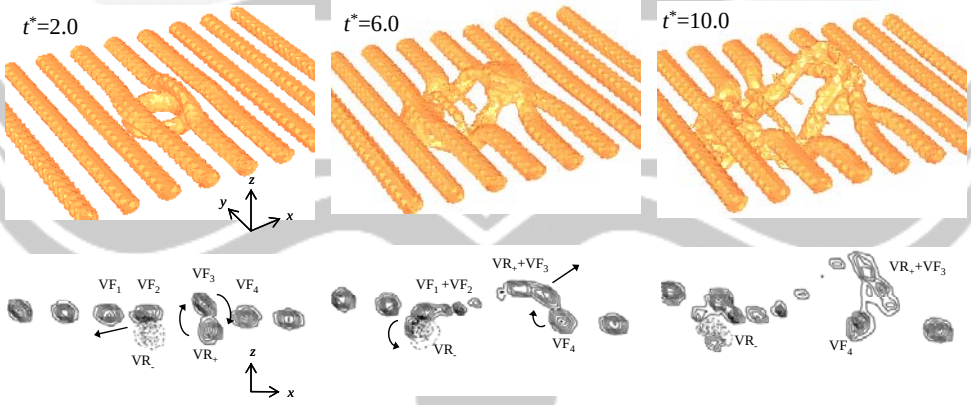


Figure 4. Vortex ring interacting with mixing later vortices. Circulation ratio λ = 1.0. Top figures are the isosurface of magnitude of vorticity. Bottoms are the cross section of contour of vorticity.

In view of enhancement of growth rate and mixing, the above results suggest that a vortex ring should have approximately the same circulation as that of mixing layer vortices.

5 Vortex interaction in impinging round jets

Simulations are made for interaction of vortices during collision of two impulsively started round jets issuing from nozzles of the same radius R, the axes of the nozzles lying in the same plane. The head-on collision and the collision at right angles are considered. The viscous diffusion of vorticity is incorporated by the particle-exchange scheme. Reynolds number based on the diameter of the nozzle and the exit velocity is 2000.

The surface of the nozzle of length $1.05R$ is constructed by rectangular panels of vortex blobs. The jet flow is produced by a source disk located inside the nozzle. Nascent vortex blobs are introduced into the flow at $0.261R$ downstream of the edge of the nozzle in such a way as to satisfy Kelvin's law.

The circular shear layer of the jets rolls up to form a series of vortex rings, the most significant one being the starting vortex. For the head-on collision the starting vortex rings approach each other to be rapidly stretched in the radial direction owing to the mutually induced velocity (Fig. 5). Periodic deformation appears along the circumferential direction of the colliding vortex rings due to a Crow type of instability [11], developing into small vortex rings (ringlets), which will travel in the radial direction as time advances as shown experimentally by Lim & Nickels [12]. The successive collision of vortex rings in the jets produces multi-scale vortex

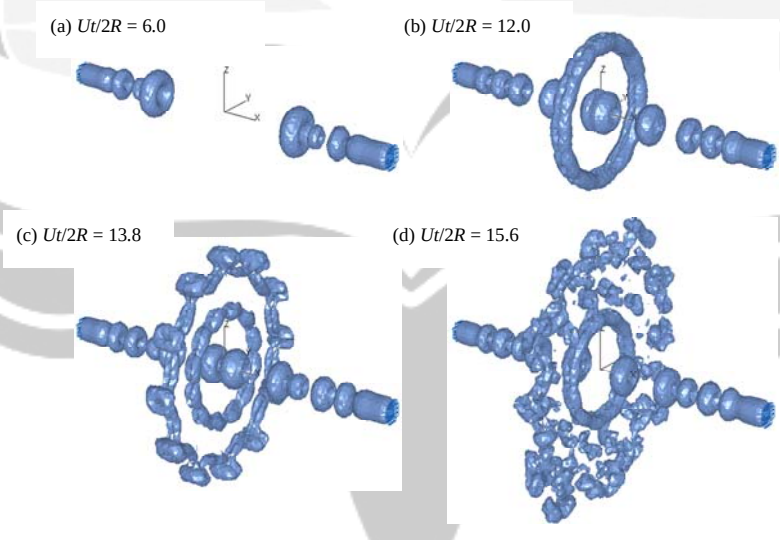


Figure 5. Head-on collision of impulsively started round jets. U = velocity at exit of nozzle, R = radius of nozzle, t = time after start of flow.

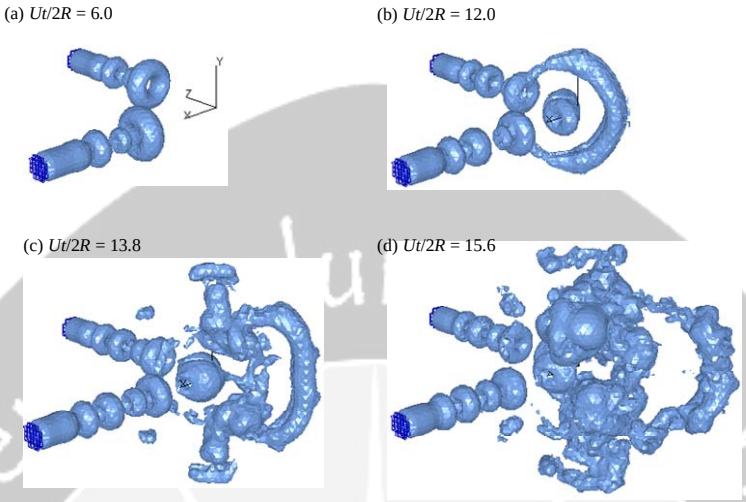


Figure 6. Collision of impulsively started round jets at right angles.

structure in a region centered in the mid plane.

It is worth noting that vortex rings in both jets have almost the same arrangements in time and space, a vortex ring in one jet having its counterpart in the other one, although no systematic forcing is made. This fact may suggest a cross talk between the jets. Moreover, the number of the ringlets is the same to the number of panels of the jet nozzle. Thus a small disturbance associated with the panels may have triggered the initial deformation.

On the other hand, when the jets impinge at right angles, the starting vortex rings collide to be stretched as a whole in the plane of bisector (Fig. 6). Those parts of the vortex rings which first meet are rapidly dissipated while the remaining parts touch to form a vortex-pair-like structure, which moves in the direction of bisector. The impingement of successive vortex rings in the jets produces complicated vortex structure of multiple scales in the region of collision. Further details of the vortex interaction are also clarified in terms of distributions of pressure and dissipation rate.

6 Conclusions

This paper presented numerical results of the interaction of multiple-scale vortices. To the purpose of enhancing the local growth and mixing, and controlling actively the turbulent flow by a vortex ring, the numerical method is made by means of a three dimensional vortex method.

The interaction of the mixing layer vortices with the vortex ring is expected to include two major results: One is the enhanced mixing archived the most

remarkable effect when the circulation ratio of the vortex ring relative to that of the mixing layer vortices is unity. The vortex ring experiences large stretching in streamwise direction. If the circulation ratio is greater than approximately 1.5, otherwise the spatial evolution by vortex interaction is weak this because the vortex ring passes through the vortex tube.

The evolution of vortex interaction in impinging jet was also simulated by vortex method. The process of generation of the small vortex ring (ringlet) in head-on collision is demonstrated by vortex motion. The deformation of vortical structure in the jets impinge at right angles, produces complicated vortex structure of multiple scales in the region of collision.

References

[1] Izawa, S. and Kiya, M. [2000], *Turbulence model for the three-dimensional vortex blob method.* JSME International Journal, **43**, No. 3, 434-442.

[2] Ishikawa, H., Izawa, S., Mochizuki, O. and Kiya, M. [2002], *Vortex ring – vortex tube interaction*, Transactions of the JSME, **68**, 2688-2694 (in Japanese).

[3] Winckelmans, G.S. and Leonard, A. [1993], *Contributions to vortex particle methods for computation of three-dimensional incompressible unsteady flows.* Journal of Computational Physics, **109**, 247-273.

[4] Kiya, M., Ohyama, M. and Hunt. J.C.R. [1986], *Vortex pairs and rings interacting with shear-layer vortices*, Journal of Fluid Mechanics, **172**, 1-15.

[5] Leonard, A. and Chua, K. [1989], *Three-dimensional interactions of vortex tubes.* Physica. D.**37**, 490-496.

[6] Degond, P. and Mas-Gallic, S. [1989], *The weight particle method for convection-diffusion equations. Part 1. The case of an isotropic viscosity*. Math. Compt., **52**,485-507.

[7] Winckelmans, G.S. and Leonard, A. [1993], *Contributions to vortex particle methods for computation of three-dimensional incompressible unsteady flows.* Journal of Computational Physics, **109**, 247-273.

[8] Greengard, C. [1985], *The core spreading vortex method approximates the wrong solution.* Journal of Computational Physics **61**, 345-347.

[9] Nakajima, T. and Kida, T. [1990], *A remark on discrete vortex method (Derivation from Navier-Stokes equations).* Transactions of the Japan Society of Mechanical Engineers, **56**, 3284-3291 (in Japanese).

[10] Kiya, M., Mochizuki, O. and Suzuki, N. [1999], *Separation Control by Vortex Projectiles.* AIAA Paper 99-3400.

[11] Crow, S.C. [1970], *Stability of theory for a pair of trailing vortices.* AIAA Journal, **8**, 2172-2179.

[12] Lim, T.T. and Nickels, T.B. [1992] *Instability and reconnection in the head-on collision of two vortex rings.* Nature, **357**, 225-227.

H. ISHIKAWA: Department of Mechanical Engineering, Tokyo University of Science, Tokyo 162-8601, Japan. E-mail: ishi@rs.kagu.tus.ac.jp

S. IZAWA: Department of Mechanical Systems and Design, Tohoku University, Sendai 980-8579, Japan. E-mail: izawa@fluid.mech.tohoku.ac.jp

M. KIYA: President, Kushiro National College of Technology, Kushiro 084-0916, Japan. E-mail: kiya@office.kushiro-ct.ac.jp

Illuminating the Sumatra Great Earthquake Sequences by Employing Tomographic Inversions of Seismic Data

Sri Widiyantoro

Geophysics Study Program, Department of Geophysics and Meteorology,
Institut Teknologi Bandung, Indonesia

Abstract: Earthquakes generate seismic waves that acquire characteristics related to the properties of the regions through which they have traveled. Therefore, seismic wave data can be used to image structures of the Earth's interior. For this purpose a tomographic inversion technique represents a powerful tool to delineate the internal structure of the three-dimensional (3D) Earth.

We have conducted inversions of P- and S-wave travel time data for the Sumatra region and its vicinity in order to illuminate the great earthquake with its attendant devastating tsunami of December 26, 2004. The resulting seismic tomograms clearly depict that the earthquake initiated where the dip of the subducting oceanic lithosphere of the Indo-Australian Plate is most gentle along the Andaman, Sumatra and Java trenches. This may have caused that coupling between the subducting and overriding plates is exceptionally strong so that it could generate such a great earthquake. The gently dipping lithospheric slab may be due to the relatively young age of the incoming plate combined with the oblique subduction below Sumatra.

In addition, we have extracted elastic parameters (incompressibility and rigidity) from the resulting P- and S-wave models. The pattern of the elastic parameters gradient is surprisingly striking, in which locations of high incompressibility gradient agree remarkably well with locations of recent great earthquakes e.g. the December 26, 2004 and March 28, 2005 events. So it is suggestive that incompressibility represents a sensitive parameter to predict the potential location of great earthquake in region that has undergone severe compression.

Tsunami Wave Mathematics

E. Pelinovsky

Laboratory of Hydrophysics and Nonlinear Acoustics, Institute of Applied Physics, Nizhny Novgorod, Russia

Abstract: The giant tsunami occurred in the Indian Ocean on 26th December 2004 draws attention to this natural phenomenon. The given course of lectures deals with the physics of the tsunami wave propagation from the source to the coast. Briefly, the geographical distribution of the tsunamis is described and physical mechanisms of their origin are discussed. Simplified robust formulas for the source parameters (dimension and height) are given for tsunamis of different origin. It is shown that the shallow-water theory is an adequate model to describe the tsunamis of the seismic origin; meanwhile for the tsunamis of the landslide or explosion (volcanoes, asteroid impact) origin various theories (from linear dispersive to nonlinear shallow-water equations) can be applied. The applicability of the existing theories to describe the tsunami wave propagation, refraction, transformation and climbing on the coast is demonstrated. Nonlinear-dispersive effects including the role of the solitons are discussed. The practical usage of the tsunami modeling for the tsunami forecasting and tsunami risk evaluation is described. The results of the numerical simulations of the two global tsunamis in the Indian Ocean induced by the catastrophic Krakatau eruption in 1883 and the strongest North Sumatra earthquake in 2004 are given.

Simulating Flood Waves for Accurate Warning Systems

G.S. Stelling

TUDelft, The Netherlands

Abstract: Floods are an increasing hazard to the populations of many countries. The hazards range from only some local, not life-threatening, inconvenience such as flooding of streets, to the recent tsunami disaster. In particular this tsunami disaster has lead to an increased interest for flood warning systems. Only accurate simulations can be applied as part of an on-line operational warning system. Inaccurate predictions might cause panic among populations for no reason. Tsunamis, for instance, propagate at high speeds so there is or only a limited amount of time available for accurate warnings. This seems to be drawback of these systems. Off-line analysis of the impact of flooding can be useful as well. It yields information about the potential effects on certain areas that can be used for evacuation planning. In this contribution both aspects of flood wave simulations will be dealt with. Examples will be given of these types of systems for forecasting and analyzing flood waves in the Netherlands and in the Bay of Bangal. A numerical model, based upon a combination of 1D channels for normal drainage, and a 2D unstructured grid model for overland flow, will be described in some detail.

KINEMATICAL CONSERVATION LAWS, RAY THEORIES AND APPLICATIONS

Phoolan Prasad

Indian Institute of Science, Bangalore, India

Abstract. In this article we describe a general form of ray equations and prove the extended lemma on bicharacteristics. We also prove equivalence of the ray equations to the differential form of the kinematical conservation laws (KCL) in two space dimensions. We mention some applications of KCL and discuss briefly its application to sonic boom by a maneuvering aerofoil.

Key-words: ray theory, bicharacteristics, hyperbolic equations, conservation laws, sonic boom.

1 Introduction

Rays have been used in the construction of successive positions of a wavefront Ω_t since a very long time. A general definition of rays requires formulation of the equation of the wavefront in terms of an eikonal equation, which is a first order nonlinear partial differential equation. The ray velocity $\boldsymbol{\chi}$ at any point of Ω_t in a continuum medium depends not only on the position $\mathbf{x}$ but also on the orientation of Ω_t given by the unit normal $\mathbf{n}$ of the Ω_t. This formulation shows that not every vector $\boldsymbol{\chi}$ qualifies for being a ray velocity but must satisfy a set of compatibility conditions.

Kinemtical conservation laws (KCL) describing the time-evolution of a wavefront has only a recent origin in 1992 [8]. We shall prove that ray equations are equivalent to KCL for a wavefront Ω_t in two space-dimensions.

KCL has been found to be very useful in solving many practical problems in nonlinear wave propagation, where singularities in the form of kinks appear on Ω_t. We shall mention some of these applications and describe in some detail its application to sonic boom problem which has been described in two articles [2,3].

2 Ray equations and extended lemma on bicharacteristics

Let $\Omega_t : \phi(\boldsymbol{x}, t) = 0$ be a wavefront in $\mathbb{R}^m$, which evolves according to the eikonal equation

$$Q(\boldsymbol{x}, t, \nabla\phi, \phi_t) \equiv \phi_t + \langle \boldsymbol{\chi}, \nabla\rangle\phi = 0 \tag{1}$$

and where the ray velocity $\boldsymbol{\chi}$ depends not only on the position $\boldsymbol{x} \in I\!R^m$, time $t \in I\!R$ but also on the unit normal $\boldsymbol{n} = \nabla\phi/|\nabla\phi|$ of Ω_t. The Hamilton's canonical equations of the first order partial differential equation $Q(\boldsymbol{x}, t, \boldsymbol{p}, q) = 0,\ \boldsymbol{p} = \nabla\phi,\ q = \phi_t$ give

$$\frac{dx_\alpha}{dt} = \chi_\alpha + p_\gamma \frac{\partial \chi_\gamma}{\partial p_\alpha}, \quad \frac{dp_\alpha}{dt} = -p_\beta \frac{\partial \chi_\beta}{\partial x_\alpha} \tag{2}$$

where we use summation convention for a repeated suffix. But

$$\frac{\partial}{\partial p_\alpha} = \frac{\partial n_\beta}{\partial p_\alpha}\frac{\partial}{\partial n_\beta} = \frac{1}{|\boldsymbol{p}|}\left(\delta_{\alpha\beta} - n_\alpha n_\beta\right)\frac{\partial}{\partial n_\beta}$$

where $\delta_{\alpha\beta}$ are Kronecker deltas, so that

$$\begin{aligned} \frac{dx_\alpha}{dt} &= \chi_\alpha + n_\gamma\left(\frac{\partial \chi_\gamma}{\partial n_\alpha} - n_\alpha n_\beta \frac{\partial \chi_\gamma}{\partial n_\beta}\right), \\ &= \chi_\alpha + n_\beta n_\gamma\left(n_\beta \frac{\partial}{\partial n_\alpha} - n_\alpha \frac{\partial}{\partial n_\beta}\right)\chi_\gamma, \text{ using } 1 = n_\beta n_\beta \end{aligned} \tag{3}$$

This equation is inconsistent with assumption that $\boldsymbol{\chi}$ is the ray velocity (i.e., $\frac{d\boldsymbol{x}}{dt} = \boldsymbol{\chi}$) unless for each α

$$n_\beta n_\gamma\left(n_\beta \frac{\partial}{\partial n_\alpha} - n_\alpha \frac{\partial}{\partial n_\beta}\right)\chi_\gamma = 0 \tag{4}$$

The equation for p_α can be written in terms of an equation for $n_\alpha = p_\alpha/|\boldsymbol{p}|$. Thus, when (4) is satisfied, the ray equations (2) of (1) take the form

$$\frac{dx_\alpha}{dt} = \chi_\alpha, \quad \frac{dn_\alpha}{dt} = -n_\beta n_\gamma\left(\frac{\partial}{\partial \eta^\alpha_\beta}\right)\chi_\gamma \tag{5}$$

where

$$\frac{\partial}{\partial \eta^\alpha_\beta} = n_\beta \frac{\partial}{\partial x_\alpha} - n_\alpha \frac{\partial}{\partial x_\beta} \tag{6}$$

represents a derivative in the direction of a tangent to the surface Ω_t.

A ray associated with a wavefront Ω_t is defined with the help of an eikonal equation. When the ray velocity $\boldsymbol{\chi}$ is given, the eikonal equation is of the form (1). The ray velocity $\boldsymbol{\chi}$ at a point P of Ω_t depends not only on the position $\boldsymbol{x}$ of P and the orientation (given by the unit normal $\boldsymbol{n}$) of Ω_t at P but also on the state of the medium at P, specified by an amplitude function $w(\boldsymbol{x}, t)$. The eikonal equation (1) is not a linear or quasilinear equation but a more general nonlinear equation as $\boldsymbol{\chi}$ depends on $\boldsymbol{n} = \nabla\phi/|\nabla\phi|$. The condition (4) shows that $\boldsymbol{\chi}(\boldsymbol{x}, t, \boldsymbol{n}) \equiv \boldsymbol{\chi}(\boldsymbol{x}, t, \boldsymbol{n}, w(\boldsymbol{x}, t))$ can not be prescribed arbitrarily but when $\boldsymbol{\chi}$ is prescribed satisfying (4), the first equation in (5) implies the second equation in it through the eikonal equation (1).

Consider now a hyperbolic system of first order quasilinear equations

$$A(\boldsymbol{u},\boldsymbol{x},t)\boldsymbol{u}_t + B^{(\alpha)}(\boldsymbol{u},\boldsymbol{x},t)\boldsymbol{u}_{x_\alpha} + \boldsymbol{C}(\boldsymbol{u},\boldsymbol{x},t) = 0 \tag{7}$$

where $\boldsymbol{u} \in I\!R^n, A \in I\!R^{n\times n}$, $B^{(\alpha)} \in I\!R^{n\times n}$ and $\boldsymbol{C} \in I\!R^n$. Then the velocity C of a wavefront Ω_t, across which $\boldsymbol{u}$ is continuous, is equal to an eigenvalue of (7) and Ω_t is the projection on $\boldsymbol{x}$-space of a section of the characteristic surface Ω by t=constant plane. Note that Ω is a surface in space-time i.e., $(\boldsymbol{x},t)$-space. The ray velocity components corresponding to the eigenvalue C are given by the lemma on bicharacteristics directions [4, p.597]

$$\chi_\alpha = \frac{\boldsymbol{l}B^{(\alpha)}\boldsymbol{r}}{\boldsymbol{l}A\boldsymbol{r}} \tag{8}$$

where $\boldsymbol{l}$ and $\boldsymbol{r}$ are the left and right null vectors satisfying

$$\boldsymbol{l}(n_\alpha B^{(\alpha)} - CA) = 0, \quad (n_\alpha B^{(\alpha)} - CA)\boldsymbol{r} = 0 \tag{9}$$

and

$$C = n_\alpha \chi_\alpha = \frac{(\boldsymbol{l}n_\alpha B^{(\alpha)}\boldsymbol{r})}{(\boldsymbol{l}A\boldsymbol{r})} \tag{10}$$

is an eigenvalue.

We know that $\boldsymbol{l}$ and $\boldsymbol{r}$ depend on $\boldsymbol{n}$ but A and $B^{(\alpha)}$ do not. Hence

$$\begin{aligned} n_\gamma \frac{\partial \chi_\gamma}{\partial n_\alpha} &= n_\gamma \frac{\partial}{\partial n_\alpha}\left(\frac{\boldsymbol{l}B^{(\gamma)}\boldsymbol{r}}{\boldsymbol{l}A\boldsymbol{r}}\right) = \frac{1}{(\boldsymbol{l}A\boldsymbol{r})^2}\left[n_\gamma\left(\frac{\partial}{\partial n_\alpha}\boldsymbol{l}\right)\left\{(B^{(\gamma)}\boldsymbol{r})(\boldsymbol{l}A\boldsymbol{r})\right.\right. \\ &\quad \left.\left. - (A\boldsymbol{r})(\boldsymbol{l}B^{(\gamma)}\boldsymbol{r})\right\} + n_\gamma\{(\boldsymbol{l}B^{(\gamma)})(\boldsymbol{l}A\boldsymbol{r}) - (\boldsymbol{l}A)(\boldsymbol{l}B^{(\gamma)}\boldsymbol{r})\}\left(\frac{\partial}{\partial n_\alpha}(\boldsymbol{r})\right)\right] \\ &= \frac{1}{(\boldsymbol{l}A\boldsymbol{r})}\left[\left(\frac{\partial}{\partial n_\alpha}\boldsymbol{l}\right)\left(n_\gamma B^{(\gamma)} - CA\right)\boldsymbol{r}\right. \\ &\quad \left. + \boldsymbol{l}(n_\gamma B^{(\gamma)} - CA)\left(\frac{\partial}{\partial n_\alpha}(\boldsymbol{r})\right)\right] = 0 \end{aligned} \tag{11}$$

By replacing α by β, we also get $n_\gamma \dfrac{\partial}{\partial n_\beta}\chi_\gamma = 0$. Therefore, the condition (4) is satisfied and we have proved that:

Theorem 2.1. *If $\boldsymbol{\chi}$ be a ray velocity of the hyperbolic system (7), then the condition (4) is satisfied and the rays are given by (5).*

In the second equation in (5), the derivatives $\dfrac{\partial}{\partial \eta^\alpha_\beta}$ and hence $\frac{\partial}{\partial x_\alpha}$ operates on χ_γ in which $\boldsymbol{n}$ also appears through $\boldsymbol{l}$ and $\boldsymbol{r}$. We shall simplify it in such a way that the derivatives $\frac{\partial}{\partial x_\alpha}$ appear only on the on A and $B^{(\gamma)}$. If we follow the procedure of differentiation in (11), now with respect to x_α instead of n_α, we get

$$n_\gamma \frac{\partial \chi_\gamma}{\partial x_\alpha} = \frac{1}{(\boldsymbol{l}A\boldsymbol{r})}\boldsymbol{l}\left[n_\gamma \frac{\partial B^{(\gamma)}}{\partial x_\alpha} - C\frac{\partial A}{\partial x_\alpha}\right]\boldsymbol{r}$$

The ray equations (5) now become

$$\frac{dx_\alpha}{dt} = \chi_\alpha \tag{12}$$

$$\frac{dn_\alpha}{dt} = -\frac{1}{(\boldsymbol{l}A\boldsymbol{r})}\boldsymbol{l}\left\{n_\beta\left(n_\gamma\frac{\partial B^{(\gamma)}}{\partial \eta^\alpha_\beta} - C\frac{\partial A}{\partial \eta^\alpha_\beta}\right)\right\}\boldsymbol{r} = \psi_\alpha, \text{ say.} \tag{13}$$

This exactly is the form of the extended lemma on bicharacteristics in [5]. Here is a complete proof. We take the expression (8) for $\boldsymbol{\chi}$ from the lemma on bicharacteristic directions in [4] and then use the present derivation for (13). The proof of (4) can also be extended to the shock ray velocity $\boldsymbol{\chi_s}$ when the shock is weak [10, section 9.2]. The proof for a shock of arbitrary strength will be a little more complex but we believe that it should be possible to complete the proof.

Though the first part $\dfrac{dx}{dt} = \boldsymbol{\chi}$ of the ray equations (5) determines the second part (with the help of the eikonal equation (1)) i.e., the equation for $\boldsymbol{n}$, the first part alone is an under determined set due to the presence of $\boldsymbol{n}$ in $\boldsymbol{\chi}$. For a linear wave propagation, the full set i.e., equations (5) are sufficient for ray tracing which of course, is well known in many applications such as geophysics. For a nonlinear wave propagation or shock propagation the eikonal equation also contains the wave amplitude w as an unknown quantity. Therefore, to the ray equations (5) we need to add some more equations. These additional equations have been obtained in our weakly nonlinear ray theory (WNLRT) and shock ray theory (SRT) [10, Chapters 4 and 9].
In terms of the normal velocity $C = -\phi_t/|\nabla\phi| = \langle \boldsymbol{n}, \boldsymbol{\chi}\rangle$, the equation (1) becomes,

$$\phi_t + C|\nabla\phi| = 0 \tag{14}$$

which is used in the level set method (LSM). Since the vector $\boldsymbol{\chi}$ can not be obtained from the scalar C, the equation (1) is more general than (14). In the LSM, the transport equation for the front intensity w, is not available - this makes our WNLRT and SRT more powerful.

3 Equivalence of ray equations and kinematical conservation laws

We have been able to develop a complete theory of kinematical conservation laws for propagating curves Ω_t only in two space dimensions, hence we restrict our discussion only to two space dimensions where we denote the spatial coordinates by x and y. The unit normal (n_1, n_2) is expressed in terms of θ by $n_1 = \cos\theta,\ \ n_2 = \sin\theta$.

We define a ray coordinate system (ξ, t) such that $\xi =$ constant is a ray and $t =$ constant is the wavefront Ω_t at time t. Let g be the metric associated with ξ i.e.,

$g\,d\xi$ is an element of length along Ω_t, then

$$g = \sqrt{x_\xi^2 + y_\xi^2} \tag{15}$$

The normal and tangential components of $\boldsymbol{\chi}$, denoted by C and T respectively, are

$$C = n_1\chi_1 + n_2\chi_2, \quad T = -n_2\chi_1 + n_1\chi_2 \tag{16}$$

Following [9] and [10], we can derive a pair of kinematical conservation laws,

$$(g\sin\theta)_t + (C\cos\theta - T\sin\theta)_\xi = 0 \tag{17}$$

$$(g\cos\theta)_t - (C\sin\theta + T\cos\theta)_\xi = 0 \tag{18}$$

In two dimensions, the ray equations (5) reduce to

$$\frac{dx}{dt} = \chi_1, \quad \frac{dy}{dt} = \chi_2, \quad \frac{d\theta}{dt} = -\frac{1}{g}\left(n_1\frac{\partial\chi_1}{\partial\xi} + n_2\frac{\partial\chi_2}{\partial\xi}\right) \tag{19}$$

Theorem 3.1. *Given $\boldsymbol{\chi}$ as a known C^1 function of $\boldsymbol{x}, t$ and $\boldsymbol{n}$ satisfying (4), the ray equations (19) for a wave propagation in two space dimensions, are equivalent to the KCL (17) and (18) for smooth solutions.*

Proof: The proof that the ray equations imply KCL is too simple. Given $\boldsymbol{\chi}$ as a function of x, y, t and θ, and an arbitrarily prescribed Ω_0, we can construct the rays and the family of curves Ω_t. Then we can choose a variable ξ and construct the ray coordinate system (ξ, t). g is given by (15). As long as a caustic or a focus does not appear, the mapping from (x, y) plane to (ξ, t)-plane for a given Ω_t is well defined and one to one. Now we can derive KCL in just few steps ([9] or [10, section 3.3.2]). Alternately, we shall show that it is simple to deduce the differential form of KCL

$$\theta_t = -\frac{1}{g}C_\xi + \frac{1}{g}T\theta_\xi, \quad g_t = C\theta_\xi + T_\xi \tag{20}$$

from the ray equations (19). Using $\chi_1 = n_1C - n_2T$ and $\chi_2 = n_2C + n_1T$ and noting that d/dt becomes $\partial/\partial t$ in (ξ, t)-plane, we find that the third equation in (19) reduces to the first equation in (20). We now differentiate the relation $g^2 = x_\xi^2 + y_\xi^2$ with respect to t and use $n_1 = y_\xi/g$, $n_2 = -x_\xi/g$ and also use $x_t = \chi_1$, $y_t = \chi_2$ to get the second equation in (20).

To show the converse, we note that (17) and (18) imply existence of two functions $x(\xi, t), y(\xi, t)$ such that

$$\begin{pmatrix} x_\xi & x_t \\ y_\xi & y_t \end{pmatrix} = \begin{pmatrix} -g\sin\theta & C\cos\theta - T\sin\theta \\ g\cos\theta & C\sin\theta + T\cos\theta \end{pmatrix} \tag{21}$$

The mapping from (ξ, t) to (x, y)-plane is one to one as long as the Jacobian

$$\frac{\partial(x, y)}{\partial(\xi, t)} = -gC \tag{22}$$

does not vanish. Image of the lines t-constant in (ξ, t)-plane is a curve in (x, y)-plane, let us denote it by Ω_t, along which ξ-varies. The first column of (21) gives $x_\xi = -g\sin\theta,\ \ y_\xi = g\cos\theta$, which show that g is a metric associated with ξ and the normal to Ω_t makes an angle θ with the x-axis. Propagation of the curve Ω_t in (x, y)-plane is governed according to the second column of (21) with a ray velocity $\boldsymbol{\chi} = (\chi_1, \chi_2) := (C\cos\theta - T\sin\theta, C\sin\theta + T\cos\theta)$. This shows that C and T satisfy (16) and so they are the normal and tangential components of the ray velocity $\boldsymbol{\chi}$. Using this relation between (χ_1, χ_2) and (C, T) we get the third equation in (19) from the first equation in (20). Thus, we have derived the ray equations from KCL. However, the quantities C and T appearing in KCL must satisfy the consistency condition (4) through χ_1 and χ_2.

This completes the proof of the theorem. ■

KCL being in conservation form, it also admits solutions with shock type of discontinuities in (ξ, t)-plane. These discontinuities, when mapped onto (x, y)-plane with the help of $x_t = \chi_1, y_t = \chi_2$, they give rise to kinks across which the directions of the tangent to a ray and that to the front Ω_t change discontinuosly, ([10, section 3.3.3]). The KCL are physically realistic, they represent conservation of distance in (x, y)-plane. The concept of kink was first introduced by Whitham in 1957, [10], intuitively as he did not have KCL. He called it *Shock-Shock*.

4 Two ray theories: weakly nonlinear ray theory (WNLRT) and shock ray theory (SRT)

KCL, being only two equations in four quantities g, θ, C and T, is an under determined system. This is expected as KCL is a purely mathematical result and the dynamics of a particular moving curve has not been taken into account in their derivation. We describe here two sets of closure equations. Both of these belong to the case of an isotropic wave propagation, where $T = 0$ i.e., the rays are normal to the front. When a small amplitude curved wave front (across which the physical variables are continuous) or a shock front propagates into a medium at rest and in equilibrium with density $\bar\rho = \bar\rho_0$, fluid velocity $\bar{\boldsymbol{q}} = \mathbf{0}$ and gas pressure $\bar p = \bar p_0$, the perturbation on the wavefront or behind the shock front is given by

$$\bar\rho = \bar\rho_0 + \bar\rho_0 w, \bar{\boldsymbol{q}} = a_0(n_1 w, n_2 w), \bar p = \bar p_0 + \bar\rho_0 a_0^2 w \tag{23}$$

where w is the non-dimensional amplitude of the perturbation and a_0 is the dimensional sound velocity in the ambient medium [10, section 6.1], note that w here is w/a_0 of [10, section 6.1]. The Mach number m of a weakly nonlinear wavefront and M of a shock front are given by

$$m = 1 + \frac{\gamma+1}{2}w, M = 1 + \frac{\gamma+1}{4}w|_s \tag{24}$$

where $w|_s$ the value of w on a suitable side (behind a shock for a shock propagating in to the constant state $(\rho_0, \boldsymbol{q} = \mathbf{0}, p_0)$). The nondimensional value of C in (1.4) is m or M as the case may be.
The evolution equations of Ω_t when it is a weakly nonlinear wavefront, are

$$(g\sin\theta)_t + (m\cos\theta)_\xi = 0, (g\cos\theta)_t - (m\sin\theta)_\xi = 0, \\ \{g(m-1)^2e^2(m-1)\}_t = 0 \tag{25}$$

These are the equations of weakly nonlinear ray theory (WNLRT). The mapping from (ξ, t)-plane to (x, y)-plane is given by the first part of [5] i.e., $x_t = m\cos\theta, y_t = m\sin\theta$.
When we choose Ω_t to be a shock front, the closure equations for KCL form an infinite system of equations ([10], Chapters 7 and 9). This infinite system is exact unlike the third equation in [25] derived under the high frequency approximation. However, not only the derivation of the infinite system is too complex and it is too difficult to solve numerically, its solution is non-unique for many interesting problems . By taking the shock to be weak and by truncating the system at a suitable stage, we can construct an approximate shock ray theory, which forms an efficient system of equations for calculation of successive positions of a curved shock in two space dimensions [10, section 10.2]. Denoting the unit normal to the shock front Ω_t to be $\boldsymbol{N} = (\cos\Theta, \sin\Theta)$, a system of conservation form of the equations for a weak shock Ω_t are two KCL and two additional closure equations, [1]:

$$(G\sin\Theta)_t + (M\cos\Theta)_\xi = 0, (G\cos\Theta)_t - (M\sin\Theta)_\xi = 0 \tag{26}$$

$$(G(M-1)^2e^{2(M-1)})_t + 2M(M-1)^2e^{2(M-1)}GV = 0 \tag{27}$$

$$(GV^2e^{2(M-1)})_t + GV^3(M+1)e^{2(M-1)} = 0 \tag{28}$$

where G is the metric associated with the variable ξ and

$$V = \frac{\gamma+1}{4}\{\langle \boldsymbol{N}, \nabla\rangle w\}|_s \tag{29}$$

where the normal derivative $\langle \boldsymbol{N}, \nabla\rangle w$ is first obtained in the region behind the shock if the shock is moving into the undisturbed region and in the region ahead of the shock if it is moving into the disturbed region and then the limit is taken as we approach the shock. The mapping from (ξ, t)-plane can be obtained by integrating the first part of the shock ray equations $x_t = M\cos\Theta, y_t = M\sin\Theta$. (26)-(28) form the equations of our SRT, which is ideally suited in dealing with many practical problems involving propagation of a curved shock since (i) it has been shown that it gives results which agree well with known exact solutions and experimental results, [4], (ii) it gives sharp geometry of the shock and many finer details of geometrical features of the shock ([10], Chapter 10 or [7]) , (iii) results obtained by it agree well with those obtained by numerical solutions of full Euler's equations, [1] and [6], (iv) it takes considerably less computational time (say less than 10%) compared to the Euler's numerical solution and (v) for a problem like sonic boom, it is difficult to get information in a long narrow region away from the aircraft by Euler's numerical solution, SRT and WNLRT are most suited.

Before we take up the sonic boom problem with some details, we mention two important results [10] obtained by WNLRT and SRT: (i) the genuine nonlinearity in the original system causes a strong diffraction of the rays and does not allow rays from a converging wavefront to form a caustic so that the caustic is resolved and (ii) again the genuine nonlinearity significantly accelerates a non-circular shock to evolve into a circular shock [1].

5 Formulation of the problem of sonic boom by a maneuvering aerofoil as a one parameter family of Cauchy problems

Consider a two dimensional unsteady flow produced by a thin maneuvering aerofoil moving with a supersonic velocity along a curved path. We are interested in calculating the sonic boom produced by the aerofoil, the point of observation being far away say at a distance L, from the aerofoil. We use coordinates x, y and time t nondimensionalized by L and the sound velocity a_0 in the ambient medium. In a local rectangular coordinate system (x', y') with origin O' at the nose of the aerofoil and $O'x'$ axis tangential to the path of the nose, which moves along a curve $(X_0(t), Y_0(t))$, let the upper and lower surfaces of the aerofoil be given by

$$(x' = \zeta, y' = b_u(\zeta)) \text{ and } (x' = \zeta, y' = b_l(\zeta)), -d < \zeta < 0 \tag{30}$$

respectively. Here d is the nondimensional camber length. We assume that $b_u'(-d) > 0$, $b_u'(0) < 0$, $b_l'(-d) < 0$ and $b_l'(0) > 0$, so that the nose and the tail of the aerofoil are not blunt. We further assume that

$$d = \frac{\bar{d}}{L} = O(\epsilon), O\left\{\frac{max_{-d<\zeta<0} b_u(\zeta)}{d}\right\} = O\left\{\frac{max_{-d<\zeta<0}(-b_l(\zeta))}{d}\right\} = O(\epsilon)$$

where ϵ is a small positive number. Then the amplitude w of the perturbation in the sonic boom also satisfies $w = O(\epsilon)$.

In Fig. 1, we show the geometry of the aerofoil and the sonic boom produced by it at a time t. The sonic boom produced either by the upper or lower surface consists of a leading shock LS: $\Omega_t^{(0)}$ and a trailing shock TS: $\Omega_t^{(-d)}$ and since high frequency approximation is satisfied by the flow between the two shocks, a one parameter family of nonlinear wavefronts $\Omega_t^{(\zeta)}(-d < \zeta < 0, \zeta \neq G)$ originating from the points P_ζ on the aerofoil in between the two shocks. The nonlinear wavefronts produced from the points on the front portion of the aerofoil start interacting with the LS $\Omega_t^{(0)}$ and those from the points near the trailing edge do so with the TS $\Omega_t^{(-d)}$, and after the interaction they keep on disappearing continuously from the flow. These two sets, one interacting with LS and another interacting with TS are

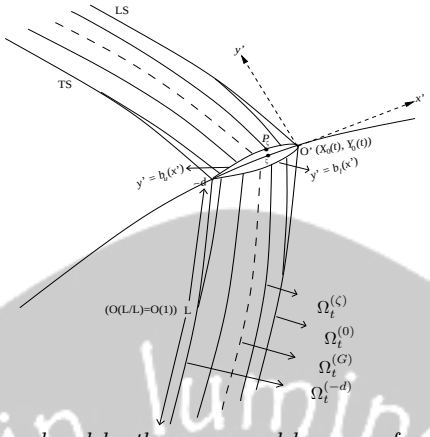


Figure 1: *Sonic boom produced by the upper and lower surfaces:* $y' = b_u(x')$ *and* $y' = b_l(x')$ *respectively. The boom produced by either surface consists of a one parameter family of nonlinear wavefronts*

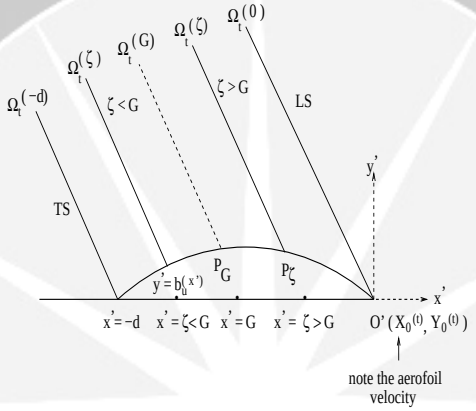


Figure 2: *It An enlarged version of the upper part of the Figure 1 near the aerofoil.*

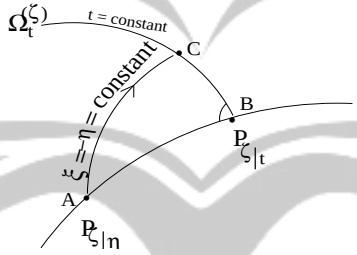


Figure 3: *A formulation of the ray coordinate system* (ξ, t) *for* $\Omega_t(\zeta)$*. AB represents the path of a fixed point* P_ζ *on the aerofoil. A is the position of* P_ζ *at time* η *and B that at time* t*.*

separated by a linear wavefront $\Omega_t^{(G)}$, which originates from a point P_a where the function $b_u(\zeta)$ $(b_l(\zeta))$ are maximum (minimum). Fig. 2 shows an enlarged

version of the upper part of the Fig. 1 near the aerofoil. This is simply an enlarged version of Fig. 1 as high frequency approximation is not valid near the aerofoil.
separated by a linear wavefront $\Omega_t^{(G)}$, which originates from a point P_G where the function $b_u(\zeta)$ ($b_l(\zeta)$) are maximum (minimum). Fig. 2 shows an enlarged version of the upper part of the Fig. 1 near the aerofoil. This is simply an enlarged version of Fig. 1 as high frequency approximation is not valid near the aerofoil.

Let us introduce a ray coordinate system (ξ, t) for $\Omega_t^{(\zeta)}$. The front $\Omega_t^{(\zeta)}$ at a given time t can be obtained as the locus of the tip of the rays (at time t) in (x, y)-plane starting from all positions $P_\zeta|_\eta$ of P_ζ at times $\eta < t$ as shown in Fig. 3. Therefore, a value of η identifies a ray and we choose

$$\xi = -\eta, \eta \leq t \tag{31}$$

for $\Omega_t^{(\zeta)}$ from the upper surface (for lower surface we need to choose $\zeta = \eta, \eta \leq t$). When $\xi = -\eta = t$, the points A, B and C in the Fig.3 coincide. Hence the base point P_ζ of $\Omega_t^{(\zeta)}$, which lies on the upper surface of the aerofoil, corresponds to a point, which lies on the line $\xi + t = 0$ in the (ξ, t)-plane.

The nonlinear wavefront $\Omega_t^{(\zeta)}, (-d < \zeta < 1, \zeta \neq G)$ satisfies the system (1.9)-(1.10). The Cauchy data on $\xi + t = 0$ to solve this system, can be determined from the inviscid flow condition on the surface of the aerofoil. Retaining only the leading order terms, this is [2].

$$m(\xi, -\xi) = m_0(\xi) := 1 - \frac{(\gamma+1)(\dot{X}_0^2 + \dot{Y}_0^2) b_u'(\zeta)}{2(\dot{X}_0^2 + \dot{Y}_0^2 - 1)^{\frac{1}{2}}} \tag{32}$$

$$g(\xi, -\xi) = g_0(\xi) := (\dot{X}_0^2 + \dot{Y}_0^2 - 1)^{\frac{1}{2}} \tag{33}$$

$$\theta(\xi, -\xi) = \theta_0(\xi) := \frac{\pi}{2} + \psi - \sin^{-1}\{1/(\dot{X}_0^2 + \dot{Y}_0^2)^{\frac{1}{2}}\} \tag{34}$$

where $\psi = \tan^{-1}\left\{\dot{Y}_0/\dot{X}_0\right\}$. Since $b_u'(\zeta) < 0$ and $b_u'(\zeta) > 0$ for $G < \zeta < 1$ and $-d < \zeta < G$ respectively, $m_0 > 1$ on P_ζ for $G < \zeta < 1$ and $m_0 < 1$ on P_ζ for $-d < \zeta < G$. This can be used to argue that

$$m > 1 \text{ on } \Omega^{(\zeta)}, G < \zeta < 0 \text{ and } m < 1 \text{ on } \Omega^{(\zeta)}, -d < \zeta < G \tag{35}$$

Since the eigenvalues of the system (1.9)-(1.10) are

$$\lambda_1 = \sqrt{(m-1)/(2g^2)}, \lambda_2 = 0, \lambda_3 = \sqrt{(m-1)/(2g^2)} \tag{36}$$

we get a Cauchy problem for a hyperbolic system for $\Omega_t^{(\zeta)}, G < \zeta < 0$ and an elliptic system for $\Omega_t^{(\zeta)}, -d < \zeta < G$ (we call it elliptic even though $\lambda_2 = 0$ is real).

The derivation of the Cauchy data on $\xi + t = 0$ for the system (1.11)-(1.13) governing the evolution of the shock fronts $\Omega_t^{(-d)}$ and $\Omega_t^{(0)}$ is far more complex.

We quote from [2]and [3], the leading order terms in this Cauchy data

$$M(\xi,-\xi) = M_0(\xi) := 1 - \frac{(\gamma+1)(\dot{X}_0^2+\dot{Y}_0^2)b_u'(\xi)}{4(\dot{X}_0^2+\dot{Y}_0^2-1)^{\frac{1}{2}}} \tag{37}$$

$$G(\xi,-\xi) = G_0(\xi) := (\dot{X}_0^2+\dot{Y}_0^2-1)^{\frac{1}{2}} \tag{38}$$

$$\Theta(\xi,-\xi) = \Theta_0(\xi) := \frac{\pi}{2} + \psi - \sin^{-1}\{1/(\dot{X}_0^2+\dot{Y}_0^2)^{\frac{1}{2}}\} \tag{39}$$

$$V(\xi,-\xi) = V_0(\xi) := \frac{\gamma+1}{4}\{\Omega_{p_{(-d)}} w_0(\xi) - \mathcal{F}(-d,t)\} \tag{40}$$

where

$$\begin{aligned}
\Omega_{P_{(-d)}} &= \frac{(\dot{X}_0\ddot{X}_0+\dot{Y}_0\ddot{Y}_0)}{2g(\dot{X}_0^2+\dot{Y}_0^2)(\dot{X}_0^2+\dot{Y}_0^2-1)^{1/2}} + \frac{\dot{X}_0\ddot{Y}_0-\dot{Y}_0\ddot{X}_0}{g\ddot{X}_0^2} \\
\mathcal{F}(\zeta,t) &= \frac{(\dot{X}_0^2+\dot{Y}_0^2)b_u''(\zeta)}{(\dot{X}_0^2+\dot{Y}_0^2-1)^{1/2}}\{\dot{\mathcal{X}}_0(t)\} - \frac{(\dot{X}_0^2+\dot{Y}_0^2-2)(\dot{X}_0\ddot{X}_0+\dot{Y}\ddot{Y}_0)}{(\dot{X}_0^2+\dot{Y}_0^2-1)^{3/2}} b_u'(\zeta) \\
\mathcal{X} &= X_0\cos\psi + Y_0\sin\psi
\end{aligned}$$

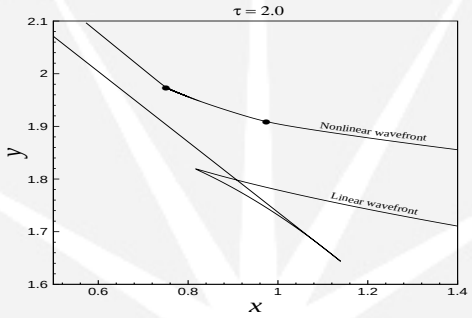


Figure 4: *Sonic boom wavefront at $t = 2$ from the leading edge of an accelerating aerofoil moving in a straight path. Kinks on the nonlinear wavefront are shown by dots. The initial Mach number is 1.8 and the acceleration is 10 in the line interval (1, 1/2).*

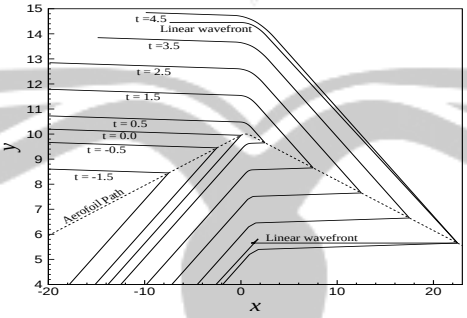


Figure 5: *The nonlinear wavefront from the leading edge of an aerofoil moving with a constant Mach number 5 along a path concave downwards with $b_u'(0) = -0.01$.*

Again $M > 1$ on $\Omega_t^{(0)}$ and $M < 1$ on $\Omega_t^{(-d)}$ so that the two eigenvalues $\Lambda_1 = \sqrt{(M-1)/2G^2}, \Lambda_2 = -\sqrt{(M-1)/2G^2}$ of (26)-(28) are real for $\Omega_t^{(0)}$ and purely imaginary for $\Omega_t^{(-d)}$. The other two eigenvalues are $\Lambda_{11} = 0, \Lambda_{12} = 0$. Thus, for the LS we get a Cauchy problem for a system which is hyperbolic and for the TS we get it for a system which has elliptic nature.

We have numerically solved the system (25) with Cauchy data (32)-(34) for $\zeta = 0$. This nonlinear wavefront from the leading edge is immediately annihilated by the shock $\Omega_t^{(0)}$. We have solved the system (26)-(28) with data (37)-(40) and we find that the geometric shape of nonlinear wavefront is not only topologically same as that of $\Omega_t^{(0)}$ but is very close to it. Hence the nonlinear wavefront from the leading edge gives valuable information about $\Omega_t^{(0)}$. We present some results in Fig 4 and Fig 5.

We note that for an accelerating aerofoil along a straight line, the linear wavefront from the nose develops fold in the caustic region but the nonlinear wavefront does not fold and has a pair of kinks. For a supersonic aerofoil moving on a highly curved path (curved downwards), the nonlinear wavefront from the upper surface is smooth but that from the lower surface has a pair of kinks. The most interesting result seen from our new formulation of the sonic boom problem is the elliptic nature of

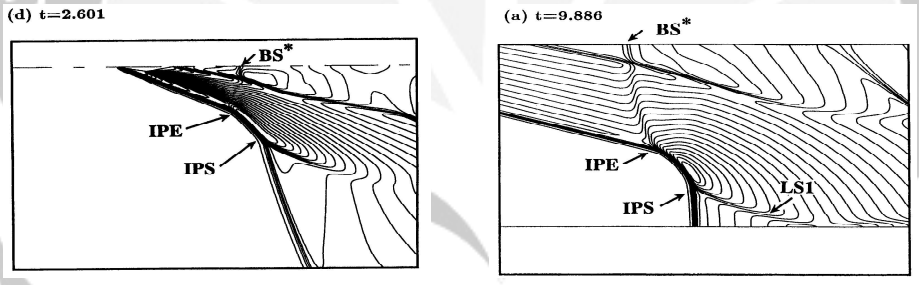


Figure 6: *Numerical solutions of Euler equations for the flow over a diamond-shaped projectile moving from right to left with Mach number sarting from 1.2 to 4. The Figures are reproduced from Inoue, Sakai and Nishida (1997) with the permission of the authors.*

the equations governing $\Omega_t^{(-d)}$. This implies that whatever may be the flight path and acceleration of the aerofoil, the trailing shock $\Omega_t^{(-d)}$ must be free from kinks. All these features, which we obtain from our theory are seen in the Euler's numerical solution in [5].

Acknowledgement: We thank AR&DB, Ministry of Defence, Govt. of India for financial support through the project "Nonlinear Hyperbolic Waves in Multi-Dimensions with Special Reference to Sonic Booms"(No. DARO/08/1031199/M/I).

References

[1] Baskar, S., and Phoolan Prasad (2005), Propagation of curved shock fronts using shock ray theory and comparison with other theories, *J. Fluid Mech.*, **523**, 171-198.

[2] Baskar, S., and Phoolan Prasad (2005), *Formulation of the sonic boom problem by a manuevering aerofoil as a one parameter family of Cauchy problems*, Preprint / 2005, Department of Mathemtatics, IISc, Bangalore.

[3] Baskar, S., and Phoolan Prasad (2005), KCL, ray theories and applications to sonic boom, to appear in *Proceedings of 10th International Conference on Hyperbolic Problems,* Osaka, Japan, September 13-17, 2004.

[4] Courant, R., and D. Hilbert (1962), *Methods of mathematical physics, vol II: partial differential equations*, John Wiley & Sons, New York.

[5] Inoue, O., T. Sakai and M. Nishida (1997), *Focusing shock wave generated by an accelerating projectile*, Fluid Dynamics Research, **21**, 403 - 416.

[6] Kevlahan, N. K. R., (1996), *The propagation of weak shocks in non-uniform flow*, *J. Fluid Mech.* **327**, 167-197.

[7] Morton, K.W., Phoolan Prasad and Renuka Ravindran (1992), Conservation forms of the nonlinear ray equations, Tech. Report No.2, Dept. Math, Indian Institute of Science, Bangalore.

[8] Monica, A., and Phoolan Prasad (2001), *Propagation of a curved weak shock*, J. Fluid Mech., **434**, pp. 119 - 151.

[9] Prasad, Phoolan (1995), Formation and propagation of singularities on a nonlinear wavefront and shockfront. J.Indian Institute of Science, Bangalore (Special volume of Fluid Mechanics), 75, 517-535.

[10] Prasad, Phoolan (2001), *Nonlinear hyperbolic waves in multi-dimensions*, Chapman and Hall/CRC, Monographs and Surveys in Pure and Applied Mathematics - **121**.

[11] Whitham, G., (1974), *Linear and nonlinear waves*, John Wiley & Sons, New York.

Dynamical Systems Method for Solving Operator Equations

Alexander G. Ramm

Mathematics Department, Kansas State University, USA

Abstract: Consider an operator equation $F(u)=0$ in a Hilbert space **H** and assume that this equation is solvable. Let us call the problem of solving this equation ill-posed if the operator $F'(u)$ is not boundedly invertible, and well-posed otherwise. A general method, Dynamical Systems Method (DSM), for solving linear and nonlinear ill-posed problems in **H** is presented. This method consists of the construction of a dynamical system, that is, a Cauchy problem, which has the following properties:

1) it has a global solution,
2) this solution tends to a limit as time tends to infinity,
3) the limit solves the original linear or non-linear problem.

The DSM is justified for

a) an arbitrary linear solvable equations with bounded operator,
b) for well-posed nonlinear equations with twice *Fréchet* differentiable operator F,
c) for ill-posed nonlinear equations with monotone operators,
d) for ill-posed nonlinear equations with non-monotone operators from a wide class of operators,
e) for operators such that $A := F'(u)$ satisfies the spectral assumption:
$\left\|(A+sI)^{-1}\right\| \le c/s$, where $c>0$ is a constant, and $s \in (0, s_0)$, $s_0 > 0$ is a fixed number, arbitrarily small, c does not depend on s and u,
f) for some monotone operators which are not *Fréchet* differentiable, and
g) for some unbounded, closed, densely defined F.

In Newton-type schemes the main difficulty is to invert the derivative of the operator. A novel scheme, based on the DSM, allows one to avoid this inversion.

A global convergence theorem is obtained for the regularized continuous analog of Newton's method for monotone operators. Global convergence means that convergence is established for an arbitrary initial approximation, not necessarily the one which is sufficiently close to the solution.

A general approach to constructing convergent iterative schemes for solving well-posed nonlinear operator equations is described and convergence theorems are obtained for such schemes. Stopping rules for stable solution of ill-posed problems with noisy data are given.

References:

[1] A.G.Ramm, Global convergence for ill-posed equations with monotone operators: the dynamical systems method, J. Phys A, 36, (2003), L249-L254.

ON GENERATION OF UNIDIRECTIONAL SINGLE STEEP WAVES IN TANKS

L. Shemer, K. Goulitski, E. Kit

Tel-Aviv University, Israel

Abstract. Very steep waves constitute an essentially nonlinear and complicated phenomenon. Inter-related experimental and theoretical efforts are thus required to gain a better understanding of their generation and propagation mechanisms. A nonlinear focusing process in which a single unidirectional steep wave emerges from an initially wide amplitude- and frequency-modulated wave group at a predicted position in the laboratory wave tank is studied both theoretically and experimentally. The spatial version of the Zakharov equation was applied in the numerical simulations. Experiments were carried out in the 330 m long Large Wave Channel in Hanover and in the 18 m long Tel-Aviv University wave tank. Quantitative comparison between the experimental and the corresponding numerical results is carried out. Good agreement is obtained between experiments and computations.
Key-words: Nonlinear water waves, rogue waves, freak waves, Zakharov equation, spatial evolution, bound (locked) waves

1 Introduction

Generation of very steep waves in wave tanks enables experimental study of the wave damage potential and is thus of great importance. Excitation of a single steep wave at a prescribed location in a laboratory wave tank of constant depth is also often required for model testing in coastal and ocean engineering. It is well known that such waves can be generated by focusing a large number of waves at a given location and instant. Dispersive properties of deep or intermediate-depth surface gravity waves can be utilized for this purpose. Since longer gravity waves propagate faster, a wave group generated at the wave maker in which wave length increases from front to tail may be designed to focus the wave energy at a desired location. Such a wave sequence can be seen as a group that is modulated both in amplitude and in frequency. One-dimensional theory describing spatial and temporal focusing of various harmonics of dispersive gravity waves based on the linear Schrödinger equation was presented by Pelinovsky & Kharif (2000). They suggested such a focusing as a possible mechanism for generation of extremely steep singular waves. However, the experiments of Brown & Jensen (2001) demonstrated that nonlinear effects are essential in the evolution of those waves. An extensive review of field observations of those waves, as well as of the relevant theoretical, numerical and experimental studies was recently presented by Kharif and Pelinovsky (2003)

The essentially nonlinear behavior of wave groups with high maximum wave steepness has been demonstrated in a number of studies. Attempts were made to describe the propagation of deep or intermediate depth gravity water-wave groups with a relatively narrow initial spectrum by a cubic Schrödinger equation (CSE). Shemer et al. (1998) demonstrated that while CSE is adequate for description of the global properties of the group envelope evolution, it is incapable to capture more subtle features such as the emerging front-tail asymmetry observed in

experiments. For the weakly-dispersive wave groups in shallow water, application of the Korteweg – deVries equation provided results that were in very good agreement with the experiments (Kit et al. 2000). In the case of stronger dispersion in deeper water, models that are more advanced than the CSE are required, since due to nonlinear interactions, considerable widening of the initially narrow spectrum can occur. The modified Schrödinger equation (Dysthe 1979) is a higher (4^{th}) order extension of the CSE. Application of this model indeed provided good agreement with experiments on narrow-band wave groups (Shemer et al. 2002). An alternative theoretical model that is free of band-width constraints is the Zakharov (1968) equation. Unidirectional spatial version of this equation was derived in Shemer et al. (2001) and applied successfully to describe the evolution of nonlinear wave groups in the tank. Kit & Shemer (2002) showed the relation between the spatial versions of the Dysthe and the Zakharov equations.

An attempt to check the limits of applicability of the Dysthe equation to describe evolution of wave groups with wider spectrum has been carried out by Shemer et al. (2002). Numerical solutions of the wave group evolution problem were carried out using both Dysthe and Zakharov equations. The obtained results demonstrated that while the Dysthe model performed in a satisfactory fashion for not too wide spectra, it failed for wave groups with initially very wide spectra.

The focusing is more effective when the number of free wave harmonics generated at the wavemaker is large. Excitation of single wave with extreme amplitude thus requires wide spectrum of the initial wave group generated at the wavemaker. Extremely steep (freak) wave therefore can be seen as wave groups with very narrow envelope and correspondingly wide spectrum. In the current study we perform an experimental investigation of propagation of steep wave groups with wide spectrum in two wave tanks that differ in size by an order of magnitude, i.e. in the 18 m long Tel-Aviv University (TAU) wave tank, and in the 330 m long Large Wave Channel (GWK) in Hanover, Germany. The experiments are accompanied by numerical simulations based on modification of the spatial version of the Zakharov equation. Some preliminary results of this study were presented in Goulitski et al. (2004) for measurements carried out in the TAU wave tank, and in Shemer et al. (2005) for the experiments performed in the Hanover experimental facility.

2 Theoretical background

The purpose of the present study is to obtain at a prescribed distance from the wavemaker, $x = x_0$, a steep unidirectional wave group with a narrow, Gaussian-shaped envelope with the surface elevation variation in time, $\zeta(t)$, given by

$$\zeta(t) = \zeta_0 \exp(-t/mT_0)^2 \cos(\omega_0 t) \tag{1}$$

where $\omega_0=2\pi/T_0$ is the carrier wave frequency, ζ_0 is the maximum wave amplitude in the group, and the parameter m defines the width of the group. The small parameter representing the magnitude of nonlinearity ε is the maximum wave steepness $\varepsilon = \zeta_0 k_0$. The wave number k is related to the frequency ω by the finite depth dispersion relation

$$\omega^2 = kg \tanh(kh), \tag{2}$$

g being the acceleration due to gravity. The parameter m determines the width of the group; higher values of m correspond to wider groups and narrower spectra. The spectrum of the surface elevation given by (1) is also Gaussian.

The wave field at earlier locations, $x < x_0$ is obtained from the computed complex surface elevation frequency spectrum at this location. To this end, the unidirectional discretized spatial version of the Zakharov equation derived by Shemer et al. (2001) can be used:

$$i\frac{dB_j(x)}{dx} = \sum_{\omega_j+\omega_l=\omega_m+\omega_n} \alpha_{j,l,m,n} B_l^* B_m B_n e^{-i(k_j+k_l-k_m-k_n)x} \tag{3}$$

where * denotes complex conjugate and the interaction coefficient $\alpha_{j,l,m,n}$ is given by

$$\alpha_{j,l,m,n} = V(k(\omega_j), k(\omega_l), k(\omega_m), k(\omega_n))/c_{g,j} \tag{4}$$

In (4), the values of V represent the quartet interaction coefficient in the temporal Zakharov equation as given by Krasitskii (1994), and $c_{g,j}$ is the group velocity of the j-th spectral component. Equations (3) and (4) accurately describe the slow evolution along the tank of each free spectral component $B_j = B(\omega_j)$ of the surface elevation spectrum in inviscid fluid of constant (infinite or finite) depth, as long as the quartet nonlinear interactions considered occur among components that are relatively close. When the spectrum considered is wide, this limitation can be removed by modifying (4) for the interaction coefficient. The modified expression is

$$\alpha_{j,l,m,n} = V(\kappa, k(\omega_l), k(\omega_m), k(\omega_n))\frac{k(\omega_j)-\kappa}{\chi-\omega(\kappa)}; \tag{4a}$$

$$where \quad \kappa = k(\omega_m)+k(\omega_n)-k(\omega_l), \quad \chi = \omega_m+\omega_n-\omega_l$$

The dependent variables $B(\omega_j, x)$ in (3) are related to the generalized complex 'amplitudes' $b(\omega_j, x)$ composed of the Fourier transforms of the surface elevation $\hat{\zeta}(\omega_j, x)$ and of the velocity potential at the free surface $\hat{\phi}^s(\omega_j, x)$:

$$b(\omega,x) = \left(\frac{g}{2\omega}\right)^{1/2}\hat{\zeta}(\omega,x) + i\left(\frac{\omega}{2g}\right)^{1/2}\hat{\phi}^s(\omega,x) \tag{5}$$

The "amplitudes" b consist of a sum of free and the bound waves:

$$b(\omega_j, x) = [\varepsilon B(\omega_j, x_2) + \varepsilon^2 B'(\omega_j, x, x_2) + \varepsilon^3 B''(\omega_j, x, x_2)]\exp(ikx) \tag{6}$$

The higher order bound components B' and B'' can be computed at each location once the free wave solution $B_j(x)$ is known. The phase velocity of these components depends of the parent free waves and can not be determined using (2). The corresponding formulae, as well as the kernels necessary for their computations

are given in Krasitskii (1994) and in Stiassnie and Shemer (1984, 1987). In (6), the scaled slow coordinate $x_2 = \varepsilon^2 x$. Inversion of (5) allows computing the Fourier components of the surface elevation $\tilde{\zeta}(\omega, x)$. Inverse Fourier transform then yields the temporal variation for the surface elevation $\zeta(x,t)$.

In this paper, the spatial Zakharov equation (3) is used with the modified interaction coefficient (4a). The spectrum corresponding to (1) is integrated from the planned focusing location x_0 backwards up to the wavemaker at x = 0. The waveforms derived from the computed spectra serve as a basis for computations of the wavemaker driving signals that take into the account the theoretical wavemaker transfer function for a given wavemaker shape (piston in Hanover and paddle in TAU) with corrections that account for the actual wavemaker response.

3 Experimental facilities and procedure

The TAU wave tank is 18m long, 1.2m wide and has the water depth of 0.6m. A paddle-type wavemaker hinged near the floor is located at one end of the tank. The instantaneous surface elevation is measured simultaneously by four resistance-type wave gauges made of blackened platinum for better sensitivity. The probes are mounted on a bar parallel to the side walls of the tank and fixed to a carriage which can be moved along the tank. Focusing location in different experiments varied from 5 m to 10 m from the wavemaker. The Hanover tank has a length of 330 m, width of 5 m and depth of 7 m. Water depth in the present experiments was set to be 5 m. At the end of the wave tank there is a sand beach starting at the distance of 270 m with slope of 30°. The piston-type wavemaker is equipped with the reflected wave energy absorption system. The focusing location in all Hanover experiments was set at 120 m from the wavemaker. The instantaneous water height is measured using 25 wave gauges of resistance type, which are placed along the tank wall; higher concentration of the wave gauges is in the region of expected focusing of the wave group.

The Gaussian energy spectrum of (1) has a shape with the relative width at the energy level of ½ of the spectrum maximum that depends on the value of the parameter m in (1) and is given by

$$\Delta\omega / \omega_0 = 1 / m\pi \sqrt{1/2 \, ln\, 2} \qquad (7)$$

The value of the group width parameter in all experiments was selected to be m=0.6, so that (7) yields the relative spectrum width $\Delta\omega/\omega_0$=0.312, which is beyond the domain of applicability of the narrow spectrum assumption of the cubic Schrödinger and Dysthe models.

In the Hanover experiments, the carrier wave period adopted in (1) is T_0 = 2.8 s, corresponding to the wavenumber k_0= 0.52 m^{-1}, so that k_0h = 2.59 and thus deep-water dispersion relation is only approximately satisfied. Therefore, in all expressions for the interaction coefficients finite depth versions were used. The focusing location in Hanover experiments is located at the distance of about 10 carrier wave lengths from the wavemaker.

TAU experiments were carried out with two carried wave periods, T_0 = 0.85 s, k_0 = 0.056 cm^{-1}; k_0h = 3.35, corresponding to the intermediate depth conditions, and T_0 = 0.60 s, k_0 = 0.112 cm^{-1}; k_0h = 6.71, with deep water conditions satisfied even for the low harmonics in the spectrum. The driving amplitudes in the cases considered here are selected so that at the focusing location, the resulting carrier wave has the maximum wave amplitudes ζ_0 corresponding to the steepness $\varepsilon = k_0\zeta_0 = 0.3$.

For each set of the carrier wave period T_0, and the focusing location x_0, the solution of the system of N ODEs (3), N being the total number of wave harmonics considered, was obtained for distances from the wavemaker up to x_0 and beyond. The number of free wave harmonics considered is N = 120. The wavemaker[-driving signal was adjusted to get as good as possible agreement between the calculated and the measured wave field at a location close to the wavemaker, but beyond the range of existence of evanescent modes (see, e.g. Dean and Dalrymple 1991).

4 Results

A representative selection of the accumulated in this study results is discussed in this Section. Results obtained in Hanover are shown first. The computed and the measured temporal variations of the surface elevation at different locations along the tank are presented in Figs. 1a and 1b, respectively. The selected value of m =0.6 in (1) yields a narrow wave group with a single steep wave at the focusing location. Closer to the wavemaker the group becomes notably wider, and the maximum wave amplitudes decrease accordingly. Modulation of the amplitude and the frequency within the group is clearly seen. The experimental results presented in Fig. 1b demonstrate good agreement with the computations, although the wave shape measured at the focusing location is not exactly symmetric.

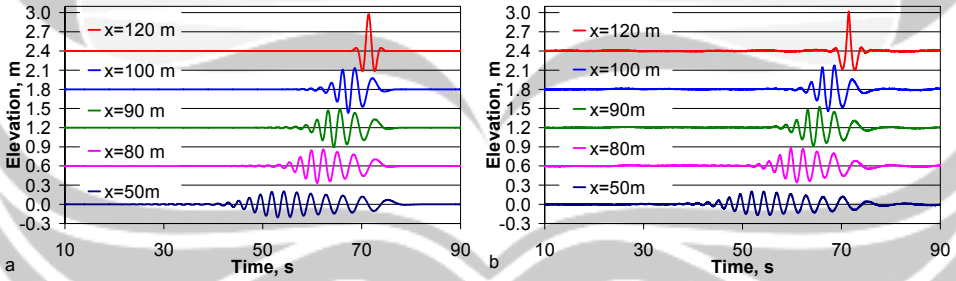


Figure 1. Calculated and measured in the Hanover wave tank surface elevation within the group at different distances x from the wavemaker (T_0=2.8s, ε=0.3).

The computed and the measured spectra for the experimental parameters of Fig. 1 are presented in Fig. 2 at various locations along the tank. The variation of the spectral shape along the tank is evident and indicates that wave evolution is essentially nonlinear even at this relatively low amplitude of forcing. The agreement between experiments and computations is quite satisfactory and both the numerical simulations and the measurements exhibit similar features. The spectral shapes shown in Fig. 2b indicate that the spectrum becomes wider with the distance from the wavemaker and at the prescribed distance (x_0 = 120 m) approaches the Gaussian shape assumed in the numerical simulations. The peak

frequency at x = 50 m is shifted to the right relative to the carrier frequency $f_0=1/T_0$. Note that the peak values within the group appear to be somewhat different in those figures. The low frequency part of the spectrum remains unaffected during the evolution process. It should be stressed that the computed surface elevation is obtained here by taking into account free modes only, while in the experiments the effect of the bound waves can be significant.

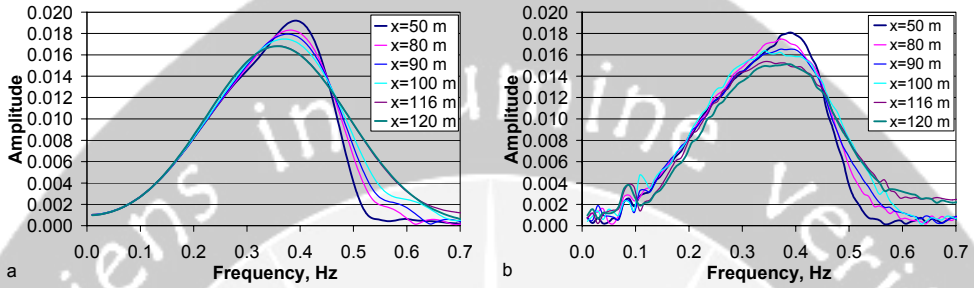


Figure 2. Frequency spectra of the surface elevation variation with time along the Hanover tank for x_0=120 m and $\varsigma_0 k_0$=0.3: a) computed; b) measured

Careful analysis of the extensive data sets accumulated in Hanover and TAU experiments clearly indicate that in addition to accounting for the contribution of the bound waves to the generalized "amplitudes" b, see (6), the effect of viscous dissipation has to be considered. Since the dissipation in the boundary layers at the tank walls and bottom is relatively weak, it is sufficient to account for the wave energy loss along the tank by adding a linear term, $-i\gamma_j B_j$, to the r.h.s. of (3). The dissipation coefficient γ_j is calculated following Kit and Shemer (1989).

The substantially smaller dimensions of the TAU tank as compared to the Hanover facility make it possible to perform numerous experiments and to attain a better agreement of the computed and the actually obtained waveforms near the wavemaker. The detailed comparison of the theoretical predictions and experiments carried out in sequel is based therefore on the TAU-derived results.

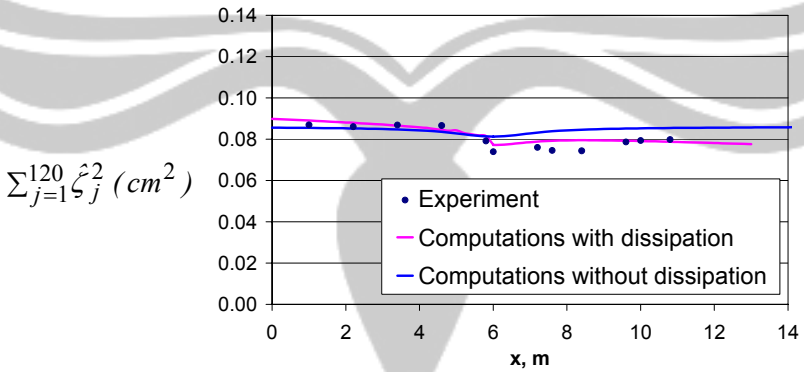


Figure 3. Variation of the total energy of free spectral modes along the TAU tank. T_0 = 0.6 s, $\varsigma_0 k_0$=0.3 and x_0 = 6 m.

Experiments in the TAU tank for the carrier wave period of 0.6 s (carrier wave length λ_0 = 0.56 cm) were designed for the focusing distance from the wavemaker x_0 = 6 m, i.e. about 10 carrier wave lengths, similar to conditions in Hanover. The results of Fig. 3 clearly show that the non-linear contribution to the total wave field energy is essential mainly in the vicinity of the focusing location. As a result of dissipation along the tank, the amplitude of the waves generated by the wavemaker should be somewhat higher than that computed for a purely Hamiltonian case.

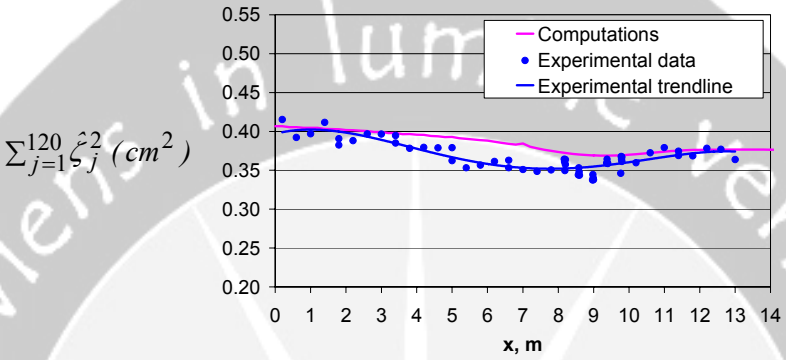


Figure 4. As in Fig. 3, for T_0= 0.85 s, $\varsigma_0 k_0$=0.3 and x_0 = 9 m.

Further experiments in TAU were carried out for a longer carrier wave with the period T_0=0.85 s and length λ_0 = 1.12 m. In this case, focusing occurred at x_0 =9 m, about 8 carrier wave lengths from the wavemaker. Notable decay of wave energy along the tank is visible. Away from focusing, the sum of squared amplitudes of all free waves adequately represents the total wave field energy, and excellent agreement between measurements and computations indicates that dissipation is properly accounted for. Around the focusing locations contribution of energy contained in bound waves is essential.

The effect of bound waves on both surface elevation and frequency spectrum is further investigated in Fig. 5. Since the effect of bound waves is mostly visible for very steep waves, those waves are computed here at the focusing location. In Fig. 5a, the experimentally measured temporal variation of the surface elevation is compared with computations performed both with and without contribution of the 2nd order bound waves, denoted by *B*' in (6). As expected, bound waves contribute to steeper crest and flatter trough of the wave, and result in a better agreement with the measured wave shape.

Comparison of the corresponding amplitude spectra in Fig. 5b demonstrates that when the contribution of bound waves is accounted for, the agreement of theoretical predictions with the experiments is improved drastically, in particular in the high frequency region. In this frequency domain, bound waves can be seen as the 2nd harmonic of the dominant free waves. Certain improvement of the agreement between experiment and computations is also obtained for lower frequencies. The remaining discrepancies between experiments and computations can be attributed to difficulties in exact reproduction of the computed wave forms by the wavemaker.

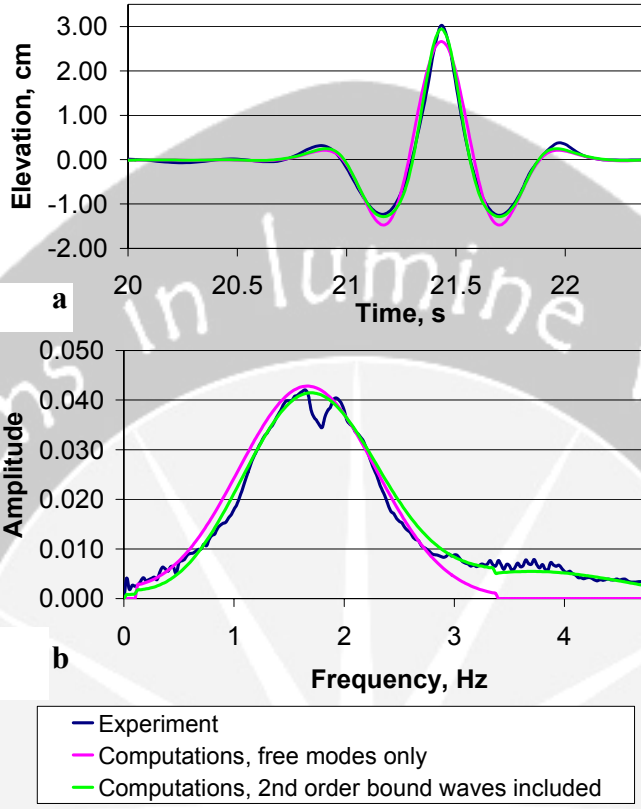


Figure 5. The effect of 2nd order bound waves, wave parameters as in Fig. 3.
a) Computed and measured surface elevation at the focusing location;
b) the corresponding amplitude spectra.

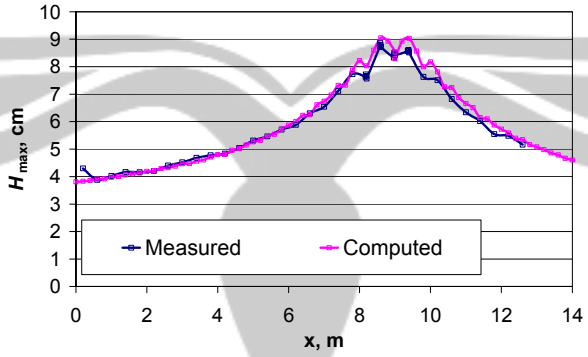


Figure 6. Maximum wave height variation along the tank, wave parameters as in Fig. 4.

Wave height H is defined as the difference between the consecutive minimum and maximum surface elevation. The evolution of the maximum wave height, H_{max}, within the group along the tank for $\varsigma_0 k_0$ = 0.3 is shown in Fig. 6. Very good agreement is observed, and the measured and computed rates of increase of the maximum wave height during the focusing process and the following decrease in the maximum wave height during defocusing for $x > x_0$ are practically identical.

5 Conclusions

The ability to excite focused steep waves at any desired location along the tank is demonstrated in two very different experimental facilities. Large number of wave harmonics is required to generate very steep wave at the focusing location. It is shown that the focusing process is accompanied by a notable change of the spectral shape and is thus essentially nonlinear. The modified unidirectional spatial discrete version of the Zakharov equation as given by (3) and (4a) is adequate to describe nonlinear evolution of steep wave groups with wide spectrum propagating in water of constant intermediate depth. To achieve not only qualitative but also quantitative agreement between the model predictions and the experiments, it is insufficient, however, to consider the nonlinear evolution process of the free wave components only. At least two additional effects have to be accounted for. First, dissipation along the tank is essential and can be adequately described by an additional linear term in (3) that represents the decay of amplitude of each spectral mode as a result of viscous boundary layers at the bottom and side walls of the tank. Secondly, effects related to the bound waves can not be neglected. These effects strongly depend of the wave steepness and become important mainly in the vicinity of focusing. Second-order bound are accounted for in the present study, and the appropriate corrections are introduced. The effect of the 3rd order bound waves will be investigated in future. With both dissipation and 2nd order bound waves accounted for, very good agreement between experiments and numerical simulations is achieved.

Acknowledgement

The authors acknowledge the European Community support under the Access to Research Infrastructures Action of the Human Potential Programme (contract HPRI-CT-2001-00157) that made possible experiments in the Large Wave Channel (GWK) of the Coastal Research Center (FZK) in Hanover.

References

[1] Brown, M.G. and Jensen, A. (2001), Experiments on focusing unidirectional water waves, *J. Geophys. Res.,* **106**, 16,917-16,928.

[2] Dean, R.G. and Dalrymple, R.A. (1991), *Water wave mechanics for engineers and scientists*, World Scientific, Singapore.

[3] Dysthe, K.B. (1979), Note on a modification to the nonlinear Schrödinger equation for application to deep water waves, *Proc. Roy. Soc. London,* **A369**, 105-114.

[4] Goulitski, K, Shemer, L. and Kit, E. (2004), Steep unidirectional waves: experiments and modeling, *Izvestiya VUZ. Applied Nonlinear Dynamics* **12,** 122-131.

[5] Kit, E., Shemer, L (1989), On dissipation coefficients in a wave tank. *Acta Mechanica* **77**, 171 - 180.

[6] Kit, E., Shemer, L., Pelinovsky, E., Talipova, T., Eitan, O. and Jiao, H.-Y. (2000), Nonlinear wave group evolution in shallow water, *J. Waterway, Port, Coastal & Ocean Eng.*,**126**, 221-228.

[7] Kit, E. and Shemer, L. (2002), Spatial versions of the Zakharov and Dysthe evolution equations for deep water gravity waves, *J. Fluid Mech.,* **450**, 201-205.

[8] Kharif, C. and Pelinovsky, E. (2003), Physical mechanisms of the rogue wave phenomenon, *Europ. J. Mech. B/Fluids,* **22**, 603-634.

[9] Krasitskii, V.P. (1994), On the reduced equations in the Hamiltonian theory of weakly nonlinear surface waves, *J. Fluid Mech.,* **272**, 1-20.

[10] Pelinovsky, E. and Kharif, C. (2000), Simplified model of the freak wave formation from the random wave field, in *Proc. 15th Int. Workshop on Water Waves and Floating Bodies*, Editors. T. Miloh, G. Zilman, Caesaria, 142-145.

[11] Shemer, L., Kit, E., Jiao, H.-Y., and Eitan, O. (1998), Experiments on nonlinear wave groups in intermediate water depth, *J. Waterway, Port, Coastal & Ocean Eng.,* **124**, 320-327.

[12] Shemer, L., Jiao, H.-Y., Kit, E., and Agnon, Y. (2001), Evolution of a nonlinear wave field along a tank: experiments and numerical simulations based on the spatial Zakharov equation, *J. Fluid Mech.,* **427**, 107-129.

[13] Shemer, L., Kit, E., and Jiao, H.-Y. (2002), An experimental and numerical study of the spatial evolution of unidirectional nonlinear water-wave groups, *Phys. Fluids,* **14,** 3380-3390.

[14] L. Shemer, K. Goulitski, E. Kit, J. Gruene, R. Schmidt-Kopenhagen (2005), On generation of single steep waves in tanks, WAVES 2005 - Madrid, Spain.

[15] Stiassnie, M and Shemer, L (1984), On modification of Zakharov equation for surface gravity waves, *J. Fluid Mech*., **143**, 47 - 67.

[16] Stiassnie. M. and Shemer, L. (1987), Energy computations for coupled evolution of Class I and Class II instabilities of Stokes waves, *J. Fluid Mech*., **174**, 299 - 312.

[17] Zakharov, V.E. (1968), Stability of periodic waves of finite amplitude on the surface of deep fluid, *J. Appl. Mech. Tech. Phys. (English transl.),* **2,** 190-194.

Optimization Modeling Technology: Past, Present and Future

J. Bisschop

University of Twente, The Netherlands

Abstract: During the last fifty years the field of mathematical programming has evolved into a mature discipline of mathematics. Starting with the invention of the simplex method for linear programming in the late forties, a wealth of theory and algorithms has been developed since then. At the same time, a large number of planning and scheduling applications were developed for a variety of industries and government agencies that have led to improved decision making and better use of resources. Optimization technology in the form of advanced computer implementations of solution algorithms and modeling systems has played a major role in bridging the gap between theory and applications.

The presentation will give an overview of how optimization technology has evolved over time, and will sketch some of the accomplishments that have been reached using this technology. The major portion of the presentation will emphasize likely future developments. In particular, it will motivate the extension of modeling systems to provide new algorithmic capabilities in addition to the current modeling capabilities. One illustration will be the description of a decomposition algorithm that solves a job shop scheduling problem with the use of both mathematical programming and constraint logic programming. A second illustration will be the description of the new open framework in the modeling system AIMMS to solve mixed-integer nonlinear programming models using variants of the outer approximation method. The talk will conclude with a brief software demonstration.

Embedding Graphs into Super Edge Magic Graphs

F.A. Muntaner-Batle

Universidad Internacional de Cataluña, Spain

Abstract: The area of graph labeling has experienced a great development during the last three decades, and many applications of this area have been found and studied in other branches of science. For instance we can find graph labelings showing up in coding theory, X-ray crystallography, radar, astronomy, circuit design, communication network addressing and data base management. Also a close relationship exists between graph decompositions and graph labelings. Due to this close relationship, many problems involving labelling and trees have shown up and have proven to be very hard. In this talk, we will discuss some classical applications involving graph labelling and we will study how close is a tree to be super edge magic by finding for any given tree T, a small super edge magic tree T', that contains T as a subgraph.

Mathematical Models for Hydrodynamic Laboratories: How to Make Them Fit

R.H.M. Huijsmans

Maritime Research Institute Netherlands (MARIN)

Abstract: The purpose of hydrodynamic research is to support the hydrodynamics aspects of the design and safe operation of economical offshore structures. For deep water developments an integrated application of model tests, numerical simulations and full scale measurements is necessary to achieve this. The development of reliable numerical simulation tools requires research in the field of viscosity, Vortex Induced
Vibrations (VIV), wave drift forces in survival conditions, non-linear relative wave motions, hydrodynamic interaction of structures and the performance of DP systems. Reliable model testing for deep water requires the controlled modelling of wind, waves and current in time and space to achieve a realistic and well defined offshore environment in the model basin. Therefore adequate knowledge is required of the propagation of waves on current in a hydrodynamic laboratory is of prime importance. Nowadays a large part of model test of floating structures in waves is devoted to validation experiments of CFD codes. Essential requirement for these tests is that spurious tank effects should be identified and taken into account in the CFD analysis. In this paper a validation study is presented for an LNG offloading operation in shallow water waves in the Offshore Basin of MARIN.

Complex Semidefinite Programming and Applications

Shuzhong Zhang

Chinese University of Hong Kong

Abstract: Semidefinite Programming (SDP) has played a similar role in the past decade as Linear Programming (LP) did a half century ago. The excitement was caused mainly due to two reasons: (1) SDP comes with a beautiful theory, not as an oddity; (2) SDP has an immense modeling power to solve practical problems. In this talk, we shall present our recent results on complex-valued SDP (CSDP). We shall discuss novel properties of CSDP, and discuss their applications in solving problems from combinatorial optimization and problems from signal processing.

Unexpected Outcomes

N. Fowkes

University of Western Australia

Abstract: I work primarily on problems that arise out of an industrial context where the objectives are clear cut, and the results are usually of limited general interest. Sometimes, however, the investigations lead to results that are both unexpected and of general interest. I will describe two such investigations; one arising out of the defense industry, the other out of the electronics industry.

Quasi-Periodicity in a Historical Perspective

H.W. Broer

Department of Mathematics, University of Groningen, The Netherlands

Abstract: Kolmogorov-Arnol'd-Moser (or KAM) theory was developed for conservative dynamical systems that are nearly integrable. Integrable systems in their phase space usually contain lots of invariant tori and KAM theory establishes persistence results for such tori, which carry quasi-periodic motions. We sketch this theory which begins with Kolmogorov's pioneering work in 1954.

Dynamics of a Predator-Prey Model with Non-Monotonic Response

Khairul Saleh

Department of Mathematics, University of Groningen, The Netherlands

Abstract: We discuss the dynamics of a family of planar vector fields that models certain populations of predators and their prey. This model, which depends on five parameters, is obtained from the standard Volterra-Lotka system by a non-monotonic response function that takes into account group defense, competition between prey and competition between predators. Also we initiate research on the time-periodic perturbations, which model seasonal dependence. Partly our results are computer assisted. We are interested in persistent features. For the planar autonomous model this amounts to structurally stable phase portraits. We focus on the attractors, where it turns out that multi-stability occurs thereby giving rise to several basins of attraction. Further, we investigate the bifurcations between the various domains of structural stability. It is convenient to fix the values of two of the parameters and study the bifurcations in terms of the remaining three. Here we find several codimension 3 bifurcations that form organizing centers for the global bifurcation set. Studying the time-periodic system, our interest is in the chaotic dynamics. Numerically, we show the existence of strange attractors near homoclinic bifurcations.

Dynamics and Bifurcations of a 3-Dimensional Piecewise-Linear Integrable Map

J.M. Tuwankotta

Departemen Matematika, Institut Teknologi Bandung, Indonesia

Abstract: In this paper we consider a four-parameter family of piecewise-linear ordinary difference equations in 3-dimensional linear space. This system is obtained as a limit of another family of three-dimensional integrable systems of ordinary difference equations. We prove that the limiting procedure sends integrals of the original system to integrals of the limiting system. We derive some results for the solutions such as boundedness of solutions and the existence of periodic solutions. We describe all topologically different shapes of the integral manifolds and present all possible scenarios of transitions as we vary the natural parameters in the system, i.e. the values of the integrals.

Fluxon Dynamics in a Long Josephson Junction

S.A. van Gils

Applied Analysis and Mathematical Physics Group, University of Twente, The Netherlands

Abstract: A long Josephson junction consists of two superconducting layers, separated by a thin insulating layer. The important physical quantity is the difference between the phase of the electrons in the two layers. This phase difference is described by a perturbed sine-Gordon equation. Waves, called *fluxons* in the physics literature, travel in the junction when an external current is applied. Existence of these fluxons can be analyzed analytically for small applied bias currents. The fate of the fluxons for larger currents can be studied numerically.

We determine the linearised stability of the fluxons by calculating the Evans function. Surface resistance corresponds to a singular perturbation term in the governing equation, which specifically complicates the computation of the corresponding Evans function. Both the flow of quasi-particles across and along the junction stabilise the waves. The parameter values for which the fluxons exist appear to lie on a spiral. We show that this is a consequence of the presence of a heteroclinic solution, which lies at the centre of the spiral.

In high temperature superconductors there is the possibility for discontinuity points in the phase. We investigate analytically a long Josephson junction with one π-discontinuity point characterized by a jump of π in the phase difference of the junction. This system is described by a perturbed-combined sine-Gordon equation. Via phase-portrait analysis, it is shown how the existence of static semifluxons localized around the discontinuity point is influenced by the applied bias current. In junctions with more than one corner, there is a minimum-facet-length for semifluxons to be spontaneously generated.

Three Dimensional Modeling of Cohesive Suspended Sediment Transport in Estuary of Mahakam Delta

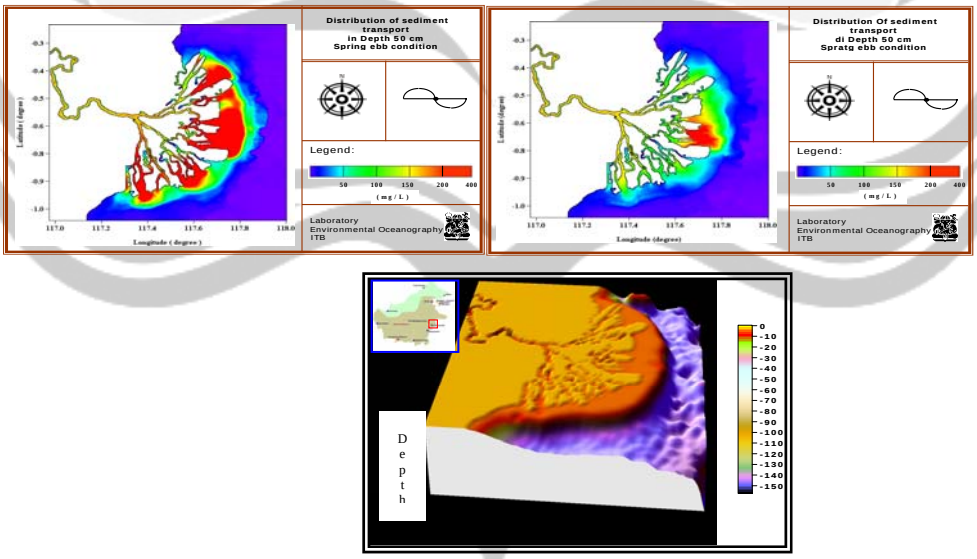
Safwan Hadi, Nining Sari Ningsih, Ayi Tarya

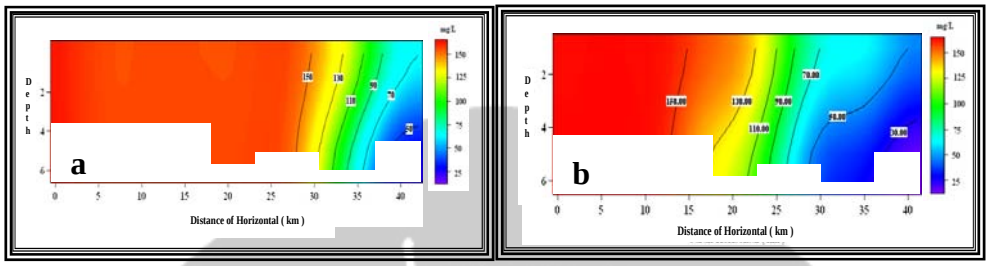
Study Program of Oceanography, Dept. of Geophysics and Meteorology, Institut Teknologi Bandung, Indonesia

Abstract: A coupled three-dimensional hydrodynamics and sediment transport model of HydroQual, Inc., (2002), ECOMSED, has been used to simulate variation of suspended cohesive sediment transport in Estuary of Mahakam Delta. The simulation results indicate that tides and seasonal variation of river discharges are the main causes of variations in the suspended sediment concentration in this area.

A one-year simulation of suspended sediment distribution shows that the suspended cohesive sediment discharge to the Makassar Strait is mainly transported southward, namely through locations of Muara Jawa and Muara Pegah and seems to reach a maximum distance of distribution in January and a minimum one in October. The simulation results also show that river discharges less influence the suspended sediment concentration at Tanjung Bayur, which is located in the tip of the channel in the middle, compared to the other locations.

Keywords: *Three-dimensional hydrodynamics and sediment transport models, Estuary of Mahakam Delta, Suspended cohesive sediment concentration, Tides, River discharges, Turbidity front.*





ON THE BREAKING PARAMETERS OF SIGNALLING PROBLEM

W.M. Kusumawinahyu[1,2], Andonowati[1,3,4], E. van Groesen[4]

[1)] Departemen Matematika, Institut Teknologi Bandung, Indonesia
[2)] Jurusan Matematika, Universitas Brawijaya Malang, Indonesia
[3)] Centre for Mathematical Modelling and Simulation, Institut Teknologi Bandung, Indonesia
[4)] Applied Analysis and Mathematical Physics Group, University of Twente, The Netherlands

Abstract: This research is motivated by the requirement of hydrodynamic laboratories to generate extreme waves for testing ships in steep, large amplitude wave fields. It is also desired that such a wave will not break in its spatial evolution before reaching the tested ship position. For this purpose, finding criteria that determine if wave breaking will occur is important.

In the study of wave breaking, Banner et.al. [1] proposed a non-dimensional quantity that can be interpreted as the dynamic of the maximal square of wave steepness over the spatial domain. The investigation uses a simulation model to calculate the evolution of ocean waves for a given initial profile that depends on certain parameters. A threshold value for the quantity that marks the breaking of waves was found.

Different from Banner's initial value problems, in this contribution we will consider the *signalling problem*: a time signal is prescribed to a wave maker in a wave tank that produces propagating waves running in initially still water. The aim is to observe the resulting nonlinear effects on the waves and to study in which cases the waves will or will not break. This also leads to a threshold value for the steepness of signal at wavemaker and for adjusted Banner's quantity as the breaking parameter of signalling problem. In this observation we consider similar classes of waves as in [1], namely Bichromatic waves and Benjamin Feir-waves, and investigate the evolution by using a numerical simulation code HUBRIS developed by Westhuis [2]. The validity of this code has been tested against laboratory experiments. The result of our investigations is that for both classes the parameters of wave breaking are more extreme in the signaling case than in the case of Banner's initial value problem.

References

[1] Jin-Bao Song and Banner, M. L., On Determining the Onset and Strength of breaking for Deep Water Waves, Part 1: Unforced Irrotational Wave Groups, *Journal of Physical Oceanography*, **32**, 2541 – 2558, 2002

[2] Groesen, E. W. C. van, and Westhuis, J. H., Modelling and simulation of surface water waves, *Mathematics and computers in simulation*, **59(4)**: 341 – 360, 2002

Stationary Optical Solitons In One Dimensional Deep Nonlinear Bragg Grating And Their Potential Applications

H. Alatas[1], A. Iskandar[1], M.O. Tjia[1], T.P. Valkering[2]

[1)] Dept. of Physics, Institut Teknologi Bandung, Indonesia
[2)] Computational Materials Science Group, Univ. of Twente, The Netherlands

Abstract: In this talk, a brief review of our investigation on the stationary soliton and its related phenomena in one dimensional deep nonlinear Bragg grating system will be given in the framework of a model extended from previous one based on an asymptotic formalism introduced earlier. The modified model takes into account the possibility of strong nonlinear modulation produced by large refractive index contrast or induced by high intensity illumination. The discussion covers both the case of infinite and finite length grating systems. For the infinite-length deep grating system, three important results were found. Firstly, all admitted solitonic solutions of the model were identified and classified, including the existence of new types of solutions in the system, namely the in-gap dark and antidark soliton solutions, along with their exact expressions. Secondly, a simple scheme has been developed for the study of bifurcation processes due to frequency variation. Thirdly, the existence of the dark-antidark rational soliton solutions is established on the bifurcation point. For the finite-length deep Bragg grating system, a simple device model equipped with a continuous wave laser source and a movable metallic mirror was investigated in connection with the possible existence of stationary soliton-like solutions. Taking advantage of the source intensity and the mirror distance as independent control parameters, a numerical calculation was carried out on the model with specific parameters for the study of its potential device applications. Three important results were also found in this case. Firstly, in response to either control parameter, the system was shown to exhibit new types of hysteretic relation between the fields on the opposite grating ends. Secondly, soliton-like profiles could be generated by proper tuning of the source intensity following the hysteretic curve. Thirdly, a transition between dark to antidark soliton-like profiles can be induced by changing the mirror distance as well as the optical intensity. This device offers possible applications for sub-micron displacement sensing and optical switching.

Band Structure Design Of A Finite 1D Optical Grating

A. Iskandar[1], W. Yonan[1], M. O. Tjia[1], I. van de Voorde[2], E. van Groesen[2]

[1)] Department of Physics, Institut Teknologi Bandung, Indonesia
[2)] University of Twente, The Netherlands

Abstract: A finite one-dimensional optical grating with possible tailoring of its transmission characteristics can serve as a functional building block for various optical devices ranging from filters (pass filter and stop filters), Distributed Bragg Reflectors (DBR), Wavelength Division Multiplexing (WDM) filters, and optical sensors. Two basic requirements to be met in such a system are optimal transmittance achieved in the bandpass region and a good quality bandgap at the band stop region.

In order to meet the first requirement above, a commonly adopted approach is to use the Anti Reflection Coating (ARC) in the front and the back of the original grating. The use of this coating is to give extra parameters that can be tuned to produce the desired response. With this coating, the geometrical parameters of the ARC is optimized without changing the physical and geometrical parameters of the original grating. While for the second requirement, one resort to either designing the grating structure with a large number of unit cells or a large index contrast between the dielectric layers in an unit cell.

In this work, we present a structured and simplified method to formulate a 1D multilayer system or grating as a bulk effective medium. In this effective formulation, a frequency dependent effective boundary condition is shown to determine the envelope function of the transmittance curve. Using this function, the optimization of the transmittance curve in a certain passband can be achieved without adding extra ARC layers. Furthermore, through this formulation, we find the minimum requirement of the index contrast for achieving a good bandgap.

Keywords: Multilayer System, Transfer Matrix, Band Structure

Pulse Loading and Radiative Unloading of an Optical Defect Grating Structure: Low Dimensional Modeling and Numerical Simulations

A. Sopaheluwakan, E. van Groesen

Applied Analysis & Mathematical Physics Group, MESA + Research Institute, University of Twente, The Netherlands

Abstract: We present a low dimensional model for pulse loading and radiative unloading of an optical defect grating structure. This model describes a phenomenon, when a light pulse with spectral components covering the band gap is incident towards a defect grating structure [1]. Identifying two phases of the phenomenon, the loading and the radiative unloading of light, a low dimensional model is used to describe each of these phases as a separate independent process. The qualitative and quantitative aspects of the dynamics of each phases are given in terms of the defect states [2, 3] of the defect grating structure. For a given structure, the low dimensional model is compared with direct calculation using FETD method [4, 5].

Keywords: Defect grating structure, band gap, defect states, loading, radiative unloading, low dimensional model, FETD.

References

[1] A. Sopaheluwakan and E. van Groesen, Calculation of pulse loading and radiative unloading of an optical defect grating structure, submitted.

[2] E. van Groesen, A. Sopaheluwakan and Andonowati, Direct characterization of states and modes in defect grating structures, J. Nonl. Opt. Phys. And Mat. 13, 155 – 173 (2004).

[3] E. van Groesen, A. Sopaheluwakan and Andonowati, in Proc. of the IEEE-LEOS Benelux Symposium, University of Twente, 273 (2003).

[4] J. F. Lee and R. Lee, Time – domain finite – element methods, IEEE Trans. Ant. Prop.3, 430 – 442 (1997).

[5] V. F. Rodriguez – Esquerre and H. E. Hernandez – Figueroa, Novel time – domain step – by – step scheme for integrated optical applications, IEEEP hot. Tech. Lett. 13, 311 – 313 (2001).

Long Josephson Junctions with Phase-Shifts

H. Susanto

Applied Analysis and Mathematical Physics Group, University of Twente, The Netherlands

Abstract: We consider a Josephson junction with one or more phase-shifts. Eventhough this system was theoretically known in the late 70's, it is just recently that it is possible to fabricate and manipulate it experimentally. I will present some results that have been obtained in the last five years, during which period I have been working on my PhD on this subject, and that describe the characteristics of this system from a theoretical point of view. Some interesting and important problems for future study will be addressed as well.

Keywords: Josephson junctions, sine-Gordon equations, phase-shifts.

Spatially Chaotic Trajectories in a Kerr Grating

A. Irman, T.P. Valkering

MESA+ Research Institute, University of Twente, The Netherlands

Abstract: Monochromatic waves near the first band gap in a 1-dimensional Kerr grating are usually described by the Coupled Mode equations [1] for the, slowly varying, mode amplitudes A_+ and A_- of two counter-propagating modes $\exp(\pm ik_B z)$. Here $k_B = \pi / d$ is the Bragg wavenumber of a grating with period *d*. In this paper we compare the analytical results of this CM model, with numerical calculations on the basis of the full Helmholtz equations.

We consider a specific case: standing waves in a grating consisting of units of two layers with index contrast ε, and with frequency in the band gap (mid-gap). One conclusion is that waves with period (of the envelope) longer than (the order of) $d \times \varepsilon^{-2}$ show irregular features both in phase and in wavelength, that are not covered by the CM equations. A second conclusion is that for

$$\varepsilon \gtrsim 0.2$$

values the CM equations do not apply at all.

These results are interpreted as follows: the Helmholtz equations can be transformed exactly into a set of coupled differential equations of the first order for the amplitudes A_+ and A_-. These equations depend explicitly and periodically on the propagation direction *z*. The vector field consists of two parts, a *z*-independent part and a part that depends on *z* and that has zero average. The CM equations are obtained if one neglects the latter part. The use of these approximate equations can be justified for a shallow grating by the Averaging Theorem [2], which provides an estimate of the error in terms of the small parameters in the problem.

The CM equations are Hamiltonian and autonomous, and their trajectories are equilibria, or periodic. The separatrices represent solitons. The trajectories of the full equations however show irregular, *chaotic* [3] behavior. Consequences of this property will be explored.

Keywords: grating, band gap, coupled mode equations, chaos.

References

[1] CM de Sterke and JE Sipe, 'GapSolitons', in Prog in Opt XXXIII, E. Wolf (ed), Elsev 1994.

[2] JA Sanders and F Verhulst, Averaging Methods in Nonlinear Dynamical Systems, Springer, 1985.

[3] AJ Lichtenberg and MA Liebermann, 'Regular and StochasticMotion', Springer (1983)

A Study Of Slow Light In 1D Photonic Crystals

D. Yudistira, H.J.W.M. Hoekstra, M. Hammer, D.A.I. Marpaung

MESA+ Research Institute, University of Twente, The Netherlands

Abstract: Slow light (SL) states corresponding to wavelength regions near the bandgap edge of grating structure are known to show strong field enhancement. Such states may be excited efficiently by well-optimised adiabatic transitions in such structures, e.g., by slowly turning on the modulation depth. To study adiabatic excitations, a detailed research in 1D is performed to obtain insight into the relation between the device parameters and properties like enhancement and modal reflection. The results enable the design of an adiabatic device for efficient excitation of SL states in 1D. In addition of that, the effects of small wavelength variations as well as that of small fluctuations in the modulation depth of the grating have been investigated.

Keywords: Slow light, Photonic crystals, Field enhancement.

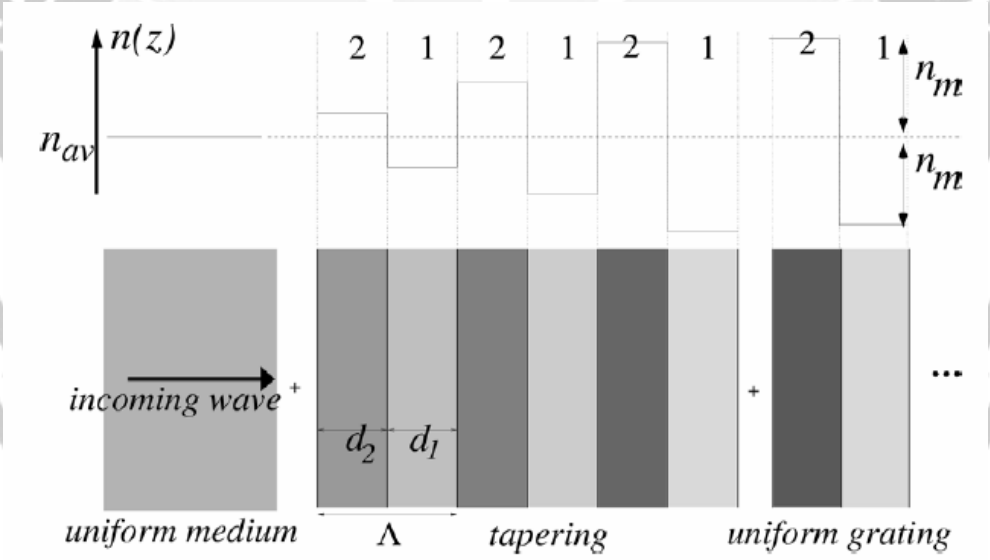


Figure 1: Refractive index as a function of z of 1D grating with a tapered index modulation.

Distance Geometry via Semidefinite Optimization

C. Roos

Algorithm Group, Faculty EWI, TU Delft, The Netherlands

Abstract: One of the most studied problems in distance geometry is the *Graph Realization* problem. In this problem, we are given a graph $G = (V; E)$ and a set of nonnegative weights $\{d_{ij} : \{i, j\} \in E\}$ on its edges. The goal is to compute a realization of G in the Euclidean space R^d for a given dimension d. Thus we want to place the vertices of G in R^d such that the Euclidean distance between every pair of vertices coincides with the given weight for that pair. This problem and its variants arise from applications in various areas, such as molecular confirmation, dimensionality reduction, Euclidean ball packing, and more recently, wireless sensor network location. Following an approach recently proposed by Y. Ye (Stanford Univ.), we discuss how this problem can be tackled by using a semidefinite optimization model.

Snake-in-the-Box Codes and Covers of Hypercubes with Snakes

L. Haryanto

Delft Institute of Applied Mathematics, TU Delft, The Netherlands

Abstract: Let $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \ldots \mathbf{b}_k)$ be an ordered basis of the Reed-Muller code $R(m-2, m)$, which is a linear $[n, k, 4]$-code with $n = 2^m$ and $k = \sum_{i=0}^{m-2} \binom{m}{i}$. One can choose the basis vectors b_i such that $\|\mathbf{b}_i\| = 4$, $1 \le i \le k$, i.e. B is a minimal-weight basis. For each b_i, there is a corresponding block $B_i = (i_1, i_2, i_3, i_4)$, where the set of integers i_j is the support of b_i and $i_j \in \{0,1,\ldots n-1\}$, for all i and j. The order of these integers within the blocks B_i, $1 \le i \le k$ can be chosen such that any integer $a \in \{0,1,\ldots n-1\}$ always occurs on the same position in the blocks (*fixed position property*).
If the basis B is chosen properly and ordered well, then by applying the transition sequence T_k of the standard Gray code $G(k)$ to the indices of the blocks of the block list
$B = (B_1, B_2, \ldots B_k)$,
we obtain a transition sequence $T_k(B)$ which generates a snake-in-the-box code S (or a snake in the hypercube Q_n).

A similar result can be obtained for n-values satisfying $2^{m-1} < n < 2^m$ by considering special subcodes of $R(m-2, m)$. Moreover, by exploiting the linear structure of the underlying code, it appears that we can translate such a snake S over some properly chosen translation vector $\mathbf{a}$ such that S and the translated snake $\mathbf{a} + S$ are disjoint. We proved that for $8 < n \le 16$, there exists a cover of Q_n by eight disjoint translations of such a snake S. This is a stronger result than the one presented by Wojciechowski in [Combinatorica, 14(4), 1994, 491-496]

SURVEY ON RAMSEY NUMBERS OF WHEEL GRAPH

Surahmat

Universitas Islam Malang, Indonesia

Abstract. For two given graphs G and H, the *Ramsey number* $R(G, H)$ is the smallest positive integer N such that for every graph F of order N the following holds: either F contains G as a subgraph or the complement of F contains H as a subgraph. In this paper, we give a brief survey on Ramsey numbers, we have working on during the last four years. In particular, we present the determination of a wheel W_m versus some fixed graph G. Open problems and conjectures are also listed.
Key-words: Ramsey number, wheel

1 Introduction

One of the fundamental problems in mathematical logic is the problem of finding a regular procedure to determine the consistency of any given logical formula. F.P. Ramsey (1928) in [42] studied such a problem and in his investigation he made the following famous theorem as follow:

For given any positive numbers r, n, and μ, then there exist a positive number M_0 such that if $m \geq M_0$ and the r-subsets of any m-set Γ_m are divided in any manner into μ mutually exclusive classes $\mathcal{C}_i$ $(i = 1, 2, ..., \mu)$, then Γ_m must contain a sub-class Δ_n such that all r-subset of members of Δ_n belong to the same $\mathcal{C}_i$.

The information about how large the value of M_0 that can be obtained for given positive integers r, n and μ, has been receiving a lot of attention. Next, F.P. Ramsey [42] showed that for $r = 2$ and $\mu \neq 2$, the value of M_0 is $n!!!...!$, where the process of taking the factorial is performed $\mu - 1$ times. But, this numbers is still too large.

The theorem above, then, is called the *Ramsey Theory* and it became famous after Paul Erdos and George Szekeres (1935) applied it in graph theory for $r = 2$ and $\mu = 2$, see [21]. The idea behind *classical Ramsey numbers* is basically the following.

For any positive integers n and m, we would like to determine the smallest integer $R = R(n, m)$ such that for every graph F of R vertices will satisfy the following condition: either F contain a subgraph K_n or the complement of F contains a subgraph K_m.

The research on finding the exact value of classical Ramsey numbers $R(a, b)$ has been receiving a lot of attention. However, the results are still far from satisfactory. Since firstly introduced, there are only nine exact Ramsey numbers known so far. Greenwood and Gleason (1955) in [22] showed that $R(3, 3) = 6$, $R(3, 4) = 9$, $R(3, 5) = 14$, and $R(4, 4) = 18$. Kery (1964) in [32], then, proved that $R(3, 6) = 18$, and followed by Kalbfeisch [30] who showed that $R(3, 7) = 23$. Grinstead and Roberts (1982), in [23], showed that $R(3, 8) = 28$ and $R(3, 9) = 36$. The latest result is that $R(4, 5) = 25$, due to McKay and Radziszowski (1995) in [34]. While $R(5, 5)$ is the smallest case which is still open (see [28]). For any other values of $n, m \geq 3$, determining the exact value Ramsey number is a difficult problem. However, some *non-trivial* lower and upper bounds for these numbers have been obtained, see [41], for more details.

Graph Ramsey theory has grown enormously in the last two decades to become presently of the most active areas in Ramsey theory. A major impetus behind the early development of graph Ramsey theory was the hope that it eventually lead to methods for determining larger values of the classical Ramsey numbers $R(n, m)$. However, as often happens in mathematics, this hope has not been realized; rather, the field has blossomed into a discipline of its own [21].

Let G and H be two graphs. Basically, the Ramsey number $R(G, H)$ is defined as the smallest integer N such that for every graph F of N vertices will satisfy the following condition: either F contain a subgraph G or the complement of F contains a subgraph H.

One of the most general results in graph Ramsey theory is the following. For a graph G (with no isolated vertices), let $\chi(G)$ denote the chromatic number of G and let $c(G)$ denote the cardinality of the largest connected component of G. Then, Chvátal and Harary [12] showed: $R(G, H) \geq (\chi(G) - 1)(c(H) - 1) + 1$. In particular, Chvátal [10] obtained $R(T_n, K_m) = (n-1)(m-1)+1$, where T_n is an arbitrary tree on n vertices and K_m is a complete graph on m vertices.

For wheels versus triangles, Radziszowski and Xia [38] showed that $R(G, C_3) = 2m + 1$ for $m \geq 5$ where G is a path P_{m+1}, a cycle C_{m+1}, or a wheel W_m. In particular for the combination of wheel and cycles, Zhou [51] has a general result, namely $R(W_m, C_n) = 2m + 1$ if n odd and $m \geq 5n - 7$. However for the case of n even remains open. For more information about the development of graph Ramsey number theory can be seen in a nice and regularly updated survey by S.P. Radziszowski in [39], [40] and [41].

In this paper, we shall give a survey on determination of Ramsey numbers $R(G, W_m)$ for some fixed G. In particular, consider G as a path, star, starlike, forest and cycle.

Throughout the paper, all graphs are finite and simple. Let G be such a graph. We write $V(G)$ or V for the vertex set of G and $E(G)$ or E for the edge set of G. The graph $\overline{G}$ is the *complement* of the graph G, i.e., the graph obtained from the

complete graph $K_{|V(G)|}$ on $|V(G)|$ vertices by deleting the edges of G. For any vertex v in a graph G, the *neighborhood* of vertex v, $N_G(v)$, is the set of all vertices of G which are adjacent to v, and also be denoted $N_G[v] = N_G(v) \cup \{v\}$. For each $x \in V(G)$ and $B \subset V(G)$, define $N_B(x) = \{y \in B : xy \in E(G)\}$. The *minimum degree* of G is the minimum degree among the vertices of G and is denoted $\delta(G)$. A path of n vertices will be denoted by P_n. A cycle and a star of n vertices are denoted by C_n and S_n respectively. A wheel with m vertices will be denoted by W_m.

2 Paths

The Ramsey numbers for a combination a path and wheel, we have information that $R(P_n, W_1)$, $R(P_n, W_2)$ and $R(P_n, W_3)$ can be saw in [20], [38] and [37], respectively. In the previous AWOCA 2001, we showed the following Ramsey numbers for path versus wheels W_4 and W_5 in [45] .

Theorem 2.1 *For all* $n \geq 3$, $R(P_n, W_4) = 2n - 1$.

Theorem 2.2 *For all* $n \geq 3$, $R(P_n, W_5) = 3n - 2$.

The Ramsey numbers still remain the same when replace W_4 and W_5 by W_6 and W_7, respectively, in [1]

Theorem 2.3 *For all* $n \geq 6$, $R(P_n, W_6) = 2n - 1$.

Theorem 2.4 *For all* $n \geq 7$, $R(P_n, W_7) = 3n - 2$.

By employing a generalized version of the methods in [45, 1], we could show the Ramsey numbers in [3] as the following.

Theorem 2.5 *If* $n \geq \frac{m}{2}(m-2)$ *and* $m \geq 4$ *even then* $R(P_n, W_m) = 2n - 1$.

Theorem 2.6 *If* $n \geq \frac{m-1}{2}(m-3)$ *and* $m \geq 5$ *odd then* $R(P_n, W_m) = 3n - 2$.

This result has been refined by Yoojun Chen et.al [9] by showing that:

Theorem 2.7 *If* m *even and* $n \geq m - 1 \geq 3$ *then* $R(P_n, W_m) = 2n - 1$.

Theorem 2.8 *If* m *odd and* $n \geq m - 1 \geq 2$ *then* $R(P_n, W_m) = 3n - 2$.

Problem 2.9 *Determine the Ramsey numbers* $R(P_n, W_m)$ *for* $n < m - 1$ *general.*

3 Star

The Ramsey numbers for stars versus wheels (W_4 and W_5) have been determined in [45].

Theorem 3.1 $R(S_4, W_4) = 9$.

Theorem 3.2 $R(S_n, W_4) = 2n - 1$ *if* $n \geq 3$ *odd or* $R(S_n, W_4) = 2n + 1$ *if* $n \geq 4$ *even.*

Theorem 3.3 $R(S_n, W_5) = 3(n-1) + 1$ *if* $n \geq 3$.

Theorem 3.4 *For all* $n \geq 2m - 4$, $m \geq 5$ *and* m *odd,* $R(S_n, W_m) = 3n - 2$.

With a *star-like tree* we mean a subdivided star (which is not a path), i.e., a tree with exactly one vertex of degree exceeding two. A star-like tree in which only one of the edges of the star has been subdivided, is sometimes called a *comet* in literature; it is usually denote by $Y_{n,l}$, and consists of a path P_n and l additional vertices of degree one, all adjacent to the same end vertex of the P_n. For this reason, and because of the series of results we will present below, we denote by $Y_{n,l_1,l_2,\ldots,l_k}$ the star-like tree consisting of a P_n, and k additional mutually disjoint paths $P_{l_1}, P_{l_2}, \ldots, P_{l_k}$ all attached by one edge from one of their end vertices to the same end vertex of the P_n. If all l_i are equal to 1, we use the shorter notation $Y_{n,\underline{k}}$ to denote $Y_{n,l_1,l_2,\ldots,l_k}$. We have been determined in [47] as follows.

Theorem 3.5 $R(Y_{n,1,1}, W_m) = 3(n+2) - 2$ *for* $n \geq m \geq 5$ *and* m *odd.*

Theorem 3.6 $R(Y_{n,r,\underline{k}}, W_m) = 3(n + r + k) - 2$ *for* $n \geq 2m - 4, n \geq r, m \geq 5$, m *odd, and* $k + r \geq \lfloor \frac{m}{2} \rfloor + 1$.

Theorem 3.7 $R(Y_{n,l_1,l_2,\ldots,l_k}, W_m) = 3(n + \sum_{i=1}^{k} l_i) - 2$ *for* $n \geq 2m - 4, n \geq l_i$ *for each* $i = 1, 2, ..., k$, $m \geq 5$ *odd, and* $\lfloor \frac{m}{2} \rfloor + 1 \leq \sum_{i=1}^{k} l_i$.

Problem 3.8 *Determine the Ramsey numbers* $R(S_n, W_m)$ *for* $n < m - 1$ *and* $n \geq m$ *in general, respectively.*

4 Tree

We have the Ramsey Graph numbers of any tree $T_n \neq S_n$ versus W_m for $n \geq m-1$ in [2]. LetH_t be a cocktail-party graph,i.e., a graph obtained by removing t disjoint edges from K_{2t}. The following lemmas and theorems must hold.

Lemma 4.1 *For odd $n \geq 3$, $n = 2t + 1$, the graph $H_t + K_1$ contains all trees T_n on n vertices.*

Lemma 4.2 *For even $n \geq 4$, $n = 2t$, the graph H_t contains all trees T_n on n vertices other than a star.*

Theorem 4.3 *Let $n \geq 4$ and assume that we are given a particular tree T_n of n vertices other than a star. Then the Ramsey number $R(T_n, W_4) = 2n - 1$.*

Theorem 4.3 *Let $n \geq 3$ and assume that we are given a particular tree T_n of n vertices other than a star. Then the Ramsey number $R(T_n, W_5) = 3n - 2$.*

Problem 4.4 *Determine the Ramsey numbers $R(T_n, W_m)$ for $n < m - 1$ and $n \geq m$ in general, respectively.*

5 Linear forest versus wheels

In this section we have been found a Ramsey number for combination a linear forest versus wheel in [46]. Let F_p be a linear forest on p vertices and W_m a wheel of $m + 1$ vertices.

Theorem 5.1 *If $l \geq \frac{m}{2}(m-2)$, $m \geq 4$, and m is even then $R(F_l, W_m) \leq 2l - 1$.*

Theorem 5.2 *If $l \geq \frac{m-1}{2}(m-3)$, $m \geq 5$ and m is odd then $R(F_l, W_m) \leq 3l - 2$.*

Furthermore, we have the following theorem for odd m.

Theorem 5.3 *If $l \geq m \geq 5$ and m is odd then $R(F_l, W_m) \leq 3l - 2$.*

Theorem 5.4 *(1) If $n \geq m-1$, m is even and $t \geq 2$ then $R(tP_n, W_m) = tn+n-1$. (2) If $n \geq m - 1$, m is odd and $t \geq 2$ then $R(tP_n, W_m) = tn + 2n - 2$.*

The following two theorems generalize the above results.

Theorem 5.5 *Let $n_1 \leq n_2 \leq \cdots \leq n_t$, with $t \geq 2$, $n_i \geq m$ for each i and m be even. If $n_t \leq 2n_1$ then $R(\bigcup_{i=1}^{t} P_{n_i}, W_m) = (\sum_{i=1}^{t} n_i) + n_1 - 1$.*

Theorem 5.6 *Let $n_1 \leq n_2 \leq \cdots \leq n_t$, with $t \geq 2$, $n_i \geq m$ for each i and m be* **odd**. *If $n_t \leq 2\lfloor \frac{3}{2} n_1 \rfloor$ then $R(\bigcup_{i=1}^{t} P_{n_i}, W_m) = (\sum_{i=1}^{t} n_i) + 2n_1 - 1$.*

Problem 5.7 *Determine the Ramsey numbers $R(T_n, W_m)$ for $n < m - 1$ and $n \geq m$ in general, respectively.*

6 Cycles

In this section is to present the Ramsey number of a cycle C_n versus W_m as citation in [49].

Theorem 6.1 $R(C_4, W_4) = 9$.

Theorem 6.2 $R(C_4, W_6) = 9$.

Theorem 6.3 $R(C_4, W_5) = 10$.

In this section is to show the Ramsey number of a cycle C_n versus W_4 or W_5 in [48].

For given graphs G and H, Chvátal and Harary [12] established the lower bound $R(G, H) \geq (c(G) - 1)(\chi(H) - 1) + 1$, where $c(G)$ is the number of vertices of the largest component of G and $\chi(H)$ is the chromatic number of H. In particular, if $n \geq 5$, $G = C_n$ and $H = W_4$ or W_5, then we have $R(C_n, W_4) \geq 2n - 1$ and $R(C_n, W_5) \geq 3n - 2$, respectively.

For the upper bounds we will present proofs by induction. In order to prove the main results of this paper, we need the following known results and lemmas.

Theorem 6.3 (Ore [35]).
If G is a graph of order $n \geq 3$ such that for all distinct nonadjacent vertices u and v, $d(u) + d(v) \geq n$, then G is hamiltonian.

Theorem 6.4 (Faudree and Schelp [18]; Rosta [43]).

$$R(C_n, C_m) = \begin{cases} 2n - 1 \text{ for } 3 \leq m \leq n, \ m \textbf{ odd}, \ (n, m) \neq (3, 3). \\ n + \frac{m}{2} - 1 \text{ for } 4 \leq m \leq n, \ m \textbf{ even} \text{ and } n \textbf{ even}, \ (n, m) \neq (4, 4). \\ \max\{n + \frac{m}{2} - 1, 2m - 1\} \text{ for } 4 \leq m < n, \ m \textbf{ even} \text{ and } n \textbf{ odd}. \end{cases}$$

Lemma 6.5 (Chvátal and Erdős [11]; Zhou [51]).
If $H = C_s \subseteq F$ for a graph F, while $F \not\supseteq C_{s+1}$ and $\overline{F} \not\supseteq K_r$, then $|N_H(x)| \leq r - 2$ for each $x \in V(F) \backslash V(H)$.

Lemma 6.6 *Let F be a graph with $|V(F)| \geq R(C_n, C_m) + 1$. If there is a vertex $x \in V(F)$ such that $|N_F[x]| \leq |V(F)| - R(C_n, C_m)$ and $F \not\supseteq C_n$, then $\overline{F} \supseteq W_m$.*

Lemma 6.7 *Let F and G be graphs with $2n - 1$ and $3n - 2$ vertices without a C_n, respectively. If $\overline{F}$ and $\overline{G}$ contain no W_m, then $\delta(F) \geq n - \frac{m}{2}$ for* **even** *$m \geq 4$ and $n \geq \frac{3m}{2}$, and $\delta(G) \geq n - 1$ for* **odd** *$m \geq 5$ and $n \geq m$.*

Before we deal with the general case of a cycle and W_4, we have the results $R(C_6, W_4) = 11$ and $R(C_7, W_4) = 13$ as follows.

Theorem 6.8 $R(C_6, W_4) = 11$.

Theorem 6.9 $R(C_7, W_4) = 13$.

Lemma 6.10 *Let F be a graph on $2n-1$ vertices with $n \geq 8$, and suppose $\overline{F}$ contains no W_4. If $C_{n-1} \subseteq F$ and $F \not\supseteq C_n$, then $|N_{\mathcal{A}}(x)| \leq 2$ for each $x \in V(F) \backslash \mathcal{A}$, where $\mathcal{A} = V(C_{n-1})$.*

Proof. Let $\mathcal{A} = \{x_1, x_2, ..., x_{n-1}\}$ be the set of vertices of a cycle C_{n-1} in F in a cyclic ordering, and let $\mathcal{B} = V(F) \backslash \mathcal{A}$. Suppose there exists a vertex $b_1 \in \mathcal{B}$ with $|N_{\mathcal{A}}(b_1)| \geq 3$. Clearly, $b_1 x_{i+1} \notin E(F)$ whenever $b_1 x_i \in E(F)$ (indices modulo $n-1$). Since $n \geq 8$, $|\mathcal{A}| \geq 7$, and hence we can choose two neighbors x_i and x_j of b_1 in $\mathcal{A}$ such that $x_{i+1} \neq x_{j-1}$ and $x_{i-1} \neq x_{j+1}$ (indices modulo $n-1$). Let $A = \{x_{i-1}, x_{i+1}, x_{j-1}, x_{j+1}\}$. Then $|A| = 4$ and $xb_1 \notin E(F)$ for each $x \in A$. Moreover, since F contains no C_n, by standard long cycle arguments $x_{i-1}x_{j-1}, x_{i+1}x_{j+1} \notin E(F)$, If $|N_A(x)| \leq 1$ for all $x \in A$, then in $\overline{F}$ all vertices of A have at least $2 = \frac{1}{2}|A|$ neighbors, implying that $\overline{F}$ contains a W_4 with hub b_1. Hence $|N_A(x)| \geq 2$ for some $x \in A$. By symmetry, considering the two possible orientations of C_{n-1}, we may assume without loss of generality that $|N_A(x_{i+1})| \geq 2$, hence $x_{i-1}x_{i+1}, x_{i+1}x_{j-1} \in E(F)$. Then $x_i x_{j-1} \notin E(F)$; otherwise we can obtain a C_n from $E(C_{n-1}) \setminus \{x_{j-1}x_j, x_i x_{i+1}, x_{i-1}x_i\} \cup \{x_j b_1, b_1 x_i, x_i x_{j-1}\}$. Similarly, $x_i x_{j+1} \notin E(F)$. Since $\delta(F) \geq n-2$ by Lemma 6.7 and $|N_{\mathcal{A}}(b)| \leq 5-2 = 3$ for each $b \in \mathcal{B}$ by Lemma 6.5, there exist distinct vertices $b_2, b_3 \in \mathcal{B}$ such that $b_1b_2, b_1b_3 \in E(F)$. This implies that x_{j-1} and x_{j+1} are not adjacent to any vertex in $\{b_2, b_3\}$ since otherwise F contains a C_n (extending the C_{n-1} by including b_1 and b_2 or b_3, while skipping x_i). Now, we will distinguish the following two cases.

Case 1: $x_{j-1}x_{j+1} \notin E(F)$.
Since $\overline{F}$ contains no W_4, $x_t b_2, x_t b_3 \in E(F)$ for each $t \in \{i-1, i+1\}$. Suppose to the contrary, e.g., that $x_{i-1}b_2 \notin E(F)$. Then $\overline{F}$ contains a W_4 with hub x_{j-1} and rim $\{x_{i-1}, b_2, x_{j+1}, b_1\}$. The other cases are symmetric. See Figure 1. Clearly then $x_i b_2, x_i b_3 \notin E(F)$ since $F \not\supseteq C_n$. Thus, we have a W_4 in $\overline{F}$ with hub x_i and rim $\{x_{j-1}, b_2, x_{j+1}, b_3\}$, a contradiction.

Case 2: $x_{j-1}x_{j+1} \in E(F)$.
If $b_2 x_{i-1} \in E(F)$, then we obtain a C_n in F with edge set
$E(C_{n-1}) \setminus \{x_{j-1}x_j, x_j x_{j+1}, x_{i-1}x_i\} \cup \{x_{i-1}b_2, b_2b_1, b_1x_i, x_{j-1}x_{j+1}\}$.
Hence $b_2 x_{i-1} \notin E(F)$. Similarly, $b_2 x_{i+1}, b_3 x_{i-1}, b_3 x_{i+1} \notin E(F)$. If $x_j x_{i-1} \in E(F)$, we obtain a C_n with edge set
$E(C_{n-1}) \setminus \{x_j x_{j+1}, x_{j-1}x_j, x_{i-1}x_i\} \cup \{x_j b_1, b_1 x_i, x_{j-1}x_{j+1}\}$.
Hence, by symmetry, $x_j x_{i-1}, x_j x_{i+1} \notin E(F)$. Since $\overline{F}$ contains no W_4 (with hub x_i and rim $\{x_{j+1}, b_2, x_{j-1}, b_3\}$), x_i is adjacent to a vertex in $\{b_2, b_3\}$. Without loss of generality, let $x_i b_2 \in E(F)$. Since $\delta(F) \geq n-2$ by Lemma 6.7, x_{i+1} must be adjacent to two vertices in $\mathcal{B} \setminus \{b_1, b_2, b_3\}$. Let $x_{i+1}b_4, x_{i+1}b_5 \in E(F)$ for $b_4, b_5 \in \mathcal{B}$. By similar arguments as before, $C_n \not\subseteq F$ implies $b_1 b, b_2 b \notin E(F)$ for each $b \in \{b_4, b_5\}$. Suppose $b_4 x_{i-1} \notin E(F)$. Then we have a W_4 in $\overline{F}$ with hub x_{i-1} and

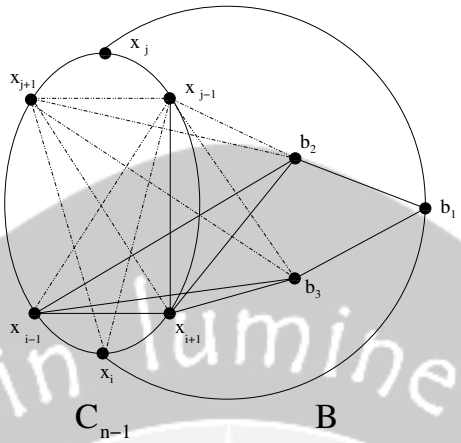


Figure 1: The proof of Lemma 6.10 for Case 1.

rim $\{b_4, b_1, x_{j-1}, b_2\}$. Similar case analysis show that $b_4x, b_5x \in E(F)$ for each $x \in \{x_{i-1}, x_{j-1}\}$. Since F contains no C_n, we clearly have $b_4b_5 \notin E(F)$, and also $x_ix_j \notin E(F)$ (otherwise consider $E(C_{n-1}) \setminus \{x_{j-1}x_j, x_{i-1}x_i\} \cup \{x_ix_j, x_{i-1}b_4, b_4x_{j-1}\}$). Since $\delta(F) \geq n-2$ by Lemma 6.7, there exists a vertex $b_6 \in \mathcal{B} \setminus \{b_1, \ldots, b_5\}$ such that $b_4b_6 \in E(F)$. This clearly implies $b_6x_i, b_6x_j, b_6b_5 \notin E(F)$. See Figure 2.

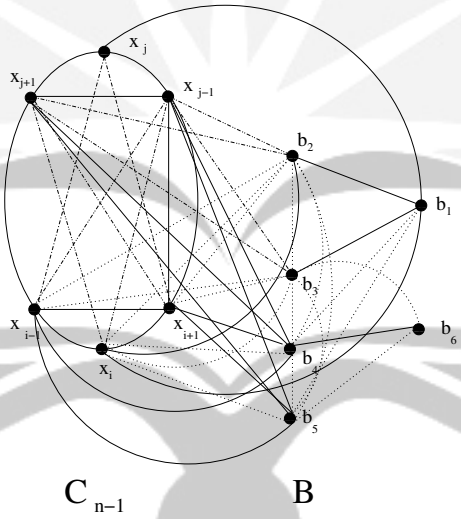


Figure 2: The proof of Lemma 6.10 for Case 2.

Thus, $\overline{F}$ contains a W_4 with hub b_5 and rim $\{x_i, b_6, x_j, b_4\}$, a contradiction. This completes the proof. ■

Theorem 6.11 *$R(C_n, W_4) = 2n-1$ for $n \geq 5$.*

Proof. We use induction on $n \geq 5$. We already know that $R(C_n, W_4) \geq 2n-1$ for $n \geq 5$. For $n = 5$, 6, and 7, we respectively know from [27], Theorem 6.8, and Theorem 6.9 that $R(C_n, W_4) = 2n-1$. Now assume that $R(C_n, W_4) = 2n-1$ for $n < k$ with $k \geq 8$ and let F be a graph on $2k-1$ vertices containing no C_k. We shall show that $\overline{F}$ contains W_4. To the contrary, assume $\overline{F}$ contains no W_4. By the induction hypothesis, we have $F \supseteq C_{k-1}$. Let $A = V(C_{k-1})$, $B = V(F)\backslash V(C_{k-1})$ and so $|B| = k$. By Lemma 6.10, we have $|N_A(x)| \leq 2$ for each $x \in B$. Since by Lemma 6.7, $\delta(F) \geq k-2$, we obtain $|N_B(x)| \geq k-2-2 = k-4 \geq \frac{1}{2}k = \frac{1}{2}|B|$ for all $x \in B$. Now $F[B]$ and hence F contains a C_k by Theorem 6.3, a contradiction. This completes the proof. ■

Theorem 6.12 *$R(C_n, W_5) = 3n-2$ for $n \geq 5$.*

Proof. We use induction on n. We already know that $R(C_n, W_5) \geq 3n-2$ for $n \geq 5$. For $n = 5$, we know from [29] that $R(C_5, W_5) = 3.5-2$. Assume the theorem holds for $n < k$ with $k \geq 6$ and let F be a graph on $3k-2$ vertices containing no C_k. We shall show that $\overline{F}$ contains W_5. To the contrary, assume that $\overline{F}$ contains no W_5. Consequently, F must contain a C_{k-1}, and we let $A = \{a_1, a_2, ..., a_{k-1}\}$ denote the set of vertices of a cycle C_{k-1} in F, in a cyclic ordering. Let $B = V(F)\backslash A$, so $|B| = 2k-1$. Then, by Theorem 6.11, the complement of the subgraph $F[B]$ of F induced by B must contain a W_4. Let x_0 be the hub and $X = \{x_1, x_2, x_3, x_4\}$ be the rim of a W_4 in $\overline{F}[B]$. We distinguish the following cases.

Case 1: *k is* **even**.
Since F contains no C_k, within F: $|N_A(z)| \leq \lfloor \frac{k-1}{2} \rfloor$ for each $z \in B$. This implies that there exist vertices $a_j, a_{j+1} \in A$ for some $j \in \{1, 2, ..., k-1\}$ such that $a_j x_0, a_{j+1} x_0 \notin E(F)$. No C_k in F also implies $N_X(a_j) \cap N_X(a_{j+1}) = \emptyset$. No W_5 in $\overline{F}$ implies in F: $|N_X(a_j)| \geq 2$ and $|N_X(a_{j+1})| \geq 2$, and without loss of generality we may assume a_j is adjacent to x_1 and x_3, and a_{j+1} is adjacent to x_2 and x_4. This implies $x_1x_3, x_2x_4, x_0a_{j+2}, x_0a_{j-1} \in E(F)$ since otherwise $\overline{F} \supseteq W_5$ (Note that $F \not\supseteq C_k$ implies none of a_{j-1} and a_{j+2} is adjacent to a vertex in X). Since F contains no C_k, it is not difficult to check $x_0a_{j-2}, a_{j-2}x_1, a_{j+1}a_{j-2} \notin E(F)$. This implies $\overline{F} \supseteq W_5$ with hub x_0 and rim $\{x_3, a_{j+1}, a_{j-2}, x_1, x_2\}$, a contradiction.

Case 2: *k is* **odd**.
We may assume $a_i x_0 \in E(F)$ for each odd $i \in \{1, 2, ..., k-1\}$, since otherwise we can use the same arguments as in the first case. Since F contains no C_k, $a_j a_h \notin E(F)$ for all even $j, h \in \{1, 2, ..., k-1\}$. If $k \geq 11$, we have K_6 in $\overline{F}$ which implies $\overline{F} \supseteq W_5$, a contradiction. Now assume $7 \leq k < 11$. In F we have $|N_X(a_j)| \geq 2$ for all even $j \in \{1, 2, ..., k-1\}$, since otherwise $\overline{F} \supseteq W_5$. By the same token, we may assume without loss of generality that a_j is adjacent to x_1 and x_3 for some even $j \in \{1, 2, ..., k-1\}$. We distinguish two subcases.

Subcase 2.1: *x_1 is adjacent to x_3.*

Then x_1 and x_3 are not adjacent to any vertex in $\{a_{j-1}, a_{j-2}, a_{j+1}, a_{j+2}\}$, since otherwise F clearly contains a C_k. Thus, we get $\overline{F} \supseteq W_5$ with hub x_0 and rim $\{x_3, a_{j+2}, a_{j-2}, x_1, x_2\}$, a contradiction.

Subcase 2.2: *x_1 is not adjacent to x_3.*
This implies x_2 and x_4 are adjacent to all vertices in $\{a_{j-1}, a_{j+1}\}$, since otherwise $\overline{F} \supseteq W_5$. Suppose, e.g., $x_2a_{j-1} \notin E(F)$. Then $\overline{F} \supseteq W_5$ with hub x_1 and rim $\{a_{j-1}, x_2, x_0, x_3, a_{j+1}\}$; the other cases are similar. Thus, we get $x_2a_j, x_4a_{j+2} \notin E(F)$, otherwise a C_k in F is immediate. Thus, we get $\overline{F} \supseteq W_5$ with hub x_0 and rim $\{x_4, a_{j+2}, a_j, x_2, x_3\}$, our final contradiction.
This completes the proof. ■

Problem 6.13 *Find the Ramsey number $R(C_n, W_m)$ for $n \geq m \geq 6$.*

Conjecture 6.14 *The Ramsey number $R(C_n, W_m) = 2n - 1$ for m even and $n \geq m \geq 4$, and $R(C_n, W_m) = 3n - 2$ for m odd and $n \geq m \geq 3$.*

Acknowledgment

We would like to thank to the Knaw-Epam Project for giving us support for research on Ramsey numbers. In particular, our big thank is to Dr. Edy Tri Baskoro, Prof. Dr. H.J. Broersma and Prof. Dr. Mirka Miller for to nice research activity and nice corporation.

References

[1] Baskoro, E.T.(2002), The Ramsey number of paths and small wheels, *Majalah Ilmiah Himpunan Matematika Indonesia*, MIHMI, **Vol. 8, No. 1** 13-16.

[2] Baskoro, E.T. & Surahmat & Nababan S.M. & Miller M.(2002), On Ramsey graph numbers for all trees versus W_4 or W_5, *Graphs and Combinatorics* **Vol. 18 Number 14** 712-721 .

[3] Baskoro, E.T. & Surahmat (2005), The Ramsey number of paths with respect to wheels, *Discrete Mathematics*, **Vol. 294** 275 - 277.

[4] Bollobás B. & Yang Jian Sheng & Huang Yi Ru & Rousseau C.C. & Zhang Ke Min, On a conjecture involving cycle- complete Ramsey numbers, *to appear.*

[5] Bondy J.A. & Erdös P. (1973), Ramsey numbers for cycles in graphs, *Journal of Combinatorial Theory*, Series B, **14** 46-54.

[6] Bondy J.A. & Murty U.S.R. (1976), *Graph Theory with Applications*, Macmillan.

[7] Burr S.A. & Erdös P. (1983), Generalizations of a Ramsey-theoretic result of Chvatal, *Journal of Graph Theory* **7** 39-51.

[8] Chartrand G. & Lesniak L. (1986), *Graph and Digraph*, Pacific Grove, California.

[9] Chen Y. & Zhang Y. & Zhang K.M. (2002), The Ramsey number of paths with versus wheels, *preprint*, .

[10] Chvátal V.(1977), Tree-complete graph Ramsey numbers, *Journal of Graph Theory* **1** 93.

[11] Chvátal V. & Erdös P. (1972), A note on Hamiltonian circuits, *Discrete Math.* **2** 111-113.

[12] Chvátal V. & Harary F. (1972), Generalized Ramsey theory for graphs III: small off diagonal numbers, *Pas. Journal Math.* **41** 335-345.

[13] , Clancy M. (1977) , Some small Ramsey numbers, *Journal of Graph Theory*, **1** 89-91.

[14] Dirac G.A. (1952), Some theorems on abstract graphs, Proc. London Math. Soc. **2** 69-81.

[15] Erdös P. & Szekeres G. (1935), A combinatorial problem in graph, *Compositio Math.*, **2**, 463-470.

[16] Faudree R.J. & Lawrence S.L. & Parsons T.D. & Schelp R.H.(1974), Path-cycle Ramsey numbers, Discrete Mathematics, **10** 269-277.

[17] Faudree R. J. & Schelp R. H. (1974), All Ramsey number for cycles in graph, *Discrete Mathematics*, **8** 313-329.

[18] Faudree R. J. & Schelp R.L. (1981) , Pancyclic graph connected, *Ars Combinatoria*, **11** 37-49.

[19] Faudree R. J. & McKay B. D. (1993) , A conjecture of Erdös and the Ramsey number $R(W_6)$, *Journal of Combinatorial Mathematics and Combinatorial Computing*, **13** 23-31.

[20] Geréncser L. & Gyárfás A. (1993), On Ramsey-type problems, *Annales Universitatis Scientiarum Budapestinensis*, **13** 23-31.

[21] Graham R. L. & Rothschild B. L. & Spencer J. H. (1990), *Ramsey Theory*, John Wiley and Sons.

[22] Greenwood R.E. & Gleason A.M. (1955), Combinatorial relations and chromatic graph, *Canadian Journal of Mathematics*, **7** 1-7.

[23] Grinstead C. & Roberts S. (1982), On the Ramsey numbers $R(3,8)$ and $R(3,9)$, *Journal of Combinatorial Theory*, **33: B** 27-51.

[24] Guantao C. (1997), A result on C_4-star Ramsey numbers, *Discrete Mathematics*, **163** 243-246.

[25] Gupta S.K.& Gupta L. & Sudan A. (1997), On Ramsey numbers for Fan-Fan Graphs, *Journal of Combinatorics, Information & System Sciences* **22** 85-93.

[26] Harborth H. & I. Mengersen (1988/89), All Ramsey number for five vertices and seven or eight edges, *Discrete Mathematics* **73** 91-98.

[27] Henry G.R.T. (1992) , The Ramsey numbers $R(K_2 + \overline{K}_3, K_4)$ and $R(K_1 + C_4, K_4)$, *Utilitas Mathemathica*, **41**, 181-203.

[28] Jason C. S. (2000), Utilizing a cancellation algorithm to improve the bound of $R(5, 5)$, *The Electronic Journal of Combinatorics*, **Oktober**.

[29] Jayawardene C. J. & Rousseau C. C. (2000), Ramsey number $R(C_5, G)$ for all graphs G of order six, *Ars Combinatoria* **57**, 163-173.

[30] Kalbfleish J.G. (1965), Construction of special edge-chromatic graphs, *Canadian Mathematical Bulletin*, **8**, 575-584.

[31] Karolyi G. & Rosta V. (2001), Generalized and geometric Ramsey numbers for cycles, *Theoretical Computer science* **263** 87-98.

[32] Kéry G. (1964), On a theorem of Ramsey (in hungarian), *Matematikai Lapok*, **15**, 204-224.

[33] Li Y. & Rousseau C.C. (1996), Fan-complete graph Ramsey numbers, *Journal of Graph Theory* **23** 413-420.

[34] Mckay B. D. & Radziszowski S. P. (1995), $R(4, 5) = 25$, *Journal of Graph Theory*, **19:3**, 309-322.

[35] Ore O. (1960), Note on hamilton circuits, *American Mathematics Montly* **67** 55.

[36] Parson T.D. (1973, The Ramsey numbers $R(P_m, K_n)$, *Discrete Mathematics*, **6**) 159-162.

[37] Parson T.D. (1974), Path-star Ramsey numbers, *Journal of Combinatorial Theory*, Series B, **17** 51-58.

[38] Radziszowski S.P. & Xia J.(1994), Paths, cycles and wheels without antitriangles, *Australasian Journal of Combinatorics* **9** 221-232.

[39] Radziszowski S.P. (2000), Small Ramsey numbers, *The Electronic Journal of Combinatorics* **July** DS1, http://www.combinatorics.org.

[40] Radziszowski S.P.(2001), Small Ramsey numbers, *The Electronic Journal of Combinatorics* **July** DS1, http://www.combinatorics.org.

[41] Radziszowski S.P. (2002), Small Ramsey numbers, *The Electronic Journal of Combinatorics* **July** DS1, http://www.combinatorics.org.

[42] Ramsey F. P. (1928), On a problem of formal logic, *Proceeding of The London Mathematical Society*, **30**, 264-286.

[43] Rosta V. (1973), On a Ramsey type problem of J.A. Bondy and P. Erdos, I,II, *Journal Combinatorics Theory* **15B**, 94-120.

[44] Sheng Y.J. & Ru H.Y. & Min Z.K.,$R(C_n, K_4) = 3(n-1)+1, (n \geq 4)$,*Preprint.*

[45] Surahmat & Baskoro E.T. (2001), On the Ramsey number of a path or a star versus W_4 or W_5, *Proceedings of the 12-th Australasian Workshop on Combinatorial Algorithms*, Bandung, Indonesia, July 14-17 174-179.

[46] Surahmat & Baskoro E.T. (2002), The Ramsey number of a linear forest versus wheel, *Proceedings of the 13-th Australasian Workshop on Combinatorial Algorithms*, Fraser Island, Queensland, Australia, July 7-14 127-131.

[47] Surahmat & Baskoro E.T. & Broersma H.J. (2002), The Ramsey number of large star-like trees versus large odd wheels, *preprint*, .

[48] Surahmat & Baskoro E.T. & Broersma H.J. (2004), The Ramsey number of large cycles versus large small wheels, *The Electronics Journal of Combinatorics and Nuber Theory*, **Vol. 4**.

[49] Surahmat & Baskoro E.T. & Nababan S.M. (2002),The Ramsey numbers for a cycle of length four versus a small wheel, *Proceedings of the 11-th Conference Indonesian Mathemathics*, State of University Malang, Indonesia, July 22-25 .

[50] Wilson R.J. & Watkins J.J. (1990), *Graphs: An Introductory Approach*, John Wiley & Sons, New York .

[51] Zhou H. L.(1995), The Ramsey number of an odd cycles with respect to a wheel (in chinese),*Journal of Mathematics, Shuxu Zazhi (Wuhan)*,**15** , 119-120.

[52] Zhou H.L. (2000), On book-wheel Ramsey number, *Discrete Mathematics*, **224** 239-249.

SURAHMAT: Department of Mathematics Education, Universitas Islam Malang, Jl. MT Haryono 193, Malang 65144, Indonesia.
Phone/Fax: +62 +341 571950
E-mail: caksurahmat@yahoo.com

The Robust Maximum Flow Problem

D.Chaerani[a,b], C.Roos[a]

[a] Delft University of Technology, The Netherlands
[b] Universitas Padjadjaran, Bandung, Indonesia

Abstract

We consider the robust maximum flow problem (RMFP) using the robust linear optimization methodology. We discuss two types of uncertainty on the arc capacities: box uncertainty and ellipsoidal uncertainty. In both cases, the RMFP is the usual maximum flow problem with modified arc capacities. In the case of box uncertainty the flow of each arc is bounded by the lower bounds of the box. In the case of ellipsoidal uncertainty, the capacity of each arc a is replaced by $c_a^n - \|\mathcal{Q}_a\|$ where c_a^n is the nominal arc capacity and $\mathcal{Q}_a$ is the corresponding column of the ellipsoidal scaling matrix $\mathcal{Q}$. We present some examples.

Key-words:robust optimization, robust counterpart, uncertainty set,maximum flow problem

1 Introduction

Let $G = (\mathcal{V}, \mathcal{A})$ be a directed graph, let $r, s \in \mathcal{V}$ and let $c : \mathcal{A} \to \mathrm{Q}_+$ be a capacity function. The objective in the maximum flow problem is to find an $r - s$ flow of maximum value under c. Adding an arc from s to r with $c_{sr} = \infty$, the maximum $r - s$ flow problem can be formulated as

$$\max\{x_{sr} \ : \ Ax = 0, 0 \leq x \leq c\}, \tag{1}$$

where A is the node-arc incidence matrix and x is the vector of flow variables.
The maximum flow problem arises in a wide variety of situations and in several forms. For example, determining the maximum steady state flow of petrol in a pipeline network, cars in a road network, messages in a telecommunication network and electricity in an electrical network. In such examples, uncertainty in the value of the arc capacities may occur.
In this paper, we propose a model to handle the maximum flow problem with uncertain arc capacities. We assume that the arc capacities belong to a so-called uncertainty set $\mathcal{U}$. We then have to deal with a whole family of maximum flow problems, namely

$$\mathcal{H} = \{\max\{x_{sr} \ : \ Ax = 0, 0 \leq x \leq c\} \ : \ c \in \mathcal{U}\}. \tag{2}$$

The major challenge is when and how we can reformulate (2) as a computationally tractable optimization problem. We approach this problem by applying the Robust Linear Optimization (RLO) methodology as proposed by Ben-Tal and Nemirovski

(see [1, 2, 3]). We require the flow to be feasible under all possible values of $c \in \mathcal{U}$, and we seek to maximize the flow value under this condition.
We call the problem the robust maximum flow problem (RMFP) and the flow of maximum value under the uncertain arc capacities the robust maximum flow (RMF) value.
The paper is organized as follows. Section 2 briefly introduces the theory of RLO. Section 3 is devoted to the RMFP. We derive the robust counterpart of the given uncertain maximum flow problem for the cases of box and ellipsoidal uncertainty sets. In Section 3.3.1 we consider a special case of box uncertainty, namely when the uncertainty is relative to the nominal arc capacities values c^n. In Section 3.5, we discuss a parametric variant of the above RMFP, where the sizes of the uncertainty perturbation in c are controlled by a nonnegative scaling parameter. In Section 3.6, we discuss a special case when the RMFP optimal solution for the both cases of box and ellipsoidal uncertainty set have the same solution. Some examples are presented. Conclusions can be found in Section 4.

2 Robust Linear Optimization

In this section, we briefly recall some definitions and main results from [3]. Consider a linear optimization problem

$$\min_x \{c^T x \ : \ Ax \geq b\}, \qquad (\mathcal{LO})$$

and let $\mathcal{U}$ be the set of all possible realizations of (A, b, c). So the set $\mathcal{U}$ models the uncertainty in the data of $(\mathcal{LO})$. We call $\mathcal{U}$ the uncertainty set. As a consequence, we have a whole family of $\mathcal{LO}$ problems, for each $(c, A, b) \in \mathcal{U}$ one $\mathcal{LO}$ problem. This family is given by

$$\left\{ \min_x \{c^T x \ : \ Ax \geq b\} \ : \ (c, A, b) \in \mathcal{U} \right\}. \qquad (3)$$

Following [3] we consider instead the so-called robust counterpart of (3), namely

$$\min \left\{ \ell \ : \ \ell \geq c^T x, Ax \geq b, \forall (c, A, b) \in \mathcal{U} \right\}. \qquad (4)$$

The formulation (4) is a linear optimization problem with (usually) infinitely many constraints, depending on the uncertainty set $\mathcal{U}$. Hence, in general this problem may be very hard to solve. Special cases for $\mathcal{U}$ make (4) computationally tractable. In [1], it has been shown that the robust counterpart (4) is equivalent to an explicit computationally tractable problem provided that $\mathcal{U}$ has a simple structure. This becomes clear from the following theorem which presents three examples of computationally tractable uncertainty sets.

Theorem 2.1 (cf. [3]) *Assume that the uncertainty set $\mathcal{U}$ in (4) is given as the affine image of a bounded set $\mathcal{Z} = \{\zeta\} \subset \mathbf{R}^N$, and $\mathcal{Z}$ is given either*

1. *by a system of linear inequalities $P\zeta \leq p$, or*

2. *by a system of conic quadratic inequalities* $\|P_i\zeta - p_i\|_2 \leq q_i^T\zeta - r_i, i = 1,\ldots,M$, *or*

3. *by a system of linear matrix inequalities* $P_0 + \sum_{i=1}^{dim\zeta} \zeta_i P_i \succeq 0$.

In the cases 2 *and* 3 *assume also that the system of constraints defining* $\mathcal{U}$ *is strictly feasible. Then the robust counterpart* (4) *of* (3) *is equivalent to a linear optimization problem in case* 1*, a conic quadratic problem in case* 2*, and a semidefinite problem in case* 3*. In all cases, the data of the resulting robust counterpart problem are readily given by* M, N *and the data specifying the uncertainty set. Moreover, the size of the resulting problem is polynomial in the size of the data specifying the uncertainty set.*

In the following section we discuss how the general result of RLO as stated in Theorem 2.1 can be applied to the RMFP. We make the natural assumption that the network $G = (\mathcal{V}, \mathcal{A})$ is fixed, as well as the nodes r and s. So the uncertainty occurs only in the vector c of arc capacities.

3 The robust maximum flow problem

3.1 Preliminaries

In this subsection, we briefly recall some well known definitions and results about the maximum flow problem including the max flow-min cut theorem (see [5]).

Definition 3.1 *For a given network* $G = (\mathcal{V}, \mathcal{A})$ *and* $r, s \in \mathcal{V}$*, a function* $x : \mathcal{A} \to \mathbf{R}$ *is called an* $r - s$ *flow if*

(i)

$$x_a \geq 0 \quad \textit{for each} \quad a = (u, v) \in \mathcal{A}, \tag{5}$$

(ii)

$$\sum_{a \in \delta^+(v)} x_a = \sum_{a \in \delta^-(v)} x_a \quad \textit{for each} \quad v \in \mathcal{V} \setminus \{r, s\}, \tag{6}$$

where $\delta^+(v)$ *and* $\delta^-(v)$ *denote the sets of arcs leaving* v *and entering* v*, respectively. Condition* (6) *is called the flow conservation law, i.e., the amount of flow entering a node* $v \neq r, s$ *is equal to the amount of the flow leaving* v*.*

The value of an $r - s$ flow is, by definition, the net flow entering the network, i.e.,

$$x_{sr} = \sum_{a \in \delta^+(r)} x_a - \sum_{a \in \delta^-(r)} x_a. \tag{7}$$

Let $c : \mathcal{A} \to \mathbf{R}_+$ be a capacity function. We say that a flow x is under c (or subject to c) if

$$x_a \leq c_a \quad \text{for each} \quad a \in \mathcal{A}. \tag{8}$$

Let $X \subseteq \mathcal{V}$ with $r \in X$ and $s \notin X$. Then the set $\delta^+(X)$

$$\delta^+(X) = \{(u,v) \ : \ u \in X, v \notin X\} \tag{9}$$

is called an $r-s$ cut. Note that when removing the edges in an $r-s$ cut $\delta^+(X)$ from the network then there is no longer a path from r to s, and hence no flow can be sent through the network.
The capacity of a cut $\delta^+(X)$ is defined by

$$c(\delta^+(X)) = \sum_{a \in \delta^+(X)} c_a. \tag{10}$$

The following theorem holds.

Theorem 3.2 *(cf. [5]) For every flow x and every cut $\delta^+(X)$ one has:*

$$x_{sr} \leq c(\delta^+(X)). \tag{11}$$

Equality holds if and only if $x_a = c_a$ for each $a \in \delta^+(X)$ and $x_a = 0$ for each $a \in \delta^-(X)$.

This implies that

$$\max x_{sr} \leq \min_{r \in X \subseteq V \setminus \{s\}} c(\delta^+(X)). \tag{12}$$

Theorem 3.2 is called the weak duality theorem.
Now, consider the dual problem of (1):

$$\min\{c^T y \ : \ A^T \pi + y \geq 0, \pi_s - \pi_r \geq 1\}. \tag{13}$$

From the total unimodularity of A, it follows that there exist an optimal solution of (13) that is integer. Also as the dual is unchanged if we replace π_j by $\pi_j + \alpha$ for all $j \in \mathcal{V}$, we can set $\pi_r = 0$. Given such a solution, let

$$X = \{j \in \mathcal{V} \ : \ \pi_j \leq 0\} \ \text{ and } \ \overline{X} = V \setminus X = \{j \in \mathcal{V} \ : \ \pi_j \geq 1\}.$$

Then one may easily see that $r \in X \subseteq \mathcal{V} \setminus \{s\}$. So any integer optimal solution of (13) gives rise to the cut $\delta^+(X)$ with X as just defined. We now show that the capacity of this cut is equal to the maximal flow value. The definition of X implies that $y_a \geq \pi_j - \pi_i \geq 1$ for $a \in \delta^+(X)$. Hence we have

$$c^T y = \sum_{a \in \mathcal{A}} c_a y_a \geq \sum_{a \in \delta^+(X)} c_a y_a \geq \sum_{a \in \delta^+(X)} c_a = c(\delta^+(X)). \tag{14}$$

However, the lower bound $c(\delta^+(X))$ for the optimal value of (13) is attained by the solution

$$\pi_j = \begin{cases} 0, & j \in X, j \in \mathcal{V}, \\ 1, & j \in \overline{X}, j \in \mathcal{V}. \end{cases} \quad \text{and} \quad y_a = \begin{cases} 1, & a \in \delta^+(X), a \in \mathcal{A}. \\ 0, & a \notin \delta^+(X), a \in \mathcal{A}. \end{cases} \tag{15}$$

So (15) is an optimal $0-1$ solution and $\{a \ : \ y_a = 1\}$ is the set of arcs of the $r-s$ cut $\delta^+(X)$. Thus the problem (13) can be restated as follows:

$$\min_X \left\{ \sum_{a \in \delta^+(X)} c_a \ : \ r \in X \subset \mathcal{V} \setminus \{s\} \right\}, \tag{16}$$

which is the minimum $r-s$ cut problem. The following theorem, which is the max flow-min cut theorem, is a special case of the (strong) duality theorem for linear optimization.

Theorem 3.3 *A strong dual to the maximum $r-s$ flow problem* (1) *is the minimum $r-s$ cut problem* (16).

3.2 The uncertain maximum flow problem

In this subsection we discuss the uncertain maximum flow problem. The natural assumption is that the network $G = (\mathcal{V}, \mathcal{A})$ is fixed, as well as the nodes r and s. Thus, the uncertainty occurs only in the vector c of arc capacities. We assume $c \in \mathcal{U}$, when $\mathcal{U}$ is the uncertainty set for c.
By RLO methodology, the robust counterpart of the RMFP can be stated as

$$\max\{x_{sr} \ : \ Ax = 0, 0 \leq x \leq c, \forall c \in \mathcal{U}\}. \tag{17}$$

Thus, the objective is to find the maximum value of a flow that satisfy $x \leq c$ for all $c \in \mathcal{U}$ where $c_{sr} = \infty$. Of course this robust counterpart depends on how we choose the uncertainty set $\mathcal{U}$. We consider two different uncertainty sets, namely box and ellipsoidal uncertainty sets, as described in the following subsections.

3.3 Box uncertainty

Assume that the uncertainty set $\mathcal{U}$ is a box, i.e., $\mathcal{U}$ is defined as follows

$$\mathcal{U} = \{c \ : \ \ell \leq c \leq u\}, \tag{18}$$

where ℓ and u are two vectors in $\mathbf{R}^{\mathcal{A}}$ with $\ell \leq u$. In this case, the following holds

$$x \leq c, \forall c \in \mathcal{U} \Leftrightarrow x \leq \ell, \tag{19}$$

so that (17) reduces to

$$\max\{x_{sr} \ : \ Ax = 0, 0 \leq x \leq \ell\}. \tag{20}$$

This implies that the RMFP with box uncertainty is the usual maximum flow problem with the arc capacity vector c replaced by ℓ. So we have proved the next result.

Theorem 3.4 *The RMFP with box uncertainty set given by* (18) *is*

$$\max\{x_{sr} \ : \ Ax = 0, 0 \leq x \leq \ell\}. \tag{21}$$

3.3.1 A special case of box uncertainty

It is interesting to consider a special case of box uncertainty, namely when the uncertainty is relative to the nominal arc capacities values c^n, i.e. when the uncertainty set has the form

$$\mathcal{U} = \{c \,:\, |c - c^n| \leq \lambda c^n\} = \{c \,:\, (1-\lambda)c^n \leq c^n \leq (1+\lambda)c^n\}, \text{ for some } \lambda \geq 0. \tag{22}$$

In this case the solution of the RMFP can be easily expressed in terms of the optimal solution of the original problem. This can be shown as follows.
Let x^{opt} be an optimal flow of the original maximum flow problem (1). Now let $0 \leq \lambda \leq 1$, then $x^{opt} \leq c^n$ implies

$$x := (1-\lambda)x^{opt} \leq (1-\lambda)c^n. \tag{23}$$

Hence x is a feasible flow. On the other hand, if $\delta^+(X)$ is a minimal cut for the original problem, then $c^n(\delta^+(X)) = x^{opt}_{sr}$. Hence

$$c(\delta^+(X)) = (1-\lambda)c^n(\delta^+(X)) = (1-\lambda)x^{opt}_{sr} = x_{sr} \tag{24}$$

So x is optimal. Thus, the original minimal cut capacity and maximum flow value is multiplied by $(1-\lambda)$. We have proved the the following theorem.

Theorem 3.5 *For the RMFP with relative uncertainty set* (22) *the maximal flow value is* $(1-\lambda)$ *times the nominal maximum flow value.*

3.4 Ellipsoidal uncertainty

In this subsection we consider the case of ellipsoidal uncertainty. We assume that $\mathcal{U}$ has the form

$$\mathcal{U} = \{c \,:\, c = c^n + \mathcal{Q}w, \|w\| \leq 1\}, \tag{25}$$

where $\mathcal{Q}$ is a fixed matrix of size $|\mathcal{A}| \times p$ and $w \in \mathbf{R}^p$ for some p.

Lemma 3.6 *The flow x is robust feasible if and only if*

$$0 \leq x_a \leq c^n_a - \|\mathcal{Q}_a\|, \qquad \forall a \in \mathcal{A} \tag{26}$$

where c^n_a is the nominal capacity on arc a and $\mathcal{Q}_a$ is the column of $\mathcal{Q}$ corresponding to arc a.

Proof: The flow x_a on arc a must satisfy

$$x_a \leq c^n_a + (\mathcal{Q}_a)^T w, \qquad \forall w \,:\, \|w\| \leq 1, \tag{27}$$

where $\mathcal{Q}_a$ is the column of $\mathcal{Q}$ corresponding to arc a. This means that

$$x_a \leq c^n_a + \min_w \left\{ \mathcal{Q}_a^T w \,:\, \|w\| \leq 1 \right\}. \tag{28}$$

The minimum at the right hand side is attained when

$$w = -\frac{\mathcal{Q}_a}{\|\mathcal{Q}_a\|}, \tag{29}$$

whence the capacity of arc a becomes

$$c_a^n - (\mathcal{Q}_a)^T \frac{\mathcal{Q}_a}{\|\mathcal{Q}_a\|} = c_a^n - \frac{\|\mathcal{Q}_a\|^2}{\|\mathcal{Q}_a\|} = c_a^n - \|\mathcal{Q}_a\|. \tag{30}$$

□

This implies that also in this case, the RMFP is a usual maximum flow problem, with the nominal capacities c_a^n replaced by $c_a^n - \|\mathcal{Q}_a\|$, $a \in \mathcal{A}$. So we have proved the next result.

Theorem 3.7 *The RMFP with ellipsoidal uncertainty as given by* (25)*, is equivalent to*

$$\max\{x_{sr} \ : \ Ax = 0, 0 \le x_a \le c_a^n - \|\mathcal{Q}_a\|, a \in \mathcal{A}\}. \tag{31}$$

In the next subsection, we discuss a parametric variant of the above RMFP, where the sizes of the uncertainty perturbation in c are controlled by a nonnegative scaling parameter.

3.5 Parametric uncertainty

Let the uncertainty set $\mathcal{U}_\alpha$ be defined by

$$\mathcal{U}_\alpha = \{c \ : \ c = c^n + \alpha\mathcal{Q}w, \|w\| \le 1\}, \tag{32}$$

where α is a nonnegative scaling parameter. Note that $c^n + \alpha\mathcal{Q}w$ must be nonnegative for $\forall w \ : \ \|w\| \le 1$ to ensure feasibility. Thus, we assume that

$$0 \le \alpha \le \alpha_{\max} := \min\left\{\frac{c_a^n}{\|\mathcal{Q}_a\|} \ : \ \|\mathcal{Q}_a\| > 0, a \in \mathcal{A}\right\}. \tag{33}$$

Theorem 3.8 *Let $\mathcal{U}_\alpha$ be the ellipsoidal uncertainty set given by* (32) *with $0 \le \alpha \le \alpha_{\max}$ and let $x_{sr}(\alpha)$ denote the optimal flow value for the robust counterpart. Then $x_{sr}(\alpha)$ is a piecewise monotonically decreasing linear concave function.*

Proof: By Theorems 3.3 and 3.7, the maximum flow of the RMFP with ellipsoid set $\mathcal{U}_\alpha$ satisfies

$$x_{sr}(\alpha) = \min_X \left\{ \sum_{a \in \delta^+(X)} (c_a^n - \alpha \|\mathcal{Q}_a\|) \ : \ r \in X \subseteq \mathcal{V} \setminus \{s\} \right\}. \tag{34}$$

We shall show that $x_{sr}(\alpha)$ is a piecewise linear concave function of α by proving that it is the minimum of a finite family of linear functions. To this end, it is convenient to introduce

$$\mathcal{X} = \{X : r \in X \subseteq \mathcal{V} \setminus \{s\}\} \tag{35}$$

such that (34) can rewritten as follows

$$x_{sr}(\alpha) = \min_X \left\{ c^\alpha(\delta^+(X)) \ : \ X \in \mathcal{X} \right\}, \tag{36}$$

where

$$c^\alpha(\delta^+(X)) = \sum_{a \in \delta^+(X)} c_a^n - \alpha \sum_{a \in \delta^+(X)} \|\mathcal{Q}_a\|. \tag{37}$$

Fixing $X \in \mathcal{X}$ and since $\sum_{a \in \delta^+(X)} \|\mathcal{Q}_a\| \geq 0$, $c^\alpha(\delta^+(X))$ is a monotonically decreasing linear function of α. If $|\mathcal{V}| = n$, then the number of $r - s$ cuts $\mathcal{X}$ is 2^{n-2}. Hence $\mathcal{X}$ is finite. We conclude that $x_{sr}(\alpha)$ is the minimum of a finite set of monotonically decreasing linear functions. This implies that $x_{sr}(\alpha)$ is continuous, concave and monotonically decreasing piecewise linear function. □

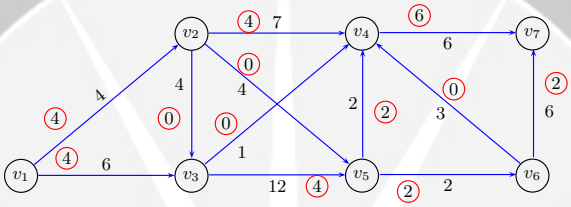


Figure 1: A maximum flow problem: the nominal arc capacities are un-circled number, a maximum flow is given by the circled number with $x_{71}^n = 8$.

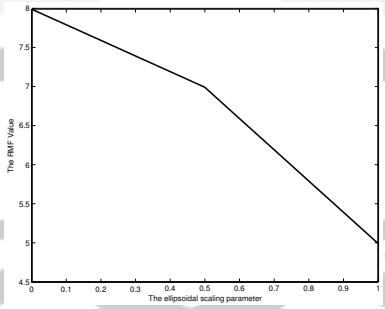


Figure 2: The RMF as a piecewise linear concave function α.

Example 3.9 *Consider the network of Figure 1. Taking $\mathcal{Q} = I$, the ellipsoid $\mathcal{U}_\alpha$ becomes*

$$\mathcal{U}_\alpha = \{c \ : \ c = c^n + \alpha w, \|w\| \leq 1\} \tag{38}$$

where α satisfy $0 \leq \alpha \leq 1$ by (33). The robust arc capacities are then

$$c_a = c_a^n - \alpha, \qquad \forall a \in \mathcal{A}, \tag{39}$$

hence the RMFP for this example is

$$\max\{x_{71} \ : \ Ax = 0, 0 \leq x_a \leq c_a^n - \alpha, \forall a \in \mathcal{A}\}. \tag{40}$$

In Table 1, we present the RMF for $0 \leq \alpha \leq 1$. *In Figure 2 we see that the RMF value function* $x_{71}(\alpha)$ *is a piecewise monotonically decreasing linear concave function of* α *with two different intervals and three break points.*

Table 1: The RMF value for $\alpha \in [0, 1]$

Arcs	c^n	α										
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
x_1	4	3.7929	3.7322	3.6745	3.6163	3.5585	3.5000	3.4000	3.3000	3.2000	3.1000	3.0000
x_2	6	4.2071	4.0678	3.9255	3.7837	3.6415	3.5000	3.2000	2.9000	2.6000	2.3000	2.0000
x_3	4	0.1999	0.1562	0.1111	0.0683	0.0290	0	0	0	0	0	0
x_4	7	3.3709	3.4123	3.4533	3.4899	3.5067	3.5000	3.4000	3.3000	3.2000	3.1000	3.0000
x_5	4	0.2221	0.1636	0.1101	0.0581	0.0229	0	0	0	0	0	0
x_6	1	0.7753	0.7011	0.6320	0.5680	0.5190	0.5000	0.4000	0.3000	0.2000	0.1000	0
x_7	12	3.6317	3.5230	3.4045	3.2840	3.1514	3.0000	2.8000	2.6000	2.4000	2.2000	2.0000
x_8	6	6.0000	5.9000	5.8000	5.7000	5.6000	5.5000	5.2635	5.0197	4.7710	4.5150	4.2567
x_9	2	1.8539	1.7866	1.7146	1.6421	1.5743	1.5000	1.4000	1.3000	1.2000	1.1000	1.0000
x_{10}	2	2.0000	1.9000	1.8000	1.7000	1.6000	1.5000	1.4000	1.3000	1.2000	1.1000	1.0000
x_{11}	3	0	0	0	0	0	0	0.0635	0.1197	0.1710	0.2150	0.2567
x_{12}	6	2.0000	1.9000	1.8000	1.7000	1.6000	1.5000	1.3365	1.1803	1.0290	0.8850	0.7433
x_{13}	∞	8.0000	7.8000	7.6000	7.4000	7.2000	7.0000	6.6000	6.2000	5.8000	5.4000	5.0000
max-flow		8.0000	7.8000	7.6000	7.4000	7.2000	7.0000	6.6000	6.2000	5.8000	5.4000	5.0000

In the next subsection we discuss some properties of the RMF value function $x_{sr}(\alpha)$.

3.5.1 The minimal cuts on a linearity interval and at a breakpoint

The values of α where the slope of $x_{sr}(\alpha)$ changes are called breakpoints of $x_{sr}(\alpha)$ and any interval between two successive break points of $x_{sr}(\alpha)$ is called a linearity interval of $x_{sr}(\alpha)$. For any α in the domain of $x_{sr}(\alpha)$ we denote the set of minimal cut sets by

$$\mathcal{X}_\alpha = \{X \in \mathcal{X} \ : \ x_{sr}(\alpha) = c^\alpha(\delta^+(X))\}. \tag{41}$$

The following theorem shows that the set $\mathcal{X}_\alpha$ is constant on the interior of a linearity interval.

Theorem 3.10 *If* $x_{sr}(\alpha)$ *is linear on the interval* $[\alpha_1, \alpha_2]$, *where* $\alpha_1 < \alpha_2$ *then* $\mathcal{X}_\alpha$ *is constant for* $\alpha \in (\alpha_1, \alpha_2)$.

Proof: Rewrite the RMF value as

$$x_{sr}(\alpha) = \tau - \alpha\sigma, \qquad \alpha \in [\alpha_1, \alpha_2], \tag{42}$$

where

$$\tau = \sum_{a \in \delta^+(X)} c_a^n \text{ and } \sigma = \sum_{a \in \delta^+(X)} \|\mathcal{Q}_a\|. \tag{43}$$

Consider that for $\beta \in (\alpha_1, \alpha_2)$ such that $X \in \mathcal{X}_\beta$ we have that τ and σ are independent of β. This implies that $\mathcal{X}_\beta$ is independent of β. Since β is arbitrary

on the open interval (α_1, α_2), then for any $\alpha \in (\alpha_1, \alpha_2)$ we conclude that $\mathcal{X}_\alpha$ is constant. □

At a break point $(\alpha, x_{sr}(\alpha))$, the following holds.

Theorem 3.11 *Let $\mathcal{X}_{\alpha_1}$ and $\mathcal{X}_{\alpha_2}$ be the minimal cuts on two neighboring intervals (α_1, α) and (α, α_2) respectively. Then the minimal cut set at the breakpoint $(\alpha, x_{sr}(\alpha))$ satisfies*

$$\mathcal{X}_\alpha \supseteq \mathcal{X}_{\alpha_1} \bigcup \mathcal{X}_{\alpha_2}. \tag{44}$$

Proof: By Theorem 3.10, the minimal cuts $\mathcal{X}_{\alpha_1}$ and $\mathcal{X}_{\alpha_2}$ are constant on the interval (α_1, α) and (α, α_2) respectively. This implies that at the breakpoint $(\alpha, x_{sr}(\alpha))$, the minimal cuts $\mathcal{X}_a$ contains $\mathcal{X}_{\alpha_1}$ and $\mathcal{X}_{\alpha_2}$. Thus the proof is followed. □

In the following example, we show that it is possible to have $n+1$ different linearity interval and n breakpoints.

Example 3.12 *Consider a network as shown in Figure 3 with $n \geq 2$.*

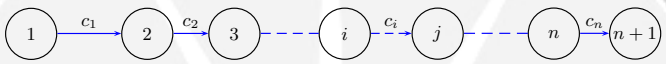


Figure 3: A simple network

We show that it is possible to have $n+1$ different linearity intervals and n breakpoints. For a given R, define

$$c_k = \frac{R}{\sin \gamma_k}, \qquad \text{where} \qquad \gamma_k = k(\frac{\pi}{2n}), k = 1, 2, \ldots, n. \tag{45}$$

The matrix $\mathcal{Q}$ is a diagonal matrix with

$$\mathcal{Q}_{kk} = -\cot \gamma_k. \tag{46}$$

Then

$$\alpha_{\max} = \min_k \left\{ R \frac{\cos \gamma_k}{\sin^2 \gamma_k} \right\}. \tag{47}$$

As an example, for a case with $R = 1$ and $n = 4$, the nominal arc capacities c^n and the matrix $\mathcal{Q}$ are

$$c^n = \begin{pmatrix} 1.0353 \\ 1.1547 \\ 1.4142 \\ 2.0000 \\ 3.8637 \\ \infty \end{pmatrix}, \mathcal{Q} = \begin{pmatrix} -0.2679 & 0 & 0 & 0 & 0 & 0 \\ 0 & -0.5774 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1.0000 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1.7321 & 0 & 0 \\ 0 & 0 & 0 & 0 & -3.7321 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{48}$$

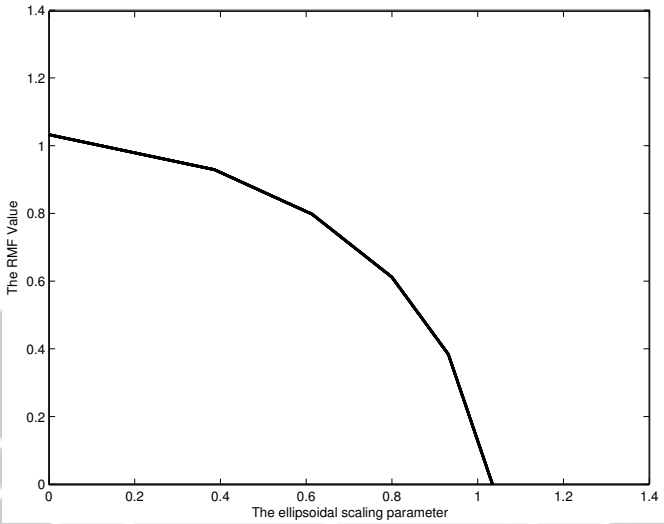


Figure 4: Optimal value function.

In Figure 4, we see that there are four breakpoints and five linearity intervals for $\alpha \in [0, 1.0353]$.

In the next subsection we discuss a special case of ellipsoid, i.e the maximum volume ellipsoid inscribing a box.

3.6 A special case: the maximum volume ellipsoid inscribing a box

In this subsection we discuss a special case of ellipsoid, i.e., the maximum volume ellipsoid inscribing a box. The box uncertainty set in (18) and the ellipsoidal uncertainty set in (25) will be denoted by $\mathcal{I}$ and $\mathcal{E}$.

Theorem 3.13 *The ellipsoid* $\mathcal{E}$ *given by*

$$\mathcal{E} = \{c \,:\, c = c^n + \mathcal{Q}w, \|w\| \leq 1\} \tag{49}$$

is the maximum volume ellipsoid inscribing a box $\mathcal{I}$ *of form*

$$\mathcal{I} = \{c \,:\, \ell \leq c \leq u\}, \tag{50}$$

if

$$c^n = \left(\frac{u+\ell}{2}\right), \quad \textit{and} \quad \mathcal{Q} = \operatorname{diag}\left(\frac{u-\ell}{2}\right). \tag{51}$$

Proof: The ellipsoid $\mathcal{E}$ is contained in $\mathcal{I}$ if and only if

$$\ell_a \leq (c^n)_a + (\mathcal{Q}_a)^T w \leq u_a, \forall a \in \mathcal{A}, \forall w \,:\, \|w\| \leq 1. \tag{52}$$

This is equivalent to

$$\ell_a \leq c_a^n - \|\mathcal{Q}_a\| \quad \text{and} \quad c_a^n + \|\mathcal{Q}_a\| \leq u_a, \tag{53}$$

hence the following holds

$$\|\mathcal{Q}_\alpha\| \leq \frac{u_a - \ell_a}{2}, \forall a \in \mathcal{A}. \tag{54}$$

Consider that from (53), we have that

$$\|\mathcal{Q}_a\| \leq u_a - c_a^n, \tag{55}$$
$$\|\mathcal{Q}_a\| \leq c_a^n - \ell_a. \tag{56}$$

This shows us that $\|\mathcal{Q}_\alpha\|$ is maximal if $u_a - c_a^n = c_a^n - \ell_a$, in which case we have

$$c_a^n = \frac{u_a + \ell_a}{2}. \tag{57}$$

Now, to show that volume $\mathcal{E}$ is maximal when $\mathcal{Q} = \text{diag}\left(\frac{u_a - \ell_a}{2}\right)$, let $\mathcal{I}$ be a full dimensional box uncertainty. This implies that $\mathcal{E}$ is full dimensional as well. Thus, $\text{rank}(\mathcal{Q}) = |\mathcal{A}|$. This means that $\mathcal{Q}^{-1}$ exists. This implies that $\forall c \in \mathcal{E}$ the following holds

$$w = (c - c^n)\mathcal{Q}^{-1}. \tag{58}$$

Thus, $c \in \mathcal{E}$ if and only if $\|w\| \leq 1$, which is equivalent to

$$\left\|(c - c^n)\mathcal{Q}^{-T}\mathcal{Q}^{-1}(c - c^n)\right\| \leq 1. \tag{59}$$

The volume of this ellipsoid is inverse proportional to $\det(\mathcal{Q}^{-T}\mathcal{Q}^{-1})$, so it is proportional to $\det \mathcal{Q}^2$. By the *Hadamard inequality* for the determinant, we have that

$$\det(\mathcal{Q}) \leq \|\mathcal{Q}_1\|_2 \|\mathcal{Q}_2\|_2 \|\mathcal{Q}_3\|_2 \cdots \left\|\mathcal{Q}_{|\mathcal{A}|}\right\|_2. \tag{60}$$

The equality hold if and only if $\mathcal{Q}_i \perp \mathcal{Q}_j$ for $i \neq j$. This implies that $\mathcal{E}$ is the maximum-volume ellipsoid contained in $\mathcal{I}$ if and only if

$$\mathcal{Q} = \text{diag}\left(\frac{u - \ell}{2}\right). \tag{61}$$

□

For the RMFP with ellipsoid $\mathcal{E}$, we have the following result.

Theorem 3.14 *Let $\mathcal{E}$ be the maximum volume ellipsoidal inscribing the box $\mathcal{I}$ with c^n and $\mathcal{Q}$ as stated in* (51)*. Then the RMFP with ellipsoid uncertainty set $\mathcal{E}$ and the RMFP with box uncertainty set $\mathcal{I}$ have the same solution.*

Proof: Consider the RMFP with ellipsoidal uncertainty set $\mathcal{E}$ with c^n and $\mathcal{Q}$ as stated in (51). In this case the flow on arc a has form

$$x_a \leq c_a^n - \|\mathcal{Q}_a\| = \frac{u_a + \ell_a}{2} - \frac{u_a - \ell_a}{2} = \ell_a. \tag{62}$$

Thus, the RMFP with ellipsoid $\mathcal{E}$ merely is the RMFP with box $\mathcal{I}$. □

4 Conclusions

In all considered cases, the RMFP merely is the usual maximum flow problem with modified arc capacities. In the case of box uncertainty the capacity of each arc is its lowest value in the box. In the case of ellipsoidal uncertainty, the capacity of each arc a is $c_a^n - \|\mathcal{Q}_a\|$ where c_a^n is the nominal arc capacity and $\mathcal{Q}_a$ is the column of $\mathcal{Q}$ corresponding to a. We showed that in the parametric case the RMF value is a piecewise monotonically decreasing linear concave function of nonnegative parameter α.

Acknowledgement

We are grateful to Arkadi Nemirovski for his valuable suggestion. The financial support was provided by the Royal Netherlands Academy of Arts and Sciences in the framework of the Scientific Programme Indonesia - Netherlands (SPIN).

References

[1] A. Ben-Tal and A. Nemirovski. (1999) Robust solutions of uncertain linear programs. Oper. Res. Lett., **25(1)**1–13.

[2] A. Ben-Tal and A. Nemirovski. (2001) Lectures on Modern Convex Optimization. Analysis, Algorithms and Engineering Applications. Volume 1 of MPS/SIAM Series on Optimization. SIAM, Philadelphia,USA.

[3] A. Ben-Tal and A. Nemirovski. (2002) Robust optimization—methodology and applications. Math. Program, **92**(3, Ser. B): 453–480.

[4] Laurence A.Wolsey (1998). Integer Programming. ISBN 0-471-28366-5. John Wiley & Sons.

[5] Alexander Schrijver (2003). Combinatorial Optimization : Polyhedra and Efficiency. ISBN 3-540-44389-4. Springer Verlag Berlin Heidelberg.

[6] C. Roos, T. Terlaky and J. Ph. Vial (2005). Interior Point Methods for Linear Optimization. Willey & Son New York.

D. Chaerani: PhD student at Department of Software Technology, Delft University of Technology, The Netherlands.
Department of Mathematics, Universitas Padjadjaran, Indonesia.
E-mail: d.cherani@ewi.tudelft.nl

C. Roos: Department of Software Technology, Delft University of Technology, Mekelweg 4 2628 CD Delft, The NetherlandsE-mail: c.roos@ewi.tudelft.nl

TESTING FOR STOCHASTIC DOMINANCE

R. Zitikis

University of Western Ontario, London, Canada

Abstract. Researchers and practitioners are frequently interested in comparing various random variables. These can, for example, be loss variables in Actuarial Science, incomes in Econometrics, survival times in Medical Sciences. Various notions of stochastic comparison, usually called stochastic dominance, have been proposed in the literature and used in practice. Among them are, for example, marginal conditional stochastic dominance, Lorenz dominance, second and higher order stochastic dominance. In this paper we discuss a number of such notions and also show how to develop test statistics and the corresponding statistical inferential theory for them. Non-parametric and parametric approaches are considered in detail, including bootstrap based techniques for estimating critical values.

Key-words: marginal conditional stochastic dominance, concentration curve, Lorenz curve, Gini index, L-statistic, confidence interval, hypothesis test, non-parametric statistics, parametric statistics, bootstrap, asymptotic distribution.

1 Introduction

Let $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_n\}$ be a portfolio with n assets. Let $\sum_{i=1}^n \alpha_i = 1$, which means 100%. Let r_i be the rate of return on asset i. Then the portfolio rate of return is $P = \sum_{i=1}^n \alpha_i r_i$. Denote the distribution function of P by F_P, and define (cf. [33]) the absolute concentration curve (ACC) by

$$A_i(t) = \mathbf{E}\left[r_i \mathbf{1}\left\{P \le F_P^{-1}(t)\right\}\right], \quad 0 < t < 1.$$

We are interested if the ith asset is dominated by the jth asset (cf. [33], [34]). This introduces the notion of marginal conditional stochastic dominance (MCSD) that can, for example, be formulated in terms of the hypotheses

$$\begin{aligned} H_0 &: \ A_i(t) \le A_j(t) \quad \forall\, t \in (0,1), \\ H_1 &: \ A_i(t) > A_j(t) \quad \exists\, t \in (0,1). \end{aligned}$$

Naturally, testing the above hypotheses is based on empirical estimators of the two ACCs. We shall next discuss a non-parametric approach to solving this problem.

Let X be a random variable with cdf F; we write this as $X \sim F$. Furthermore, let $Y \sim G$ be a random variable with finite first moment $\mathbf{E}[Y]$. The ACC is

$$A_{F,G}(t) = \mathbf{E}\left[Y \mathbf{1}\left\{X \le F^{-1}(t)\right\}\right], \quad 0 < t < 1.$$

We shall later use the fact that if the cdf F is continuous, then (cf., e.g., [30])

$$A_{F,G}(t) = \int_0^t \mu_{Y|X}(F^{-1}(u))\, d\, u. \tag{1}$$

where $\mu_{Y|X}$ is the regression function defined by $\mu_{Y|X}(x) = \mathbf{E}[Y|X = x]$. In the special case when $Y = X$, we have $\mu_{X|X}(x) = x$ and thus

$$A_{F,F}(t) = \int_0^t F^{-1}(u)\, d\, u = ALC_F(t).$$

The function ALC_F, defined by the right-most equality above, is called the absolute Lorenz curve (ALC). Note in passing that since the ALC is a special case of the ACC, the following set of hypotheses (defining the absolute Lorenz dominance) is a special case of those in the MCSD context: the null hypothesis is $ALC_F(t) \le ALC_G(t)\ \forall t \in (0,1)$, and the alternative is $ALC_F(t) > ALC_G(t)\ \exists t \in (0,1)$. We next discuss how to construct empirical estimators for the ACC and then explain an econometric meaning of the ALC. This will in turn give us an insight into the econometric meaning of the ACC.

The idea of constructing an empirical ACC can be presented as follows (cf., e.g, [30], [31])

$$A_{F,G}(t) \approx \frac{1}{n}\sum_{i=1}^{n} Y_i \mathbf{1}\{X_i \le F^{-1}(t)\} \approx \frac{1}{n}\sum_{i=1}^{n} Y_i \mathbf{1}\{X_i \le F_n^{-1}(t)\} = \widehat{A}_{F,G}(t)$$

with the right-most equality defining the estimator $\widehat{A}_{F,G}$ of $A_{F,G}$. For the sake of practical implementation it is useful to note that $\widehat{A}_{F,G}(t)$ equals $\frac{1}{n}\sum_{i=1}^{k} Y_{(i)}$ for all $t \in ((k-1)/n, k/n]$, where $Y_{(1)}, \dots, Y_{(n)}$ are the induced order statistics of $Y_1, \dots, Y_n$ corresponding to $X_{1:n} \le \cdots \le X_{n:n}$. Next, we discuss the promised econometric meaning of the ALC.

Let $X_1,\ X_2,\ \cdots,\ X_n$ be the incomes of n individuals. Order the incomes, which gives the order statistics $X_{1:n} \le X_{2:n} \le \cdots \le X_{n:n}$. Connect the points $(k, \sum_{i=1}^{k} X_{i:n})$, $k = 0, 1, \dots, n$, using straight lines, which gives the function $\sum_{i=1}^{[x]} X_{i:n} + (x - [x])X_{[x]+1:n}$. The function has the domain of definition $[0, n]$, and it is therefore convenient to re-define it so that its domain of definition would be $[0, 1]$. This gives the function $\sum_{i=1}^{[nt]} X_{i:n} + (nt - [nt])X_{[tn]+1:n}$. Divide the latter formula by n and note the equalities

$$\frac{1}{n}\sum_{i=1}^{[nt]} X_{i:n} + \left(\frac{nt - [nt]}{n}\right) X_{[nt]+1:n} = \int_0^t F_n^{-1}(s)\, d\, s = ALC_n(t),$$

where the right-most equality defines the absolute Lorenz curve ALC_n corresponding to the (empirical) cdf F_n. Hence, the ALC can be used to describe the distribution of incomes in a population, or in a sample, depending on the problem at hand (cf., e.g., [7], [8], [9], [30], and references therein). On the topic, we also refer to [3] and [13] for a number of other interesting applications and theoretical results concerning absolute and relative LCs.

2 Estimating the ACC and testing for the MCSD

Before constructing a test for the MCSD, we need to investigate asymptotic properties of the empirical ACC $\widehat{A}_{F,G}$ over the entire domain of definition $(0,1)$. As the first step toward this goal, we work out our intuition on the topic by looking at the asymptotic behaviour of $\widehat{A}_{F,G}(t)$ for any fixed $t \in (0,1)$. Assume that the cdf F be continuous. Furthermore, assume that $t \in (0,1)$ is a continuity point of the function $\mu_{Y|X}(F^{-1}(t))$. Then (cf. [5], [29])

$$\sqrt{n}\big(\widehat{A}_{F,G}(t) - A_{F,G}(t)\big) \to_d \mathcal{N}(0, v^2_{F,G}(t)),$$

where the asymptotic variance $v^2_{F,G}(t)$ is given by

$$v^2_{F,G}(t) = \mu^2_{Y|X}(F^{-1}(t))t(1-t) + \int_0^t \sigma^2_{Y|X}(F^{-1}(u))du + \int_0^t \mu^2_{Y|X}(F^{-1}(u))du - \left(\int_0^t \mu_{Y|X}(F^{-1}(u))du\right)^2$$

with the notation $\sigma^2_{Y|X}(x) = \mathbf{Var}[Y|X = x]$. In a standard way we now derive the asymptotic $(1-\alpha)100\%$ confidence interval (CI) for $A_{F,G}(t)$:

$$\widehat{A}_{F,G}(t) \pm z_{\alpha/2}\frac{v_{F,G}(t)}{\sqrt{n}}$$

with the notation $z_{\alpha/2} = \Phi^{-1}(1 - \frac{\alpha}{2})$. Of course, the above CI is not readily applicable in practice because $v_{F,G}(t)$ depends on the (unknown) cdfs F and G. Since the formula of $v_{F,G}(t)$ is fairly complex, instead of finding a plug-in estimator we suggest using a bootstrap approximation. This leads to the following CI for $A_{F,G}(t)$:

$$\widehat{A}_{F,G}(t) \pm \frac{z^*_\alpha}{\sqrt{n}},$$

where z^*_α is defined as follows. Given the original n observations, which are independent pairs $(Y_1, X_1), \ldots, (Y_n, X_n)$, we sample with replacement from the n pairs and obtain n new pairs, which we denote by $(Y^*_1, X^*_1), \ldots, (Y^*_n, X^*_n)$. Using these new pairs, we construct the corresponding ACC, which we denote by $\widehat{A}^*_{X,X}(t)$. Next, we calculate the quantity (note the absolute value)

$$\sqrt{n}\big|\widehat{A}^*_{X,X}(t) - \widehat{A}_{F,G}(t)\big| \tag{2}$$

and then repeat the whole procedure M times. Finally, we define z^*_α as the smallest real number z such that the proportion of the obtained M values of the quantity in (2) is at least $1-\alpha$.

Now we construct confidence bands (CBs) for the ACC. Naturally, instead of pointwise limit results, we now need the corresponding ones that hold uniformly over

the entire interval $(0,1)$. That is, we need to establish the asymptotic behaviour of the quantity

$$\sqrt{n}\sup_{0<t<1}|\widehat{A}_{F,G}(t)-A_{F,G}(t)|.$$

This can be achieved by proving weak convergence of the stochastic process $\sqrt{n}(\widehat{A}_{F,G}-A_{F,G})$, and it appears (cf. [5], [29]) that it converges to the Gaussian process $\Gamma_{F,G}$ defined by

$$\Gamma_{F,G}(t)=-\int_{(0,t]}\mathcal{B}(u)\,d\,\mu_{Y|X}(F^{-1}(u))+\int_{(0,t]}\sigma^2_{Y|X}(F^{-1}(u))\,d\,\mathcal{W}(u),$$

where the Brownian bridge $\mathcal{B}$ and the Wiener process $\mathcal{W}$ are independent. Of course, keeping in mind the conditions that we have imposed for the validity of the point-wise results, we now require the continuity of F and $\mu_{Y|X}(F^{-1}(\cdot))$ over their respective domains of definition. In addition, we also assume that there are constants $0<a<\frac{1}{2}$ and $c<\infty$ such that $t^a(1-t)^a|\mu_{Y|X}(F^{-1}(t))|\le c$ for all $t\in(0,1)$. Under these conditions, we formulate the following CB for $A_{F,G}$:

$$\widehat{A}_{F,G}(t)\pm\frac{z^*_\alpha}{\sqrt{n}},\quad 0<t<1,$$

where the bootstrap-based z^*_α is defined along the corresponding lines above but this time with $\sqrt{n}\sup_t|\widehat{A}^*_{X,X}(t)-\widehat{A}_{F,G}(t)|$ instead of quantity (2).

We are now ready to discuss testing the hypotheses (MCSD)

$$\begin{aligned}H_0:&\ A_{F,G}(t)\le A_{F,H}(t)\quad\forall\,t\in(0,1),\\ H_1:&\ A_{F,G}(t)>A_{F,H}(t)\quad\exists\,t\in(0,1).\end{aligned}$$

Note that under the null hypothesis H_0 the supremum of the difference $A_{F,G}(t)-A_{F,H}(t)$ over all $t\in(0,1)$ is non-positive, whereas under the alternative H_1 the supremum is strictly positive. Based on these observations, we now construct a test statistic. Let $(Z_1,Y_1,X_1),\dots,(Z_n,Y_n,X_n)$ be n independent triplets with $X_i\sim F$, $Y_i\sim G$, and $Z_i\sim H$. Furthermore, let $\widehat{A}_{F,G}(t)$ be the empirical ACC based on the pairs $(Y_1,X_1),\dots,(Y_n,X_n)$, and let $\widehat{A}_{F,H}(t)$ be the empirical ACC based on $(Z_1,X_1),\dots,(Z_n,X_n)$. The test statistic for H_0 vs H_1 is given by

$$S_{n,m}=\sqrt{\frac{nm}{n+m}}\sup_{0<t<1}\big(\widehat{A}_{F,G}(t)-\widehat{A}_{F,H}(t)\big).$$

Note that under the null hypothesis H_0 we have the bound

$$S_{n,m}\le\sqrt{\frac{nm}{n+m}}\sup_{0<t<1}\Big(\big(\widehat{A}_{F,G}(t)-A_{F,G}(t)\big)-\big(\widehat{A}_{F,H}(t)-A_{F,H}(t)\big)\Big)$$

with the quantity on the right-hand side having a non-degenerate limiting distribution under both the null and the alternative hypotheses when the sample sizes n and m tend to infinity so that $m/(n+m)\to\eta$ for a constant $\eta\in(0,1)$. Under

the alternative H_1, we have $S_{n,m} \to \infty$, which gives the desired performance of the test statistic $S_{n,m}$. For calculating the critical values of the test, we employ bootstrap along the lines above but with the following quantity instead of (2):

$$\sqrt{\frac{nm}{n+m}} \sup_{0<t<1} \left| \left(\widehat{A}^*_{F,G}(t) - \widehat{A}_{F,G}(t)\right) - \left(\widehat{A}^*_{F,H}(t) - \widehat{A}_{F,H}(t)\right) \right|$$

In some circumstances it might be reasonable to assume certain parametric families and then develop the corresponding statistical tests about the ACC either pointwise or uniformly over the interval $(0,1)$. As an example, assume that $Y = G_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X = G_1 + G_2$, where $G_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Since the cdf of X is continuous, we use formula (1) to calculate the ACC and notice that $\mu_{Y|X}(x) = a + bx$ with the coefficients

$$a = \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X \quad \text{and} \quad b = \rho\frac{\sigma_Y}{\sigma_X}.$$

It is easy to check that $b = (\sigma_Y/\sigma_X)^2$ and $\int_0^t \Phi^{-1}(u)du = -\phi(\Phi^{-1}(t))$, which are the two formulas needed to make the derivation of the the following equality straightforward:

$$A_{F,G}(t) = \mu_1 t - h(\sigma_1^2, \sigma_2^2)\phi\left(\Phi^{-1}(t)\right) \quad \text{with} \quad h(x,y) = \frac{x}{\sqrt{x+y}}.$$

The maximum likelihood estimators of μ_1, σ_1^2, and σ_2^2 are, respectively,

$$\overline{G}_1 = \frac{1}{n}\sum_{i=1}^n G_{1,i}, \quad s_1^2 = \frac{1}{n}\sum_{i=1}^n \left(G_{1,i} - \overline{G}_1\right)^2, \quad s_2^2 = \frac{1}{n}\sum_{i=1}^n \left(G_{2,i} - \overline{G}_2\right)^2.$$

Hence, the parametric estimator of $A_{F,G}(t)$ is

$$\widehat{A}_{F,G}(t) = \overline{G}_1 t - h(s_1^2, s_2^2)\phi\left(\Phi^{-1}(t)\right).$$

Construction of CIs is now based on establishing the limiting distribution of

$$\sqrt{n}\left(\widehat{A}_{F,G}(t) - A_{F,G}(t)\right), \tag{3}$$

whereas construction of CBs can be based - depending on desired results - on any of the following two statements:

$$\sqrt{n}\sup_{0<t<1}\left|\widehat{A}_{F,G}(t) - A_{F,G}(t)\right|, \quad \sqrt{n}\sup_{0<t<1} w(t)\left|\widehat{A}_{F,G}(t) - A_{F,G}(t)\right|. \tag{4}$$

This program of research can be developed with the help of the delta method or, simply, using the Taylor formula. Take also into account that $s_1^2 - \sigma_1^2$ equals $n^{-1}\sum_{i=1}^n \xi_{1,i} - (\overline{G}_1 - \mu_1)^2$, where we have used the notation $\xi_{1,i} = (G_{1,i} - \mu_1)^2 - \sigma_1$. Analogous representation holds for $s_2^2 - \sigma_2^2$ with the obvious definition of $\xi_{2,i}$. The following asymptotic result is now easy:

$$\sqrt{n}\left(\widehat{A}_{F,G}(t) - A_{F,G}(t)\right) \to_d \mathcal{N}(0, \mathbf{Var}[\eta(t)]),$$

where $\eta(t) = (G_{1,1} - \mu_1)t - (\xi_{1,1}h'_x(\sigma_1^2, \sigma_2^2) + \xi_{2,1}h'_y(\sigma_1^2, \sigma_2^2))\phi(\Phi^{-1}(t))$. Write the variance $\mathbf{Var}[\eta(t)]$ as $H(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ and in this way define the function H. We obtain the following CI for $A_{F,G}(t)$:

$$\widehat{A}_{F,G}(t) \pm z_{\alpha/2}\frac{\sqrt{H(\overline{G}_1, \overline{G}_2, s_1^2, s_2^2)}}{\sqrt{n}}.$$

Instead of the above plug-in estimator of the variance, we can use bootstrap to estimate the margin of error, which we do next. Namely, using simple random sampling, we generate $G^*_{1,1}, \ldots, G^*_{1,n}$ from the original sample $G_{1,1}, \ldots, G_{1,n}$. Likewise, from $G_{2,1}, \ldots, G_{2,n}$ we generate $G^*_{2,1}, \ldots, G^*_{2,n}$. Next we calculate

$$\sqrt{n}|\widehat{A}^*_{F,G}(t) - \widehat{A}_{F,G}(t)| \tag{5}$$

and then repeat the procedure M times. Having thus obtained M values of quantity (5), we define z^*_α as the smallest number z such that at least $100(1-\alpha)\%$ of the proportion of those M values that are at or below z is at least $1-\alpha$. This gives the asymptotic $100(1-\alpha)\%$ CI for $A_{F,G}(t)$:

$$\widehat{A}_{F,G}(t) \pm z^*_\alpha\frac{1}{\sqrt{n}}.$$

The above arguments can easily be adapted for deriving limiting distributions of the two suprema in (4) as well.

3 Absolute Lorenz dominance and related notions

Let $X \geq 0$ and $Y \geq 0$ be two random variables with cdfs F and G respectively. If X tends to take on larger values than Y, then the cdf F should tend to be below the cdf Y. This leads to the notion of first order stochastic dominance: $F(x) \leq G(x)\ \forall\ x \geq 0$. Comparing the corresponding areas under the two cdfs F and G leads to the notion of second order stochastic dominance (SSD), meaning that $\int_0^x F(y)dy \leq \int_0^x G(y)dy\ \forall\ x \geq 0$ (cf., e.g., [2], [4], [10], [11], [16], [20], [21], [22], [35], and references therein). The latter inequality can written as

$$\Delta_F(x) \leq \Delta_G(x) \quad \forall\ x \geq 0,$$

where we have used the notation $\Delta_F(x) = \int_0^x F(y)dy$. Simple geometrical arguments show that the SSD is equivalent to the absolute Lorenz dominance (ALD), which is defined as $\int_0^t F^{-1}(s)ds \geq \int_0^t G^{-1}(s)ds\ \forall\ 0 < t < 1$ or, equivalently,

$$ALC_F(t) \geq ALC_G(t) \quad \forall\ 0 < t < 1,$$

with the already introduced notation for the ALC. Since ALD and SSD are equivalent, and since the latter notion is somewhat easier to tackle from the technical point of view, we formulate the following hypotheses

$$\begin{aligned} H_0 &: \ \Delta_F(x) \leq \Delta_G(x) \quad \forall\ x \geq 0, \\ H_1 &: \ \Delta_F(x) > \Delta_G(x) \quad \exists\ x \geq 0. \end{aligned}$$

Define a test statistic by

$$T_{n,m} = \sup_{x \geq 0} V_{n,m}(x) \quad \text{with} \quad V_{n,m}(x) = \sqrt{\frac{nm}{n+m}}\big(\widehat{\Delta}_F(x) - \widehat{\Delta}_G(x)\big),$$

where $\widehat{\Delta}_F(x) = \int_0^x \widehat{F}(y)dy$ with the empirical cdf $\widehat{F}$ based on n observations of X, and with $\widehat{G}$ based on m observations of Y. We assume that the Xs and Ys are all independent. Under the null hypothesis H_0 and when $m/(n+m) \to \eta \in (0,1)$, we have

$$T_{n,m} \leq \widetilde{T}_{n,m} \to_d \sup_{x \geq 0} \Gamma(x).$$

In the statement above,

$$\widetilde{T}_{n,m} = \sup_{x \geq 0} \left(V_{n,m}(x) - \sqrt{\frac{nm}{n+m}}\big(\Delta_F(x) - \Delta_G(x)\big) \right),$$

and Γ is a Gaussian process. It is worthwhile noting that since we are later going to use a bootstrap approximation to estimate critical values, we do not really need to know the Gaussian process Γ explicitly, except that we want to make sure that the process is well defined and non-degenerate. The validity of these facts can be derived from the hint

$$\sqrt{\frac{nm}{n+m}}\big(\widehat{\Delta}_F(x) - \Delta_F(x)\big) \Rightarrow \sqrt{\eta} \int_0^x \mathcal{B}(F(y))dy,$$

where we assume of course that the second moment of X is finite, as well as the second moment of Y. Finally we note that under the alternative H_1, the test statistic $T_{n,m}$ tends to infinity when n and m tend to infinity in the fashion specified above. Hence, the test statistic $T_{n,m}$ separates the null hypothesis from the alternative. We shall now modify the null hypothesis H_0 and in this way introduce a new 'twist' into our discussion.

Suppose that we are interested in testing whether the distribution of incomes during a year "F" changed if compared to a year "G" so that $\Delta_F(x) > \Delta_G(x)$ for some $x \geq 0$. This leads to the formulation of the hypotheses (cf. [16])

$$\begin{aligned} H_0^{\text{eq}} &: \; F(x) = G(x) \quad \forall\, x \geq 0, \\ H_1 &: \; \Delta_F(x) > \Delta_G(x) \quad \exists\, x \geq 0. \end{aligned}$$

Under the null hypothesis H_0^{eq} we have

$$T_{n,m} \to_d \sup_{x \geq 0} \Gamma^{\text{eq}}(x),$$

where the Gaussian process Γ^{eq} is defined by $\Gamma^{\text{eq}}(x) = \int_0^x \mathcal{B}(F(y))dy$. In order to see why the latter limiting process appears under H_0^{eq}, we write the equality (in distribution)

$$\sqrt{\eta} \int_0^x \mathcal{B}_1(F(y))dy + \sqrt{1-\eta} \int_0^x \mathcal{B}_2(F(y))dy =_d \int_0^x \mathcal{B}(F(y))dy.$$

Coming back to our main discussion, we note that since the limiting distribution of $T_{n,m}$ depends on the (unknown) cdf F, we use a bootstrap approximation to estimate critical values of the test (cf. [16]). The null hypothesis $H_0^{\rm eq}$ allows us to pool the Xs and the Ys, which we do and arrive at the pooled cdf $F_{\rm pool} = \frac{n}{n+m}\widehat{F} + \frac{m}{n+m}\widehat{G}$. Next, we generate two samples, $X_1^*, \ldots, X_n^*$ and $Y_1^*, \ldots, Y_m^*$, from the (same) pooled cdf $F_{\rm pool}$. Denote the empirical cdfs based on the first and the second new samples by $\widehat{F}^*$ and $\widehat{G}^*$ respectively. Define

$$T_{n,m}^* = \sqrt{\frac{nm}{n+m}} \sup_{x\ge 0} \big(\widehat{\Delta}_F^*(x) - \widehat{\Delta}_G^*(x)\big).$$

With the conditional cdf $L^*(z) = \mathbf{P}[T_{n,m}^* \le z \,|\, \text{original } X\text{s and } Y\text{s}\,]$ we define an estimator of z_α by $z_\alpha^* = \inf\{z \ge 0 :\ L^*(z) \ge 1-\alpha\}$. It now remains to note that $\mathbf{P}[T_{n,m} > z_\alpha^*] \to \alpha$ under $H_0^{\rm eq}$, and $\mathbf{P}[T_{n,m} > z_\alpha^*] \to 1$ under the alternative H_1.

4 Gini-type indices and their estimation

We have noted already that there is an interest in comparing ALCs and, more generally, ACCs. The comparison, however, might in a sense not be as strict as we have considered above. Indeed, we might only wish to compare areas under the curves, which is in the spirit of our reasoning encountered earlier when making a transition from the first to the second order SD. To continue the discussion, choose as an example ALC_F. From the definition of the curve we see that ALC_F is convex, with $ALC_F(0) = 0$ and $ALC_F(1) = \mu_F$. Hence, we can consider the area between the ALC_F and $t \mapsto t\mu_F$, as a measure of inequality among the X values. To see this more transparently, note that if $X = c$ for a constant $c > 0$, then ALC_F is exactly the straight line $t \mapsto t\mu_F$, and thus the area between the two curves is zero. This leads to the definition of the *absolute* weighted Gini index (cf., e.g., [42], [43], and references therein)

$$AG_F(w) = \frac{1}{b(w)} \int_0^1 (t\mu_F - ALC_F(t))\, w(t)dt,$$

where $w : (0,1) \to (0,\infty)$ is a function chosen by the researcher and such that the constant $b(w) = \int_0^1 tw(t)dt$ is finite and strictly positive. Note that if $w(t) \equiv 1$, then $AG_F(w)$ is the (classical) absolute Gini index AG_F, and if $w(t) = (1-t)^{\nu-2}$, then $AG_F(w)$ is the absolute S-Gini index $AG_{F,\nu}$ (cf., e.g., [44], and references therein). Note also that if we divide $AG_F(w)$ by μ_F, then we get the relative weighted Gini index $RG_F(w) = \mu_F^{-1} AG_F(w)$. Certainly, we assume that $\mu_F > 0$. Note that $RG_F(w) \in [0,1]$, with $RG_F(w) = 0$ in the case of "perfect equality" and $RG_F(w) = 1$ in the case of "extreme inequality".

In the definitions of absolute and relative (weighted) Gini indices we have two ingredients that make the indices dependent on the population cdf F: first, the mean μ_F, and second, the absolute Lorenz curve ALC_F. Hence, in order to construct empirical estimators for the indices, we need to construct estimators for μ_F

and ALC_F. It is a standard problem to estimate the mean, and we have already discussed how to estimate ACCs. This essentially solves the problem. There is, however, a better way to approach the problem, and it employs L-statistics. To see this, we define a function ψ by the equation $\psi(t) = \int_t^1 w(s)ds$, notice that $w(t)$ equals $-\psi'(t)$, then integrate by parts, and finally arrive at the equality

$$AG_F(w) = \mu_F - \frac{1}{b(w)} \int_0^1 F^{-1}(t)\psi(t)dt.$$

Here again we have two quantities to estimate: the first one is the mean $\mu_F = \int_0^1 F^{-1}(t)dt$ and second one is $\int_0^1 F^{-1}(t)\psi(t)dt$, which we denote by L_F. (Note that the mean μ_F is L_F with $\psi(t) \equiv 1$.) We estimate L_F using the L-statistic (i.e., linear combination of order statistics)

$$\widehat{L}_F = \sum_{i=1}^{n} \left(\int_{(i-1)/n}^{i/n} \psi(s)ds \right) X_{i:n},$$

which has a well understood asymptotics (cf. [14], [32], [37], [38], and references therein). In addition to the econometric context that we have hinted at above, the quantity L_F also appears in the actuarial literature (cf. [17], [18], [19], [41]).

Assumption. Let ψ be continuous on $(0, 1)$, and let there exist α, $\beta > 1/2$ and $c < \infty$ such that $|\psi(s)| \le cs^{\alpha-1}(1-s)^{\beta-1}$ for all $0 < s < 1$. Furthermore, assume that $\mathbf{E}[|X|^\gamma] < \infty$ for some $\gamma > 1/(\alpha - 1/2)$ and $\gamma > 1/(\beta - 1/2)$.

Under Assumption, it is well known (cf., e.g., [14], [32], [37], [38], [39], [42]) that

$$\sqrt{n}(\widehat{L}_F - L_F) \to_d \mathcal{N}(0, Q_{F,F}(\psi)),$$

where

$$Q_{F,F}(\psi) = \iint \left(F(x \wedge y) - F(x)F(y)\right) \psi(F(x))\psi(F(y))dxdy.$$

Replacing F by its empirical estimator F_n in the formula above, we have under Assumption that

$$\widehat{Q}_{F,F}(\psi) = \sum_{j,k=1}^{n-1} \left(\frac{j}{n} \wedge \frac{k}{n} - \frac{j}{n}\frac{k}{n} \right) \psi\left(\frac{j}{n}\right) \psi\left(\frac{k}{n}\right) (X_{j+1:n} - X_{j:n})(X_{k+1:n} - X_{k:n})$$

is a consistent estimator of $Q_{F,F}(\psi)$. Hence, using Slutsky's arguments, we obtain

$$\frac{\sqrt{n}\ (\widehat{L}_F - L_F)}{\sqrt{\widehat{Q}_{F,F}(\psi)}} \to_d \mathcal{N}(0, 1),$$

which can be used for testing hypothesis about L_F and also for constructing asymptotic CIs for L_F.

Since quadratic forms usually converge slowly, instead of the above empirical variance we can use a bootstrap approximation to construct CIs. Namely, we have the following $100(1-\alpha)\%$ CI for L_F

$$\widehat{L}_F \pm z^*_\alpha \frac{1}{\sqrt{n}},$$

where z^*_α is defined as follows. From the original sample $X_1, \ldots, X_n$ we sample with replacement n values $X^*_1, \ldots, X^*_n$. Then we calculate the corresponding risk measure and denote it by $\widehat{L}^*_F$. Next we calculated the quantity

$$\sqrt{n}|\widehat{L}^*_F - \widehat{L}_F| \tag{6}$$

and repeat the procedure M times. Finally, we define z^*_α as the smallest value of z such that at least $100(1-\alpha)\%$ of the obtained M values of the quantity (6) are at or below z.

Because of various reasons the researcher might want to assume parametric families for modeling population distributions. As an example, consider the following three distributions (cf. [1], [18]):

- Exponential $F_1(x) = 1 - e^{-(x-x_0)/\theta}$, $x > x_0$, with parameter $\theta > 0$;
- Pareto $F_2(x) = 1 - (x_0/x)^\gamma$, $x > x_0$, with parameter $\gamma > 0$;
- Log-normal $F_3(x) = \Phi(\log(x - x_0) - \mu)$, $x > x_0$, with parameter $\mu \in \mathbf{R}$.

We also assume that $\psi(t) = r(1-t)^{r-1}$ for some parameter $r > 0.5$. Given the information above, the following formulas hold (cf. [1])

$$L_{F_1} = x_0 + \frac{\theta}{r}, \quad L_{F_2} = x_0 + \frac{x_0}{\gamma r - 1}, \quad L_{F_3} = x_0 + c_r e^\mu,$$

where we have assumed $\gamma > 1/r$ and defined $c_r = \int (1 - \Phi(z))^r e^z dz$. The corresponding empirical estimators $\widehat{L}_{F_1}$, $\widehat{L}_{F_2}$, and $\widehat{L}_{F_3}$ are defined by replacing the parameters θ, γ, and μ by their respective maximum likelihood estimators (MLEs) (cf., e.g., [1])

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n (X_i - x_0), \quad \widehat{\gamma} = \frac{1}{\frac{1}{n}\sum_{i=1}^n \log(X_i/x_0)}, \quad \widehat{\mu} = \frac{1}{n}\sum_{i=1}^n \log(X_i - x_0).$$

Using the delta method and the classical CLT, we obtain the statement

$$\frac{\sqrt{n}\,(\widehat{L}_F - L_F)}{\sqrt{Q_{F,F}(\psi)}} \to_d \mathcal{N}(0,1)$$

with the expressions

$$Q_{F_1,F_1}(\psi) = \frac{\theta^2}{r^2}, \quad Q_{F_2,F_2}(\psi) = \frac{r^2 x_0^2 \gamma^2}{(\gamma r - 1)^4}, \quad Q_{F_3,F_3}(\psi) = c_r^2 e^{2\mu}.$$

The empirical estimators $\widehat{Q}_{F_1,F_1}(\psi)$, $\widehat{Q}_{F_2,F_2}(\psi)$, and $\widehat{Q}_{F_3,F_3}(\psi)$ of the quantities above are obtained by replacing the parameters θ, γ, and μ by their respective MLEs (cf. [1], [18]). We can now construct CIs for, and also test various hypotheses about, L_F.

Parametric bootstrap based CIs for L_F can be constructed as follows. First we calculate $\widehat{\theta}$, $\widehat{\gamma}$, and $\widehat{\mu}$ using the original data. Then we simulate from the exponential distribution with $\widehat{\theta}$, Pareto with $\widehat{\gamma}$, and log-normal with $\widehat{\mu}$. From the new samples we calculate, respectively, $\widehat{\theta}^* = n^{-1}\sum_{i=1}^{n}(X_i^* - x_0)$, $\widehat{\gamma}^* = \ldots$, and $\widehat{\mu}^* = \ldots$ Then we calculate $\widehat{L}^*_{F_1} = x_0 + \widehat{\theta}^*/r$, $\widehat{L}^*_{F_2} = \ldots$, and $\widehat{L}^*_{F_3} = \ldots$. For any $F \in \{F_1, F_2, F_3\}$, we calculate

$$\sqrt{n}|\widehat{L}^*_F - \widehat{L}_F|. \tag{7}$$

Then we repeat the procedure M times and define z^*_α as the smallest value z such that at least $100(1-\alpha)\%$ of the obtained M values of quantity (7) are at or below z. The $100(1-\alpha)\%$ CI for L_F is

$$\widehat{L}_F \pm z^*_\alpha \frac{1}{\sqrt{n}}.$$

5 Comparing Gini-type indices

In addition to estimating indices individually, it is also of interest to compare them (cf. [18], [19], [26], [27], [28]). Consider comparing *two* Gini-type indices (cf. [18]).

Scenario I (independent samples; cf. [18]) Assume that: first, $X_1, X_2, \ldots, X_n$ are i.i.d. r.v.'s with cdf F and finite second moments; second, $Y_1, Y_2, \ldots, Y_m$ are i.i.d. r.v.'s with cdf G and finite second moments; and third, the r.v.'s $X_1, \ldots, X_n$, $Y_1, \ldots, Y_m$ are independent.

We already know that $\sqrt{n}(\widehat{L}_F - L_F) \to_d \mathcal{N}(0, Q_{F,F}(\psi))$. With this asymptotic results and also an analogous one for G, we obtain that

$$\frac{(\widehat{L}_F - \widehat{L}_G) - (L_F - L_G)}{\sqrt{\frac{Q_{F,F}(\psi)}{n} + \frac{Q_{G,G}(\psi)}{m}}} \to_d \mathcal{N}(0,1)$$

when $m/(n+m) \to \eta \in (0,1)$. The result holds in both non-parametric and parametric setups, and desired formulas for estimators and their asymptotic variances have already been derived above, when discussing the estimation of individual indices. Hence, we omit further details concerning Scenario I and concentrate on the case when the random variables Xs and Ys are paired.

Scenario II (paired samples; cf. [18]) Assume that: first, $X_1, X_2, \ldots, X_n$ are i.i.d. r.v.'s with cdf F and finite second moments; second, $Y_1, Y_2, \ldots, Y_n$ are i.i.d. r.v.'s with cdf G and finite second moments; and third, the pairs (X_i, Y_i), $i = 1, \ldots, n$, are independent but their coordinates X_i and Y_i might be dependent.

Due to dependence, the individual asymptotic results for $\widehat{L}_F$ and $\widehat{L}_G$ do not imply desired results for the difference $\widehat{L}_F - \widehat{L}_G$. Hence, we make a step back in our earlier considerations and start with the asymptotic representation

$$\sqrt{n}(\widehat{L}_F - L_F) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Delta_{F,\psi}(X_i) + o_{\mathbf{P}}(1)$$

and also with an analogous one for $\sqrt{n}(\widehat{L}_G - L_G)$. It is instructive to see how the above representation is derived:

$$\sqrt{n}(\widehat{L}_F - L_F) = \sqrt{n} \int_{-\infty}^{\infty} x \, d\left(\Psi(F_n(x)) - \Psi(F(x))\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Delta_{F,\psi}(X_i) + o_{\mathbf{P}}(1),$$

where we have used the integration-by-parts and the Taylor formulas, and also the notation

$$\Psi(t) = \int_0^t \psi(s)ds \quad \text{and} \quad \Delta_{F,\psi}(x) = -\int \left(\mathbf{1}\{x \le z\} - F(z)\right) \psi(F(z))dz.$$

Note that the variance of $\Delta_{F,\psi}(x)$ is the earlier defined quantity $Q_{F,F}(\psi)$. Hence, with the notation $\zeta_i = \Delta_{F,\psi}(X_i) - \Delta_{G,\psi}(Y_i)$ we have the representation

$$\sqrt{n}\big((\widehat{L}_F - \widehat{L}_G) - (L_F - L_G)\big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \zeta_i + o_{\mathbf{P}}(1).$$

We see that $\mathbf{E}[\zeta_i] = 0$ and $\mathbf{Var}[\zeta_i] = Q_{F,F}(\psi) + Q_{G,G}(\psi) - 2Q_{F,G}(\psi)$. Hence, the classical CLT implies

$$\frac{\sqrt{n}\big((\widehat{L}_F - \widehat{L}_G) - (L_F - L_G)\big)}{\sqrt{Q_{F,F}(\psi) + Q_{G,G}(\psi) - 2Q_{F,G}(\psi)}} \to_d \mathcal{N}(0,1).$$

Because of practical considerations, the denominator needs to be estimated. We already know that $\widehat{Q}_{F,F}(\psi)$ and $\widehat{Q}_{G,G}(\psi)$ are such estimators of $Q_{F,F}(\psi)$ and $Q_{G,G}(\psi)$, respectively. We are therefore left to discuss only the estimation of

$$Q_{F,G}(\psi) = \iint \left(\mathbf{P}\{X \le x, Y \le y\} - F(x)G(y)\right) \psi(F(x))\psi(G(y))dxdy.$$

Replacing the population quantities by their empirical counterparts and then subdividing the two integration regions into subintervals using order statistics, we arrive at the consistent estimator for $Q_{F,G}(\psi)$:

$$\begin{aligned}&\widehat{Q}_{F,G}(\psi)\\ &= \sum_{j,k=1}^{n-1} \int_{X_{j:n}}^{X_{j+1:n}} \int_{Y_{k:n}}^{Y_{k+1:n}} \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \le x, Y_i \le y\} - \frac{j}{n}\frac{k}{n}\right) \psi\left(\frac{j}{n}\right)\psi\left(\frac{k}{n}\right)dxdy.\end{aligned}$$

The presence of integrals in the formula above is not convenient from the practical point of view, but we can rewrite it without using any integrals. Toward this end, we first write the sum $\sum_{i=1}^{n} \mathbf{1}\{X_i \le x, Y_i \le y\}$ as $\sum_{i=1}^{n} \mathbf{1}\{X_{i:n} \le x, Y_{(i)} \le y\}$, where $Y_{(1)}, \ldots, Y_{(n)}$ are the induced order statistics corresponding to $X_{1:n} \le \cdots \le X_{n:n}$. Then we note that when $x \in (X_{j:n}, X_{j+1:n})$ and $y \in (Y_{k:n}, Y_{k+1:n})$, then the sum $\sum_{i=1}^{n} \mathbf{1}\{X_{i:n} \le x, Y_{(i)} \le y\}$ equals $\sum_{i=1}^{j} \mathbf{1}\{Y_{(i)} \le Y_{k:n}\}$, and we denote the latter one by $\kappa_n(j,k)$. Summarizing the discussion, we have the equation

$$\widehat{Q}_{F,G}(\psi) = \sum_{j,k=1}^{n-1} \left(\frac{\kappa_n(j,k)}{n} - \frac{j}{n}\frac{k}{n} \right) \psi\left(\frac{j}{n}\right) \psi\left(\frac{k}{n}\right) (X_{j+1:n} - X_{j:n})(Y_{k+1:n} - Y_{k:n}).$$

Using Slutsky's arguments, we arrive at the statement

$$\frac{\sqrt{n}\big((\widehat{L}_F - \widehat{L}_G) - (L_F - L_G)\big)}{\sqrt{\widehat{Q}_{F,F}(\psi) + \widehat{Q}_{G,G}(\psi) - 2\widehat{Q}_{F,G}(\psi)}} \to_d \mathcal{N}(0,1),$$

which we can use for constructing CIs for the difference $L_F - L_G$ and also for testing various hypotheses about L_F and L_G (cf. [18]).

Instead of the plug-in estimator for the asymptotic variance, we can employ a bootstrap approximation. This can be done as follows. Starting with the original pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, we calculate $\widehat{L}_F$ and $\widehat{L}_G$ using the first and the second coordinates respectively. Next, we sample with replacement from the original pairs and obtain the new ones $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$. Using the first and the second coordinates of the new pairs, we calculate $\widehat{L}_F^*$ and $\widehat{L}_G^*$, respectively. Then we calculate the quantity

$$\sqrt{n}|(\widehat{L}_F^* - \widehat{L}_G^*) - (\widehat{L}_F - \widehat{L}_G)| \tag{8}$$

and repeat the above procedure M times. Finally, we define z_α^* as the smallest value z such that at least $100(1-\alpha)\%$ values of quantity (8) are at or below z. The CI for $L_F - L_G$ is (cf. [18])

$$\widehat{L}_F - \widehat{L}_G \pm z_\alpha^* \frac{1}{\sqrt{n}}.$$

We can also work out the corresponding asymptotic results in the case of parametric families (cf. [18]). These are based on the statement

$$\frac{\sqrt{n}\big((\widehat{L}_F - \widehat{L}_G) - (L_F - L_G)\big)}{\sqrt{Q_{F,F}(\psi) + Q_{G,G}(\psi) - 2Q_{F,G}(\psi)}} \to_d \mathcal{N}(0,1),$$

where $\widehat{L}_F$ and $\widehat{L}_G$ are parametric estimators of L_F and L_G. As to the three quantities in the denominator, from our earlier results we know how $Q_{F,F}(\psi)$ looks like when $F \in \{F_1, F_2, F_3\}$. As to the third quantity, $Q_{F,G}(\psi)$, some additional calculations are needed. Assume, for example, that F is F_1 (i.e., exponential with

parameter $\theta > 0$) and G is F_2 (i.e., Pareto with parameter $\gamma > 0$). We already have formulas for L_F and L_G, as well as for $\widehat{L}_F$ and $\widehat{L}_G$ with the MLEs $\widehat{\theta}$ and $\widehat{\gamma}$. Using the Taylor formula, we derive the asymptotic expansion

$$\sqrt{n}\big((\widehat{L}_F - \widehat{L}_G) - (L_F - L_G)\big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{1}{r}\big((X_i - x_0) - \theta\big) - \frac{x_0 r \gamma^2}{(\gamma r - 1)^2} \left(\log\left(\frac{Y_i}{x_0}\right) - \frac{1}{\gamma} \right) \right\} + o_{\mathbf{P}}(1).$$

Hence,

$$Q_{F,G}(\psi) = \frac{x_0 \gamma^2}{(\gamma r - 1)^2} \mathbf{Cov}\left[X_1 - x_0, \log(Y_1/x_0)\right] = \frac{x_0 \gamma^2}{(\gamma r - 1)^2} \left(\mathbf{E}\big[G_\theta^{-1}(U) G_{1/\gamma}^{-1}(V)\big] - \frac{\theta}{\gamma} \right),$$

where U, V are (dependent) $(0,1)$-uniform r.v.'s, and $G_\theta(x) = 1 - e^{-x/\theta}$. If desired, the joint distribution of U and V can be specified by choosing a copula (cf. [12], [18], [25], [40]).

6 Summary and concluding remarks

In this paper we have discussed statistical inferential methods for estimating and comparing concentration curves and Gini-type indices that play fundamental roles when measuring economic inequality as well as in other areas of application. Both parametric and non-parametric approaches have been discussed in various contexts, including the cases of independent and paired samples. Plug-in and bootstrap based approaches for estimating critical values have been discussed extensively. For a more complete picture of diverse areas of application with numerous notions of stochastic dominance we refer, for example, to monographs [23], [24], [36], as well as to references therein. We also note a few recent references where quantities of the present paper are analyzed in the context of dependent observations: [6], [7], [8], [9], [15].

Acknowledgments

This paper is a reworked version of my lecture notes prepared for a workshop at the Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia (UII), Jogjakarta, in cooperation with the Doctoral Program in Economics, Faculty of Economics, UII, Jogjakarta. The paper has been prepared for a special EPAM (Extended Programme in Applied Mathematics between Indonesia and The Netherlands) symposium at the Institut Teknologi Bandung (ITB). My sincere thanks are due to Rohmatul Fajriyah (UII, Jogjakarta), Roelof Helmers (CWI, Amsterdam), Manon Post (CICAT/TUDelft, Delft), as well as to the ITB, the UII, and the Royal Netherlands Academy of Arts and Sciences (KNAW) for arrangement and support of my visit in Indonesia in August 2005.

References

[1] Brazauskas, V. & T. Kaiser (2004), Discussion of the paper "Empirical estimation of risk measures and related quantities" by Jones and Zitikis, *North American Actuarial Journal*, **8**, 117 – 118.

[2] Barrett, G.F. & S.G. Donald (2003), Consistent tests for stochastic dominance, *Econometrica*, **71**, 71 – 104.

[3] Csörgő, M., Csörgő, S. & L. Horváth (1986), *An Asymptotic Theory for Empirical Reliability and Concentration Processes*, Springer, Berlin.

[4] Davidson, R. & J.-Y. Duclos (2000), Statistical inference for stochastic dominance and for the measurement of poverty and inequality, *Econometrica*, **68**, 1435 – 1464.

[5] Davydov, Y. & V. Egorov (2000), Functional limit theorems for induced order statistics, *Math. Methods Statist.*, **9**, 297 – 313.

[6] Davydov, Y. & R. Zitikis (2002), Convergence of generalized Lorenz curves based on stationary ergodic random sequences with deterministic noise, *Statistics and Probability Letters*, **59**, 329 – 340.

[7] Davydov, Y. & R. Zitikis (2003), Generalized Lorenz curves and convexifications of stochastic processes, *Journal of Applied Probability*, **40**, 906 – 925.

[8] Davydov, Y. & R. Zitikis (2004), Convex rearrangements of random elements, in *Asymptotic Methods in Stochastics*, Editors: L. Horváth and B. Szyszkowicz, American Mathematical Society, Providence, RI, 141 – 171.

[9] Davydov, Y., Khoshnevisan, D., Shi, Z. & R. Zitikis (2005), Convex rearrangements, generalized Lorenz curves, and correlated Gaussian data, *Journal of Statistical Planning and Inference* (to appear).

[10] Deshpande, J.V. & H. Singh (1985), Testing for second-order stochastic dominance, *Communications in Statistics – Theory and Methods*, **14**, 887 – 893.

[11] Eubank, R., Schechtman, E., & S. Yitzhaki (1993), A test for second order stochastic dominance., *Communications in Statistics – Theory and Methods*, **22**, 1893 – 1905.

[12] Frees, E. W., & E. A. Valdez (1998), Understanding relationships using copulas, *North American Actuarial Journal*, **2**, 1 – 25.

[13] Goldie, C. M. (1977), Convergence theorems for empirical Lorenz curves and their inverses, *Advances in Appl. Probability*, **9**, 765 – 791.

[14] Helmers, R. (1982), *Edgeworth Expansions for Linear Combinations of Order Statistics*, Mathematisch Centrum, Amsterdam.

[15] Helmers, R. & R. Zitikis (2005), Strong laws for generalized absolute Lorenz curves when data are stationary and ergodic sequences, *Proceedings of the American Mathematical Society* (to appear).

[16] Horváth, L., Kokoszka, P. & R. Zitikis (2005), Testing for stochastic dominance using the weighted McFadden type statistic, *Journal of Econometrics* (to appear).

[17] Jones, B. L., & R. Zitikis (2003), Empirical estimation of risk measures and related quantities, *North American Actuarial Journal*, **7**, 44 – 54.

[18] Jones, B. L. & R. Zitikis (2005), Testing for the order of risk measures: an application of L-statistics in actuarial science, *Metron* (to appear).

[19] Jones, B. L., Puri, M. & R. Zitikis (2005), Testing hypotheses about the equality of several risk measure values with applications in insurance. *Insurance: Mathematics and Economics* (to appear).

[20] Kaur, A., Prakasa Rao, B. L. S. & H. Singh (1994), Testing for second-order stochastic dominance of two distributions, *Econometric Theory*, **10**, 849 – 866.

[21] Linton, O., Maasoumi, E. & Y. Whang (2003), *Consistent testing for stochastic dominance under general sampling schemes.* London School of Economics, London, UK.

[22] McFadden, D. (1989), Testing for stochastic dominance, in *Studies in the Economics of Uncertainty*, Editors: T. B. Fomby and T.K. Seo, Springer, New York, 113 – 134.

[23] Mosler, K. & M. Scarsini (1991), *Stochastic Orders and Decision under Risk*, Institute of Mathematical Statistics, Hayward, CA.

[24] Müller, A. & D. Stoyan (2002), *Comparison Methods for Stochastic Models and Risks*, Wiley, Chichester.

[25] Owzar, K. & P. K. Sen (2003), Copulas: concepts and novel applications. *Metron*, **61**, 323 – 353.

[26] Puri, M. L. (1965), On the combination of independent two sample tests of a general class, *Review of the International Statistical Institute*, **33**, 229 – 241.

[27] Puri, M. L. (1965), Some distribution-free k-sample rank tests of homogeneity against ordered alternatives, *Communications on Pure and Applied Mathematics*, **18**, 51 – 63.

[28] Puri, M. L. (1967), Combining independent one-sample tests of significance, *Annals of the Institute of Statistical Mathematics*, **19**, 285 – 300.

[29] Rao, C. R. & L. C. Zhao (1995), Convergence theorems for empirical cumulative quantile regression functions, *Math. Methods Statist.*, **4**, 81 – 91.

[30] Schechtman, E. & S. Yitzhaki (2004), The Gini instrumental variables, or "the double IV estimator", *Metron*, **62**, 1 – 27.

[31] Seiler, E. J. (2001), A nonparametric test for marginal conditional stochastic dominance, *Applied Financial Economics*, **11**, 173 – 177.

[32] Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, Wiley, New York.

[33] Shalit, H. & S. Yitzhaki (1994), Marginal conditional stochastic dominance, *Management Science*, **40**, 670 – 684.

[34] Shalit, H. & S. Yitzhaki (2003), Solving the portfolio allocation puzzle, *American Economic Review*, **93**, 1002 – 1008.

[35] Schmid, F. & M. Trede (1998), A Kolmogorov-type test for second-order stochastic dominance, *Statistics and Probability Letters*, **37**, 183 – 193.

[36] Shaked, M. & J. G. Shanthikumar (1994), *Stochastic Orders and their Applications*, Academic Press, Boston, MA.

[37] Shorack, G. R. (2000), *Probability for Statisticians*, Springer, New York.

[38] Shorack, G. R. & Wellner, J. A. (1986), *Empirical Processes with Applications to Statistics*, Wiley, New York.

[39] Tarsitano, A. (2004), A new class of inequality measures based on a ratio of L-statistics, *Metron*, **62**, 137 – 160.

[40] Venter, G.G. (2002), Tails of copulas, *Proceedings of the Casualty Actuarial Society*, **89**, 68 – 113.

[41] Wang, S. S. (1998), An actuarial index of the right-tail risk, *North American Actuarial Journal*, **2**, 88 – 101.

[42] Zitikis, R. (2002), Large sample estimation of a family of economic inequality indices, *Pakistan J. Statist.*, **18** (special issue in honour of Dr. S. Ejaz Ahmad), 225 – 248.

[43] Zitikis, R. (2002), Analysis of indices of economic inequality from a mathematical point of view (Lecture at the 11th Indonesian Mathematics Conference, State University of Malang, Indonesia), *Matematika*, **8**, 772 – 782; also available as *LRSP Technical Report*, **366**, Carleton University and the University of Ottawa, Ottawa.

[44] Zitikis, R. & J. L. Gastwirth (2002), The asymptotic distribution of the S-Gini index, *Aust. N. Z. J. Stat.*, **44**, 439 – 446.

R. Zitikis: Department of Statistical and Actuarial sciences, University of Western Ontario, London, Ontario N6A 5B7, Canada.
E-mail: zitikis@stats.uwo.ca

STATISTICAL ESTIMATION OF A CYCLIC POISSON INTENSITY FUNCTION

I Wayan Mangku

IPB, Bogor, Indonesia

Abstract. We will survey some recent results on estimating the intensity function of a cyclic Poisson process. It is assumed that only a single realization of the Poisson process is observed in a bounded window. We prove that a nonparametric kernel type estimator of the intensity function of the cyclic Poisson process is consistent, when the size of the window expands. We also compute the asymptotic bias, variance and the mean-squared error of our estimator.
Next, we consider the problem to estimate a cyclic Poisson intensity function in the presence of linear trend. For this slightly more complicated situation a new kernel type estimator of the cyclic part is proposed and investigated in detail.
This is joint work with R. Helmers (CWI, Amsterdam) and R. Zitikis (UWO, London Ont., Canada).
Key-words: cyclic Poisson process, intensity function, linear trend, nonparametric estimation, consistency, bias, variance, mean-squared error.

1 Introduction

Let X be a Poisson point process on $[0,\infty)$ with (unknown) locally integrable intensity function λ which is assumed to consist of two components, namely a periodic or cyclic component with (unknown) period $\tau > 0$ and a (unknown) linear trend component. In other words, for any point $s \in [0,\infty)$, we can write the intensity function λ as

$$\lambda(s) = \lambda_c(s) + as \tag{1}$$

where $\lambda_c(s)$ is a periodic function with period τ and a denotes the slope of the linear trend. We do not assume any (parametric) form of λ_c except that it is periodic. That is we assume that the equality

$$\lambda_c(s + k\tau) = \lambda_c(s) \tag{2}$$

holds for all $s \in [0,\infty)$ and $k \in \mathbf{Z}$. Here we consider a Poisson point process on $[0,\infty)$ instead of, for instance, on $\mathbf{R}$ because λ has to satisfy (1) and must be non negative. For the same reason we also restrict our attention to the case $a > 0$.

Let $W_1,\ W_2, \ldots$ be a sequence of intervals of $[0,\infty)$, called windows, such that the size or the Lebesgue measure $|W_n|$ of W_n is finite for each fixed $n \in \mathbf{N}$, but

$$|W_n| \to \infty,$$

as $n \to \infty$. Furthermore, let h_n be a sequence of positive real numbers such that

$$h_n \downarrow 0 \tag{3}$$

as $n \to \infty$.

We will assume throughout that s is a Lebesgue point of λ, which automatically means that s is a Lebesgue point of λ_c as well. This assumption is a mild one since the set of all Lebesgue points of λ is dense in $\mathbf{R}$, whenever λ is assumed to be locally integrable.

2 Purely cyclic Poisson process

In this section we present some recent results on estimating the intensity of a purely cyclic Poisson process, that is Poisson process with intensity function $\lambda = \lambda_c$ (cf. (1) with $a = 0$). These results are special case of those in [4] and [5].

Suppose now that, for some $\omega \in \Omega$, a single realization $X_c(\omega)$ of the Poisson process X_c defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ with intensity function $\lambda = \lambda_c$ is observed, though only within a bounded interval, called 'window' $W \subset [0, \infty)$. Our goal is to construct a consistent non-parametric estimator of λ_c at a given point $s \in [0, \infty)$ from a single realization $X_c(\omega)$ of the Poisson process X_c observed in $W := W_n$. We also compute the asymptotic bias, variance, and the mean-squared error of the proposed estimator.

Let $\hat{\tau}_n$ be any consistent estimator of the period τ, that is, $\hat{\tau}_n \xrightarrow{p} \tau$, as $n \to \infty$. One may use the estimators constructed in [1], [2], or [6].

With these notations, we now define the estimator of $\lambda_c(s)$ as

$$\bar{\lambda}_{c,n}(s) = \frac{\hat{\tau}_n}{|W_n|} \sum_{k=-\infty}^{\infty} \frac{X_c([s + k\hat{\tau}_n - h_n, s + k\hat{\tau}_n + h_n] \cap W_n)}{2h_n}. \tag{4}$$

The idea behind the construction of the estimator $\bar{\lambda}_{c,n}(s)$ is as follows. Since there is only one realization of the Poisson process X_c available, we have to combine information about the (unknown) value of $\lambda_c(s)$ from different places of the window W_n. For this reason, the periodicity of λ_c (cf. (2)), plays a crucial role.

Let $N_n = \#\{k : \ s + k\tau \in W_n\}$ and $B_h(x) = [x - h, x + h]$. Then we have the

following string of (approximate) equations

$$\begin{aligned}
\lambda_c(s) &= \frac{1}{N_n}\sum_{k=-\infty}^{\infty}\lambda_c(s+k\tau)\mathbf{I}\{s+k\tau\in W_n\} \\
&\approx \frac{1}{N_n}\sum_{k=-\infty}^{\infty}\frac{1}{|B_{h_n}(s+k\tau)|}\int_{B_{h_n}(s+k\tau)\cap W_n}\lambda_c(x)dx \\
&\approx \frac{1}{N_n}\sum_{k=-\infty}^{\infty}\frac{1}{2h_n}X_c(B_{h_n}(s+k\tau)\cap W_n) \\
&\approx \frac{\tau}{|W_n|}\sum_{k=-\infty}^{\infty}\frac{1}{2h_n}X_c(B_{h_n}(s+k\tau)\cap W_n). \qquad (5)
\end{aligned}$$

We note that, in order to make the first $\approx$ in (5) works, we require the assumptions that s is a Lebesgue point of λ_c and (3) holds true.

Thus, from (5) we conclude that the quantity

$$\lambda_{c,n}(s) := \frac{\tau}{|W_n|}\sum_{k=-\infty}^{\infty}\frac{1}{2h_n}X_c(B_{h_n}(s+k\tau)\cap W_n), \qquad (6)$$

can be viewed as an estimator of $\lambda_c(s)$, provided that the period τ is known. When we do not know the period, we modify the quantity in (6) by replacing the unknown period τ by its estimator $\hat{\tau}_n$ and obtain (4)

Here are the main results.

Theorem 2.1 *Let the intensity function λ_c be periodic and locally integrable. Furthermore, let the bandwidth h_n be such that $h_n \downarrow 0$ and $h_n|W_n| \to \infty$ as $n \to \infty$. If*

$$|W_n||\hat{\tau}_n - \tau|/h_n \xrightarrow{p} 0$$

as $n \to \infty$, then

$$\bar{\lambda}_{c,n}(s) \xrightarrow{p} \lambda_c(s) \qquad (7)$$

as $n \to \infty$, provided s is a Lebesgue point of λ_c. In other words, $\bar{\lambda}_{c,n}(s)$ is a consistent estimator of $\lambda_c(s)$.

Under, naturally, stronger assumptions than those of Theorem 2.1, we also have the complete convergence of the estimator $\bar{\lambda}_{c,n}(s)$ which, in turn, gives a rate of consistency of the estimator $\bar{\lambda}_{c,n}(s)$.

Theorem 2.2 *Let the intensity function λ_c be periodic and locally integrable. Furthermore, let the bandwidth h_n be such that $h_n \downarrow 0$ as $n \to \infty$, and*

$$\sum_{n=1}^{\infty}\exp\{-\epsilon\sqrt{|W_n|h_n}\} < \infty$$

for any $\epsilon > 0$. *If*

$$|W_n||\hat{\tau}_n - \tau|/h_n \xrightarrow{c} 0,$$

as $n \to \infty$, *then*

$$\bar{\lambda}_{c,n}(s) \xrightarrow{c} \lambda_c(s), \tag{8}$$

as $n \to \infty$, *provided* s *is a Lebesgue point of* λ_c.

Next we present statistical properties of our estimator under minimal conditions on the intensity function, the estimator of the period, and other parameters involved.

In order to be able to employ a weaker condition on the estimator of the period, it is required to modify our estimator of λ_c. Here, instead of using the estimator $\bar{\lambda}_{c,n}(s)$ as given in (4), we derive some statistical properties of its modification which is given by

$$\bar{\lambda}^{\diamond}_{c,n}(s) \quad = \quad \mathbf{I}\left\{\bar{\lambda}_{c,n}(s) \le D_n\right\}\bar{\lambda}_{c,n}(s) + \mathbf{I}\left\{\bar{\lambda}_{c,n}(s) > D_n\right\}D_n, \tag{9}$$

where "truncating" constants D_n are deterministic and converging to infinity when $n \to \infty$.

Theorem 2.3 *Suppose that* λ_c *is periodic and bounded in a neighborhood of* s, $h_n \downarrow 0$, $h_n|W_n| \to \infty$, *and the sequence* D_n *such that for some* $c > 0$ *and* $\epsilon > 0$ *we have* $D_n \ge c(h_n|W_n|)^{\epsilon}$ *holds for all sufficiently large* n. *If, in addition, for any* $\delta > 0$ *we have*

$$\mathbf{P}\left(\frac{|W_n|^{3/2}}{h_n^{1/2}}\,|\hat{\tau}_n - \tau| \ge \delta\right) = o\left(\frac{1}{D_n^2|W_n|h_n}\right)$$

as $n \to \infty$, *then we have*

$$Var\left(\bar{\lambda}^{\diamond}_{c,n}(s)\right) \quad = \quad \frac{\tau\lambda_c(s)}{2|W_n|h_n} + o\left(|W_n|^{-1}h_n^{-1}\right) \tag{10}$$

as $n \to \infty$, *provided* s *is a Lebesgue point of* λ_c.

Theorem 2.4 *Suppose that* λ_c *is periodic and locally integrable,* $h_n \downarrow 0$, $h_n^2|W_n| \to \infty$, *the sequence* D_n *such that for some* $c > 0$ *and* $\epsilon > 0$ *we have* $D_n \ge c(h_n)^{-\epsilon}$ *holds for all sufficiently large* n, *and for any* $\delta > 0$ *we have*

$$\mathbf{P}\left(\frac{|W_n|}{h_n^3}\,|\hat{\tau}_n - \tau| \ge \delta\right) = o\left(\frac{h_n^2}{D_n}\right)$$

as $n \to \infty$. *If, in addition,* λ_c *has finite second derivative* λ_c'' *at* s, *then*

$$\mathbf{E}\bar{\lambda}^{\diamond}_{c,n}(s) \quad = \quad \lambda_c(s) + \frac{\lambda_c''(s)}{6}h_n^2 + o(h_n^2) \tag{11}$$

as $n \to \infty$.

We note that, without assuming $h_n^2|W_n| \to \infty$, we can only prove that the remainder term on the r.h.s. of (11) is of order $o(h_n^2) + \mathcal{O}(|W_n|^{-1})$, as $n \to \infty$. Since the second term on the r.h.s. of (11) is exactly of the order $\mathcal{O}(h_n^2)$, it is therefore natural to have $|W_n|^{-1} = o(h_n^2)$, as $n \to \infty$.

By Theorems 2.3 and 2.4 (i.e. (10) and (11)), we can compute the asymptotic approximation to the mean-squared error of $\bar{\lambda}_{c,n}^{\diamond}(s)$, that is

$$MSE\left(\bar{\lambda}_{c,n}^{\diamond}(s)\right) = \frac{\tau\lambda_c(s)}{2|W_n|h_n} + \frac{1}{36}\left(\lambda_c''(s)\right)^2 h_n^4 + o\left(|W_n|^{-1}h_n^{-1}\right) + o\left(h_n^4\right) \qquad (12)$$

as $n \to \infty$. Now, we consider the r.h.s. of (12). By minimizing the sum of its first and second terms (the main terms for the variance and the squared bias), we then obtain the 'optimal' choice of h_n, which is given by

$$h_n = \left[\frac{9\tau\lambda_c(s)}{2\left(\lambda_c''(s)\right)^2}\right]^{\frac{1}{5}} |W_n|^{-\frac{1}{5}}. \qquad (13)$$

With this choice of h_n, the 'optimal' rate of decrease of $MSE(\bar{\lambda}_{c,n}^{\diamond}(s))$ is of order $\mathcal{O}(|W_n|^{-4/5})$ as $n \to \infty$.

3 Cyclic Poisson Process in the presence of linear trend

In this section we present some recent results on estimating the intensity of a cyclic Poisson process in the presence of linear trend, that is Poisson process with intensity function λ given by (1). This is a preview of a more complete paper (cf. [3]), which is submitted for publication elsewhere. We consider the case that both the period τ and a in (1) are unknown.

Suppose now that, for some $\omega \in \Omega$, a single realization $X(\omega)$ of the Poisson process X defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ with intensity function λ (cf. (1)) is observed, though only within a bounded interval, called 'window' $W \subset [0, \infty)$. Our goal is to construct a consistent non-parametric estimator of λ_c at a given point $s \in [0, \infty)$ from a single realization $X(\omega)$ of the Poisson process X observed in $W := W_n$. We also compute the asymptotic bias, variance, and the mean-squared error of the proposed estimator.

Let $\hat{\tau}_n$ be a consistent estimator of τ. Now we may define estimators of respectively a and λ_c, at a given point s, as follows:

$$\hat{a}_n := \frac{2X(W_n)}{|W_n|^2}, \qquad (14)$$

and

$$\hat{\lambda}_{c,n}(s) := \frac{1}{\ln|W_n|}\sum_{k=-\infty}^{\infty}\frac{1}{k}\frac{X([s+k\hat{\tau}_n-h_n,s+k\hat{\tau}_n+h_n]\cap W_n)}{2h_n} - \hat{a}_n\left(s+\frac{|W_n|}{\ln|W_n|}\right). \tag{15}$$

To obtain the estimator $\hat{a}_n$ of a it suffices to note that

$$\mathbf{E}X(W_n) = \frac{a}{2}|W_n|^2 + \mathcal{O}(|W_n|),$$

as $n\to\infty$, which directly yields the estimator given in (14). While the construction of the kernel-type estimator $\hat{\lambda}_{c,n}(s)$ of $\lambda_c(s)$ is using a similar idea to the one given in (5).

The construction of estimators $\hat{\tau}_n$ of the period τ of a cyclic Poisson process with desired accuracy (cf. (16), (19) or (21)), using only a single realization from X, is outside the scope of the present paper.

Here are the main results:

Theorem 3.1 *Suppose that the intensity function λ satisfies (1) and is locally integrable. Furthermore, let $h_n\downarrow 0$ and $h_n\ln|W_n|\to\infty$. If, in addition, for any $\delta>0$ we have*

$$\mathbf{P}\left(\frac{|W_n|^2}{h_n\ln|W_n|}\,|\hat{\tau}_n-\tau|>\delta\right)=o(1) \tag{16}$$

as $n\to\infty$, then

$$\hat{\lambda}_{c,n}(s)\ \xrightarrow{p}\ \lambda_c(s), \tag{17}$$

as $n\to\infty$, provided s is a Lebesgue point of λ_c. In other words, $\hat{\lambda}_{c,n}(s)$ is a consistent estimator of $\lambda_c(s)$.

Next we present statistical properties of our estimator under minimal conditions on the intensity function, the estimator of the period, and other parameters involved.

In order to be able to derive asymptotic approximations to respectively bias and variance of the estimator of $\lambda_c(s)$ under weak assumptions on the estimator $\hat{\tau}_n$ of the period, it is required to modify our estimator $\hat{\lambda}_{c,n}(s)$ of $\lambda_c(s)$ given in (15) slightly as follows:

$$\hat{\lambda}^{\diamond}_{c,n}(s) := \mathbf{I}\left(\hat{\lambda}_{c,n}(s)\le D_n\right)\hat{\lambda}_{c,n}(s)+\mathbf{I}\left(\hat{\lambda}_{c,n}(s)>D_n\right)D_n \tag{18}$$

where the non-random D_n will approach infinity when $n\to\infty$. The truncation level D_n in the definition of $\hat{\lambda}^{\diamond}_{c,n}(s)$ is needed to avoid accumulation of errors due

to estimation of τ in estimating $\lambda_c(s)$.

In our two final theorems we show that one can estimate the period τ without affecting the statistical properties of our estimate of the intensity function $\lambda_c(s)$, provided the rate of consistency of the estimator of the period τ is sufficiently fast.

Theorem 3.2 *Suppose that the intensity function λ satisfies (1) and is locally integrable. Furthermore, let $h_n \downarrow 0$, $h_n \ln |W_n| \to \infty$, and the sequence D_n be such that, for some $c > 0$ and $\epsilon > 0$, the bound $D_n \geq c(h_n \ln |W_n|)^{\epsilon}$ holds for all sufficiently large n. If, in addition, for any $\delta > 0$ we have*

$$\mathbf{P}\left(\frac{|W_n|^2}{h_n^{1/2}(\ln |W_n|)^{1/2}} |\hat{\tau}_n - \tau| > \delta\right) = o\left(\frac{1}{D_n^2 h_n \ln |W_n|}\right) \tag{19}$$

as $n \to \infty$, then

$$Var\left(\hat{\lambda}_{c,n}^{\diamond}(s)\right) = \frac{a\tau}{2h_n \ln |W_n|} + o\left(\frac{1}{h_n \ln |W_n|}\right) \tag{20}$$

as $n \to \infty$, provided s is a Lebesgue point of λ_c.

Theorem 3.3 *Suppose that the intensity function λ satisfies (1) and is locally integrable. Furthermore, let $h_n \downarrow 0$, $h_n^2 \ln |W_n| \to \infty$, and the sequence D_n be such that, for some $c > 0$ and $\epsilon > 0$, the bound $D_n \geq c(h_n)^{-\epsilon}$ holds for all sufficiently large n. If, in addition, for any $\delta > 0$ we have*

$$\mathbf{P}\left(\frac{|W_n|^2}{h_n^3 \ln |W_n|} |\hat{\tau}_n - \tau| > \delta\right) = o\left(\frac{h_n^2}{D_n}\right) \tag{21}$$

as $n \to \infty$ and λ_c has finite second derivative λ_c'' at s, then

$$\mathbf{E}\hat{\lambda}_{c,n}^{\diamond}(s) = \lambda_c(s) + \frac{\lambda_c''(s)}{6} h_n^2 + o(h_n^2) \tag{22}$$

as $n \to \infty$.

By Theorems 3.3 and 3.2 (i.e. (20) and (22)), we can compute the MSE of $\hat{\lambda}_{c,n}^{\diamond}(s)$ as follows

$$MSE\left(\hat{\lambda}_{c,n}^{\diamond}(s)\right) = \frac{a\tau}{2h_n \ln |W_n|} + \frac{(\lambda_c''(s))^2}{36} h_n^4 + o\left(\frac{1}{h_n \ln |W_n|}\right) + o(h_n^4) \tag{23}$$

as $n \to \infty$. Now, we consider the r.h.s. of (23). By minimizing the sum of the first and second term (the leading term for the variance and the squared bias), we then get the optimal choice of h_n, which is given by

$$h_n = \left[\frac{9a\tau}{2(\lambda_c''(s))^2}\right]^{\frac{1}{5}} (\ln |W_n|)^{-\frac{1}{5}}. \tag{24}$$

With this choice of h_n, the optimal rate of decrease of $MSE(\hat{\lambda}^{\diamond}_{c,n}(s))$ is of order $\mathcal{O}((\ln|W_n|)^{-4/5})$ as $n \to \infty$.

Remark: If we compare the statistical properties of the estimator $\bar{\lambda}^{\diamond}_{c,n}(s)$ of $\lambda_c(s)$ given by (9) (for the purely cyclic Poisson process) and the estimator $\hat{\lambda}^{\diamond}_{c,n}(s)$ of $\lambda_c(s)$ given by (18) (for the cyclic Poisson process in the presence of linear trend), we have the followings. From (11) and (22) we see that both $\bar{\lambda}^{\diamond}_{c,n}(s)$ and $\hat{\lambda}^{\diamond}_{c,n}(s)$ are having the same asymptotic bias. However, (10) and (20) show that $\bar{\lambda}^{\diamond}_{c,n}(s)$ and $\hat{\lambda}^{\diamond}_{c,n}(s)$ are having different asymptotic variance in two ways. First, the role of $\lambda_c(s)$ on the r.h.s. of (10) is replaced by a on the r.h.s. of (20). This because, in the case of cyclic Poisson process in the presence of linear trend, the trend component dominate the variability of our estimator. Second, the role of $|W_n|$ on the r.h.s. of (10) is replaced by $\ln|W_n|$ on the r.h.s. of (20). This is a consequence of using weight $1/k$ in the construction of the estimator $\hat{\lambda}^{\diamond}_{c,n}(s)$.

References

[1] Bebbington, M., and R. Zitikis (2004), A robust heuristic estimator for the period of a Poisson intensity function. *Methodology and Computing in Applied Probability*, **6**, 441-462.

[2] Helmers, R., and I W. Mangku (2003), On estimating the period of a cyclic Poisson process. *Mathematical Statistics and Applications: Festschrift in honor of Constance van Eeden*. (Editors: Marc Moore, Sorana Froda and Christian Leger), IMS Lecture Notes Series - Monograph Series, Volume 42, 345-356.

[3] Helmers, R., and I W. Mangku (2005), Estimating the intensity of a cyclic Poisson process in the presence of linear trend, CWI Report, submitted for publication.

[4] Helmers, R., I W. Mangku, and R. Zitikis (2003), Consistent estimation of the intensity function of a cyclic Poisson process. *J. Multivariate Anal.* **84**, 19-39.

[5] Helmers, R., I W. Mangku, and R. Zitikis (2005), Statistical properties of a kernel-type estimator of the intensity function of a cyclic Poisson process. *J. Multivariate Anal.*, **92**, 1-23.

[6] Vere-Jones, D. (1982). On the estimation of frequency in point-process data. *J. Appl. Prob.* 19A, 383-394.

COMPOUND SUMS: A SURVEY OF SOME RECENT DEVELOPMENTS

Roelof Helmers
CWI, Amsterdam
The Netherlands

Abstract. Compound sums play an important role in accountancy (statistical auditing) and in insurance (total claim size of a portfolio). Accurate statistical inference will typically be based in a Studentized compound sum. We shall discuss some recent results on Edgeworth/saddlepoint approximations for these statistics and indicate their relevance in statistical applications.
This is joint work in progress with Bing-Yi Jing (Hong Kong University of Science and Technology) and Wang Zhou (National University of Singapore).
Key-words: Studentized compound sums, Edgworth expansions, saddlepoint approximations, insurance applications, statistical auditing.

1 Introduction and main results

Compound sums $S_N = \Sigma_{i=1}^N X_i$, where $X_1, X_2, \ldots$ are independent and identically distributed (i.i.d.) random variables (r.v.) with common distribution function $(df) F$, i.e. $F(x) = P(X_i \leq x), i = 1, 2, \ldots$ and N denotes a non-negative integer valued r.v., independent of the X_i's. For instance, in typical applications in accountancy and insurance, N is assumed to be Poisson distributed with parameter $\nu > 0$, i.e.

$$P(N = n) = e^{-\nu} \frac{\nu^n}{n!}, n = 0, 1, 2, \ldots \tag{1}$$

Mixtures of Poisson distributions are also of interest - a specific example is the negative binomial distribution - but outside the scope of this short review paper (cf.[1])).
Compound Poisson sums S_N, with N as in (1), play an important role in statistical auditing and in insurance (total claim size of a portfolio). The compound Poisson process

$$\left\{ \sum_{i=1}^{N(t)} X_i, 0 < t < \infty \right\} \tag{2}$$

arises in insurance mathematics, with claim sizes $X_i, i = 1, 2, \ldots$, which are assumed to be i.i.d. with common $df F$ (with support in $\mathbb{R}^+ \cup \{0\}$, the nonnegative real numbers), and where $N(t)$, the number of claims occuring in $(0, t]$, is supposed to be Poisson distributed with parameter ν, where $\nu = \lambda t$; here $\lambda > 0$ denotes the constant intensity of the homogeneous Poisson process

$$\{N(t), 0 < t < \infty\} \tag{3}$$

For any fixed t, the random variable $S_{N(t)} = \sum_{i=1}^{N(t)} X_i$ denotes the total claim size in a portfolio in $[0, t)$.
A second application of compound Poisson sums S_N occurs in accountancy (statistical auditing); (cf. [4]), where an auditor attemps to check the validity of financial statements of a firm or a government agency. In these accountancy applications $S_N = \sum_{i=1}^{N} X_i$ denotes the total error amount in a random sample of size n drawn without replacement from an audit population of bookamounts, the X_i's represent now the non-zero errors observed by the auditor in n recorded bookvalues, N is nothing but the random number of bookvalues in the sample of size n with error. In typcial applications error are rare, that is the probability that errors are non-zero is close to zero, and the Poisson distribution for N (cf. (1)) works well. Clearly $\frac{T}{N} S_N$ is an unbiased estimator of the total error amount in a finite audit population of bookvalues of size T; T is nothing but the size of the finite audit population in a given period of time, a given year, say, in other words, T denotes the population size, i.e. the total number of recorded book values in a year. In [4] a conservative confidence upperbound for the total error amount in an audit population, the parameter of interest, is constructed using Edgeworth expansions and bootstrap calibration.
Accurate statistical inference, for instance aiming at the construction of confidence upperbounds for a parameter of interest, e.g. the total claim size in a portfolio of an insurance company or the total error amount in an audit population of a government agency, is typically based on a Studentized compound Poisson sum T_N, which is given by

$$T_N = \frac{S_N - \nu\mu}{V_N} \tag{4}$$

where S_N as before and $V_N^2 = \sum_{i=1}^{N} X_i^2$. Note that

$$ES_N = EN.EX_1 = \nu\mu \tag{5}$$

where $\mu = \int x dF(x)$, and

$$\sigma^2(S_N) = E_N \sigma^2(S_N|N) + \sigma_N^2(E(S_N|N)) =$$
$$= \nu\sigma^2 + \nu\mu^2 = \nu\mu_2 \tag{6}$$

where $\mu_2 = \int x^2 dF(x)$. The empirical counterpart of $\sigma^2(S_N)$ is therefore given by

$$\hat{\nu}\hat{\mu}_2 = \quad N \int x^2 d\hat{F}_N(x) =$$
$$= N \frac{1}{N} \sum_{i=1}^{N} X_i^2 = V_N^2 \tag{7}$$

where $\hat{F}_N$ denotes the empirical *df* corresponding to a sample from F with random (Poisson) sample size N:

$$\hat{F}_N(x) = \frac{1}{N} \sum_{i=1}^{N} I(X_i \le x) \tag{8}$$

for real x; here $I(A)$ denotes the indicator of a set A; if $N = 0$, which happens with probability $e^{-\nu}$, we set arbitrary $\hat{F}_N(x) = 0$.
The aim of this short survey paper (cf., also [6]) is to discuss some recent results on Edgeworth/saddlepoint approximations for T_N. Define, for any $\nu > 0$,

$$G_\nu(x) = P(T_N \leq x) \tag{9}$$

with T_N as in (4). It is well-known that

$$\sup_x |G_\nu(x) - \Phi(x)| \to 0, \text{ as } \nu \to \infty \tag{10}$$

where Φ denotes the standard normal *df*, provided $0 < EX_1^2 < \infty$. The rate of convergence towards normality in (10) is fairly slow: if moreover $E|X_1|^3 < \infty$, then

$$\sup_\nu |G_\nu(x) - \Phi(x)| = \mathcal{O}(\nu^{-1/2}), \text{ as } \nu \to \infty \tag{11}$$

One way to improve upon the normal approximation to G_ν is to establish an Edgeworth expansion for T_N:

$$\begin{aligned} G_\nu(x) = \quad & \Phi(x) + \frac{1}{6\sqrt{\nu}} \frac{EX_1^3}{(EX_1^2)^{3/2}}(2x^2+1)\phi(x) + \\ & + o(\nu^{-1/2}), \text{ as } \nu \to \infty \end{aligned} \tag{12}$$

where ϕ denotes the standard normal density.
The Edgeworth expansion (12) has an absolute error of smaller order than the normal approximation, namely, $o(\nu^{-1/2})$ or $\mathcal{O}(\nu^{-1})$ under somewhat stronger conditions, instead of $\mathcal{O}(\nu^{-1/2})$ the 'normal error' in (11). The reason for this improvement is that the Edgeworth expansion picks up the skewness which is typically present in the distribution of T_N, the normal *df* Φ of course fails to do this.
An interesting open problem at present is to investigate what happens when $EX_1^2 = \infty$, but X_1 is assumed to be in the domain of attraction of a normal law. We conjecture that the exact rate of convergence and the leading term in the central limit theorem for T_N can be determined in this more general setting as well, using a method recently developed in [10] where such results were established for the old and famous Student t-statistic.
We note here in passing that, like T_N, the Student t-statistic can also be expressed as a simple function of $\sum_{i=1}^n X_i$ and $\sum_{i=1}^n X_i^2$, n being the fixed sample size in this case.
In practical applications one will need an empirical Edgeworth expansion for T_N, replacing the skewness coefficient $\nu^{-\frac{1}{2}} EX_1^3|(EX_1^2)^{3/2}$ in (12) by its empirical counterpart

$$N^{-\frac{1}{2}} \frac{\int x^3 \hat{F}_N(x)}{(\int x^2 d\hat{F}_N(x))^{3/2}} = \frac{\sum_{i=1}^N X_i^3}{(\sum_{i=1}^N X_i^2)^{3/2}} \tag{13}$$

with $\hat{F}_N$ as in (8). It is easily checked that replacing $\nu^{-\frac{1}{2}} EX_1^3/(EX_1^2)^{3/2}$ in (12) by (13) will not affect the order of magnitude of the remainder term in (12).
Saddlepoint approximations – first introduced in mathematical statistics by Henry Daniels in a famous paper ([2]) as early as in 1954 – provide us with a completely different way to approximate the *df* of a statistic under consideration. Indeed, in a forthcoming paper ([7]) a saddlepoint approximation of classical Lugannani-Rice form for T_N, properly normalized – i.e. instead of T_N one considers $T_N/\sqrt{N}$, the appropriate statistic to look at when one aims at accurately approximating (small) tail probabilities (large deviations) – is established:

$$P(T_N/\sqrt{N} \geq x) = 1 - \Phi(\sqrt{\nu}w) - \\ -\frac{\phi(\sqrt{\nu}w)}{\sqrt{\nu}}(\frac{1}{w} - \frac{1}{v} + \mathcal{O}(\frac{1}{\nu})) \tag{14}$$

Here w and v are given by fairly complicated formulas for which we refer the interested reader to [7].
Saddlepoint approximations of the form (14) are well-known in the classical theory of saddlepoint approximations. We refer to the monograph [9], [11] and to the recent paper [13]. In [13] a saddlepoint approximation of the form (14) (though with different expressions for w and v) is obtained under minimal conditions for the important case of the Student t-statistic. The proof of (14) is closely related to the method of proof given in [13]. In a way the only thing we do in [7] is to extend the proof in [13] to $T_N/\sqrt{N}$, i.e. to Studentized compound Poisson sums. In the next section we will sketch some of the basic ideas occuring in these proofs. We also refer to the recent PhD thesis of W. Zhou [14] for a more complete account of all this.

2 Sketch of proof of (14)

Our sketch of proof is based on [13] and [7].
To begin with let us consider the density $f_{(\bar{X}_n, \bar{V}_n^2)}(\cdot,\cdot)$ of $(\bar{X}_n, \bar{V}_n^2)^T$, where $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ and $\bar{V}_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2$. We shall assume that $EX_1 = 0$ and also that the p-th power of the characteristic function of (X_1, X_1^2) is summable, for some $p > 1$; the latter smoothness condition will imply that the density f exist.
Note that $\bar{X}_n = n^{-1}S_n, \bar{V}_n^2 = n^{-1}V_n^2$, so that (cf. (4)) $T_n = S_n/V_n = \sqrt{n}\bar{X}_n/\bar{V}_n$, whenever $N = n$, for any integer $n \geq 0$. Moreover we remark that

$$\bar{X}_n = a, \bar{V}_n^2 = \frac{a^2}{b^2} \Longleftrightarrow \bar{X}_n = a, \bar{X}_n/\bar{V}_n = b \tag{15}$$

with $ab > 0$. It is now easily seen that the density of $\bar{X}_n/\bar{V}_n$ can be written as:

$$f_{\bar{X}_n/\bar{V}_n}(b) = \int_{-\infty}^{\infty} f_{(\bar{X}_n, \bar{V}_n^2)}(a, \frac{a^2}{b^2}) J(a,b) I(ab > 0\, da \tag{16}$$

where $J(a,b) = 2a^2/|b|^3$ denotes the Jacobian of the transformation (15), while $I(ab > 0)$ is the indicator of the set $\{(a,b) : ab > 0\}$. The next step is – and here a crucial idea of the saddlepoint methodology is used – to rewrite (16) in terms of socalled 'tilted' r.v.'s (X_{st}, Y_{st}), instead of the original r.v.'s (X,Y), where X is distributed as X_1 and $Y = X^2$. Define the cumulant generating function

$$K(s,t) = \ln Ee^{sX+tY} \tag{17}$$

for $(s,t) \in \mathbb{R}^2$, and note that, with (X_{st}, Y_{st}) chosen to be such that

$$f_{(X_{st},Y_{st})}(x,y) = \frac{e^{sx+ty}}{Ee^{sX+tY}} f_{(X,Y)}(x,y) \tag{18}$$

we can rewrite (16) as follows:

$$f_{\bar{X}_n/\bar{V}_n}(b) = \int_{-\infty}^{\infty} e^{-n[sa+t\frac{a^2}{b^2}-K(s,t)]} f_{(\bar{X}_{st},\bar{Y}_{st})}(a\,,\frac{a^2}{b^2}) J(a,b) I(ab>0) da \tag{19}$$

where $\bar{X}_{st} = n^{-1}\sum_{1=1}^{n} X_{st,i}$ and $\bar{Y}_{st} = n^{-1}\sum_{1=1}^{n} Y_{st,i}$. At this point the choice of s and t is still free and the idea is now to select s and t so that

$$\frac{d}{ds}K(s,t) = a \tag{20}$$

$$\frac{d}{dt}K(s,t) = \frac{a^2}{b^2} \tag{21}$$

Let us denote the roots of (20) and (21) by

$$\hat{s} = \hat{s}(a,b), \hat{t} = \hat{t}(a,b) \tag{22}$$

Defining now

$$\Lambda(a,b) = \hat{s}a + \hat{t}\frac{a^2}{b^2} - K(\hat{s},\hat{t}) \tag{23}$$

(which is only a function of a, for each fixed b), one easily checks that one can maximize the exponential factor in the integrand of (19) by solving the equation

$$\frac{d}{da}\Lambda(a,b) = 0 \tag{24}$$

and we get

$$\hat{s} + \frac{2a}{b^2}\hat{t} = 0 \tag{25}$$

which together with (20) and (21) leads us to the following three 'saddlepoint' equations for s, t and a:

$$s + \frac{2ta}{b^2} = 0$$

$$\frac{EXe^{sX+tX^2}}{Ee^{sX+tX^2}} = a$$

$$\frac{EX^2e^{sX+tX^2}}{Ee^{sX+tX^2}} = \frac{a^2}{b^2} \tag{26}$$

where E denotes expectation w.r.t. the distribution of X.
It was proved in [13] that there exists solutions $\hat{s}_0, \hat{t}_0$ and $\hat{a}_0$ of (26) such that

$$\hat{s}_0 > 0, \hat{t}_0 < 0 \text{ and } \hat{\mathrm{a}}_0 > 0 \tag{27}$$

provided the support of X contains at least three points. This important result will imply that Ee^{sX+tX^2} automatically exists (i.e. is finite) in a neighbourhood of $(\hat{s}_0, \hat{t}_0)$. This means that the strong moment condition which is typically required for saddlepoint approximations (cf. [9]), becomes superfluous for the 'self-normalized' statistic $\bar{X}_n/\bar{V}_n$ (cf. [13]). The same result holds true for the well-known Student t-statistic (cf. [13]) – which can also be written as $n^{\frac{1}{2}}\bar{X}_n/\bar{V}_n((n-1)/(n-(\sqrt{n}\bar{X}_n/\bar{V}_n)^2))^{\frac{1}{2}}$ – and for the studentized compound Poisson sum T_N (cf. [7]). The range of validity of saddlepoint approximations for these self-normalized statistics has therefore been extended from *df s*, whose densities have tails that die out at least as fast as the normal, to heavy tails like the Cauchy.
To proceed we note (cf. [7]) or [14] for complete details) that

$$\begin{aligned} P(\bar{X}_N/\bar{V}_N \geq b) &= P(N=0) + \\ &+ \sum_{n=1}^{\infty} P(\bar{X}_n/\bar{V}_n \geq b) P(N=n) \\ &= P(N=0) + \sum_{n=1}^{\infty} \iint_{\Omega_0(b)} f_{(\bar{X}_n \bar{Y}_n)}(x,y)dxdy P(N=n) \\ &+ P((\bar{X}_N, \bar{Y}_N)^T \in \Omega_1(b), N>0) \end{aligned} \tag{28}$$

Applying now the transformation $x = a, y = \frac{a^2}{b^2}$ and using a saddlepoint approximation for $f_{(\bar{X}_n \bar{Y}_n)}(a, \frac{a^2}{b^2})$ (based on an analysis similar to (19), the argument following it, and applying a Laplace approximation to the integral involved) the infinite sum in (28) reduces to

$$\sum_{n=1}^{\infty} \int_{\Omega_0(b)} \frac{n}{2\pi} \frac{\exp\{-n\Lambda(a,b)\}}{\det\{\Delta(a,b)\}^{1/2}} (1 + \frac{r_n}{n}) J(a,b) da db P(N=n) \tag{29}$$

where $\Delta(a,b)$ denotes a 2×2-matrix, with the second derivatives of $K(s,t)$ w.r.t. s and $\hat{t}$, evaluated at $\hat{s}$ and $\hat{t}$, as entries; r_n is bounded by some constant for all $n \geq 1$. The set $\Omega_0(b)$ is, for any $b \in (0,1)$, a small neighbourhood of $(a_0, \frac{a_0^2}{b^2})$, with $a_0 = \arg\inf_a \Lambda(a,b)$, on which the equations (20) and (21) have roots $\hat{s}_1, \hat{t}_1$, such that $\hat{t}_1 < 0; \Omega_1(b)$ denotes the complement of $\Omega_0(b)$. The probability corresponding to $\Omega_1(b)$, i.e. the last term on the r.h.s. of (28), is of negligible order of magnitude (cf. also [13]), while the same holds true for $P(N=0) = e^{-\nu}$, the first term on the r.h.s. of (28).
After some further computations, involving several Laplace approximations, we obtain (14), a Lugannani and Rice formula for the tail probability of a Studentized

compound Poisson sum. Note that $T_N/\sqrt{N}$ in (14) is equal to $\bar{X}_N/\bar{V}_N$ in (28), while $x = b, 0 < b < 1$. The relative error is of order ν^{-1}, while, in contrast, the Edgeworth expansion for T_N, i.e. (12), has only an absolute error of the order ν^{-1}.

3 Final comments

Edgeworth expansions like (12) generally provide accurate approximation near the center of the distribution, but the relative error can become unacceptable large in the far tail of the distribution. The saddlepoint approximation (14) will offer an approximation whose relative error is controlled near the centre and the far tail of the distibution. We also refer to [5], where saddlepoint approximations of Lugannani-Rice type for the trimmed mean and the studentized trimmed mean were established.

A common feature of compound sums, the topic of this paper, and trimmed means, useful in robust statistics, is that both statistics cannot be viewed as a smooth function of sample means, a class of statistics for which saddlepoint approximations were derived in great generality in the literature. To establish a saddlepoint approximation for a (studentized) compound sum or a (studentized) trimmed mean a conditioning argument is needed to reduce the problem to one involving smooth functions of sample means. In the case of (studentized) compound sums the conditioning is on N, the random number of summands in the compound sum, while in the case of (studentized) trimmed means, the conditioning is on the two extreme order statistics appearing in the trimmed mean.

In modern statistical practice (cf., section 9.5 of [3]) saddlepoint approximations are an important tool in obtaining highly accurate approximations (with small relative error) to the tail probabilities of a statistic under consideration, for instance a studentized compound Poisson sum.

In practical applications, however, theoretical saddlepoint approximations like (14) cannot be used, since they will depend on F, the unknown *df* of the observations, and on ν, the unknown parameter of the Poisson distribution of N (cf. (1)). Instead one can employ an empirical saddlepoint approximation to the tail probabilities, which is obtained from (14), simply by replacing F by $\hat{F}_N$ and ν by N. In [5] it is proved that replacing (F, ν) by $(\hat{F}_N, N)$ will not affect the relative error of the resulting empirical saddlepoint approximation. A related result for the Student t-statistic can be found in [12]

References

[1] J.A. Adell and A. Lekuona (2005), Sharp estimates in signed Poisson approximations of Poisson mixtures, *Bernoulli*, **11**, 47 – 65.

[2] H. Daniels, (1954), Saddlepoint approximations in statistics, *Ann. Math. Statist*, **25**, 631 – 650.

[3] A.C. Davison, D.V. Hinkley (1997), *Bootstrap Methods and their Application*, Cambridge Series in Statistical and Probabilistic Mathematics.

[4] R. Helmers (2000), Inference on rare errors using asymptotic expansions and bootstrap calibration, *Biometrika*, **87**, 689 – 694.

[5] R. Helmers, B-Y. Jing, G. Qin, W. Zhou (2004), Saddlepoint approximations to the trimmed mean, *Bernoulli*, **10**, 465 – 501.

[6] R. Helmers and B. Tarigan (2003), Compound sums and their applications in finance, *Proceedings ITB*, **34**, 2&3, 381 – 391 (edited by Andonowati & E. van Groesen).

[7] R. Helmers, B-Y. Jing, W. Zhou (2005), Saddlepoint approximation for studentized compound Poisson sum with no moment conditions, forthcoming.

[8] R. Helmers, W. Zhou (2005), The Edgeworth expansion for a Studentized compound Poisson sum, forthcoming.

[9] J.L. Jensen (1995). *Saddlepoint Approximations*, Clarendon, Oxford.

[10] P. Hall, Q. Wang (2004), Exact convergence rate and leading term in central limit theorem for Student's t-statistic, *Ann Probab. 32*, 1419 – 1437.

[11] B-Y. Jing, J. Robinson (1994). Saddlepoint approximations for marginal and conditional probabilities of transformed variables, *Ann. Statist.* **22**, 1115 – 1132.

[12] B-Y. Jing, A. Feuerverger, J. Robinson (1994). On the bootstrap saddlepoint approximations, *Biometrika*, **81**, 211 – 215.

[13] B-Y. Jing, Q.M. Shao, W. Zhou (2004), Saddlepoint approximation for Student t-statistic with no moment conditions., *Ann. Statist.* **32**, 2679 – 2711.

[14] W. Zhou (2004), Saddlepoint approximation for Student t-statistic with no moment conditions, unpublished PhD. Thesis, Hong Kong University of Science and Technology.

Strategic Planning for Dissemination of PMRI

R. K. Sembiring

Dept. of Mathematics, Institut Teknologi Bandung, Indonesia

Abstract: Pendidikan Matematika Realistik Indonesia (PMRI) is the Indonesian version of Realistic Mathematics Education (RME) develops by the Freudenthal Institute of the University of Utrecht in the Netherlands. The PMRI team has conducted a three years trialout, started with 12 primary schools and currently 27 schools, covers grade 1 up to now grade 4. The trialout is funded by DIKTI (DGHE) and a two year grant from PBSI from the Netherlands.

The trialout uses bottom-up and top-down approach, in the sense that at the one hand teachers, school principals and parents are involved in the process, while at the other hand experiences gained in the past are used to apply corrections and enhance the quality of the program.

Demand to disseminate PMRI to other schools is high and the current team could not cope with. The team has developed a four year plan to disseminate PMRI to other schools. This talk will explain the plan: Improve the teacher training institutes (LPTK) and through them disseminate PMRI in schools. DIKTI is expected to be the main sponsor of this program and helps from other government agencies: Dikdasmen, Balitbang, Minister of Religious Affairs etc are expected. A four year grant from NUFFIC is being proposed.

The Framework for the Implementation of Realistic Mathematics Education in Indonesia

Sutarto Hadi

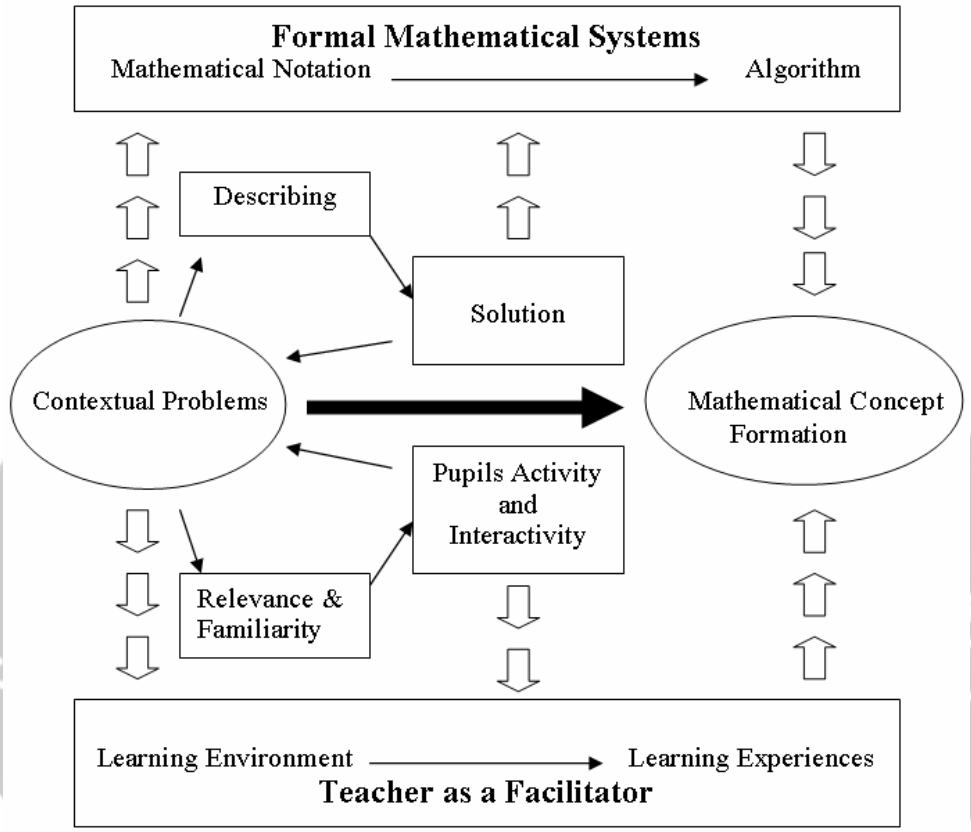
Lambung Mangkurat University, Indonesia

Abstract: Over the last few years there has been a paradigm shift in mathematics education in Indonesian. There has been a high concern particularly among the policy makers to reform mathematics teaching practices in schools. The goal of this reform is that mathematics learning to be meaningful for pupils and to give them appropriate competencies for study at the higher level or to enter work life. In the framework of the above new paradigm, the theory of realistic mathematics education (RME) is considered relevant to the mathematics education reform in the country.

However, the implementation of RME in Indonesia has some potential obstacles. First, the implementation cannot be done without the availability of RME curriculum materials that are suitable for characteristic of Indonesian contexts. Second, the obstacle stems from teachers' point of view. There are two types of teachers, namely those who support it, because they believe that RME is what they really need, and those who do not support it, because they think that RME cannot be used for all mathematics topics in the curriculum. Third, the obstacle comes from the behavior of pupils as passive learners. The change from teacher-centered to pupils-centered learning for many pupils is not easy and frustrating, because they used to being spoon-fed by the teachers.

In relation to Indonesian-contextualized RME (well-known as PMRI = *Pendidikan Matematika Realistik Indonesia*) implementation, it has been proposed the idea of didactical framework that can be used by stake-holders, especially mathematics teachers, mathematics teacher educators, script writers, and curriculum developers.

Didactical framework is a guideline that should be followed by developer (script writers) in designing PMRI exemplary curriculum materials, and a prerequisite of successful PMRI implementation in classroom lessons. Didactical framework consists of aspects: pupils, contextual problems, teachers, learning environment, and learning experiences. It is an integrated system that those aspects effectively and efficiently are intertwined each other, with the pupils as the center of instruction. The objective of the instruction is to develop pupils' understanding of mathematical concepts and ideas by using of contextual problems exploration based on reinvention process (Gravemeijer, 1994). Contextual problems should meet relevance and familiarity conditions (Sutarto Hadi, 2002). The role of the teacher as a facilitator is indicated by his/her ability to build pupils thinking process through an interactive learning environment. The didactical framework is depicted as the following figure.



The didactical framework is indicated by three types of arrows, namely bold arrow, block arrow, and line arrow. The bold arrow in the middle of the figure indicates the main concept of PMRI that the learning process should be started by giving pupils various contextual problems. By doing so they can involve immediately in the learning process meaningfuly. The purpose of the contextual problems is helping them to build their own mathematical ideas and concepts (mathematical concept formation). The bold arrow in the middle also separates the didactical framework into two parts. The above part shows the horizontal and vertical mathematizations (indicated by line arrows). Starting from contextual problems pupils begin their mathematical concept formation by describing the problems using their own symbols and notation. The next step is solving the problem. They do the same activities for other similar problems. In the long run it becoming formal mathematical procedure; it is indicated by block arrows in the direction from 'contextual problems' and 'solution' to 'formal mathematical system'. Furthermore, the below part shows the role of the teachers as facilitator and motivator in the learning processes. The role of the teacher is indicated by their ability to give pupils learning environment in order to facilitate them with rich learning expericences. However it can be done only if the contextual problems fulfill relevance and familiarity conditions. The conditions imply to pupils activity and interactivity. Teachers' knowledge of contextual problems and their ability to

develop interactivity is important for successful learning, which subsequently support pupils' mathematical concept formation.

References:

[1] Gravemeijer, K.P.E. (1994). *Developing realistic mathematics education*. Utrecht: CD-β Press.

[2] Sutarto-Hadi (2002). Effective Teacher Professional Development for the Implementation of Realistic Mathematics Education in Indonesia. Doctoral dissertation. Enschede: University of Twente.

Results on Path Like Trees

F.A. Muntaner-Batle

Unversidad Internacional de Catalua, Spain

Abstract: In this talk we will study the graceful and magic properties of the set of path like trees. Also other problems involving path like trees will be discussed during this talk.

Classification of Construction Techniques of Large Directed Graphs

Slamin

Mathematics Education Study Program, Universitas Jember, Indonesia

Abstract: There are several techniques for the construction of large digraphs, such as generalised de Bruijn digraphs, generalised Kautz digraphs, line digraphs, digon reduction, generalised digraphs on alphabets, partial line digraphs, digraphs constructed by the use of voltage assignments dan vertex deletion scheme. Some of these techniques produce new digraphs which are diregular while others produce non-diregular digraphs. Moreover, the construction techniques produce new digraphs with minimum diameter in various ranges of orders.

In this paper, we classify the construction techniques according to
(a) the general method of generating new digraphs;
(b) the diregularity of generated digraphs; and
(c) the range of orders of the generated digraphs.

λ-BACKBONE COLORINGS OF GRAPHS: KNOWN RESULTS AND OPEN PROBLEMS

A.N.M. Salman

ITB, Bandung, Indonesia

Abstract. In the application area of frequency assignment graphs are used to model the topology and mutual interference between transmitters. The problem in practice is to assign a limited number of frequency channels in an economical way to the transmitter in such a way that interference is kept at an 'acceptable level'. This has led to various different types of coloring problems in graphs. One of them is a λ-backbone coloring. Given an integer $\lambda \geq 2$, a graph $G = (V, E)$ and a spanning subgraph H of G (the backbone of G), a λ-backbone coloring of (G, H) is a proper vertex coloring $V \rightarrow \{1, 2, \ldots\}$ of G, in which the colors assigned to adjacent vertices in H differ by at least λ. In this paper we give a survey of the existing results about combinatorial and algorithmic aspects of λ-backbone coloring of graphs. Besides that, we discuss several open problems.

Key-words: backbone of a graph, λ-backbone coloring, λ-backbone coloring number, chromatic number, computational complexity.

1 Introduction

Coloring has been a central area in Graph Theory for more than 150 years. Some reasons for this are its appealingly simple definition, its large variety of open problems, and its many application areas. Whenever conflicting situations between pairs of objects can be modeled by graphs, and one is looking for a partition of the set of objects in subsets of mutually non-conflicting objects, this can be viewed as a graph coloring problem. This holds for classical settings like neighboring countries (map coloring) or interfering jobs on machines (job scheduling), as well as for more recent settings like colliding data streams in optical networks (wavelength assignment) or interfering transmitters and receivers for broadcasting, mobile phones and sensors (frequency assignment), to name just a few.

In the application area of frequency assignment graphs are used to model the topology and mutual interference between transmitters: the vertices of the graph represent the transmitters; two vertices are adjacent in the graph if the corresponding transmitters are so close (or so strong) that they are likely to interfere if they broadcast on the same or 'similar' frequency channels. The problem in practice is to assign a limited number of frequency channels in an economical way to the transmitters in such a way that interference is kept at an 'acceptable level'. This has led to various different types of coloring problems in graphs, depending on different ways to model the level of interference, the notion of similar frequency

channels, and the definition of acceptable level of interference, (See e.g. [6], [8]).

One of several different types of coloring in graphs is a λ-backbone coloring. A backbone coloring is introduced in [3] and [2]. In [3], a situation is modeled in which the transmitters form a network in which a certain substructure of adjacent transmitters (called the backbone) is more crucial for the communication than the rest of the network. This means more restrictions are put on the assignment of frequency channels along the backbone than on the assignment of frequency channels to other adjacent transmitters. The backbone could e.g. model hot spots in a (sensor) network where a very busy pattern of communications takes place (the sensors with the highest computational power and energy), whereas the other adjacent transmitters supply a more moderate service.

In this paper we present the existing results about combinatorial and algorithmic aspects of λ-backbone coloring of graphs. Besides that, we discuss several open problems. The paper is organised as follows. In the next section we present some terminologies. In Section 3 we present some known results about a relation between the λ-backbone coloring numbers and the chromatic numbers. In Section 4 we present sharp upper bounds for the λ-backbone coloring numbers of split graphs. In Section 5 we consider the λ-backbone coloring of planar graphs. In Section 6 we present the computational complexity of computing the λ-backbone coloring numbers of a graph. Finally, in the last section we present some open problems.

2 Terminology

For undefined terminology we refer to [1]. Let $G = (V, E)$ be a graph, where $V = V_G$ is a finite set of vertices and $E = E_G$ is a set of unordered pairs of two different vertices, called edges. A function $f : V \to \{1, 2, 3, \ldots\}$ is a *vertex coloring* of V if $|f(u) - f(v)| \geq 1$ holds for all edges $uv \in E$. A vertex coloring $f : V \to \{1, \ldots, k\}$ is called a *k-coloring*, and the *chromatic number* $\chi(G)$ is the smallest integer k for which there exists a k-coloring. A set $V' \subseteq V$ is *independent* if G does not contain edges with both end vertices in V'. By definition, a k-coloring partitions V into k independent sets $V_1, \ldots, V_k$.
Let H be a *spanning subgraph* of G, i.e., $H = (V_G, E_H)$ with $E_H \subseteq E_G$. Given an integer $\lambda \geq 2$, a vertex coloring f of G is a *λ-backbone coloring* of (G, H), if $|f(u) - f(v)| \geq \lambda$ holds for all edges $uv \in E_H$. The *λ-backbone coloring number* $\text{BBC}_\lambda(G, H)$ of (G, H) is the smallest integer ℓ for which there exists a λ-backbone coloring $f : V \to \{1, \ldots, \ell\}$.
A *path* is a graph P whose vertices can be ordered into a sequence $v_1, v_2, \ldots, v_n$ such that $E_P = \{v_1v_2, \ldots, v_{n-1}v_n\}$. The *distance* between two vertices u and v of a connected graph is the length of a shortest path between them. A *cycle* is a graph C whose vertices can be ordered into a sequence $v_1, v_2, \ldots, v_n$ such that $E_C = \{v_1v_2, \ldots, v_{n-1}v_n, v_nv_1\}$. A *tree* is a connected graph T that does not contain any cycles.

A *complete* graph is a graph with an edge between every pair of vertices. The complete graph on n vertices is denoted by K_n. A graph G is *complete p-partite* if its vertices can be partitioned into p nonempty independent sets $V_1, \ldots, V_p$ such that its edge set E is formed by all edges that have one end vertex in V_i and the other one in V_j for some $1 \leq i < j \leq p$.
A *star* S_q is a complete 2-partite graph with independent sets $V_1 = \{r\}$ and V_2 with $|V_2| = q$; the vertex r is called the *root* and the vertices in V_2 are called the *leaves* of the star S_q. In our context a *matching* M is a collection of pairwise disjoint stars that are all copies of S_1. We call a spanning subgraph H of a graph G

- a *tree backbone* of G if H is a (spanning) tree;
- a *path backbone* of G if H is a (Hamilton) path;
- a *star backbone* of G if H is a collection of pairwise disjoint stars;
- a *matching backbone* of G if H is a (perfect) matching.

3 A relation between the λ-backbone coloring numbers and the chromatic numbers

Obviously, $\text{BBC}_\lambda(G, H) \geq \chi(G)$ holds for any backbone H of a graph G. In order to analyze the maximum difference between these two numbers, let us consider the following values.
$\mathcal{T}_\lambda(k) = \max\{\text{BBC}_\lambda(G, T) \mid T \text{ is a tree backbone of } G, \text{ and } \chi(G) = k\}$;
$\mathcal{P}_\lambda(k) = \max\{\text{BBC}_\lambda(G, P) \mid P \text{ is a path backbone of } G, \text{ and } \chi(G) = k\}$;
$\mathcal{S}_\lambda(k) = \max\{\text{BBC}_\lambda(G, S) \mid S \text{ is a star backbone of } G, \text{ and } \chi(G) = k\}$;
$\mathcal{M}_\lambda(k) = \max\{\text{BBC}_\lambda(G, M) \mid M \text{ is a matching backbone of } G, \text{ and } \chi(G) = k\}$.

In 2003 Broersma et.al. [3] determined all the values $\mathcal{T}_2(k)$ and $\mathcal{P}_2(k)$, and observed that they roughly grow like $2k$ and $3k/2$, respectively. Their results are rewritten in Theorem 1 and Theorem 2.

Theorem 1

$$\mathcal{T}_2(k) = 2k - 1 \quad \textit{for } k \geq 1.$$

Theorem 2 *For $k \geq 1$ the function $\mathcal{P}_2(k)$ takes the following values:*

(a) for $1 \leq k \leq 4$: $\mathcal{P}_2(k) = 2k - 1$;

(b) $\mathcal{P}_2(5) = 8$ and $\mathcal{P}_2(6) = 10$;

(c) for $k \geq 7$ and $k = 4t$: $\mathcal{P}_2(4t) = 6t$;

(d) for $k \geq 7$ and $k = 4t + 1$: $\mathcal{P}_2(4t + 1) = 6t + 1$;

(e) for $k \geq 7$ and $k = 4t + 2$: $\mathcal{P}_2(4t + 2) = 6t + 3$;

(f) for $k \geq 7$ and $k = 4t + 3$: $\mathcal{P}_2(4t + 3) = 6t + 5$.

In 2004 Salman et.al considered cases where the backbone is a collection of pairwise disjoint stars or a perfect matching. In [9] was showed that for star backbones of G the number of colors needed for a λ-backbone coloring of (G, S) can roughly differ by a multiplicative factor of at most $2 - \frac{1}{\lambda}$ from the chromatic number $\chi(G)$. For the special case of matching backbones this factor is roughly $2 - \frac{2}{\lambda+1}$. Their precise behavior is summarized in Theorem 3 and Theorem 4.

Theorem 3 *For $\lambda \geq 2$ and $k \geq 2$ the function $\mathcal{S}_\lambda(k)$ takes the following values:*

(a) $\mathcal{S}_\lambda(2) = \lambda + 1$;

(b) for $3 \leq k \leq 2\lambda - 3$: $\mathcal{S}_\lambda(k) = \lceil \frac{3k}{2} \rceil + \lambda - 2$;

(c) for $2\lambda - 2 \leq k \leq 2\lambda - 1$ *with* $\lambda \geq 3$: $\mathcal{S}_\lambda(k) = k + 2\lambda - 2$; $\mathcal{S}_2(3) = 5$;

(d) for $k = 2\lambda$ *with* $\lambda \geq 3$: $\mathcal{S}_\lambda(k) = 2k - 1$; $\mathcal{S}_2(4) = 6$;

(e) for $k \geq 2\lambda + 1$: $\mathcal{S}_\lambda(k) = 2k - \lfloor \frac{k}{\lambda} \rfloor$.

Theorem 4 *For $\lambda \geq 2$ and $k \geq 2$ the function $\mathcal{M}_\lambda(k)$ takes the following values:*

(a) for $2 \leq k \leq \lambda$: $\mathcal{M}_\lambda(k) = \lambda + k - 1$;

(b) for $\lambda + 1 \leq k \leq 2\lambda$: $\mathcal{M}_\lambda(k) = 2k - 2$;

(c) for $k = 2\lambda + 1$: $\mathcal{M}_\lambda(k) = 2k - 3$;

(d) for $k = t(\lambda + 1)$ *with* $t \geq 2$: $\mathcal{M}_\lambda(k) = 2\lambda \cdot t$;

(e) for $k = t(\lambda + 1) + c$ *with* $t \geq 2$, $1 \leq c < \frac{\lambda+3}{2}$: $\mathcal{M}_\lambda(k) = 2\lambda \cdot t + 2c - 1$;

(f) for $k = t(\lambda + 1) + c$ *with* $t \geq 2$, $\frac{\lambda+3}{2} \leq c \leq \lambda$: $\mathcal{M}_\lambda(k) = 2\lambda \cdot t + 2c - 2$.

4 Sharp upper bounds for the λ-backbone coloring numbers of split graphs

In this section we consider the special case of λ-backbone colorings of split graphs with star backbones or matching backbones or tree backbones. A *split graph* is a graph whose vertex set can be partitioned into a *clique* (i.e. a set of mutually adjacent vertices) and an *independent set* (i.e. a set of mutually nonadjacent vertices), with possibly edges in between. The size of a largest clique in G and the size of a largest independent set in G are denoted by $\omega(G)$ and $\alpha(G)$, respectively. Split graphs were introduced by Hammer & Földes [7]; see also the book [5] by Golumbic.

The motivation for looking at split graphs is threefold. First of all, split graphs have nice structural properties. They form an interesting subclass of the class of perfect graphs. Hence, split graphs satisfy $\chi(G) = \omega(G)$. Secondly, every graph

can be turned into a split graph by considering any (e.g. a maximum) independent set and turning the remaining vertices into a clique. Thirdly, the number of colors needed to color the resulting split graph is an upper bound for the number of colors one needs to color the original graph. It will become clear from the results below that split graphs indeed serve us very well in this specific context, since they can provide considerably lower upper bounds on the numbers of colors we need than earlier results.

In [3] is also given sharp upper bounds for the 2-backbone coloring numbers of split graphs with tree backbones or path backbones as in the next theorem.

Theorem 5 *Let $G = (V, E)$ be a split graph with $\chi(G) = k$.*

(a) For every spanning tree $T = (V, E_T)$ of G,

$$\mathrm{BBC}_2(G, T) \leq \begin{cases} 1 & \text{if } k = 1 \\ 3 & \text{if } k = 2 \\ k + 2 & \text{if } k \geq 3. \end{cases}$$

(b) For every Hamilton path $P = (V, E_P)$ of G,

$$\mathrm{BBC}_2(G, P) \leq \begin{cases} 1 & \text{if } k = 1 \\ k + 1 & \text{if } k = 2 \text{ or } k \geq 4 \\ 5 & \text{if } k = 3. \end{cases}$$

The bounds are tight.

We can generalize the results in Theorem 5(a) for the λ-backbone coloring numbers of split graphs with tree backbones as follows.

Theorem 6 *Let $\lambda \geq 2$ and let $G = (V, E)$ be a split graph. For every tree backbone T of G,*

$$\mathrm{BBC}_\lambda(G, T) \leq \begin{cases} 1 & \text{if } \chi(G) = 1 \\ 1 + \lambda & \text{if } \chi(G) = 2 \\ \chi(G) + \lambda & \text{if } \chi(G) \geq 3. \end{cases}$$

The bounds are tight.

Proof of the upper bounds. Let $G = (V, E)$ be a split graph with a spanning tree $T = (V, E_T)$. Let C and I be a partition of V such that C with $|C| = k$ is a clique of maximum size, and such that I is an independent set. Since split graphs are perfect, $\chi(G) = \omega(G) = k$. The case $k = 1$ is trivial. If $k = 2$ then G is bipartite, and we use colors 1 and $\lambda + 1$. For $k \geq 3$, we consider the restriction of the tree T to the vertices in C, and we distinguish two cases.
In the first case, the restriction of T to C forms a star $K_{1,k-1}$. Let $v_1, \ldots, v_{k-1}$ denote the $k-1$ leaves of this star, and let v_k denote its center. For $i = 1, \ldots, k-1$ we color v_i with color i, and we color v_k with color $k + \lambda - 1$. This yields a λ-backbone coloring for the vertices in C. All vertices $u \in I$ are leaves in the tree

T. Any vertex $u \in I$ with $uv_k \notin E_T$ can be safely colored with color $k+\lambda$. It remains to consider vertices $u \in I$ with $uv_k \in E_T$. In the graph G, such a vertex u is nonadjacent to at least one of the vertices $v_1, \ldots, v_{k-1}$, say to vertex v_j (otherwise, the clique C could be augmented by vertex u and would not be of maximum size as we assumed). In this case we may color u with color j.
In the second case, the restriction of T to C does not form a star. In this case the restriction of T to C has a proper 2-coloring $C = C_1 \cup C_2$ with $|C_1| = a \geq |C_2| = b \geq 2$. Then there exist a vertex $x \in C_1$ and a vertex $y \in C_2$ for which $xy \notin E_T$. Let $v_1, \ldots, v_a = x$ be an enumeration of the vertices in C_1, and let $y = v_{a+1}, \ldots, v_{a+b}$ be an enumeration of the vertices in C_2. For $i = 1, \ldots, a$ we color vertex v_i with color $i+1$. For $i = 1, \ldots, b$ we color vertex v_{a+i} with color $a+\lambda+i-1$. This yields a λ-backbone coloring of C with colors in $\{2, \ldots, k+\lambda-1\}$. We color each vertex $u \in I$ with color

$$\begin{cases} k+\lambda & \text{if } uv \in E_T \text{ and } v \in C_1 \\ 1 & \text{if } uv \in E_T \text{ and } v \in C_2. \end{cases}$$

This yields a λ-backbone $(k+\lambda)$-coloring of (G,T), since the colors of a vertex v_i with $i \in \{1, \ldots, a\}$ and of any vertex $u \in I$ such that $uv_i \in E_T$ have distance at least $k+\lambda-(i+1) \geq k+\lambda-(k-2+1) > \lambda$, and since the colors of a vertex v_i with $i \in \{a+1, \ldots, b\}$ and of any vertex $u \in I$ such that $uv_i \in E_T$ have distance at least $a+\lambda+i-1-1 \geq k/2+\lambda-1 \geq \lambda$.

Proof of the tightness of the bounds. The cases $k = 1$ and $k = 2$ are trivial. For $k \geq 3$, we consider a split graph with a clique of k vertices $v_1, \ldots, v_k$ and with an independent set of $(k-2)(k-1)/2$ vertices $u_{i,j}$ with $1 \leq i < j \leq k-1$. Every vertex $u_{i,j}$ is adjacent to all vertices v_s with $s \neq i$. The tree backbone T contains the $k-1$ edges $v_k v_s$ with $1 \leq s \leq k-1$. The vertices $u_{i,j}$ form the leaves of T; in the tree, vertex $u_{i,j}$ is adjacent only to v_j. Clearly, $\chi(G) = k$.
Suppose to the contrary that $\text{BBC}_\lambda(G,T) \leq k+\lambda-1$, and consider such a backbone coloring. The vertices $v_1, \ldots, v_k$ in the clique must be colored with k pairwise distinct colors. Since they form a star, either vertex v_k has color 1, and colors $2, \ldots, \lambda$ are not used on the clique, or vertex v_k has color $k+\lambda-1$, and colors $k, \ldots, k+\lambda-2$ are not used on the clique. Both cases are symmetric, and we assume without loss of generality that v_k has color $k+\lambda-1$ and that colors $k, \ldots, k+\lambda-2$ are not used on the clique. Let v_i be the vertex that has color $k-2$, and let v_j be the vertex that has color $k-1$. The vertex $u_{i,j}$ is adjacent to all clique vertices except v_i; hence, it could only be colored with color $k-2$ or with a color in $\{k, \ldots, k+\lambda-2\}$. But these λ colors are forbidden for $u_{i,j}$, since in the tree backbone it is adjacent to vertex v_j with color $k-1$. Since there is no feasible color for $u_{i,j}$, we arrive at the desired contradiction.

In [4] the authors discuss the λ-backbone coloring numbers of split graphs with star backbones or matching backbones. We summarize the results in Theorem 7 and Theorem 8.

Theorem 7 *Let $\lambda \geq 2$ and let $G = (V, E)$ be a split graph with $\chi(G) = k \geq 2$. For every star backbone $S = (V, E_S)$ of G,*

$$\text{BBC}_\lambda(G, S) \leq \begin{cases} k + \lambda & \text{if either } k = 3 \text{ and } \lambda \geq 2 \text{ or } k \geq 4 \text{ and } \lambda = 2 \\ k + \lambda - 1 & \text{in the other cases.} \end{cases}$$

The bounds are tight.

Theorem 8 *Let $\lambda \geq 2$ and let $G = (V, E)$ be a split graph with $\chi(G) = k \geq 2$. For every matching backbone $M = (V, E_M)$ of G,*

$$\text{BBC}_\lambda(G, M) \leq \begin{cases} \lambda + 1 & if\ k = 2 \\ k + 1 & if\ k \geq 3 \text{ and } \lambda \leq \min\{\frac{k}{2}, \frac{k+5}{3}\} \\ k + 2 & if\ k = 9 \text{ or } k \geq 11 \text{ and } \frac{k+6}{3} \leq \lambda \leq \lceil \frac{k}{2} \rceil \\ \lceil \frac{k}{2} \rceil + \lambda & if\ k = 3, 5, 7 \text{ and } \lambda \geq \lceil \frac{k}{2} \rceil \\ \lceil \frac{k}{2} \rceil + \lambda + 1 & if\ k = 4, 6 \text{ or } k \geq 8 \text{ and } \lambda \geq \lceil \frac{k}{2} \rceil + 1. \end{cases}$$

The bounds are tight.

5 λ-Backbone colorings of planar graphs

In this section we consider the λ-backbone coloring numbers of planar graphs. The four-color theorem together with Theorem 1 and Theorem 4, implies that $\text{BBC}_2(G, T) \leq 7$ and $\text{BBC}_2(G, M) \leq 6$ holds for any planar graph G with spanning tree T and matching backbone M, respectively. However, these bounds are probably not best possible.

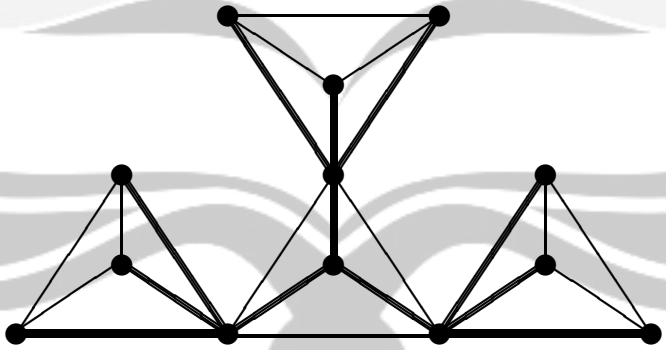


Figure 1: A planar graph G_1 with a spanning tree T (bold edges) such that $\text{BBC}(G_1, T) = 6$.

For the case planar graphs with tree backbones, can the bound be improved to 6? However, in [3], the planar graph G_1 in Figure 1 demonstrates that this bound

can not be improved to 5: Note that graph G_1 consists of four copies of K_4 that all have a $K_{1,3}$ as spanning tree. In any backbone coloring of such a K_4 with only five colors, the central vertex of the $K_{1,3}$ must either receive color 1 or color 5. With this observation, it is easy to see that $\text{BBC}(G_1, T) = 6$.

For the case planar graphs with matching backbones, can the bound be improved to 5? However, in [9], the planar graph G_1 with indicated matching backbone M consisting of edges ab', bc', cd', da' as in Figure 2 shows that one cannot improve this bound to 4. Note that we cannot find a backbone coloring of (G_1, M) with color set $\{1, 2, 3, 4\}$. First of all observe that G_1 can be obtained from a plane embedding of the K_4 induced by the vertices a, b, c, d, by putting a new vertex in each face and adding edges from this new vertex to the three vertices on the boundary of the face, and assigning the label x' to the new vertex in the triangular face bounded by the cycle $uvwu$, where $\{u, v, w, x\} = \{a, b, c, d\}$. Suppose we only use colors $1, 2, 3, 4$. Then it is clear from this construction that a, b, c and d get different colors, and that the colors of a vertex and its primed counterpart are the same. Without loss of generality assume that a and a' get color 2. Then both b' and d must get color 4, a contradiction. It is routine to check that $\text{BBC}_2(G_1, M) = 5$.

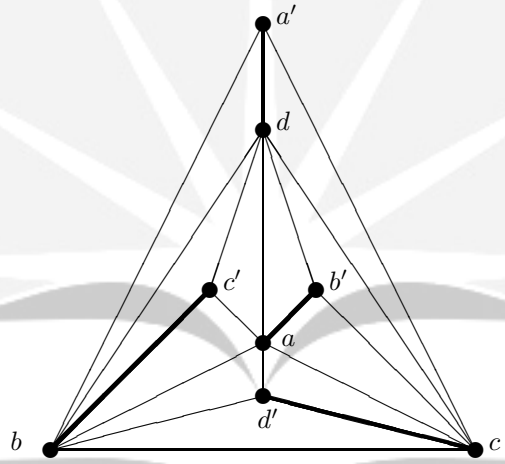


Figure 2: A graph G_2 with a matching backbone M (bold edges) such that $\text{BBC}_2(G_2, M) = 5$.

6 The computational complexity of computing the λ-backbone coloring number

We consider the computational complexity of computing the λ-backbone coloring number: *"Given a graph G, a spanning subgraph H, and an integer ℓ, is $\text{BBC}_\lambda(G, H) \leq \ell$?"* Of course, this general problem is NP-complete.

In [3] the authors considered the computational complexity of computing the 2-backbone coloring number of the graph G with a tree backbone. Then, in [10] the result is generalized for the computational complexity of computing the λ-backbone coloring number of the graph G with a tree backbone. The authors show that for this problem the complexity jump occurs between $\ell = \lambda + 2$ (easy for all tree backbones T) and $\ell = \lambda + 3$ (difficult even for path backbones P).

Theorem 9 *Let $\lambda \geq 2$.*

(a) The following problem is polynomially solvable for any $\ell \leq \lambda + 2$: Given a graph G and a spanning tree T, decide whether $\text{BBC}_\lambda(G, T) \leq \ell$.

(b) The following problem is NP-complete for all $\ell \geq \lambda + 3$: Given a graph G and a Hamiltonian path P, decide whether $\text{BBC}_\lambda(G, P) \leq \ell$.

In [4] the authors considered the computational complexity of computing the λ-backbone coloring number of the graph G with a star backbone. They show that for this problem the complexity jump occurs between $\ell = \lambda + 1$ (easy for all star backbones S) and $\ell = \lambda + 2$ (difficult even for matching backbones M).

Theorem 10 *Let $\lambda \geq 2$.*

(a) The following problem is polynomially solvable for any $\ell \leq \lambda + 1$: Given a graph G and a star backbone S, decide whether $\text{BBC}_\lambda(G, S) \leq \ell$.

(b) The following problem is NP-complete for all $\ell \geq \lambda + 2$: Given a graph G and a matching backbone M, decide whether $\text{BBC}_\lambda(G, M) \leq \ell$.

7 Open problems

In this section we present some open problems.

- For an *arbitrary* graph G with spanning tree T, $\text{BBC}_2(G, T)$ can be as large as $2\chi(G) - 1$. How about $\text{BBC}_\lambda(G, T)$ for $\lambda \geq 3$? What about *triangle-free* graphs? Does there exist a small constant c such that $\text{BBC}_\lambda(G, T) \leq \chi(G) + c$ for all triangle-free graphs G?

- What about *chordal* graphs? It can be shown that $\text{BBC}_2(G, P) \leq \chi(G) + 4$ whenever G is chordal and P is a Hamilton path of G. Does this result carry over to arbitrary spanning trees, i.e., does $\text{BBC}_\lambda(G, T) \leq \chi(G) + c$ hold for any chordal graph G with spanning tree T?

- How to prove the upper bounds for the backbone coloring number of a planar graph without using the four-color theorem?

- What is the computational complexity of the λ-coloring problem? More specially, for which graph classes and backbones is the problem polynomially solvable and for which is it NP-hard?

References

[1] J.A. Bondy & U.S.R. Murty (1976), *Graph theory with applications*, Macmillan, London.

[2] H.J. Broersma (2004), A general framework for coloring problems: old results, new results and open problems, *Lecture Notes in Computer Science*, **3330**, 65–79.

[3] H.J. Broersma, F.V. Fomin, P.A. Golovach & G.J. Woeginger (2003), Backbone colorings for networks, *Lecture Notes in Computer Science*, **2880**, 131–142.

[4] H.J. Broersma, L. Marchal, D. Paulusma & A.N.M. Salman (2005), New upper bounds for λ-backbone colorings along pairwise disjoint stars and matchings, *Preprint*.

[5] M.C. Golumbic (1980), *Algorithmic graph theory and perfect graphs*, Academic Press, New York.

[6] W.K. Hale (1980), Frequency assignment: Theory and applications, *Proceedings of the IEEE 68*, 1497–1514.

[7] P.L. Hammer & S. Földes (1977), Split graphs, *Congressus Numerantium*, **19**, 311–315.

[8] R.A. Leese (1999), Radio spectrum: a raw material for the telecommunications industry, *Progress in Industrial Mathematics at ECMI 98*, Teubner, Stuttgart, 382–396.

[9] A.N.M. Salman, H.J. Broersma, J. Fujisawa, L. Marchal, D. Paulusma & K. Yoshimoto (2004), λ-Backbone colorings along pairwise disjoint stars and matchings, *Submitted*.

[10] A.N.M. Salman, H.J. Broersma & D. Paulusma (2005), The computational comlexity of λ-backbone coloring, *Preprint*.

A.N.M. Salman: Department of Mathematics, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia.
E-mail: msalman@dns.math.itb.ac.id

Near Field Optics and Its Applications

M.O. Tjia

Department of Physics, Institut Teknologi Bandung, Indonesia

Abstract: Characteristics of electromagnetic field in the vicinity of its source are described and compared with those observed at a distance far away from the source. These characteristics are related to the so called "low dimensionality" of the wave based on the "dimension" of the wave number vector. The possibility of avoiding the Abbe's diffraction limit and hence probing finer details of the emitting object is explained. The near fields in evanescent and surface waves occurring in various configurations and structures are illustrated for their potential applications such as high spatial resolution scanning near-field optical microscopy (SNOM) and optical recording as well as nano manipulation required for fabrication of nano-optical devices.

Rational Solitons in Deep Nonlinear Optical Bragg Grating

Husin Alatas

Depatment of Physics, Institut Teknologi Bandung, Indonesia

Abstract: We have examined the rational solitons in the Generalized Coupled Mode model for a deep nonlinear Bragg grating. These solitons are a degenerate form of the ordinary solitons and appear at the transition lines in the parameter plane. A simple way is presented to investigate the bifurcations, which is applied on bifurcations due to detuning the frequency. Exact expressions for the corresponding rational solitons are presented. The analysis yields among others the appearance of in-gap dark and antidark rational solitons unknown in the nonlinear shallow grating. It is demonstrated that certain effects in the soliton energy variations are to be expected when the frequency is varied across the values where the rational solitons appear.

Numerical Modelling of Nonuniform Sinusoidal Bragg Grating

Agus Suryanto

Department of Mathematics, Brawijaya University, Malang, Indonesia

Abstract: The problem of light wave propagating in almost periodic, nonuniform and linear sinusoidal Bragg grating is considered. For this purpose we implement a Method of Single Expression. The basic feature of this method is that the Helmholtz equation is represented in the form of coupled phase-amplitude equations, from which the boundary value problem is reduced to an initial value problem of a set of ordinary differential equations. Then the Runge-Kutta integration method is applied to solve the resulting initial value problem. To demonstrate the method, we implement the method to characterize the transmission spectrum of tapered, chirped and phase-shifted grating structure.

Keywords: Mathematical Modelling, Optics

SMA – didactics for Applied Mathematics

Gerard Jeurnink

Department of Applied Mathematics, University of Twente, The Netherlands

Abstract: Educating mathematics to teenagers can be a challenging activity. Many secondary school students experience the abstraction as being an inevitable, always returning obstacle. At the level of primary school and junior high school Realistic Mathematics Education (RME) indeed has taken away a lot of aversion to this discipline. In order to recognize the usefulness of higher mathematics, we want to make senior high school students familiar with Applied Mathematics. Not only by stating the applications, but also trying to comprehend these thoroughly. We will present materials for the benefit of math lessons with interaction and cooperation between students and teacher. In classroom we are prepared to discuss student solutions with quick feed back.

The success story of RME seems to end with the change-over to (senior) high school. In fact, understanding an advanced application of mathematics is asking for a more detailed description of the mathematical model. Exactly here we find a challenge to the student for exploring the present case (and the mathematics behinds it). The operating study books all have theoretical foundations in common, and this is still very welcome to a math course. But every student gets extraordinary more motivated by going deeply into one or more real(istic) applications (taking into account student's level and theoretical knowledge).

New written small booklets can serve these additional or replacing lessons. In the Netherlands they have already introduced the so-called Zebra-booklets, to be meant for the upper grades of high school. In this lecture I'll also present some 'learner letters' for serving middle grades. Such a letter concerns one specific math application (waves, codes, etc) and can be managed within three lessons. As mentioned, the didactic is focused on interaction and cooperation, still the student has to learn fundamental mathematics.

Keywords: modeling, SMA-education

The Role of Noticing from the Teacher in Supporting Teaching and Learning Process to Promote Constructivism in Mathematics Classroom

Jozua Sabandar

Department of Mathematics Education, UPI Bandung, Indonesia

Abstract: In a mathematics learning process it is demanded from the students to actively engage in order to construct their knowledge, whereas teachers are supposed not to dominate the teaching and learning activities. In the process of the students construct their knowledge, it is imperative that the lesson should start with posing contextual problems such that the lesson will be meaningful and useful to the students. It is also expected that in order to increase the quality of the teaching and learning activities the students as well as the teacher should be responsible to their roles to help the students to reach their optimal capacity. The students should make use of what they have already had (prior knowledge, prerequisite, and their experiences) to support the active interactions among the students and between the students and the teacher. The student will have opportunity wherein they could freely express their ideas and strategies in a variety ways without being discourage of making mistakes. Therefore the teacher is demanded to be well prepared in this kind of process since they must be ready to accommodate questions, responses, comments and different solution no matter how idiosyncratic it is. Here the teacher will serve as facilitator, moderator or conductor. The teacher will not mainly provide problems and just leaves the students to work on their own and the teacher wait for the correct answer or different strategies. The teacher must be able to anticipate whatever solution that may arise from the students, as well as their failures to continue working on the problem. So, the teacher must make a good plan, and good preparation as well regarding selecting problems, and trying to solve those problems in every possible way.

In choosing this approach. It is required from the teacher to notice, observe students' works, solutions, questions, and comment and to follow it with appropriate actions and reflecting on the students' ideas at a right moment. In short, to support the teacher in a teaching and learning process emphasizing on the students construct their knowledge, the teacher should pay attention on noticing the students and make a good reflection as well on what has been developed so far in that process after the teacher set a plan, preparation/anticipation, observe/listen, and take action on a right moment.

Keywords: interaction, noticing, constructivism

Constructing Mathematical Concepts Through Activity by using Various Kinds of Representation

Yansen Marpaung

Department of Mathematics Education, University of Sanata Dharma Yogyakarta, Indonesia

Abstract: The new theories in mathematics education, like RME (realistic mathematics education), or CTL (Contextual Teaching-Learning) stressed the invention (or reinvention) of mathematical knowledge through activity. These new theories are supported by the philosophy of constructivism.

To apply these theories in the teaching-learning of mathematics in primary and secondary schools, especially in Indonesian school systems which traditionally governed by the behaviouristic approach, is not easy.

Since 2001 the department of mathematics education in the University of Sanata Dharma are involved in a project of reforming the mathematics teaching learning in the primary schools. Three other universities which play active role in that project are UPI, Bandung, UNY Yogyakarta and UNESA. From my experience in this project, I am convinced that reformation in teaching-learning of mathematics in schools depends in the first hand on the teachers' competencies. Therefore, in some lecture at our department I developed the students' consciousness about the importance of the reinvention of mathematics concepts through activity. I used the didactical material "dynamical maze or Dynamische Labyrinthe" created by Prof. Dr. E. Cohors-Fresenborg as starting point in constructing/ reinventing the concept of algorithm and connected it to how information are processed cognitively.

Keywords: Dynamical maze, algorithm, construction, representation

Extreme Wave Events in a Hydrodynamic Laboratory

Natanael Karjanto, E. van Groesen

Applied Analysis and Mathematical Physics Group, University of Twente, The Netherlands

Abstract: We adopt the spatial 'nonlinear Schrödinger' (NLS) equation as a simple mathematical model for nonlinear surface wave evolution in a wave tank of a hydrodynamic laboratory. This equation has many exact solutions, of which we only study a family of solutions which describes extreme wave events in the laboratory. This solution is known as the 'Soliton on Finite Background' (SFB) and it describes the 'modulation instability' (Akhmediev and Ankiewicz, 1997). In the context of water waves, this instability is known as the 'Benjamin-Feir instability' since Benjamin and Feir (1967) investigated the stability of a modulated wave train both theoretically and experimentally. In this presentation, we focus on the physical properties of the SFB and the comparison with experiments in a wave tank of the Maritime Research Institute Netherlands (MARIN).

The SFB in the far field is a 'continuous wave' with a constant amplitude $2r_0$ and a wave frequency ω_0, modulated with a modulation frequency ν. While running downstream, this signal becomes an extreme wave event; reaches a large amplitude at a certain position while preserving the modulation period. After passing the extreme position, it continues to evolve downstream to a situation similar to the initial signal. At the position where the SFB is an extreme wave event, there are times at which the real-valued amplitude vanishes and the phase becomes undefined, resulting in a 'phase singularity' phenomenon. In one modulation period, there is a pair of these singularities. Due to this phenomenon, the physical wave field shows a 'wavefront dislocation', when merging waves or splitting waves are observed.

The experimental setup is a wave tank of 200 m long, with 3.55 m water depth, a wavemaker on one side, and an absorbing beach on the other side. Wave gauges are installed at several positions to capture signals of the generated wave. Applying the inverse problem and using the 'maximum temporal amplitude' (Andonowati and van Groesen, 2003), we designed the initial signal of the SFB such that extreme wave events should occur at a specified position, namely at 150 m from the wavemaker. Due to the discrete positions of the wave gauges, the precise position of the extreme wave cannot be determined very well. However, the experiments showed non-breaking waves with a large amplitude. These waves have an asymmetric structure compared to the ones from the theoretical SFB. Furthermore, the experimental results also show a phase singularity phenomenon. Yet, instead of a pair of singularities as in the SFB, the experiments show only one singularity in one modulation period (van Groesen et al, 2005; Huijsmans et al, 2005). Keywords: freak wave event, Soliton on Finite Background, phase singularity and wavefront dislocation.

References:

[1] N. N. Akhmediev and A. Ankiewicz. Solitons–Nonlinear Pulses and Beams, volume 5 of Optical and Quantum Electronic Series. Chapman & Hall, first edition, 1997.

[2] T. B. Benjamin and J. E. Feir. The disintegration of wave trains on deep water. Part 1. Theory. Journal of Fluid Mechanics, 27(3):417-430, 1967.

[3] Andonowati and E. van Groesen. Optical pulse deformation in second order nonlinear media. Journal of Nonlinear Optical Physics and Materials, 12(2):221-234, 2003.

[4] E. van Groesen, Andonowati, and N. Karjanto. Deterministic aspects of nonlinear modulational instability. Proceedings of Rogue Waves 2004, Brest, France, 2005.

[5] R. H. M. Huijsmans, G. Klopman, N. Karjanto, and Andonowati. Experiments on extreme wave generation using the Soliton on Finite Background. Proceedings of Rogue Waves 2004, Brest, France, 2005.

Predicting extreme locations of propagating waves from originally tri-chromatic signals through Maximal Temporal Amplitude (MTA)

Marwan[1,2]

[1)] Department of Mathematics, Institut Teknologi Bandung, Indonesia
[2)] Department of Mathematics, Universitas Kuala, Nangroe Aceh Darrussalam, Indonesia

Abstract: Previous numerical and experimental results on the propagation of surface waves which are originally bi-chromatic signals at the wave maker show phenomena of peaking and splitting. To investigate this, in particular the peaking phenomenon, a quantity called Maximal Temporal Amplitude (MTA) is introduced. This quantity can be used to predict the position of the occurrence of maximal peaking within a wave tank for a given signal at the wave maker. In this paper, we present the MTA of propagating waves that are tri-chromatic signals at the wave maker. We predict the position of maximal peaking of the signal from the location of the maximum MTA. Here, a tri-chromatic signal is superposition of three mono-chromatic signals with different amplitudes, frequencies and initial phases. As a model of unidirectional waves, we use KdV equation with exact dispersion relation. We approximate the solution of the model using perturbation method in the power series of amplitude and wave number up to third order. We show that the initial phases influence the location of maximum MTA.

Keywords: Phase, tri-chromatic signal, Maximal Temporal Amplitude

The dispersion relation for waves above arbitrary currents

H. Margaretha, E. van Groesen

Applied Analysis and Mathematical Physics Group, University of Twente, The Netherlands

Abstract: Waves on the surface of a layer of fluid are described in the approximation of small elevations (linear wave theory) by a dispersion relation, which is an algebraic relation between the wave frequency and the wavenumber.

For waves above a uniform (depth-independent) current, the presence of a constant current simply adds a linear contribution to the frequency, reflecting the invariance of the physical phenomenon for translations with a uniform current speed.

When the current profile depends on the depth the situation becomes much more complicated. Nevertheless, for any current profile, a linear wave of (suitably) given frequency can still be found, but its wavenumber depends in a complicated way on the current profile. To find the dispersion relation in that case, the equation for the vertical fluid motion has to be studied (i.e. the so-called Rayleigh equation with impermeable bottom boundary condition). Except in the special case of a (piecewise) linear current, this equation cannot be solved explicitly and approximations have to be found.

We will show that the Rayleigh problem is actually a variational problem (i.e. has optimization properties), and that the relevant functional (or its dual formulation) is directly related to the dispersion relation. Taking trial functions that are in some sense approximations of the solutions and inserting them into the functional lead to approximations of the dispersion relation for depth-dependent currents.

A Numerical Simulation of Nonlinear Ocean Wave Focusing using the Convex-lens Like Submerged Breakwater

Erwandi

Indonesian Hydrodynamic Laboratory, Agency for the Assessment and Application of Technology
Jl. Hidrodinamika, Komplek Kampus ITS, Sukolilo, Surabaya

Abstract: This paper describes the preliminary study on ocean wave-focusing using the convex-lens like submerged breakwater. Some submerged breakwaters are set up to resemble convex-lens shape and the performance when the ocean-waves pass over them is analyzed. Since the formation of the breakwater is submerged 1 meter below the free surface, then the waves that pass over it are considered non–linear. The continuity and the momentum equation are governed in the computational domain and the free surface boundary updated at each time step by an explicit finit-difference time-marching scheme. The incident waves are generated from the inflow boundary by prescribing a simple sinusoidal wave. The numerical simulation results show that the wave can be focused. The focal-length is also fulfilled the theory of optical lens.

Keywords: Ocean wave focusing, Convex-lens,

CONSTRUCTION AND TRAINING OF RADIAL BASIS FUNCTION NEURAL NETWORK

Brodjol S[a], Subanar[b] and S. Guritno[b]

[a] Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[b] Universitas Gadjah Mada, Yogyakarta, Indonesia

Abstract: The radial basis function models are closely related to function approximation models used to perform interpolation that the performance depends on number of centers and the prototype used for training. Pototype of RBF is determined by generator function, in this case use exponential generator function. Training of RBF neural networks using gradient descent offers a solution to the tradeoff between performance and training speed. Implementation of RBF neural network use Indonesia Inflation data. According time series analysis, the data have model ARIMA(0,0,1)(0,0,1)11. RBF network does not have best model on training data or testing data.

Key-words: radial basis function, generator function, gradient descent

1. Introduction

Radial basis function (RBF) neural networks are function approximation models that can be trained by examples to implement a desired input–output mapping [1]. In fact, radial basis function models are closely related to function approximation models used to perform interpolation [6]. The performance of an RBF neural network depends on the number and centers of the RBF, their shapes, and the method used for learning the input–output mapping. Broomhead and Lowe [2] suggested that the centers of the RBF can be distributed uniformly within the region of the input space for which there is data, or chosen to be a subset of the training vectors. Moody and Darken [7] proposed a hybrid learning process for training RBF neural networks with Gaussian RBF, which employs a supervised scheme for updating the output weights, i.e., the weights that connect the RBF with the output units, and an unsupervised clustering algorithm for determining the centers of the RBF. The centers of the RBF are often determined by the K–means clustering algorithm. Poggio and Girosi [8] proposed a supervised approach for training RBF neural networks with Gaussian RBF, which updates the RBF centers together with the output weights. Chen *et al.* [3], proposed a learning procedure for RBF neural networks based on the orthogonal least squares (OLS) method, which is used as a forward regression procedure to select a suitable set of RBF centers.

The training of RBF neural networks using gradient descent offers a solution to the tradeoff between performance and training speed and can make RBF neural networks serious competitors to feedforward neural networks (FFNNs) with sigmoid

hidden units [4]. The convergence of gradient descent learning and the performance of the trained RBF neural networks are both affected rather strongly by the choice of RBF. The search for RBF functions other than the Gaussian function motivated the development of an axiomatic approach for constructing reformulated RBF neural networks suitable for gradient descent learning [4-5]. This approach reduces the development of reformulated RBF models to the selection of admissible generator functions that determine the form of the radial basis functions.

This paper presents new results on the construction and training of reformulated radial basis function neural networks. The results of the analysis presented in this paper can be used for selecting generator functions according to their suitability for gradient descent learning.

2. Reformulated Radial Basis Function

Consider the $\Re^n \to \Re^1$ mapping implemented by the model

$$\hat{y} = f\left(\omega_0 + \sum_{j=1}^{c} \omega_j g_j \left(\left\| x - v_j \right\| \right) \right) \qquad (1)$$

where $f(.)$ is a nondecreasing, continuous and differentiable function. The model (1) describes an RBF neural network with inputs from $\Re^n$, c radial basis functions, and output units y. When the RBF network is presented x as input and y response with c prototype v_j, as shown in figure 1, where $\phi_j = g_j\left(\left\| x - v_j \right\|^2 \right)$.

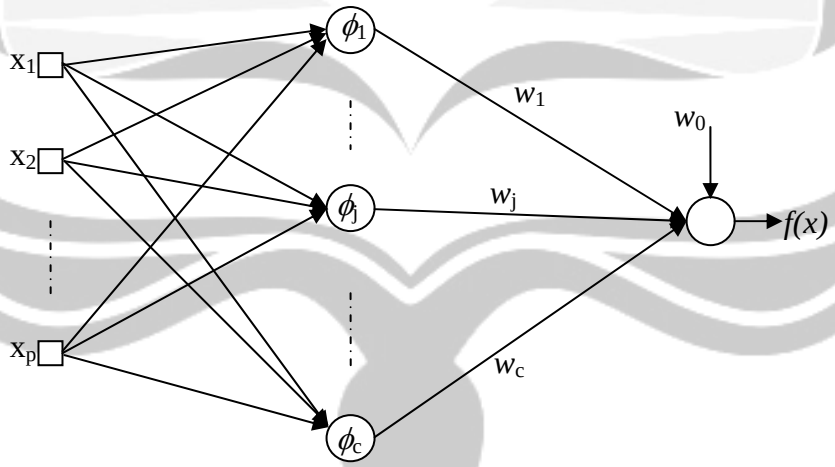


Figure 1. Architecture RBF neural network

A. Axiomatic Requirements

Reformulated RBF neural networks were developed to facilitate the training of RBF models by learning algorithms based on gradient descent [4]. In order for the model (1) to satisfy the desired properties mentioned above, any admis-sible radial basis function must satisfy the following three basic axiomatic requirements [4]:

Axiom 1: $g_j\left(\|x-v\|^2\right)>0$ for all $x,v\in\Re^n$.

Axiom 2: $g_j\left(\|x-v\|^2\right)>g_j\left(\|y-v\|^2\right)$or all $x,y,v\in\Re^n$

such that $\|x-v\|^2<\|y-v\|^2$.

Axiom 3: If $\nabla_x g_j=\nabla_x g_j\left(\|x-v\|^2\right)$ denotes the gradientwith respect to x of $g_j\left(\|x-v\|^2\right)$ at x , then

$$\frac{\|\nabla_x g_j\|^2}{\|x-v\|^2}>\frac{\|\nabla_y g_y\|^2}{\|y-v\|^2} \qquad (2)$$

for all $x,y,v\in\Re^n$ such that $\|x-v\|^2<\|y-v\|^2$

.Axiom 4: $g_j\left(\|x-v\|^2\right)<\infty$ for all $x,v\in\Re^n$

B. Admissibility Conditions for Radial Basis Functions

The selection of admissible radial basis functions can be facilitated by extending the theorem proposed in [4].

Theorem :

The model described by (1) represents an radial basis function neural network in accordance with all four axiomatic requirements if $g_j(x)$ are continuous functions on $(0,\infty)$ with continuous first-order derivatives $g_j'(x)$ such that:

1) $g_j(x)>0,\forall x\in(0,\infty)$
2) $g_j'(x)<0,\forall x\in(0,\infty)$
3) $g_j''(x)>0,\forall x\in(0,\infty)$
4) $\lim_{x\to 0_+} g_j(x)=L_j$, where L_j are finite numbers.

A radial basis function is said to be *admissible in the widesense* if it satisfies the three basic axiomatic requirements, that is, the first three conditions of Theorem [4]. If a radial basis function satisfies all four axiomatic requirements, that is, all four conditions of Theorem, then it is said to be *admissible in the strict sense*.

Theorem verifies the strong link between radial basis function neural networks and function approximation models used to perform interpolation. Such function approximation models attempt to determine a surface in a Euclidean space $\Re^n$ that providesthe best fit for the data (x_k , y_k) , $1\le k\le M$ where $x_k\in X\subset\Re^n$ and

$y_k \in \Re$ and for all k = 1, 2, 3, … , M. Micchelli [7] considered the solution of the interpolation problem $s(x_k) = y_k, 1 \le k \le M$, by functions $s : \Re^n \to \Re$ of the form

$$s(x) = \sum_{k=1}^{M} \omega_k \, g\left(\left\| x - x_k \right\|^2 \right) \qquad (3)$$

Micchelli [16] showed that the model described by eqn (3) is admissible for interpolation if the basis function is *completely monotonic* on ℓ. A function $g(x)$ is called completely monotonic on $(0,\infty)$ if it is continuous on $(0,\infty)$ and its ℓ th order derivatives $g^{(l)}(x)$ satisfy $(-1)^l g^{(l)}(x) \ge 0, \forall x \in (0,\infty)$ for ℓ = 0, 1, 2 , … .

The theorem requires that any wide-sense admissible function $g_j(x)$ be continuous on $(0,\infty)$ and its derivatives satisfy $(-1)^l g^{(l)}(x) \ge 0, \forall x \in (0,\infty)$ for ℓ = 0, 1, 2 , … . The theorem is less restrictive than Micchelli's interpolation theorem in terms of the conditions imposed on the selection of functions that are admissible in the wide sense. However, the theorem is more restrictive than Micchelli's interpolation theorem if it is used to select functions that are admissible in the strict sense.

C. Constructing Admissible Generator Functions

The search for admissible RBF can be simplified by considering basis functions of the form $\phi_j(x) = g_j(x^2)$, with each $g_j(x)$ defined in term of a *generator function* $g_{j0}(x)$ as $g_j(x) = (g_{j0}(x))^{1/(1-m)}, m \ne 1$ [4]. The construction of widesense admisible generator functions can be attempted by assuming that $g_{j0}'(x) = p_j(g_{j0}(x))$.

Generator functions admissible in the strict sense can be obtained by determining the subset of the resulting widesense admissible generator functions that satisfy the fourth condition :

if m>1 RBF neural network requirement :

1. $g_{j0}(x) > 0, \ \forall x \in (0,\infty)$
2. $g_{j0}'(x) > 0, \ \forall x \in (0,\infty)$
3. $r_{j0}(x) = [m/m-1](g_{j0}'(x))^2 - g_{j0}(x) g_{j0}''(x) > 0, \ \forall x \in (0,\infty)$
4. $\lim_{x \to 0} g_{j0}(x) = L_{1j}, \ where \ L_{1j} \in (0,\infty)$

if m<1 RBF neural network requirement :

1. $g_{j0}(x) > 0, \ \forall x \in (0,\infty)$
2. $g_{j0}'(x) < 0, \ \forall x \in (0,\infty)$
3. $r_{j0}(x) = [m/m-1](g_{j0}'(x))^2 - g_{j0}(x) g_{j0}''(x) < 0, \ \forall x \in (0,\infty)$
4. $\lim_{x \to 0} g_{j0}(x) = L_{2j}, \ where \ L_{2j} \in (0,\infty)$

The constructive approach outlined above is employed here to produce *increasing* generator functions that can be used for m>1. The same constructive approach can be extended to produce *decreasing* generator functions that can be used for m<1.

Assume that m>1 and let the function $p_j(x)$ be of the form $p_j(x) = k_j x^n$, $k_j > 0$. The function $g_{j0}(x)$ can be obtained in this case by solving the differential equation

$$g_{j0}'(x) = k_j (g_0(x))^n, k_j > 0 \quad (4)$$

According to (4), $g_{j0}''(x) = nk_j (g_0(x))^{n-1} g_{j0}'(x) = nk_j{}^2 (g_{j0}(x))^{2n-1}$. In this case

$$r_{j0}(x) = (g_{j0}'(x))^2 \left(\frac{m}{m-1} - n \right) \quad (5)$$

If m>1, it is required that $r_{j0}(x) > 0, \forall x \in (0, \infty)$, which holds for all m/(m-1)>n. For m>1, m/(m-1)>1 and the inequality m/(m-1)>n holds for all n<1. For n=1, m/(m-1)-n=1/(m-1)>0. Thus, the condition $r_{j0}(x) > 0, \forall x \in (0, \infty)$, is satisfied for all $n \le 1$.

3. Selecting Generator Functions

Consider the function $g(x) = (g_0(x))^{\frac{1}{1-m}}$, with $g_0(x) = \exp(\beta x)$, $\beta > 0$ and m>1. For all $\beta > 0$, g_0(x) is a monotonical increasing function, $g_0'(x) = \beta \exp(\beta x)$ and $g_0''(x) = \beta^2 \exp(x)$. In this case :

$$r_0(x) = \frac{1}{m-1} (\beta \exp(\beta x))^2 \quad (6)$$

If m >1, then r_0(x) > 0, thus g_0(x) is an admissible function for all $\beta > 0$. If $g_0(x) = \exp(\beta x)$, $\beta > 0$, and m>1, then g_0(x) corresponds to the Gaussian RBF $\phi(x) = g(x^2) = \exp(-x^2 / \sigma^2)$, with $\sigma^2 = (m-1)/\beta$.

Consider also the function $g_0(x) = \exp(-\beta x)$, $\beta > 0$ with m<0, g_0(x) is a monotonically decreasing function. The $g_0(x) = \exp(-\beta x)$, corresponds to the Gaussian RBF, $\phi(x) = g(x^2) = \exp(-x^2 / \sigma^2)$ with $\sigma^2 = (1-m)/\beta$

4. A Learning Algorithm Based on Gradient Descent

An RBF neural network can be trained by minimizing the error

$$E = \frac{1}{2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (7)$$

This algorithm can be summarized as follows:

1) Select m, η and ε; initialize w_{ij} with zero values; randomly initialize v_j

2) Compute the initial response $\hat{y}_i = \sum_{j=1}^{c} w_{ij} h_j$

3) Compute *E*. and set $E_{old} = E$.

4) Update the adjustable parameters:

- . $e_i^0 = y_i - \hat{y}_i$

- $. w_{ij} \leftarrow w_{ij} + \eta \sum_{i=1}^{n} e_i^0 h_i, \forall j$
- $e_{ij}^h = 2/(m-1) g_0' \left(\left\| x_i - v_j \right\|^2 \right) \left(h_{j,i} \right)^m \sum_{i=1}^{n} e_i^0 w_{ij}$
- $v_j \leftarrow v_j + \eta \sum_{j=1}^{c} e_{ij}^h \left(x_i - v_j \right), \forall j$.

6) Compute the current response: $\hat{y}_i = \sum_{j=1}^{c} w_{ij} h_j$

7) Compute E

8) If: $(E_{old} - E) / E_{old} > \varepsilon$ then: go to Step 4).

5. Result Study

The data are used in this study is Indonesia inflation data (monthly) from January 1999 until April 2005. In this study data divided into two part. First, data training, from Januari 1999 until December 2004 is used to modeling. Second, data testing is used to validation result first part, Januari 2005 until April 2005. In the first data, model is called best model if it has minimum *Mean Square Error* (MSE). The model is applied on data second to measure goodness of fit (validation). Some goodness of fit model are *Mean Percentage Error* (MPE), *Mean Absolute Deviation* (MAD), and *Mean Absolute Percentage Error* (MAPE).

Generally time series analysis use ARIMA model to get forecast model and value of the forecast. ARIMA models need stationary data and information about observation that influence to respond. ARIMA model identification is based on time series plot, ACF and PACF plot. Based on data plot (figure 1), ACF plot (figure 2) and PACF plot (appendix) show that Indonesia Inflation has nonseasonal and seasonal pattern and observation (t-1), (t-11) and (t-12) have straight influence to respond. Indonesia Inflation data has some ARIMA models. It has periodic seasonal 11 and 12. Periodic seasonal 12 base month a year and periodic seasonal 11 base on culture of Indonesia people at "Idul Fitri" is called "mudik". The best two of the models are ARIMA(0,0,1)(0,0,1)11 and ARIMA(1,0,1)(1,0,1)11. These models have MSE 0.2685 and 0.2743.

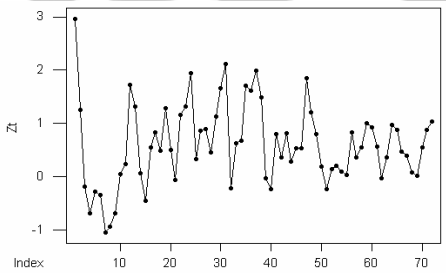


Figure 2. Time series Plot

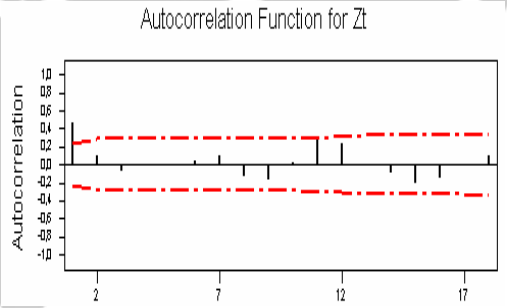


Figure 3. ACF Plot

According the plots we know that respond (t observation) depend on observation (t-1), (t-11) and (t-12), so these observation are input of radial basis function neural network. In this study are used three architecture RBFNN models :

1. Number of unit in hidden layer equal number of points (model 1).
2. Number of unit in hidden layer is 12 (model 2).
3. Number of unit in hidden layer is 12 with gradient descent algorthm (model 3).

Fist model based on standart architecture RBFNN model, that used number of unit in hidden layer equal number of point. Second and third model used 12 unit in hidden layer, it is based on number of month in a year. In third model use gradient descent algorithm to find parameter estimation. The result of RBFNN based on training and testing data for three models above is shown on table 1.

Table 1. Statistics of training and testing data

Architecture NN	MSE	MAD	MAPE	MPE
Model 1	0.000	0.000	0.000	0.000
	1.654	0.887	3.645	-1.696
Model 2	0.253	0.396	2.548	-1.058
	0.127	0.792	1.378	-1.198
Model 3	0.015	0.080	0.528	-0.030
	2.243	1.590	2.389	-0.422

□ : training model ▣ : testing model

According table 1. First model is good model in training data but worse in testing data. Second and third models show that the third model is better than second model in training data but in testing data is bad. This result show that there is not guarante result training and testing of model is good. According MSE, RBF neural network smaller than ARIMA model.

6. Conclusion

According the theory unit in hidden layer is equal number of point data, but we are possible to reduce the unit. Gradient descent learning is one of methods to get parameter estimation more efficient. In case Indonesia inflation data show that there is nonseasonal and seasonal pattern. According ARIMA model, the best model is ARIMA(0,0,1)(0,0,1)11. Function approximation base on RBFNN for the data, there is not best model, because first model is good in training data but second model is good in testing data. Third model better than second model in training data, but in testing data second model is better.

References

[1]. Bishop, C.M., 1995, *Neural Network for Pathern Recognition*, Oxford, Clarendon Perss.

[2]. Broomhead, D.S., and D.Lowe, 1988, "Multivariate function interpolation and adaptive network", *Complex Syst.*,vol 2, 321-355.

[3]. Chen, S., C.F.N. Cowan and P.M.Grand, 1991, Orthogonal Least Square Learning Algorithm for Radial Basis Function Network, *IEEE trans Neural Network*, vol 2, 302-309
[4]. Karayianis, N.B., 1999, "Reformulated radial basis neural network trained by gradient descent", *IEEE Trans. on Neural Net.*, vol 10, 657-671.
[5]. Karayianis, N.B. and G.W., Mi, 1997, "Growing radial basis neural network merging supervised and unsupervised learning with network growh techniques", *IEEE Trans. on Neural Net.*, vol 8, 1492-1506.
[6]. Micchelli, C.A., 1986,"Interpolating of scattered data : Distance matrices and conditionally positive definite function" *Constructive Approximation*, vol 2,11-12
[7]. Moody, J. and Darken, C, 1989,"Fast learning in network of locally turned processing units." Neural Com[putation, val : 1, no 2, 281-294.
[8]. Poggio, T and Girosi, F.,1990,"Network for approximation and learning", Proceedings of IEEE, vol 78, no 9, 1491-1497.

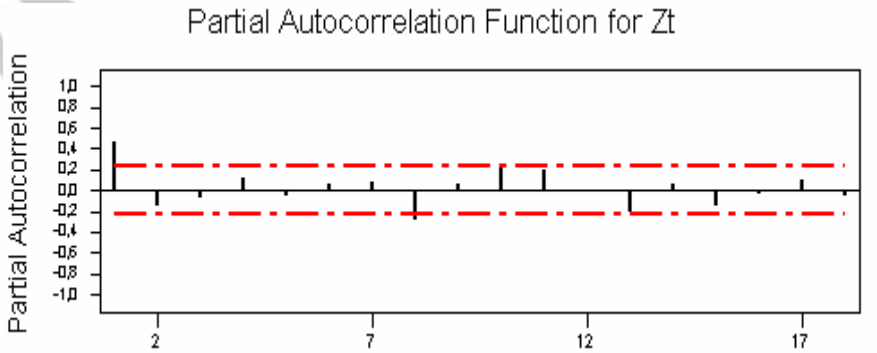
BRODJOL SUTIJO: Ph.D student at Department of Mathematics, Universitas Gadjah Mada, Bulaksumur, Yogyakarta 55281, Indonesia.
Department of Statistics, Institut Teknologi Sepuluh Nopember, Kampus ITS Keputih Surabaya 60111, Indonesia.
E-mail: sutijo_b@yahoo.com

SUBANAR: Department of Mathematics, Universitas Gadjah Mada, Bulaksumur, Yogyakarta 55281, Indonesia.
E-mail: subanar@yahoo.com

S. GURITNO: Department of Mathematics, Universitas Gadjah Mada, Bulaksumur, Yogyakarta 55281, Indonesia.
E-mail: suryoguritno@ugm.ac.id

Appendix :

Plot PACF



COMPARISON BETWEEN THE NEURAL NETWORKS (NN) AND ARIMA MODELS FOR FORECASTING THE INFLATION IN YOGYAKARTA

Dhoriva Urwatul Wutsqa

Graduate Student in Gadjah Mada University, Yogyakarta, Indonesia
UNY, Yogyakarta, Indonesia

Abstract. Neural network is a relatively new method used in many applications, such as pattern recognition, signal processing, forecasting time series data, and control processing. In this paper, the method is utilized to forecast the inflation in Yogyakarta. The backpropagation is considered as the algorithm for constructing the NN forecasting models. The NN model is compared with ARIMA model, the usual model for forecasting time series data.

It is showed that the NN model gives better performance than ARIMA model in forecasting the inflation in Yogyakarta. Both in training and in testing, the NN model obtains less MSE (Mean Squared Error) than ARIMA model.
Key-words: Neural Network model, ARIMA model, backpropagation

1. Introduction

Inflation can be interpreted as the increase of consumer price consists of commodities and services. Forecasting the inflation is important, because it is one of the macro economic indicators. Some economic plans or decisions concern the inflation as an important factor, for example in determining the interest rate of the obligation.

Predicting the inflation is related to the forecasting time series data. The most frequently used model is ARIMA model. It is developed by Box-Jenkins [1], and it becomes the standard model in forecasting time series data. Some authors discuss the ARIMA model, such as [2] and [14].

In constructing ARIMA model, it is needed the stationary assumption. In spite of that, it is only for the linear forecasting model. Recently, a relatively new technique, neural networks, is increasingly applied in performing time series prediction. Neural networks approach is more flexible, because it does not need assumptions like normality that is commonly found in statistical methods. In statistics view, NN can be considered as a non-linear model. However, some researchers find that NN model gives satisfactory performance in linear as well as in non-linear forecasting model. Results from [4] and [8] show that the NN model performs well in linear model. By using NN model, [3], [6], [7], [9], and [10] achieve accurate predictions in forecasting financial data.

This paper presents a new approach by Neural Network (NN) model to predict the inflation with case study in Yogyakarta. The result compares with the prediction from ARIMA model. Better forecasting model is determined by its less MSE (Mean Squared Error) value.

Section 2 briefly describes the ARIMA model. The next section presents the NN model, continued to the back propagation algorithm topic. Section 5 gives comparison of the empirical results of the models. The conclusion is included in the last section.

2. ARIMA model

ARIMA model is time series forecasting model developed by Box-Jenkins [1]. Application of this model is widely used in forecasting financial data. The ARIMA (p, d, q) model is composed of two processes, the autoregressive (p) known as AR(p) and moving average(q) known as MA(q). Number d denotes the differencing order needed to get stationary data. The ARIMA (p, d, q) model can be expressed as

$$\phi_p(B)(1-B)^d y_t = \theta_0 + \theta_q(B)\varepsilon_t,$$

with

$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$

$$\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q,$$

where B is the back shift operator and ε_t is a sequence of white noise with zero mean and constant variance. The estimates of autoregressive parameter ϕ and moving average parameter θ are calculated by iteration process so that the mean squared error is minimized.

3. Neural networks model

Neural network is computational process motivated by functioning biological neurons on human brain. Neural network consists of simple process elements known as neurons or units. Typically, the elements are arranged in a group or layers. They are input layer, output layer, and one or more layers between input layer and output layer called hidden layers. The type of NN is characterized by

a. its architecture, pattern connection between the neurons,
b. its training or algorithm, methods of estimating the parameter or weights on the connection,
c. its activation function, function to its net input to determine its output prediction. (see (5)).

In composing the NN model, the data set is divided into two subsets: a training set (used for weights and bias updating to get the parameter estimates of the weights) and a testing set (used to compare real and prediction out put). The following steps are done to obtain the NN model:

1. Setting of the number of hidden layers, neurons, training algorithm, initial connection weights, neuron biases and the activation function for each neurons.

2. Network training.
3. Estimation of the predicted output using the data testing.
4. Evaluation of the forecast performance of the NN model, and compare with the other model.
5. Step 1-4 are repeated if the error goal is not reached

This paper considers one popular type of NN model called Feed-Forward Neural Network (FFNN) with single hidden layer. The architecture of the FFNN model for time series forecasting with single hidden layer, is showed in Figure 1.

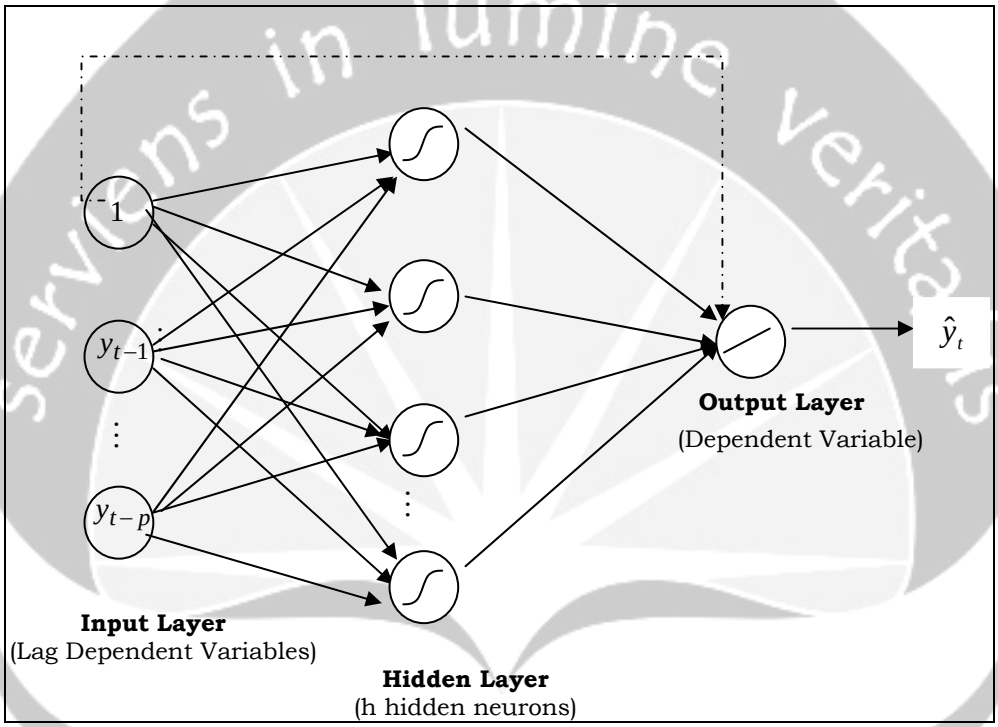


Figure 1. Architecture of FFNN model for forecasting time series data with single hidden layer

For FFNN model with single hidden layer illustrated in Figure 1, the output is predicted using past observations $y_{t-1},\ldots,y_{t-p}$, as the inputs, and it can be calculated from the equation

$$\hat{y}_t = \phi_0\left\{w_{c0} + \sum_h w_{h0}\phi_h\left(w_{ch} + \sum_i w_{ih}y_{t-i}\right)\right\},$$

with

$\{w_{ch}\}$ denote the weights for connections between the constant input and hidden neurons

w_{c0} denotes the weight of direct connection between the constant input and the output

$\{w_{ih}\}$ and $\{w_{h0}\}$ denote the weights for the other connections between he inputs and the hidden neurons and between the neurons and the output, respectively

ϕ_h and ϕ_0 denote activation functions used at the hidden layer and at the output respectively.

The commonly used activation function is logistic function

$$\phi(x) = \frac{1}{1+e^{-x}},$$

which gives values in the range (0, 1).

4. Backpropagation algorithm

The weights in NN model are estimated by process called training such that the output produced by network for a given input approximate to the output data. It is an optimization process of finding the weights by minimizing the sum squared error $E(w) = \sum_{k=1}^{K} (y_k - \hat{y}_k)^2$, where y_k and $\hat{y}_k$ is the output data and output prediction, respectively.

Backpropagation algorithm is defined as a method for determining weights w by minimizing a function $E(w) = \frac{1}{2}\sum_{k=1}^{K} (y_k - \hat{y}_k)^2$ (*).

The author (11) uses *gradient descent* form to reduce the equation (*) by *update* rule

$$\Delta w = -\eta \frac{\partial E}{\partial w}, \eta > 0,$$

where η is learning rate.

5. Empirical result

The monthly inflation data in Yogyakarta from January 1999 – March 2005 were collected from BPS (Figure 2.). The first 60 entries are used as training data, and the rest 15 are testing data.

The data processing to get ARIMA model is implemented by using MINITAB software. From Figure 2., it is obviously seen that data are stationary, so no differencing (d = 0) needed to obtain ARIMA model. In this study, different ARIMA

models are found to yield significantly model. The best model is ARMA (1,2) due to the least MSE in training data, but not in testing data. On the other hand, MA (2) performs the least MSE in testing data, so it is more preferable to be the best prediction model.

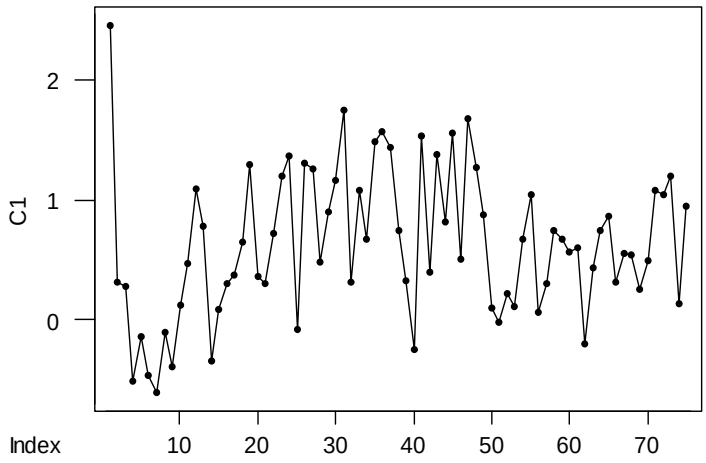


Figure 2. Inflation in Yogykarta, January 1999-March 2005

The NN model considered in this paper consists of three layers, input layer, one hidden layer, and output layer. Models with different input variables (see Table 1) are studied, each with the same number of hidden units (1 to 6 units). All the models use the back-propagation algorithm. The activation function in the hidden layer is logistic function and in the output layer is linear function. As noted in section 1, the criterion for selecting the best model uses the MSE value. By this criterion, it can be seen from many researches (see (12), (13) and (15)) that the best model in testing data does not follow the best model in training data.

The application of NN model to inflation data in Yogyakarta is consistent with their results. In this study, it is also shown that more complexity models tend to have less MSE value for training data, but not for testing data. Table1. summarizes the NN models for different lags, together with the corresponding values for ARIMA model. Not all NN models are presented, but only the selected models with least MSE in training or in testing data.

The notation NN (j_1,..., j_k; h) denotes the NN with inputs at lag j_1,..., j_k and with h neurons in the hidden layer.

Table 1. Comparison the NN models and the ARIMA model for forecasting the inflation in Yogyakarta

Lag	Models	MSE Training Data	MSE Testing data
1	NN(1;1)	0.3	0.18 *
	NN(1;4)	0.26**	0.37
1,2	NN(1,2;4)	0.24	0.12*
	NN(1,2;6)	0.2**	0.13
1,2,3	NN(1,2,3;4)	0.2	0.14*
	NN(1,2,3;6)	0.17**	0.19
	ARIMA(1, 2)	0.33**	0.15
	MA(2)	0.35	0.14*

Note: * Best model in testing data
** Best model in training data

Table 1. shows that almost all NN models have prediction accuracy better than ARIMA model in data training. But after testing the models to the rest data, NN (1,2; 4), NN(1,2; 6), and NN(1,2,3; 4) perform better than ARIMA models. NN (1,2; 4) has least MSE in testing data, 0.12, therefore, in this study the NN model with inputs lag 1 and lag 2, and with four hidden units is the best prediction model in comparison with the other NN models and the ARIMA models.

6. Conclusion

The empirical study shows that the accuracy of the various types of NN model are better than ARIMA model for training data, but not in testing data. From this study, we also can conclude that the best forecasting model tends to yield bad performance in training data. However, the best-selected forecasting model is model with the least validation of error.

Acknowledgment

I am now on leave UNY, and be a graduate student in the first year in Gadjah Mada University, Yogyakata, Indonesia.

References

[1] Box GEP. & Jenkins GM. (1976), *Time Series Analysis: Forecasting and Control*, Rev. Ed. San Francisco: Holden-day.

[2] Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (1994), *Time Series Analysis, Forecasting and Control,* 3rd edition. Englewood Cliffs: Prentice Hall.

[3] Chan M., Wong. C., and Lam C. (1999), Financial Time Series Forecasting by using Conjugate Gradient Learning Algorithm and Multiple Linier Regression Weight Initialization, *Department of Computing, The Hongkong Polytechnic University,* Hongkong.

[4] Diaz F., Borrajo L., Riverola F.F., Usero A., and Corchado J.M. (2001), Negative Feedback Network for Financial Prediction, *Artificial Intelligence Research Group. Universidad de Vigo,* Spain.

[5] Fausett, L. (1994), Fundamental f Neural Network : Architecture, Algorithms and Applications, Prentice Halls International, Inc. : New Jersey.

[6] Moody J. (1995), Economic Forecasting Challenger and Neural Network Solutions, *In Proceedings of the International Symposium on Artificial Neural Networks,* Taiwan.

[7] Nikola G and Jing Yang (2000), The Application of Artificial Neural Networks to Exchange Rate Forecasting: *The Role of Market Microstructure Variables*, Financial Markets Department Bank of Canada.

[8] Ranaweera D.K. & Hubele N.E. (1995), Application of radial Basis Function Neural Network Model for Short-Term Load Forecasting. IEE Proc.-Gener. *Transm. Distrib.*, Vol.142, No. 1.

[9] Robert R. & Jae LL. (1996), Artificial Intelligence in Finance & Investing. *Ch. 0, IRWIN.*

[10]Rumelhart, D. & McClelland, J. (1986), *Parallel Distributed Processing: Explorations in the Microstructures of Cognition,* Vol. 1., Cambridge: MIT Press.

[11] Sri Rezeki, Subanar dan Suryo Guritno (2004), Model Neural Network untuk Data Polikotomus, *International Conference in Statistics*, UNISBA, Indonesia.

(12)Suhartono, Subanar, and Sri Rezeki (2005), Feed-forward Neural Networks Model for Forecasting Trend and Seasonal Time Series, *IRCMA 2005 Proceedings*, Indonesia.

[13] Wei, W.W.S. (1990), *Time Series Analysis: Univariate and Multivariate Methods.* Addison-Wesley Publishing Co., USA.

(14) Zaiyong Tang, Chrys de Almedia, Fishwick P.A. (1991), Time series forecasting using neural networks vs. Box-Jenkins methodology, Technical Article, USA.

DHORIVA URWATUL WUTSQA : Graduate student in Department of Statistics, Gadjah Mada University, Sekip Utara Bulak Sumur 21, Yogyakarta, Indonesia..
Department of Mathematics, Universitas Negeri Yogyakarta,
Jl. Karangmalang, Yogyakarta, Indonesia.
E-mail : dhoriva@yahoo.com

FASTER TRAINING OF FEEDFORWARD NEURAL NETWORKS

Sri Rezeki[a], Suhartono[b], Subanar[c] and S. Guritno[c]

[a] Universitas Islam Riau, Pekanbaru, Indonesia
[b] Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[c] Universitas Gadjah Mada, Yogyakarta, Indonesia

Abstract. Since the discovery of the backpropagation method, many modified and new algorithms have been proposed for training of feedforward neural networks (FFNN). The objective of this paper is to compare performances of some optimization methods at FFNN backpropagation learning that were applied for regression problems. In general, there are two group of backpropagation optimization methods for updating weights and biases, those are first and second order. The first order optimization methods are gradient descent (GD) and resilient backpropagation (RP), whereas the second order are conjugate gradient (CG), quasi-Newton (QN) and Levenberg-Marquardt (LM). In this research, performances comparison is done by using MATLAB program. The simulation result shows that LM has the best performance among the others. The criterion used to compare it is convergence speed evaluated by time and the number of epoch to reach the goal of MSE.

Key-words: Feedforward neural networks, backpropagation, first and second order methods

1. Introduction

Backpropagation (BP) learning algorithm is the most commonly method used for training at FFNN [10]. Standard BP uses gradient descent (GD) technique, based on first derivatives of error function. Though this technique has been used successfully on a number of interesting problems its applicability to complex real-world problems. Standard BP has often been limited by its slow convergence to a solution [1]. In several researches that have been done, faster training algorithms such as the modified GD or the ones based on second derivatives of error function tend to be used (e.g. [5, 8]).

There has been considerable research on methods to accelerate the convergence of the algorithm. This research falls roughly into two categories. The first category involves the development of ad hoc techniques [7). This techniques include such ideas as varying the learning rate, using momentum and rescaling variables. Another category of research has focused on standar numerical optimization techniques (e.g., [1]-[4]).

The most popular approaches from the second category have used conjugate gradient or quasi-Newton (secant) methods. The quasi-Newton methods are considered to be more efficient, but their storage and computational requirements go up as the square of the size of the network. There have been some limited

memory quasi-Newton (one step secant) algorithm that speed up convergence while limiting memory requirements [3, 4].

Another area of numerical optimization that has been applied to neural networks is nonlinear least squares. The more general optimization methods were designed to work effectively on all sufficiently smooth objective functions. However, when the form of the objective function is known it is often possible to design more efficient algorithms. One particular form of objective function that is of interest for neural networks is a sum of squares other nonlinear functions. The minimization of objective functions of this type is called nonlinear least squares.

Hagan and Menhaj [9] applied a nonlinear least squares algorithm to the batch training of multilayer perceptron. For very large networks the memory requirements of the algorithm make it impractical for most current machines (as is the case for the quasi-Newton methods). However, for networks with a few hundred weights the algorithm is very efficient when compared with conjugate gradient techniques.

It is difficult to know which training technique will be the fastest for a given problem. It will be depend on many factors, including the complexity of the problem, the number of data points in the training set, the number of weights and biases in the network, and the error goal [5]. The objective of this paper is to compare performances of some optimization methods at FFNN backpropagation learning that were applied for regression problems, particularly for nonlinear regression. Section 2 and 3 briefly present neural networks architecture and the basic BP algorithm. Some faster training of FFNN are then described in Section 4. The next section shows research methodology. Afterwards, Section 5 gives comparison of convergence speed of the algorithms. The last section contains a summary and conclusions.

2. Neural Network Architecture

A neural networks consist of a number of connected nodes (in the literature nodes are also referred to as neurons, unit, or cell) each of which is capable of responding to input signals with an output signal in a predefined way. These nodes are ordered in layers. A network consist of one input layer, one output layer, and an arbitrary number of hidden layers in between. This number can be chosen by the user such that the network performs as desired. Usually, one hidden layer is used. The reason for this is that one hidden layer is sufficient to approximate any continuous function to an arbitrary precision [6].

The architecture of NN commonly used for regression problem is feedforward neural network (FFNN) with associated error backpropagation learning algorithm for minimizing the observed sum of squared errors over a given set of data. Architecture of FFNN can be seen at Figure 1, where f(.) is transfer function and W_{ij} and V_{ij} are weights.

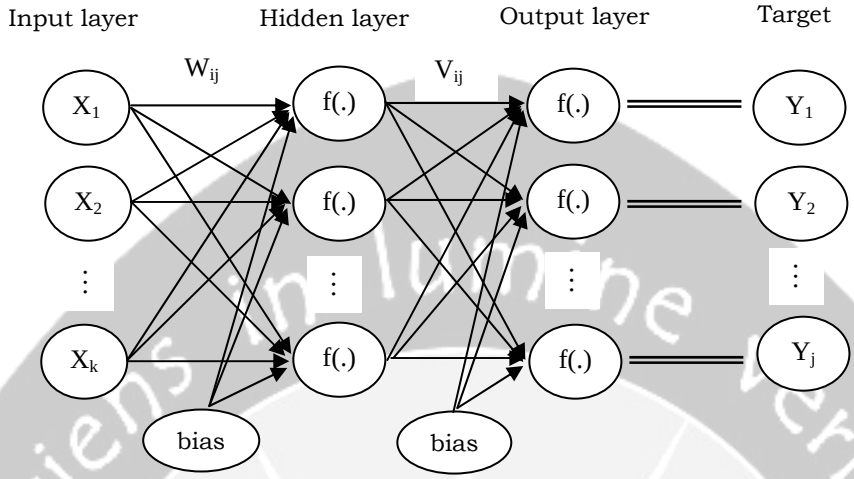


Figure 1. FFNN with single hidden layer

3. Backpropagation (BP)

BP was created by generalizing the widrow-Hoff learning rule to multilayer networks and nonlinear differentiable transfer function [13]. Standard BP uses a GD algorithm. The term backpropagation refers to the manner in which the gradient is computed for nonlinear multilayer network. The error function which is minimized is defined as follow:

$$E = \frac{1}{2}\sum_{k=1}^{n}(y_{(k)} - \hat{y}_{(k)})^2 \quad (1)$$

with $y_{(k)}$ is the target or the real value of response, $\hat{y}_{(k)}$ is an output vector at the output layer, and k is index of input-target pairs $(x_{(k)}, y_{(k)})$ used in training set where k = 1, 2, …, n. The weight update is

$$\Delta w = -\eta E'(w), \ \ \eta > 0 \quad (2)$$

The step size or learning rate η can be determined by a line search method but usually set to a small constant. In the later case the algorithm is, however, not guaranteed to converge. If η is chosen optimally in each step the method is often called the steepest descent method. The method can be used in off-line or on-line mode. The off-line mode is the one presented in (2), where the gradient vector is an accumulation of partial gradient vectors, one for each pattern in the training set. In the on-line mode, gradient descent is performed successively on each partial error function associated with one given pattern in the training set. The update formula is then given by

$$\Delta w = -\eta E_p^{'}(w), \ \ \eta > 0 \quad (3)$$

where $E_p^{'}(w)$ is the error gradient associated with pattern p. If η tends to zero over time, the movement in weight space during one epoch (one full presentation of all pattern in the training set) will be similar to the one obtained with one off-line update. However, in general the learning rate has to be large to accelerate convergence, so that the paths in weight space of the two methods differ. The on-line method is often preferable to the off-line method when the training set is large and contains redundant information.

The FFNN starts out by an initial set of weights chosen randomly. It then update the weights in such a way that given the input signals, the FFNN's output signals match the desired output signals as closely as possible (the convergence limit is specified by the user). All training pairs are presented to the FFNN and the sum of squares of the errors (SSE) over the whole training set is computed. If the SSE exceeds the specified error goal, the FFNN update the connection weights. This is called training epoch. The FFNN then begin another training epoch until either the maximum number of training epoch is reached or the SSE reaches the specified error goal. The training is said to be complete when either of this happen. How well a network is trained is measured by the MSE over the complete training dataset.

4. Some faster training methods at FFNN

GD technique is often too slow for practical problem. With standard steepest descent, the learning rate is held constant throughout training. The performance of the algorithm is very sensitive to the proper setting of the learning rate. If the learning rate is set too high, the algorithm may oscillate and become unstable. If the learning rate is too small, the algorithm will take too long to converge. It is not practical to determine the optimal setting for the learning rate before training, and, in fact, the optimal learning rate changes during the training process, as the algorithm moves across the performance surface. The performance of the steepest descent algorithm can be improved if we allow the learning rate to change during the training process.

There are several high performance algorithms which can converge faster than the previous technique. These faster algorithm fall into two groups: first and second order. A complete description of the first and second order optimization methods is given in NN toolbox [5].

A. First order optimization methods

The following methods were based on first derivatives of error function and used heuristic techniques, which were modified from an analysis of the performance of the standard steepest descent algorithm (traingd).

- Variable learning rate (traingda, traingdx)

  This technique uses an adaptive learning rate to keep the step size as large as possible while keeping learning stable. The learning rate is made responsive to the complexity of the local error surface. The procedure increases the learning

rate but only to the extent that the network can learn without large error increases. Thus a near optimal learning rate is obtained for local terrain.

- Resilient backpropagation (trainrp)

FFNN typically use sigmoid transfer functions in the hidden layers. This causes a problem when using steepest descent to train FFNN, since the gradient can have a very small magnitude, and therefore cause small changes in the weight and biases even though the weights and biases are far from their optimal values. The purpose of the resilient backpropagation training algorithm is to eliminate the harmful effects of the magnitude of the partial derivatives. Only the sign of the derivatives is used to determine the direction of the weight update.

B. Second order optimization methods

These methods use second derivatives of error function and standard numerical optimization techniques.

- Conjugate Gradient (CG) algorithms

There are four different variations of CG algorithms i.e. Fletcher-Reeves Update (traincgf), Polak-Ribiere Update (traincgp), Powell-Beale Restarts (traincgb) and Scaled Conjugate Gradient (trainscg). In most of CG algorithms the step size is adjust at each iteration. A search is made along the CG direction to determine the step size which will minimize the performance function along that line.

- Quasi-Newton (QN) Algorithms

Newton's method is an alternative to the CG methods for fast optimization. The basic step of Newton's methods is

$$w_{k+1} = w_k - H_k^{-1} g_k \tag{4}$$

where H_k is the Hessian matrix (second derivatives) of performance index at the current values of the weight and biases, and g_k is the current gradient. Newton's method often converges faster than conjugate gradient methods. Unfortunately, it is complex and expensive to compute the Hessian matrix for FFNN. There is a class of algorithms that are based on Newton's method but which don't require calculation of second derivatives. These are called quasi-Newton (or secant) methods. They update an approximate Hessian matrix at each iteration of the algorithm. The update is compute as a function of the gradient. The quasi-Newton method which has been most successful in published studies is Broyden, Fletcher, Goldfarb, and Shanno (BFGS) update. This algorithm has been implemented in the trainbfg routine. Since the BFGS algorithm require more storage and computation in each iteration than the CG algorithms, there is need for a secant approximation with smaller storage and computation requirements. The one step secant (OSS) method is an attempt to bridge the gap between the CG algorithms and the QN (secant) algorithms. This algorithm does not store the complete Hessian matrix, it assumes that at each

iteration the previous Hessian was an Identity matrix. This has the additional advantage that the new search direction can be calculated without computing a matrix inverse.

- Levenberg-Marquardt (trainlm)

  Like the QN methods, the LM algorithm was design to approach second order training speed without having to compute the Hessian matrix. When the performance function has the form of a sum of square (as is tipical in training FFNN), then the Hessian matrix can be approximated as

$$H = J^T J \tag{5}$$

and the gradient can be computed as $g = J^T e$ where **J** is Jacobian matrix, which contains first derivatives of the network errors with respect to the weights and biases, end **e** is a vector of network errors. The jacobian matris can be computed through a standard BP technique that is much less complex than computing the Hessian matrix.

5. Research Methodology

This research is a simulation study for comparison some faster training methods which were implemented for three forms of nonlinear regression models, those are sine, quadratic, and cubic models. The architecture used is FFNN 1-20-1 (the number of input, hidden neurons and output) with single hidden layer. The comparison of methods are based on NN toolbox for MATLAB programming, with value of learning rate is 0.1, error goal is 0.001, and maximum training epoch is 10000. The algorithms are variabel learning rate (traingdx), resilient backpropagation (trainrp), scaled conjugate gradient (trainscg), Fletcher-Powell Conjugate Gradient (traincgf), Polak-Ribiere Conjugate Gradient (traincgp), Powell-Beale Conjugate Gradient (traincgb), One-Step-Secant (trainoss), BFGS quasi-Newton (trainbfg), and Lavenberg-Marquardt (trainlm). The criterion used to compare it is convergence speed evaluated by time and the number of epoch to reach the goal of MSE.

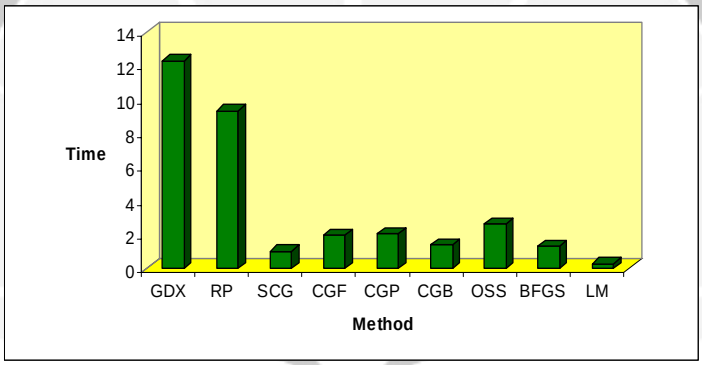
6. Results

The training set consisted of 80 input-output pairs, where the input values were scattered in interval [1,1], and the network was trained until the SSE was less than the error goal of 0.001. Programming runs were repeated 100 times for optimization methods respectively. Transfer functions used are a sigmoid nonlinearities hidden layer and a linear output layer. The algorithm was trained to approximate each model of nonlinear regression. The results can be summarized as follow:

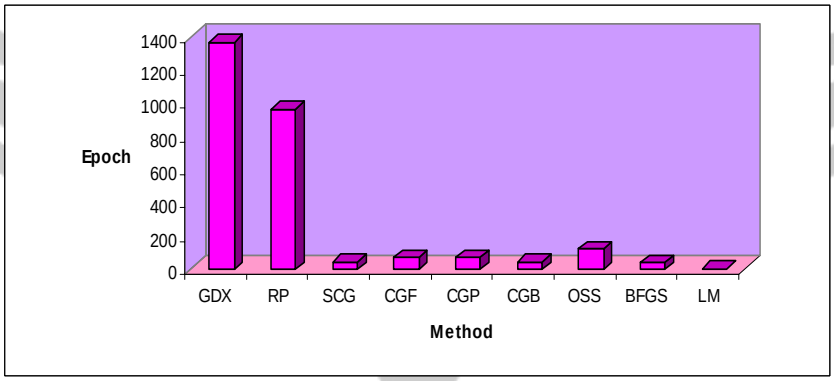
A. Sine model: $y = \sin(2\pi x) + \varepsilon$

Table 1. Comparison of time and epoch across optimization methods

Methods	Time	Stdev. time	Epoch	Stdev. epoch
Variabel Learning Rate (GDX)	12.31	11.84	1366	1340
Resilient backpropagation (RP)	9.35	17.06	964	1778
Scaled Conjugate Gradient (SCG)	1.0389	0.3098	43.07	17.46
Fletcher-Powell Conjugate Gradient (CGF)	1.9573	0.687	67.63	28.76
Polak-Ribiere Conjugate Gradient (CGP)	2.0542	0.7274	72.91	31.25
Powell-Beale Conjugate Gradient (CGB)	1.4259	0.3909	44.8	15.05
One-Step-Secant (OSS)	2.624	1.257	121.36	65.7
BFGS quasi-Newton (BFG)	1.3369	0.2888	37.47	9.21
Lavenberg-Marquardt (LM)	0.3052	0.0741	2.72	0.73



(a)



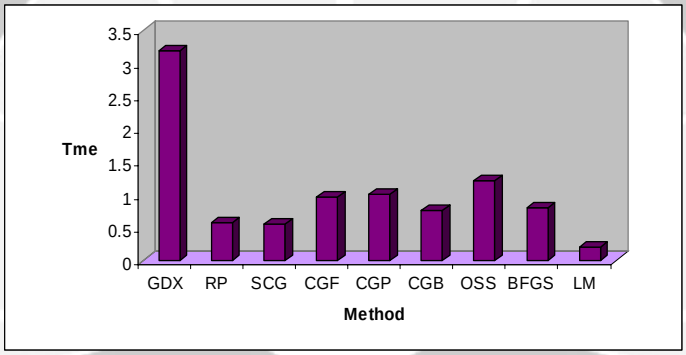
(b)

Figure 2. Comparison time (a) and epoch (b) for sine model

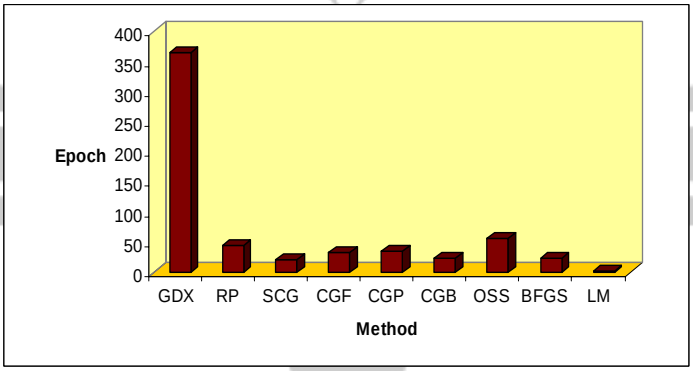
B. Quadratic model: $y = 0.5 + 0.3x^2 + \varepsilon$

Table 2. Comparison of time and epoch across optimization methods

Methods	Time	Stdev. time	Epoch	Stdev. epoch
Variabel Learning Rate (GDX)	3.198	1.111	364.7	135.2
Resilient backpropagation (RP)	0.5772	0.09585	44.84	10.75
Scaled Conjugate Gradient (SCG)	0.5582	0.08198	21.49	4.034
Fletcher-Powell Conjugate Gradient (CGF)	0.9681	0.1929	31.96	8.707
Polak-Ribiere Conjugate Gradient (CGP)	1.0089	0.2228	34.37	10.15
Powell-Beale Conjugate Gradient (CGB)	0.7628	0.1516	23.17	5.944
One-Step-Secant (OSS)	1.2168	0.244	55.95	12.99
BFGS quasi-Newton (BFG)	0.7922	0.1281	22.85	4.484
Lavenberg-Marquardt (LM)	0.2027	0.02609	1	0



(a)



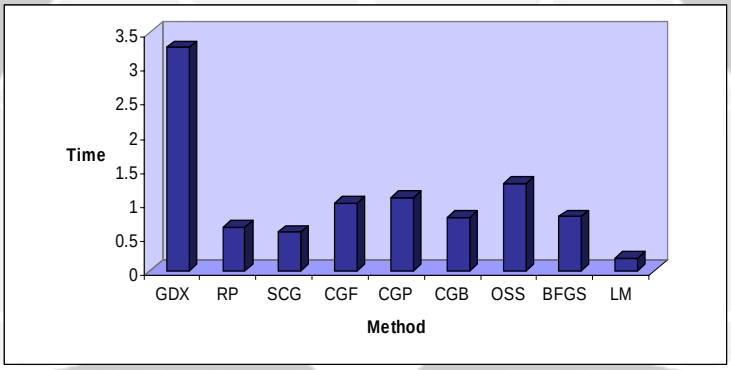
(b)

Figure 3. Comparison time (a) and epoch (b) for quadratic model

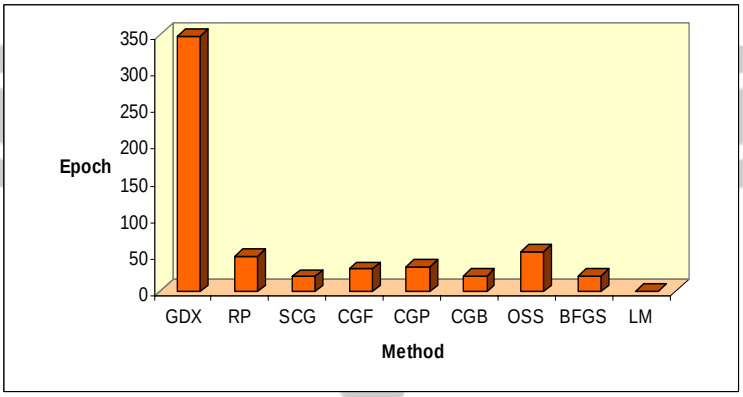
C. Cubic model: $y = 0.5 + 0.3x^3 + \varepsilon$

Table 3. Comparison of time and epoch across optimization methods

Methods	Time	Stdev. time	Epoch	Stdev. epoch
Variabel Learning Rate (GDX)	3.277	0.9823	347.2	109.9
Resilient backpropagation (RP)	0.6405	0.1329	48.15	12.91
Scaled Conjugate Gradient (SCG)	0.5733	0.08754	20.98	4.102
Fletcher-Powell Conjugate Gradient (CGF)	0.9894	0.2323	30.36	8.763
Polak-Ribiere Conjugate Gradient (CGP)	1.0823	0.302	33.63	11.37
Powell-Beale Conjugate Gradient (CGB)	0.7954	0.1724	21.87	4.644
One-Step-Secant (OSS)	1.2896	0.2657	54.62	11.24
BFGS quasi-Newton (BFG)	0.7987	0.1214	21.88	3.917
Lavenberg-Marquardt (LM)	0.1937	0.03119	1	0



(a)



(b)

Figure 3. Comparison time (a) and epoch (b) for qubic model

Based on the tables or charts for sine, quadratic and cubic models, we can see that the fastest training is Lavenberg-Marquardt (trainlm) and the slowest training is variable learning rate (traingdx). These results confirm the previous ones [5, 9]. For these cases, all of the methods can reach convergence. Variable learning rate converge after the number of epoch are 1366, 364 and 347 successively for sine, quadratic and cubic models whereas Lavenberg-Marquardt needs less than 3 epochs to converge for all given models.

7. Conclusion

Many numerical optimization techniques have been successfully used to speed up convergence of the BP learning algorithm. The speed of convergence is depend on many factors, such as the complexity of the problem, the number of data points in the training set, the number of weights and biases in the network, and the error goal. The results indicate for given regression problems, the fastest training is Lavenberg-Marquardt (trainlm). Further research should also definitely be done to other types of nonlinear regression problems which are more complex.

References

[1] Balakhrisnan, K. and V. Honavar (1992), Faster Learning in Multi-Layer Networks by Handling Flat-Spots, *Proceedings of the International Joint Conference on Neural Networks – IJCNN'92*, Beijing, China.

[2] Barnard, E. (1992), Optimization for Training Neural Nets, *IEEE Trans. Neural Net.*, **3**, no. 2, 232-240.

[3] Batiti, R. (1992), First and Second Order Method for Learning: Between Steepest Descent and Newton's Method, *Neural Computation*, **4**, no. 2, 141-166.

[4] Charalambous, C (1992), Conjugate Gradient Algorithm for Efficient Training of Artificial Neural Networks, *IEEE Proc.*, **139**, no. 3, 301-310.

[5] Demuth, H. and M. Beale (1998), *Neural Network Toolbox User's Guide*, The Mathwork, Inc.

[6] Hornik, K., M. Stinchcombe and H. White (1989), Multilayer feedforward networks are universal approximators, *Neural Networks*, ,**2**, 359-66.

[7] Jacobs, R.A. (1998), Increased Rates of Convergence Through Learning Rate Adaptation, *Neural Networks*, **1**, no. 4, 295-308.

[8] Moller, M. (1007), Efficient Training of Feed-Forward Neural Networks, Ph.D. Thesis, Aarhus University, Denmark, 1997.

[9] Hagan, M.T., and M.B. Menhaj, Training Feedforward Networks with the Marquardt Algorithm, *IEEE Transaction on Neural Networks*, **5**, no. 6, 989-993.

[10] Rumelhart, D.E., J.L. McClelland, and P.R. Group (1986), *Parallel Distributed Processing: Explorations in the Microstructures of Coginition,* **1***: Foundations*, Cambridge: MIT Press.

SRI REZEKI: Ph.D student at Department of Mathematics, Universitas Gadjah Mada
Jl. C. Simanjuntak Sekip Utara – Yogyakarta 55281, Indonesia.
Phone/Fax: +62 +274 522 443
Department of Mathematics Education, FKIP Universitas Islam Riau,
Jl. Kaharudin Nasution Marpoyan Pekanbaru, Indonesia.
E-mail: sri_rezeki_uir@yahoo.com

SUHARTONO: Ph.D student at Department of Mathematics, Universitas Gadjah Mada
Jl. C. Simanjuntak Sekip Utara – Yogyakarta 55281, Indonesia.
Phone/Fax: +62 +274 522 443
Department of Statistics, Institut Teknologi Sepuluh Nopember, Kampus ITS Keputih Surabaya 60111, Indonesia.
E-mail: suhartono@statistika.its.ac.id

SUBANAR: Department of Mathematics, Universitas Gadjah Mada,
Jl. C. Simanjuntak Sekip Utara – Yogyakarta 55281, Indonesia.
Phone/Fax: +62 +274 522 443
E-mail: subanar@yahoo.com

S. GURITNO: Department of Mathematics, Universitas Gadjah Mada,
Jl. C. Simanjuntak Sekip Utara – Yogyakarta 55281, Indonesia.
Phone/Fax: +62 +274 522 443
E-mail: suryoguritno@ugm.ac.id

THE IMPACT OF DATA PREPROCESSING ON FEEDFORWARD NEURAL NETWORKS MODEL FOR FORECASTING TREND AND SEASONAL TIME SERIES

Suhartono[a], Subanar[b], S. Guritno[c]

[a] Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[b,c] Universitas Gadjah Mada, Yogyakarta, Indonesia

Abstract: The issue of data preprocessing on the use of feedforward neural networks (FFNN) recently become one of the central topics for the neural networks (NN) community. In this paper, we will investigate this topic particularly on the use of FFNN for modeling effectively time series with both trend and seasonal patterns. Limited empirical studies on seasonal time series forecasting with neural networks show that some find neural networks are able to model seasonality directly and prior deseasonalization is not necessary, and others conclude just the opposite. In this research, we study particularly on the effectiveness of data preprocessing, including detrending and deseasonalization on FFNN modeling and forecasting performance. We use two kinds of data, simulation and real data. Simulation data are examined on multiplicative of trend and seasonality patterns. The results are compared to those obtained from the classical time series model. Our result shows that a combination of detrending and deseasonalization is the effective data preprocessing on the use of FFNN for forecasting trend and seasonal time series.

Keywords: data preprocessing, feedforward neural networks, trend, seasonality, time series, forecasting.

1. Introduction

Many business and economic time series are non-stationary time series that contain trend and seasonal variations. The trend is the long-term component that represents the growth or decline in the time series over an extended period of time. Seasonality is a periodic and recurrent pattern caused by factors such as weather, holidays, or repeating promotions. Accurate forecasting of trend and seasonal time series is very important for effective decisions in retail, marketing, production, inventory control, personnel, and many other business sectors [16]. Thus, how to model and forecast trend and seasonal time series has long been a major research topic that has significant practical implications.

There are some forecasting techniques that usually used to forecast data time series with trend and seasonality, including additive and multiplicative methods. Those methods are Winter's exponential smoothing, Decomposition, Time series regression, and ARIMA models (see e.g. [4] and [9]). Recently, Neural Networks (NN) models are also used for time series forecasting (see e.g. [7, 11, 15]). Suhartono *et al.* [21] did comparative study of these methods by using airline data and

concluded that there was no best model satisfies simultaneously in both training and testing data. They also recommended the possibility for doing further research by combining some methods.

The aim of this paper is to develop new hybrid model by combining decomposition method as data preprocessing and NN model for forecasting trend and seasonal time series. The results are compared to ARIMA models.

2. Modeling Trend and Seasonal Time Series

Modeling trend and seasonal time series has been one of the main research endeavors for decades. In the early 1920s, the decomposition model along with seasonal adjustment was the major research focus due to Persons [19, 20] work on decomposing a seasonal time series. Holt [12] and Winters [25] developed method for forecasting trend and seasonal time series based on the weighted exponential smoothing. Among them, the work by Box and Jenkins [2] on the seasonal ARIMA model has had a major impact on the practical applications to seasonal time series modeling. This model has performed well in many real world applications and is still one of the most widely used seasonal forecasting methods. More recently, NN have been widely used as a powerful alternative to traditional time series modeling (see e.g. [10, 18, 26]). While their ability to model complex functional patterns in the data has been tested, their capability for modeling seasonal time series is not systematically investigated.

In this section, we will give a brief review of these forecasting models, particularly seasonal ARIMA, decomposition method and NN model.

2.1. Seasonal ARIMA Model

The seasonal ARIMA model belongs to a family of flexible linear time series models that can be used to model many different types of seasonal as well as nonseasonal time series. The seasonal ARIMA model can be expressed as (see e.g. [3, 5, 23]):

$$\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D y_t = \theta_q(B)\Theta_Q(B^S)\varepsilon_t \ , \qquad (1)$$

where S is the seasonal length, B is the back shift operator and ε_t is a sequence of white noises with zero mean and constant variance. Box and Jenkins [2] proposed a set of effective model building strategies for seasonal ARIMA based on the autocorrelation structures in a time series.

2.2. Decomposition Method

The multiplicative decomposition model has been found to be useful when modeling time series that display increasing or decreasing seasonal variation (see [4]; chapter 7). The key assumption inherent in this model is that seasonality can be separated from other components of the series. The multiplicative decomposition model is

$$y_t = T_t \times S_t \times C_t \times I_t \quad (2)$$

where

y_t = the observed value of the time series in time period t
T_t = the trend component in time period t
S_t = the seasonal component in time period t
C_t = the cyclical component in time period t
I_t = the irregular component in time period t.

2.3. Neural Networks Model

Neural networks (NN) are a class of flexible nonlinear models that can discover patterns adaptively from the data. Theoretically, it has been shown that given an appropriate number of nonlinear processing units, NN can learn from experience and estimate any complex functional relationship with high accuracy. Empirically, numerous successful applications have established their role for pattern recognition and time series forecasting.

Feedforward Neural Networks (FFNN) is the most popular NN models for time series forecasting applications. Figure 1 shows a typical three-layer FFNN used for forecasting purposes. The input nodes are the previous lagged observations, while the output provides the forecast for the future values. Hidden nodes with appropriate nonlinear transfer functions are used to process the information received by the input nodes.

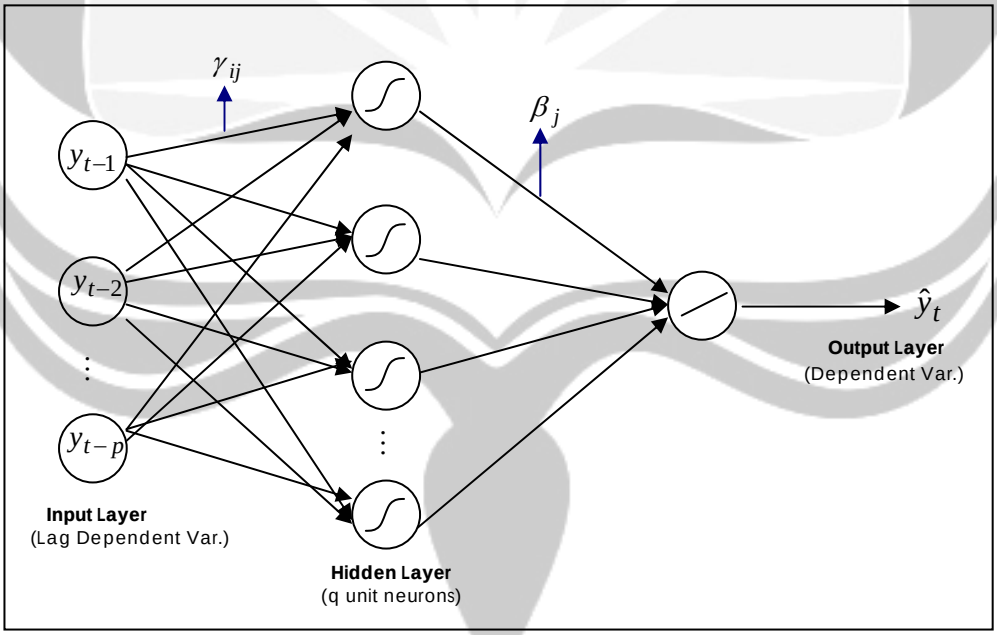


Figure 1. Architecture of neural network model with single hidden layer

The model of FFNN in figure 1 can be written as

$$y_t = \beta_0 + \sum_{j=1}^{q} \beta_j f\left(\sum_{i=1}^{p} \gamma_{ij} y_{t-i} + \gamma_{oj}\right) + \varepsilon_t \,, \tag{3}$$

where p is the number of input nodes, q is the number of hidden nodes, f is a sigmoid transfer function such as the logistic:

$$f(x) = \frac{1}{1+e^{-x}} \,. \tag{4}$$

$\{\beta_j, j = 0,1,\cdots,q\}$ is a vector of weights from the hidden to output nodes and $\{\gamma_{ij}, i = 0,1,\cdots,p; j = 1,2,\cdots,q\}$ are weights from the input to hidden nodes. Note that equation (3) indicates a linear transfer function is employed in the output node.

Functionally, the FFNN expressed in equation (3) is equivalent to a nonlinear AR model. This simple structure of the network model has been shown to be capable of approximating arbitrary function (see e.g. [6, 13, 14, 24]). However, few practical guidelines exist for building a FFNN for a time series, particularly the specification of FFNN architecture in terms of the number of input and hidden nodes is not an easy task.

3. Research Methodology

The purpose of this research is to provide empirical evidence on the comparative study of many data preprocessing method in NN model for forecasting trend and seasonal time series. The major research questions we investigate is:

- Does data preprocessing has a great impact on the accuracy of NN model for forecasting trend and seasonal time series?
- Which data preprocessing is the most effective on NN model forr forecasting model for trend and seasonal time series?

We conduct empirical study with simulation and real data, the international airline passenger data, to address these questions. This real data has been analyzed by many researchers, see for example Nam and Schaefer [17], Hill *et al.* [11], Faraway and Chatfield [7], Atok and Suhartono [1], Suhartono *et al.* [21, 22] and now become one of two data to be competed in Neural Network Forecasting Competition on June 2005 (see www.neural-forecasting.com).

3.1. Data

The simulation and real data contain 144 month observations. The first 120 data observations are used for model selection and parameter estimation (training data in term of NN model) and the last 24 points are reserved as the test for forecasting evaluation and comparison (testing data). Figure 2 plots representative time series of these data. It is clear that the series has an upward trend together with seasonal variations.

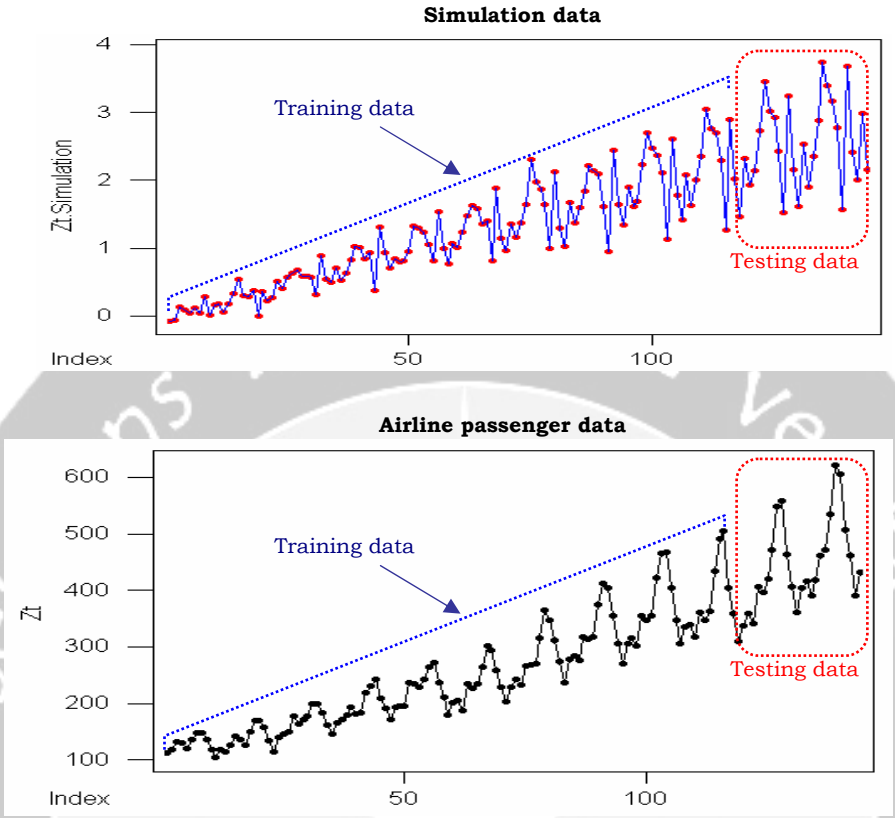


Figure 2. Time series plot of simulation and real data

3.2. Research Design

Three types of data preprocessing based on the decomposition method are applied and compare to the airline data. Those are detrend, deseasonal, and combination detrend-deseasonal. All of these data preprocessing are implemented by using MINITAB software.

To determine the best hybrid model, that is combination data preprocessing based on the decomposition method and NN model, an experiment is conducted with the basic cross validation method. The available training data is used to estimate the weights for any specific model architecture. The testing set is the used to select the best model among all models considered. In this study, the number of hidden nodes varies from 1 to 10 with an increment of 1. The lags of 1, 12 and 13 are included due to the results of Faraway and Chatfield [7], Atok and Suhartono [1], and Suhartono *et al.* [21].

The FFNN model used in this empirical study is the standard FFNN with single-hidden-layer shown in Figure 1. We use S-Plus to conduct FFNN model building and evaluation. The initial value is set to random with 50 replications in each

model to increase the chance of getting the global minimum. We also use the standard data preprocessing in NN for the airline data by transform detrend, deseasonal, and combination detrend-deseasonal data to N(0,1) scale. The performance of in-sample fit (training data) and out-sample forecast (testing data) is judged by the commonly used error measures, the mean squared error (MSE) and ratio MSE to ARIMA model.

4. Empirical Results

Table 1 summarizes the result of the impact of some data preprocessing on NN forecasting and report performance measures across training and testing samples for the simulation data. Numbers greater than one on column ratio indicate poorer forecast performance than comparable ARIMA model, and vice versa for numbers less than one.

Table 1. The result of the comparison between preprocessing data for FFNN and ARIMA models, both in training and testing data, for the simulation data.

Model and Preprocessing	IN-SAMPLE (TRAINING DATA)		OUT-SAMPLE (TESTING DATA)	
	MSE	Ratio to ARIMA	MSE	Ratio to ARIMA
▪ ARIMA model	0.0234672	1	0.0201110	1
▪ FFNN model				
(1). Original Data				
a. Model 3-1-1 (**)	0.0173123	0.738	0.0243289	1.210
b. Model 3-10-1 (*)	0.0059803	0.255	0.4041078	20.095
(2). Detrend				
a. Model 3-2-1 (**)	0.0170082	0.725	0.0252411	1.255
b. Model 3-10-1 (*)	0.0069713	0.297	0.0722953	3.595
(3). Deseasonal				
▪. Model 3-3-1 (**) (*)	0.5576327	23.762	2.951785	146.782
(4). Detrend-Deseasonal				
a. Model 3-5-1 (**)	0.0051065	0.218	0.009484	0.472
b. Model 3-10-1 (*)	0.0036444	0.155	4.308886	214.266

(*) : the best model in training data (in-sample forecast)
(**) : the best model in testing data (out-sample forecast)

The results of the impact of some data preprocessing on NN forecasting and report performance measures across training and testing samples for the airline data are summarized in table 2.

Several observations can be made from table 1 and 2. First, detrend as data processing does yield poorer result than the original data or ARIMA. It can be clearly seen from table 1 and 2 that the ratio MSE at testing samples for NN are greater than 1. Second, deseasonal as data processing gives the worst result than other data preprocessing and also compared to ARIMA. We can observe that the

best model in testing samples by using deseasonal as data preprocessing yield the greatest ratio MSE compared ti the results of the original data or the ratio of detrend as data preprocessing. Third, the combination detrend-deseasonal as data preprocessing yields the best result for forecasting the airline data. It can be shown by the least ratio of MSE at testing data.

Table 2. The result of the comparison between preprocessing data for FFNN and ARIMA models, both in training and testing data, for the airline passenger data.

Model and Preprocessing	IN-SAMPLE (TRAINING DATA)		OUT-SAMPLE (TESTING DATA)	
	MSE	Ratio to ARIMA	MSE	Ratio to ARIMA
▪ ARIMA model	88.8618	1	1527.03	1
▪ FFNN model and data transform to N(0,1)				
(1). Original Data				
a. Model 3-1-1 (**)	92.8729	1.045	1219.81	0.799
b. Model 3-10-1 (*)	26.3230	0.296	5299.06	3.470
(2). Detrend			1672.27	1.095
a. Model 3-4-1 (**)	71.0023	0.799	5630.35	3.687
b. Model 3-10-1 (*)	20.2050	0.227		
(3). Deseasonal				
a. Model 3-6-1 (**)	25.2444	0.284	4218.18	2.762
b. Model 3-10-1 (*)	12.9047	0.145	255939.30	167.609
(4). Detrend-Deseasonal				
a. Model 3-4-1 (**)	35.4608	0.399	582.93	0.382
b. Model 3-10-1 (*)	11.3842	0.128	1532.17	1.003

(*) : the best model in training data (in-sample forecast)
(**) : the best model in testing data (out-sample forecast)

In general, we can clearly see on the ratio of testing samples comparison that combination detrend-deseasonal as data preprocessing and transformation N(0,1) on FFNN with 5 unit nodes (for simulation data) and 4 unit nodes (for the airline data) in hidden layer yield the best MSE. The reduction of MSE is highly significant if compare to the result of FFNN without detrend-deseasonal as data preprocessing, those are 52.8% for simulation data and 61.8% for the airline data.

5. Conclusions

Based on the results we can conclude that the combination detrend-deseasonal (based on the decomposition method) as data preprocessing in FFNN yields a great impact on the increasing accuracy of forecasting trend and seasonal time series. Our result also shows that the best model in training data tends to yield overfitting on testing. This condition give a chance to do further research by implementing some NN model selection methods in order for the model selection process becomes efficient.

References

[1] Atok, R.M. and Suhartono [2000], *Comparison between Neural Networks, ARIMA Box-Jenkins and Exponential Smoothing Methods for Time Series Forecasting,* Research Report, Lemlit: Institut Teknologi Sepuluh Nopember.

[2] Box, G.E.P. and G.M. Jenkins (1976), *Time Series Analysis: Forecasting and Control*, San Fransisco: Holden-Day, Revised edn.

[3] Box, G.E.P., G.M. Jenkins, and G.C. Reinsel (1994), *Time Series Analysis, Forecasting and Control,* 3rd edition, Englewood Cliffs: Prentice Hall.

[4] Bowerman, B.L. and D. O'Connel (1993), *Forecasting and Time Series: An Applied Approach*, *3rd ed*, Belmont, California: Duxbury Press.

[5] Cryer, J.D. (1986), *Time Series Analysis,* Boston: PWS-KENT Publishing Co.

[6] Cybenko, G. (1989), Approximation by superpositions of a sigmoidal function, *Mathematics of Control, Signals and Systems,* **2**, 304–314.

[7] Faraway, J. and C. Chatfield (1998), Time series forecasting with neural network: a comparative study using the airline data, *Applied Statistics*, **47**, 231–250.

[8] Fildes, R. and S. Makridakis (1995), The impact of empirical accuracy studies on time series analysis and forecasting, *International Statistical Review,* **63** (3), pp. 289–308.

[9] Hanke, J.E. and A.G. Reitsch (1995), *Business Forecasting,* Prentice Hall, Englewood Cliffs, NJ.

[10] Hansen, J.V. and R.D. Nelson (2003), Forecasting and recombining time-series components by using neural networks, *Journal of the Operational Research Society,* **54** (3), pp. 307–317.

[11] Hill, T., M. O'Connor, and W. Remus (1996), Neural network models for time series forecasts", *Management Science*, **42**, pp. 1082–1092.

[12] Holt, C.C. (1957), Forecasting seasonal and trends by exponentially weighted moving averages, *Office of Naval Research*, *Memorandum* No. **52**.

[13] Hornik, K., M. Stichcombe, and H. White (1989a), Multilayer feedforward networks are universal approximators, *Neural Networks*, **2**, pp. 359–366.

[14] Hornik, K., M. Stichcombe, and H. White (1989b), Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks, *Neural Networks*, **3**, pp. 551–560.

[15] Kaashoek, J.F. and H.K. Van Dijk (2001), *Neural Networks as Econometric Tool,* Report EI 2001–05, Econometric Inst. Erasmus University Rotterdam.

[16] Makridakis, S. and S.C. Wheelwright (1987), *The Handbook of Forecasting: A Manager's Guide*, 2nd Edition, John Wiley & Sons Inc., New York.

[17] Nam, K. and T. Schaefer (1995), Forecasting international airline passenger traffic using neural networks, *Logistics and Transportation Review,* **31** (3), pp. 239–251.

[18] Nelson, M., T. Hill, T. Remus, and M. O'Connor (1999), Time series forecasting using NNs: Should the data be deseasonalized first?, *Journal of Forecasting,* **18**, pp. 359–367.

[19] Persons, W.M. (1919), Indices of business conditions, *Review of Economics and Statistics* **1**, pp. 5–107.

[20] Persons, W.M. (1923), Correlation of time series, *Journal of American Statistical Association,* **18**, pp. 5–107.

[21] Suhartono, Subanar and S. Guritno (2005), A Comparative study of forecasting models for trend and seasonal time series: Does complex model always yield better forecast than simple models?, *Proceeding National Mathematics Seminar*, UNS, Solo.

[22] Suhartono, Subanar and S. Rezeki (2005), Feedforward Neural Networks Model for Forecasting Trend and Seasonal Time Series, *Proceedings of the 1st IMT-GT Regional Conference on Mathematics, Statistics and Their applications*, North Sumatera, Indonesia.

[23] Wei, W.W.S. (1990), *Time Series Analysis: Univariate and Multivariate Methods,* Addison-Wesley Publishing Co., USA.

[24] White, H. (1990), Connectionist nonparametric regression: Multilayer feed forward networks can learn arbitrary mapping, *Neural Networks*, **3**, 535–550.

[25] Winters, P.R. (1960), Forecasting Sales by Exponentially Weighted Moving Averages, *Management Science,* **6**, pp. 324-342.

[26] Zhang, G., B.E. Patuwo, and M.Y. Hu (1998), Forecasting with artificial neural networks: The state of the art, *International Journal of Forecasting,* **14**, pp. 35–62.

SUHARTONO: Ph.D student at Department of Mathematics, Universitas Gadjah Mada, Bulaksumur, Yogyakarta 55281, Indonesia.
Department of Statistics, Institut Teknologi Sepuluh Nopember, Kampus ITS Keputih Surabaya 60111, Indonesia.
E-mail: suhartono@statistika.its.ac.id

SUBANAR: Department of Mathematics, Universitas Gadjah Mada, Bulaksumur, Yogyakarta 55281, Indonesia.
E-mail: subanar@yahoo.com

S. GURITNO: Department of Mathematics, Universitas Gadjah Mada, Bulaksumur, Yogyakarta 55281, Indonesia.
E-mail: suryoguritno@ugm.ac.id

Outlier Labeling in Multivariate Outlier Detection

D. E. Herwindiati[a], M.A. Djauhari[b], S. Darwis[b]

[a]PhD Student at Department of Mathematics, ITB,
Bandung, Indonesia
[b]ITB, Bandung, Indonesia

Abstract. The problem of detecting outliers in multivariate data has been extensively researched in recent years. Many detection procedures are available in different and various ways, some use outlier labeling approach, which is useful to separate 'suspects' from main of data. To have a better understanding about outlier labeling, this paper will investigate two approaches, i.e. projection pursuit approach for obtaining directions that maximize kurtosis coefficient and MCD approach for finding minimum ratio of scatter matrix. The performance of two approaches are analyzed by computing the value of probability of type I error.

1.Introduction

Outliers are observations which appear to be inconsistent with the remainder of set of data (Barnett and Lewis). In one or two dimensions, outlying data are easily identified from simple plot, but detection of outliers is more challenging in higher dimensions. Many detection procedures are available in different and various ways, some use outlier labeling approach. The approach is useful to separate 'suspects' from the bulk of data, the observations exceeding certain value are called 'labeled' outliers or suspects.

The different terminologies and methodologies are used for same objective. Hadi (1992) divides data set into 'basic' and 'non-basic' subset by using robust measure of outlyingness, Rousseeuw and Van Driessen (1999) use the minimum determinant of sample covariance matrix (MCD) to separate a data set into two groups, i.e., the group of 'bad' data and the group of 'good' data. With the same aim Pan et.al. (2000) separate the suspects from the bulk of data using projection pursuit. Later on, Pena and Prieto (2001) propose to separate the group of suspects from the group of 'good' data using the projection on *2p* orthogonal directions maximizing and minimizing kurtosis. This method is very tedious because we have to find all axes maximizing and all axes minimizing the kurtosis. For the reason, Herwindiati, Djauhari and Yatawara (2005) use Wilks's (1963) statistics and MCD estimator to propose 'resistant' approach for identifying the 'reliable' suspects.

To have a better understanding about outlier labeling, this paper will investigate two approaches of separation process and they will be discussed in Section 2. In the next Section, Section 3, it will be discussed outlier testing which is useful for detecting that 'labeled' outliers are really outliers. An illustrative examples will clarify the advantages of these methods are also discussed in each section.

2. Outlier Labeling Procedures

The concept of separating suspects from main data, which is known as outlier labeling, has been introduced more than twenty years. For example, Rohlf (1975) defined an inter-points distance and then construct its corresponding minimum spanning tree for a separation 'suspects' from bulk of data. Although the distribution of gap is still unknown but his 'beautiful mind' motivated such as Rousseeuw and Van Zomeren (1990), Hadi (1992), Rousseeuw and Van Driessen (1999), Pan et.al. (2000), Pena and Prieto (2001) to develop this concept with different approach.

What is outlier labeling ?

The basic principle of outlier labeling is to separate a data set into two groups, i.e., the group of 'suspected' data and the group of 'unsuspected' data. Sometimes those groups are also called the group of 'bad' or 'unclean' data and the group of 'good' or 'clean' data. We can find those terminologies, for example, in Rousseeuw & Van Driessen (1999) and Pena and Prieto (2001).

The benefit of outlier labeling in outlier detection is the potential outliers or suspects, can be separated from 'clean' observations and the process of outlier testing can be done on suspected data group only. The significant benefit can be come upon in a large data set problem.

2.1 Outlier Labeling Using Projection Pursuit Approach

Projection pursuit procedure is often used for detecting outliers in multivariate data. The basic idea of this procedure is to project the multivariate data to univariate observations and then to apply an appropriate univariate outlier identifier to identify 'candidate outliers'.

As well known, in univariate normal data, outliers have often been associated with large kurtosis values, and in multivariate normal samples each outlier is extreme point along the direction of projected data. Because of these fact, Pena and Prieto (2001) proposed projection pursuit algorithm of the fourth moment for detection outliers. All observations exceeding a cutoff value are called 'labeled' outliers or suspects. For a standard multivariate contamination model, they showed that these directions can identify a set 'candidate' outliers.

Consider a $p-$dimensional random variable X following a distribution given as a mixture of normals of the form $(1-\alpha)N(O,I)+\alpha\, N(\delta e_1, \lambda I)$, where e_1 denotes the first unit vector, δ and λ are constant. Since the kurtosis coefficient is invariant to affine transformations, the centering and the scalling variable are useful to ensure that variable has mean 0 and covariance matrix equal to the identity. The transformed variable, Y, will follow a distribution of the form $(1-\alpha)N(m_1,S)+\alpha\, N(m_2,\lambda S)$, where

$$m_1 = -\alpha\delta S^{1/2} e_1 \quad m_2 = (1-\alpha)\delta S^{1/2} e_1$$

$$v_1 = 1-\alpha(1-\lambda) \quad v_2 = \frac{\delta^2\alpha(1-\alpha)}{v_1+\delta^2\alpha(1-\alpha)} \qquad S = \frac{1}{v_1}\left(I - v_2\, e_1\, e_1^{'}\right)$$

To study the behavior of the univariate projections for the variable and the kurtosis coefficient values, it's considered an arbitrary projection direction u. Using the affine invariance of the kurtosis coefficient, it's assumed that $\|u\| = 1$. The projected univarite random variable $Z = u'Y$ has a distribution

$$(1-\alpha)N(m_1'u, u'Su) + \alpha N(m_2'u, \lambda u'Su),$$

with $\mathrm{E}(Z) = 0$ and $\mathrm{E}(Z^2) = 1$. The kurtosis coefficient of Z will be given by $\gamma_Z(\omega) = a(\alpha,\delta,\lambda) + b(\alpha,\delta,\lambda)\omega^2 + c(\alpha,\delta,\lambda)\omega^4$

where $\omega \equiv u_1 = e_1'u$, the coefficient a, b and c correspond to

$$a(\alpha,\delta,\lambda) = \frac{3}{v_1^2}\left(1-\alpha+\alpha\lambda^2\right)$$

$$b(\alpha,\delta,\lambda) = \frac{6v_2}{v_1^2}(1-\lambda)\left(\alpha^2\lambda - (1-\alpha)^2\right)$$

$$c(\alpha,\delta,\lambda) = v_2^2\left(3\frac{1-\alpha+\alpha\lambda^2}{v_1^2} - 6\frac{\alpha+\lambda(1-\alpha)}{v_1} + \frac{\alpha^3+(1-\alpha)^3}{\alpha(1-\alpha)}\right)$$

The optimization problem defining these extreme directions would be either

$$\max_{\omega} \gamma_Z(\omega)$$
$$\text{s.t} \quad -1 \le \omega \le 1$$

or the minimization problem.

The algorithm for computation of projection directions can be seen in [9].

This methodology is very tedious because one has to find all axes maximizing and all axes minimizing the kurtosis. Concerning Pena and Prieto's approach, Hubert (2001) points out that kurtosis increases rapidly when the level of contamination decreases. On the other hand, if the level of contamination increases the kurtosis decreases rapidly. This phenomenon makes Pena and Prieto's approach difficult to be implemented .

Example 1

To give a good illustration of this approach, a mixture model $(1-\varepsilon)\,N_p(\vec{\mu}_1, \Sigma) + \varepsilon\, N_p(\vec{\mu}_2, \Sigma)$ will be generated. We set $n = 100$, $p = 5$, without loss of generality, $\varepsilon = 0.04$, $\vec{\mu}_1 = \vec{0}$, $\vec{\mu}_2 = 10\vec{e}$, where $\Sigma = I_5$ and $e = (1 \;\; 1 \;\; 1 \;\; 1 \;\; 1)'$. Based on a data set, there are 4 observations which are separated from bulk of data and we wish that contaminated data can be recognized well.

In this case, outlier labeling will be done in maximize direction. For $p = 5$, the cutoff value given by simulation experiment is 4.06. According to this cutoff value, 'labeled' outliers will be determined by iteration process as follows

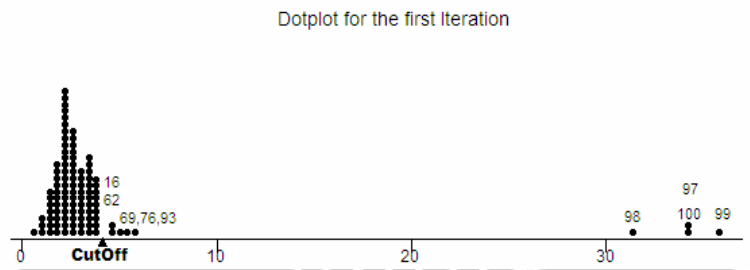


Figure 1. Dotplot kurtosis coefficient using projection pursuit, for the first iteration

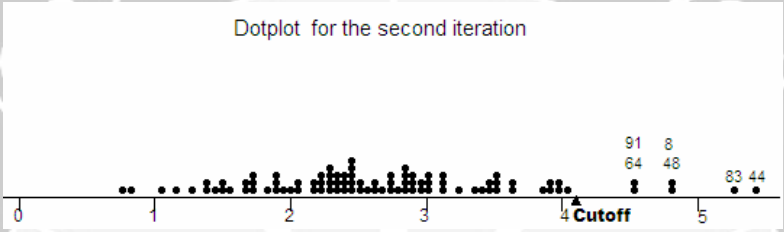


Figure 2. Dotplot kurtosis coefficient using projection pursuit, for the second iteration

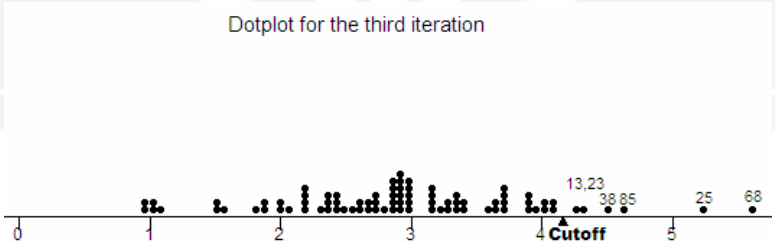


Figure 3. Dotplot kurtosis coefficient using projection pursuit, for the third iteration

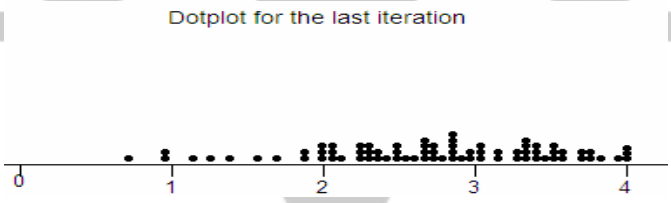


Figure 4. Dotplot kurtosis coefficient using projection pursuit, for the last iteration

Founded on the iterations, we have 21 observations which can be identified as 'labeled' outliers or 'suspects'. This result is quite 'far' from the real problem. How do we have to improve the performance of separation process? The answer is masking and swamping effects have to be avoided. The following method will be proposed to handle the problem.

2.2 Outlier Labeling Based on Wilks's Statistics and Minimum Covariance Determinant (MCD) Approach

Minimum covariance determinant (MCD) estimates which are resistant to outliers will be proposed to identify 'reliable' suspects. The various algorithms have been suggested for estimating the MCD. For MCD estimator, Rousseeuw and Van Driessen (1999) propose the fast algorithm based on C-step. The method is a highly robust estimator of multivariate location and scatter. The basic algorithm of C- step is as follows.

Consider a data set $X_n = \left\{\vec{X}_1, \vec{X}_2, \cdots, \vec{X}_n\right\}$ of $p-$variate observations.

1.Take a subset of the data of size h, H_{old}.

2.Compute $\bar{\vec{X}}_{H_{old}}$, $S_{H_{old}}$ and $d^2_{H_{old}}\left(\vec{X}_i, \bar{\vec{X}}_{H_{old}}\right) = d^2_{H_{old}}(i)$ for $i = 1,2,\cdots n$

3.Sort these distances, which yields a permutation π for which $d^2_{H_{old}}((\pi_1)) \le d^2_{H_{old}}((\pi_2)) \le \cdots \le d^2_{H_{old}}((\pi_n))$

4.Assign $\{\pi(1), \pi(2), \cdots, \pi(h)\}$ to H_{new}

5.Calculate $\bar{\vec{X}}_{H_{new}}, S_{H_{new}}$ and $d^2_{H_{new}}\left(\vec{X}_i, \bar{\vec{X}}_{H_{new}}\right)$, based on H_{new}, and it shows that $\det\left(S_{H_{new}}\right) \le \det\left(S_{H_{old}}\right)$.

6.Repeat C-step, if $\det\left(S_{H_{new}}\right) = 0$ or $\det\left(S_{H_{new}}\right) - \det\left(S_{H_{old}}\right) = 0$, the process is stopped, otherwise, the process will be continued until the sequence $\det\left(S_{H_1}\right) \ge \det\left(S_{H_2}\right) \ge \det\left(S_{H_3}\right) \ge \cdots\cdots$ is convergence.

In trying to have reliable identification of suspect, based on Wilks's concept (1963), we use MCD location estimate to construct the statistics $R^*_{(1)}, R^*_{(2)}, \cdots, R^*_k$, which is described as follows.

Let $\vec{X}_1, \vec{X}_2, \cdots, \vec{X}_n$ be a random sample drawn from $N_p(\vec{\mu}, \Sigma)$, $\bar{\vec{X}}^*$ is MCD location estimate and A^* is the scatter matrix based on $\bar{\vec{X}}^*$, A^* can be written as :

$$A^* = \sum_{i=1}^{n}\left(\vec{X}_i - \bar{\vec{X}}^*\right)\left(\vec{X}_i - \bar{\vec{X}}^*\right)^t$$

If $A^*_{(-j)}$ is the scatter matrix A^* corresponding with $\vec{X}_j$ removed from the bulk of data, $A^*_{(-j)}$ can be formulated as:

$$A^*_{(-j)} = \sum_{\substack{i=1 \\ i \neq j}}^{n} \left(\vec{X}_i - \vec{\bar{X}}^* \right) \left(\vec{X}_i - \vec{\bar{X}}^* \right)^t$$

Then the ratio of determinant of scatter matrix $R^*_{(1)}$, which will be applied to identify the first of "labeled' outlier'(index l say), is defined as

$$R^*_{(1)} = \min_j \left(\frac{\left| A^*_{(-j)} \right|}{\left| A^* \right|} \right) = \frac{\left| A^*_{(-l)} \right|}{\left| A^* \right|}$$

After the most extreme observation $\vec{X}_l$ has been identified, $R^*_{(2)}$ will be determined to investigate the second of 'labeled' outlier on $(n-1)$ observations. If m is the index of the second most extreme observation, then $R^*_{(2)}$ is defined as

$$R^*_{(2)} = \min_j \left(\frac{\left| A^*_{(jl)} \right|}{\left| A^*_{(l)} \right|} \right) = \frac{\left| A^*_{(ml)} \right|}{\left| A^*_{(l)} \right|}$$

For specified number k of potential suspects, this procedure is repeated to identify a series of potential observations as suspects $\vec{X}_l, \vec{X}_m, \ldots$ etc, so we have a series of scatter matrix ratio $R^*_{(1)}, R^*_{(2)}, \cdots, R^*_k$. The joint distribution of $R^*_{(j)}$, $j = 1,2,\cdots,k$ very complicated because they are not independent. For the reason, the critical values will be computed by simulation approach.

Let a series $R^*_{(1)}, R^*_{(2)}, \cdots, R^*_k$, compare $R^*_{(k)}, R^*_{(k-1)}, \cdots, R^*_1$ with appropriate ctitical values $\lambda_k, \lambda_{k-1}, \cdots, \lambda_1$ by sequential procedure. For declaring k suspects by specifying significant level α, inspired by Rosner (1975), the critical values λ_c can be determined such as:

$$P\left[R^*_{(j)} < \lambda_j(\beta)\right] = \beta$$

$$P\left[\bigcup_{j=1}^{k} \left\{R^*_{(j)} < \lambda_j(\beta)\right\}\right] = \alpha$$

The observations $\underbrace{\vec{X}_l, \vec{X}_m, \cdots, \vec{X}_k}_{k}$ are suspects

$$\text{if} \quad k = \max_j \left(R^*_{(j)} < \lambda_j \right)$$

Example 2

To show the advantage of our approach, we will use Mahalanobis distance for exploring data. The similar data will be applied to recognize dispersion of MCD approach.

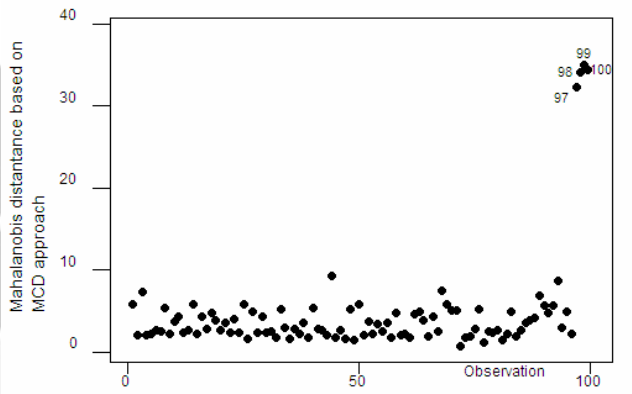


Figure 5. The scatter plot of Mahalanobis distance based on MCD approach

MCD approach clearly illustrates that there are four observations are far away from bulk of data, see Figure 6. Beginning from the farthest observation, observation 99, process of outlier labeling will be done sequentially by using ratio of determinant of scatter matrix. It has been continued until the remain of data are 'clean' and Table 1 describe the process of sequential removing. If we denote G the group of 'clean' observations and its complement G^c is the group of 'labeled' outliers or the group of suspects, by using MCD approach G^C consists of observations 99,97, 98 and 100, see Table 1 and Figure 6.

Table 1. The result of sequential separation process based on MCD approach

Observation removed	$R^*_{(j)}$	**Cutoff**
99	0.1624	0.7917
100	0.4976	0.7906
98	0.5376	0.7886
97	0.7239	0.7869
93	**0.8475**	0.7852

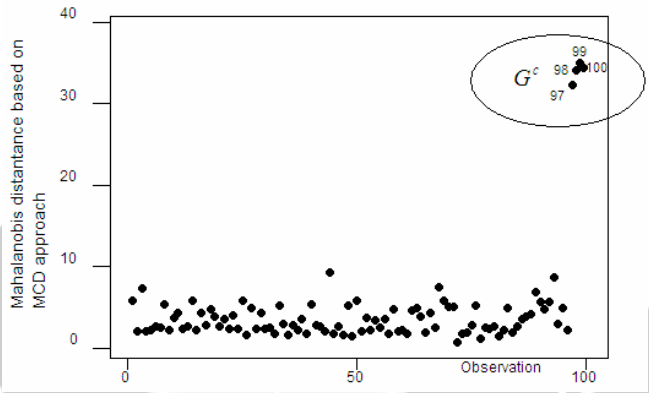


Figure 6. The group G^C of 'labeled' outliers based on Wilks's Statistics and MCD approach

To show performance of this method, the probability of type I error will be computed, see Figure 7. Regarding the value of probability, for recognizing a single suspect, we are able to say that this method has almost never failed to identify 'labeled' outliers or suspects from a data set.

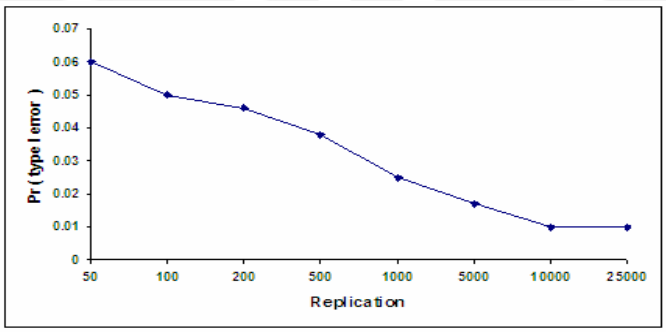


Figure 7. The probability of type I error for n=500, p=5

Compared with the previous method, projection pursuit approach, this method gives different result. Projection pursuit approach is 'always failed' to identify a number of suspects, especially for large data set. In this case, we find the probability type I error from projection pursuit approach is almost 0.99.

3. Outlier testing step

After determining suspects from main of data, we will use outlier testing step for testing whether 'labeled' outliers are outliers**.** Further analysis about the

group of labeled outliers G^c, Herwindiati, Djauhari and Yatawara (2004) proposed the theorem which can be explained as follows

Consider $\vec{X}_1, \vec{X}_2, \cdots, \vec{X}_n$ be a random sample drawn from $N_p(\vec{\mu}, \Sigma)$ where Σ is definite positive. If a data set is divided two groups, i.e. G and G^c groups. We denote G the group of 'good' observations and its complement G^c is the group of labeled outliers. Let also $\bar{\vec{X}}_G$ and S_G represent the mean vector and covariance matrix of all sample items belonging to G, and Card (G^c) = k. Thus,

$$\bar{\vec{X}}_G = \frac{1}{(n-k)} \sum_{i \in G} \vec{X}_i \qquad S_G = \frac{1}{(n-k)-1} \sum_{i \in G} \left(\vec{X}_i - \bar{\vec{X}}_G \right) \left(\vec{X}_i - \bar{\vec{X}} \right)^t .$$

Statistics d_i, for testing whether G^c really consists of all outliers, can be formulated as

$$d_i = \left(\vec{X}_i - \bar{\vec{X}}_G \right)^t S_G^{-1} \left(\vec{X}_i - \bar{\vec{X}}_G \right) \text{ for all } \vec{X}_i \in G^c$$

and the distribution of d_i can be proved as $d_i \sim \frac{p(n-k-1)(n-k+1)}{(n-k)(n-k-p)} F_{p,(n-k)-p}$

3.1 The result of outlier testing step

In this Sub-section we will discuss the result of outlier testing using our proposed in one hand, and projection pursuit approach in the other hand. Furthermore, with $\alpha = 0.05$ as in **Sub-section 2.2**, by using the exact distribution of d_i we obtain that the labeled outliers are really outliers. It is different from the projection pursuit result containing 10 outliers, i.e. observations 16, 44, 64, 69, 76, 93, 97, 98, 99 and 100.

Based on these illustrative examples, we are able to say that the result of our proposed method is close to the real situation where there are 4 contaminated observations which can be believed as outliers.

References

[1] Barnett V. and Lewis T. (1984), *Outliers in Statistical Data*, John Wiley & Sons

[2] Djauhari, M.A (1996), A Necessari and sufficient condition for the uniqueness of minimum spanning tree, *Proceedings of the 29th Institut Teknologi Bandung*, **1 / 2**, 11-18.

[3] Hadi, A.S. (1992), Identifying multivariate outlier in multivariate data, *Journal of Royal Statistical Society B*, Vol. 53, **3**, 761-771.

[4] Herwindiati, D.E., and Djauhari, M.A. (2004), Multivariate outlier labeling and testing, *Proceedings of The 12th National Symposium on Mathematical Sciences,* International Islamic University Malaysia, Dec 23-24, 2004.

[5] Herwindiati, D.E., Djauhari, M.A., and Yatawara, N. (2005), Outlier Detection in Multivariate Data. *Proceedings of The 4th International Symposium on Business*

and Industrial Statistics, Novotel Palm Cove Resort, Australia, April 13 -16, 2005.
[6] Hubert, M. (2001), Discussion, *Technometrics*, Vol. 43, **3**, 303 - 306.
[7] Mardia, K.V., Kent, J.T. and Bibby, J.M. (2000), *Multivariate Analysis*. Seventh Printing, Academic Press.
[8] Pan, J-X., Fung, W-K., and Fang, K-T. (2000), Multiple outlier detection in multivariate data using projection pursuit technique, *Journal of Statistical Planning and Inference*, **83**, 153-167.
[9] Pena, D., and Prieto, J.F. (2001), Multivariate outlier detection and robust covariance matrix estimation. *Technometrics*, **3**, 286-322.
[10] Rohlf, F.J. (1975), Generalization of the gap test for detection of multivariate outliers. *Biometrics*, **31**, 93-101.
[11] Rousseeuw, P.J., and Van Driessen, K. (1999), A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212-223.
[12] Rousseeuw, P.J., and Van Zomeren, B.C. (1990), Unmasking Multivariate Outliers and Leverage Points, *Journal of the American Statistical Association,* **85**, 633-639.
[13] Wilks, S.S. (1963), Multivariate Statistical Outliers, *Shankya A,* **25**, 407-426.

D. E. HERWINDIATI: PhD Student at Department of Mathematics, ITB,
Bandung, Indonesia
E-mail: d_erny@dns.math.itb.ac.id

M.A. DJAUHARI: Department of Mathematics, ITB,
Bandung, Indonesia

S. DARWIS: Department of Mathematics, ITB,
Bandung, Indonesia
E-mail: sdarwis@dns.math.itb.ac.id

Life Insurance with Stochastic Interest Rate

Lienda Noviyanti[1,2], Muhammad Syamsuddin[1]

[1)] Department of Mathematics, Institut Teknologi Bandung, Indonesia
[2)] Department of Statistics, Universitas Padjadjaran, Indonesia

Abstract: Pricing of insurance product is usually evaluated on a basis where interest rate is assumed to be fixed over time. To obtain a more realistic assessment of the pricing of its product it would be benefit if the interest rates are fluctuating. This paper compares actuarial quantity which calculated based on fixed interest rate to stochastic interest rate using the Vasicek model. In this case, time to maturity in financial valuation models is adjusted with *T(x),* a continuous random variable representing future lifetime of a life-aged-*x*. *T(x)* is obtained through simulation based on Gompertz Mortality Law. By means of Monte Carlo simulation we calculate actuarial quantity under different life insurance products such as whole life, term and endowment, and give empirical results. Furthermore, we perform sensitivity analysis with respect to parameters of **insurance contract.**

Keywords: life insurance pricing, interest rate derivatives, Gompertz mortality law, Monte Carlo simulation.

References:

[1] Brigo, D. and Mercurio, F., 2001, *Interest Rate Models, Theory and Practice*, Springer-Verlag, Germany.
[2] Bowers, N.L., Gerber, H., Hickman, J., Jones, D. and Nesbitt, C., 1997, *Actuarial Mathematics*, 2nd ed., Itasca, Illinois, The Society of Actuaries.
[3] Devroye, L., 1986, *Non-Uniform Random Variate Generation*, Springer-Verlag.
[4] Lamberton, D., and Lapeyre, B., 2000, *Introduction to Stochastic Calculus, Applied to Finance*, Chapman & Hall, UK.
[5] Mao, H., et.al., 2004, *Pricing Life Insurance: Combining Economic, Financial and Actuarial Approaches*, Journal of Insurance Issues, 27, 2, pp. XXX-XXX.
[6] Pai, Jeffrey S., 1997, *Generating Random Variates with a Given Force of Mortality and Finding a Suitable Force of Mortality by Theoretical Quantile-Quantile Plots*, Actuarial Research Clearing House, vol. 1, 293-312.
[7] Ross, S., 1997, Simulation, Harcourt / Academic Press, USA. Stampfli, J. and Goodman, V., 2001, *The Mathematics of Finance: Modeling and Hedging*, Brooks / Cole, Thomson Learning, USA.

University Practice of Modelling at Universitas Brawijaya Malang

Agus Suryanto

Department of Mathematics, Universitas Brawijaya Malang, Indonesia

Abstract: Mathematical modeling is one of compulsory courses in the Department of Mathematics, Brawijaya University. I have been involved in the team teaching this subject for the past two years. Here, I will present and discuss our experiences. This includes how we teach, what the contents (materials and examples) are, what we have learned from this experience, etc. Critical remarks and suggestions to improve this teaching are certainly expected and will be welcomed.

University Practice of Modelling at Technical University Eindhoven

S.W. Rienstra

Technical University Eindhoven, The Netherlands

Abstract: The most important role of mathematics as a puzzle-solving tool of applied logic is justly considered as the most difficult one, but its other role as a language to describe the real world and to pose real problems is not trivial, and needs a substantial amount of training.

At TU/e this training starts right in the first year. During the first 2 years a series of 4 basic *modelling* courses are given, consisting of 2 assignments per course, per group of 2 students, on a practical problem taken from the background of, and supervised by a member of the staff. These problems vary rather widely to cover the 3 main streams of the department: applied analysis (in physics and engineering), statistics and probability, and coding, discrete optimisation and cryptography.

In applied analysis, there is a later course on *modelling and perturbation methods*. Here a typical aspect of many models of a physical origin, viz. the hierarchy of large-smaller-small effects, is exploited by means of asymptotic methods. This course aims at an analytical approach, and shows that the inherent limited accuracy of a model is not a drawback but a source of inspiration.

University Practice of Modelling at Universitas Padjadjaran

Asep Supriatna

Department of Mathematics, Universitas Padjadjaran, Bandung, Indonesia

Abstract: In this talk I will present my experiences in teaching mathematical modelling in the Department of Mathematics, Padjadjaran University. The talk will include the materials that have been taught, the way the materials were presented, and the way the testings were done. In general, there are many aspects, both materially and methodologically, that need some improvement in teaching mathematical modelling in the department. I hope that our experience will contribute to the discussion, from which -in return- we can learn some ways to improve our teaching.

Indian Ocean Tsunami on December 26, 2004 Recorded by Indonesia Sea Level Monitoring Network

Parluhutan Manurung

National Coordinating Agency for Surveys and Mapping (BAKOSURTANAL), Indonesia

Abstract: The magnitude 9.3 earthquake centered of the west coast of Aceh Sumatra on December 26, 2004 generated a series of ocean-wide tsunami waves that devastated coastal areas throughout the Indian Ocean and this unprecedented tsunami had killed hundreds of thousands of people in several countries bordering the Indian Ocean.

Sea level monitoring stations operated on behalf of BAKOSURTANAL recorded the wave form at a number of ports along the coastal islands facing the Indian Ocean. The stations are part of national permanent sea level monitoring network consisting of 60 stations located in the main ports of Indonesian Archipelago. The early implementation was started in stages by BAKOSURTANAL in 1984 with eight stations. It was initially assigned for survey and mapping purposes such as providing mean sea level height for the national height system in the main islands and chart datum for bathymetric mapping. A significant increase of 25 stations was carried out in 1998 to support for the bathymetric mapping of exclusive economic zone and sea line passages in Indonesian waters.

The sea level stations are not currently part of a tsunami-warning network. However, lessoned learned from the tsunami in Indian Ocean, the need for real time sea level monitoring to support a warning system in the Indian Ocean region including Indonesia internal waters is increasing. In anticipation of the development of the Indian Ocean Tsunami Warning System which requires capability of both fast detection of tsunami waves and uninterrupted operability in long term, it is recommended to up grade the existing stations into real time and establish some new stations. Efforts have been made to build joint cooperation amongst related institutions and local district governments as well as international bodies to have the system in place, as expected to be operational before the end of 2006.

The first real time sea level station, located in Sibolga North Sumatra, has been operating since 22 April 2005 with data transmission via Global Telecommunication System (GTS) operated under the World Meteorological Organization (WMO). It is expected that the incoming tsunami warning system would also provide windows of opportunity to enhance our national capacity to collect high resolution sea level data which can also be used for both practical applications and research on dynamic of the surrounding Indonesian Maritime Continent.

Tsunami Early Warning System in Indonesia

Fauzi

Badan Meteorologi dan Geofisika, Jakarta, Indonesia

Abstract: Lessons learned from the tsunami in Aceh 26 of December 2004, we set up the Tsunami Early Warning System for Indonesian regional that include covering Indian and Pacific regions. Now, we have a preliminary system consist of broadband seismograph network and automatic processing system that is able to disseminate earthquake information about 10 minutes after the occurrence. This is a big change in time frame of earthquake information comparing from the previous system of more than 30 minutes. However, we need time frame speed of automatic hypocenter determination 3 minutes after the earthquake and another 2 minutes to estimate whether the earthquake generates tsunami or not. The first 3 minutes of automatic processing needs a dense seismograph network and the next 2 minutes of manual processing needs a support system such as tsunami modeling and database of historical tsunami. This kind of system will be completed in a short time plan in 2006 and followed by a long term plan.

The precursor of local tsunami can be observed without any equipment with some reasonable errors since its phenomena is physically very clear to everyone who is in alert to incoming tsunami. The equipment sets is very useful to amplify the precursor and alert the people at risk in the inundation zone of tsunami to evacuate immediately. The precursor of tsunami in the far field must be supported by several sets of equipment and the communication networks.

The physical condition of the earth that immediate changes during and a few minutes after the earthquake in a short distance are:

1. Deformation near the source up to several hundreds of kilometers
2. Strong shaking with duration more than 1 minutes
3. Abrupt change of pressure of the sea water and propagate to all body
4. Many cases, low tide of sea level
5. Some cases, explosion can be heard
6. Arrival of Tsunami

The arrival of local tsunami is only 5 to 20 minutes after the earthquake, therefore we need a quick conclusion whether earthquake will generate tsunami or not. We know that seismic wave is faster than tsunami wave, therefore we use seismic wave to predict tsunami after the earthquake. Low tide of sea level may be happened in several cases and may help to decide issuing tsunami warning. Within 5 minutes or less, we should be able to issue tsunami warning and one hour after the earthquake we should be able to confirm or to cancel the tsunami based on the detection of tsunami wave using DART-buoy or tide gauge networks.

Two important decisions are to issue the tsunami warning and to which area the warning is delivered. Both decisions need pre-computed so called modeling of tsunami run up and inundation. After the hypocenter is calculated within the first 3 minutes and the next 2 minutes, we need to visualize the earthquake map and

tsunami modeling on the screen. Looking at this map, our authority will be able to point the affected region and disseminate the warning.

Data Assimilation for Large Scale Numerical Models

Arnold Heemink

Delft Institute of Applied Mathematics, TUDelft, Netherlands.

Abstract: Data assimilation methods are used to combine the results of a large scale numerical model with the measurement information available in order to obtain an optimal reconstruction of the dynamic behavior of the model state. Data assimilation problems are inverse modeling problems and most data assimilation schemes are based on solving the Euler-Lagrange equations, a two-point boundary value problem. In our contribution we introduce two new efficient data assimilation schemes.

Iterative schemes to solve the Euler-Lagrange equations require the implementation of the adjoint model. Even with the use of the adjoint compilers that have become available recently this is a tremendous programming effort that hampers new applications of the method. Therefore we propose another approach to variational data assimilation with a comparable computational efficiency but that does not require the implementation of the adjoint of the tangent linear approximation of the original model. The approach is based on model reduction. Using an ensemble of forward model simulations, an approximation of the covariance matrix of the model variability is determined. A limited number of leading eigenvectors (EOF's) of this matrix are selected to define a model sub space. By projecting the original model onto this subspace an approximate linear model is obtained. Once this reduced model is available, its adjoint can be implemented very easily and the minimization process can be solved completely in reduced space with negligible computational costs. If necessary, the procedure can be repeated by generating new ensembles more closely to the most recent estimate.

Another approach to data assimilation is to solve the Euler Lagrange equations recursively. This results in the well-known Kalman filtering algorithm. The standard filter however would impose an unacceptable computational burden for large scale systems with a state dimension of more then, say, 100 000. In order to obtain a computationally efficient filter simplifications have to be introduced. Recently many new algorithms have been proposed in literature, all of the square root type: Ensemble Kalman filter (EnKF), Reduced Rank Square Root filter (RRSQRT), SSQRT, RRTKF, SEIK. The EnKF has been used successfully in many applications. This Monte Carlo approach is based on a representation of the probability density of the state estimate by a finite number of randomly generated system states. A serious disadvantage is that the statistical error in the estimates of the mean and covariance matrix from a sample decreases very slowly for larger sample size. Another approach to solve large scale Kalman filtering problems is to approximate the full covariance matrix of the state estimate by a matrix with reduced rank. Although they are generally more efficient then the EnKF, reduced-rank approaches often suffer from filter divergence problems. Therefore we propose the Complementary Orthogonal subspace Filter For Efficient Ensembles (COFFEE) that combines the EnKF with the reduced-rank approach. This filter does not suffer from divergence problems and is generally more accurate than the EnKF.

We will first formulate the general data assimilation problem and will discuss the model reduced variational method and a number of square root filter algorithms including the new COFFEE algorithm. For a class of filter algorithms we will present a convergence theorem. The characteristics and performance of the new data assimilation schemes will be illustrated with a number of real life data assimilation applications in storm surge forecasting and emission reconstruction problems in air pollution modeling.

Maximal Temporal Amplitude and the Design of Experiments for the Generation of Extreme Waves

Andonowati

Centre for Mathematical Modelling and Simulation & Department of Mathematics, Institut Teknologi Bandung

Abstract: This lecture concerns the down-stream propagation of waves over initially still water. Such a study is relevant to generate waves of large amplitude in wave tanks of a hydrodynamic laboratory. Input in the form of a time signal is provided at the wave-maker located at one side of the wave tank; the resulting wave then propagates over initially still water towards the beach at the other side of the tank. Experiments show that nonlinear effects will deform the wave and may lead to large waves with wave heights larger than twice the original input; the deformation may show itself as peaking and splitting. It is of direct scientific interest to understand and quantify the nonlinear distortion; it is also of much practical interest to know at which location in the wave tank, the extreme position, the waves will achieve their maximum amplitude and to know the amplitude amplification factor. To investigate this, a previously introduced concept called Maximal Temporal Amplitude (MTA) is used: at each location the maximum over time of the wave elevation. An explicit expression for the MTA cannot be found in general from the governing equations and generating signal. Using a Korteweg - de Vries (KdV) model and third order approximation theory, we approximate extreme positions for two classes of waves. The classes are the wave-groups that originate from initially bi-chromatic and Benjamin-Feir (BF) type of waves, described by superposition of two or three monochromatic waves. We further illustrate the use of the MTA to design experiments based on a non-linear extension of BF signals called Soliton on Finite Background (SFB). For a given modulation length of SFB and desired maximum amplitude at a position in a towing tank, the MTA readily shows the maximum signal that is required at the wave maker and the amplitude amplification factor of the requested signal. Some examples of the generation in realistic laboratory situations will be treated. Comparison with numerical results using a fully non-linear wave generation code will be presented.

Keywords: Nonlinear distortion, Maximal Temporal Amplitude, modulation instability, bi-chromatics waves, Benjamin-Feir instability, experimental design, extreme waves

A THRESHOLD NUMBER FOR DENGUE DISEASE ENDEMICITY IN AN AGE STRUCTURED MODEL[1]

Asep K. Supriatna[a] & Edy Soewono[b]

[a] Department of Mathematics, Universitas Padjadjaran, Indonesia
[b] Department of Mathematics, ITB, Indonesia

Abstract. In this paper we present a model for dengue disease transmission with an assumption that individuals in the under-laying populations experience a monotonically non-increasing survival rate. We show that there is a threshold for the disease transmission, below which the disease will stop (endemic equilibrium is not appearing) and above which the disease will stay endemic (endemic equilibrium is appearing). We also investigate the stability of this endemic equilibrium.
Key-words: Dengue Modeling, Threshold Number, Stability of an Equilibrium Point.

1 Introduction

Reducing the number of dengue fever disease prevalence is regarded as an important public health concern in Indonesia, and in many tropical countries, since the disease is very dangerous that may lead to fatality. To find a good management in controlling the disease, ones need to understand the dynamics of the disease. Many mathematical models have been devoted to address this issue, examples are [3],[4],[5], and [6]. However, most of the authors have ignored the presence of age structure in mortality rate of the populations in their models. In this paper we present a model for dengue disease transmission with the inclusion that individuals in the under-laying populations experience a monotonically non-increasing survival rate as their age goes by. We show that there is an endemic threshold, below which the disease will stop, and above which the disease will stay endemic.

2 Host-Vector Model with Monotonic Non-Increasing Survival Rate

The model discussed here is analogous to the following age-unstructured host-vector SI model:

[1] Presented in the *International Conference on Applied Mathematics (ICAM05)* in Bandung, August 22-26, 2005. Part of the works in this paper is funded by the Indonesian Government, through the scheme of *Penelitian Hibah Bersaing XII* (SPK No. 011/P4T/DPPM/PHB/III/2004).

$$\frac{d}{dt}S_H = B_H - \beta_H S_H I_V - \mu_H S_H, \qquad \frac{d}{dt}I_H = \beta_H S_H I_V - \mu_H I_H,$$
$$\frac{d}{dt}S_V = B_V - \beta_V S_V I_H - \mu_V S_V, \qquad \frac{d}{dt}I_V = \beta_V S_V I_H - \mu_V I_V,$$

where

S_H = the number of susceptibles in the host population

S_V = the number of susceptibles in the vector population

I_H = the number of infectives in the host population

I_V = the number of infectives in the vector population

B_H = host recruitment rate; B_V = vector recruitment rate

μ_H = host death rate; μ_V = vector death rate

β_H = the transmission probability from vector to host

β_V = the transmission probability from host to vector

The model above based on the assumption that the host population N_H and the vector population N_V each are divided into two compartments, S_H and I_H for the host, and S_V and I_V for the vector.

An analogous age-structured one for the above model is made by generalizing the model in [1]. Suppose that there exists $Q_H(a)$, a function of age describing the fraction of human population who survives to the age of a or more, such that, $Q_H(0)=1$ and $Q_H(a)$ is a non-negative and monotonically non-increasing for $0 \le a \le \infty$. If it is assumed that human life expectancy is finite, then

$$\int_0^\infty Q_H(a)da = L < \infty \text{ and } \int_0^\infty aQ_H(a)da < \infty \qquad 1$$

Let $N_H = S_H + I_H$. Further, let also assume that $N_{H(0)}(t)$, $S_{H(0)}(t)$, and $I_{H(0)}(t)$ denotes, respectively, the numbers of N_H (0) who survive at time t, the numbers of S_H (0) who survive at time t, and the numbers of I_H (0) who survive at time t. Then we have

$$N_H(t) = N_{H(0)}(t) + \int_0^t B_H Q_H(a)da\,. \qquad 2$$

Since the per capita rate of infection in human population at time t is $\beta_H I_V(t)$, the number of susceptibles at time t is given by

$$S_H(t) = S_{H(0)}(t) + \int_0^t B_H Q_H(a) e^{-\int_{t-a}^t \beta_H I_V(s)ds} da\,. \qquad 3$$

See also [2]. The number of human infectives is $I_H(t) = N_H(t) - S_H(t)$, given by

$$I_H(t) = I_{H(0)}(t) + \int_0^t B_H Q(a)\left[1 - e^{-\int_{t-a}^{t} \beta_H I_V(s)ds}\right]da. \quad 4$$

It is clear that

$$\lim_{t\to\infty} N_{H(0)}(t) = 0, \lim_{t\to\infty} S_{H(0)}(t) = 0, \text{ and } \lim_{t\to\infty} I_{H(0)}(t) = 0. \quad 5$$

Analogously, we can derive similar equations for the mosquitoes, which are

$$N_V(t) = N_{V(0)}(t) + \int_0^t B_V Q_V(a)da, \quad 6$$

$$S_V(t) = S_{V(0)}(t) + \int_0^t B_V Q_V(a)e^{-\int_{t-a}^{t} \beta_V I_H(s)ds}da, \quad 7$$

$$I_V(t) = I_{V(0)}(t) + \int_0^t B_V Q_V(a)\left[1 - e^{-\int_{t-a}^{t} \beta_V I_H(s)ds}\right]da. \quad 8$$

It is also clear that

$$\lim_{t\to\infty} N_{V(0)}(t) = 0, \lim_{t\to\infty} S_{V(0)}(t) = 0, \text{ and } \lim_{t\to\infty} I_{V(0)}(t) = 0. \quad 9$$

Hence, equations (3), (4), (7), and (8) constitute an age-structured of a host-vector SI model.

3 The existence of a threshold number

In this section we will show that there is a threshold number for the model discussed above. Let us consider the limit values of equations (2) and (4). Whenever $t \to \infty$, and by considering (5) holds, the equations (2) and (4) can be written as

$$N_H(t) = \int_0^\infty B_H Q_H(a)da, \quad 10$$

$$I_H(t) = \int_0^\infty B_H Q_H(a)\left[1 - e^{-\int_{t-a}^{t} \beta_H I_V(s)ds}\right]da. \quad 11$$

Similarly, equations (6) and (8) can be written as

$$N_V(t) = \int_0^\infty B_V Q_V(a)da. \quad 12$$

$$I_V(t) = \int_0^\infty B_V Q_V(a)\left[1 - e^{-\int_{t-a}^{t} \beta_V I_H(s)ds}\right]da. \quad 13$$

Equations (10) and (12) show that the value of $N_H(t)$ and $N_V(t)$ are constants, hence the equations for the age-structured host-vector SI model reduce to two equations, (11) and (13).

The Equilibrium of the system is given by (I_H^*, I_V^*) satisfying

$$I_H^* = \int_0^\infty B_H Q_H(a)\left[1 - e^{-\beta_H I_V^* a}\right]da = F_1(I_V^*), \quad 14$$

$$I_V^* = \int_0^\infty B_V Q_V(a)\left[1 - e^{-\beta_V I_H^* a}\right]da = F_2(I_H^*). \quad 15$$

The last equations can be reduced as

$$I_H^* = F_1(F_2(I_H^*)) = \int_0^\infty B_H Q_H(a)\left(1 - e^{-\beta_H\left[\int_0^\infty B_V Q_V(a)(1-e^{-\beta_V I_H^* a})da\right]a}\right)da, \quad 16$$

Note that $F_1 \circ F_2$ is bounded. It is easy to see that $(I_H^*, I_V^*) = (0,0)$ is the *disease-free equilibrium*. To find a non-trivial equilibrium (an *endemic equilibrium*), we could observe the following

$$\frac{dF_1(F_2(I_H))}{dI_H} > 0 \quad \text{and} \quad \frac{d^2F_1(F_2(I_H))}{dI_H^2} < 0. \quad 17$$

Therefore, a unique non-trivial value of I_H^* occurs if and only if

$$\frac{dF_1 \circ F_2(0)}{dI_H} = B_H B_V \beta_H \beta_V \int_0^\infty aQ_H(a)\left(\int_0^\infty aQ_V(a)da\right)da > 1. \quad 18$$

The existence of the corresponding non-trivial value of I_V^* follows immediately. The LHS of (18) will be refereed as a threshold number R_0 of the model. We conclude that an *endemic equilibrium* $(I_H^*, I_V^*) \neq (0,0)$ occurs if and only if $R_0 > 1$.

4 The Stability of the Equilibria

To investigate the stability of the equilibria we use the method in [1] and use the lemma therein.

Lemma 1 (Brauer, 2001). Let $f(t)$ be a bounded non-negative function which satisfies an estimate of the form

$$f(t) \leq f_0(t) + \int_0^t f(t-a)R(a)da,$$

where $f_0(t)$ is a non-negative function with $\lim_{t\to\infty} f_0(t) = 0$ and $R(a)$ is a non-negative function with $\int_0^\infty R(a)da < 1$. Then $\lim_{t\to\infty} f(t) = 0$.

Proof. See [1]. It is also showed in [1] that the lemma is still true if the inequality in the lemma is replaced by

$$f(t) \leq f_0(t) + \int_0^t \sup_{t-a\leq s\leq t} f(s)R(a)da. \quad 19$$

Further, we generalize Lemma 1 using a similar argument as in [1] as follows.

Lemma 2. Let $f_j(t),\ j=1,2$ be bounded non-negative functions satisfying

$$f_1(t) \le f_{10}(t) + \int_0^t \sup_{t-a\le s\le t} f_2(s)R_1(a)da,$$

$$f_2(t) \le f_{20}(t) + \int_0^t \sup_{t-a\le s\le t} f_1(s)R_2(a)da$$

where $f_{j0}(t)$ is non-negative with $\lim_{t\to\infty} f_{j0}(t)=0$ and $R_j(a)$ is non-negative with $\int_0^\infty R_j(a)da<1.$ Then $\lim_{t\to\infty} f_j(t)=0,\ j=1,2.$

4.1 The stability of the disease-free equilibrium

We investigate the stability of the disease-free equilibrium for the case of $R_0<1.$ Consider the following inequalities.

$$1-e^{-\int_{t-a}^t \beta_H I_V(s)ds} \le \int_{t-a}^t \beta_H I_V(s)ds \le a\beta_H \sup_{t-a\le s\le t} I_V(s). \quad 20$$

$$1-e^{-\int_{t-a}^t \beta_V I_H(s)ds} \le \int_{t-a}^t \beta_V I_H(s)ds \le a\beta_V \sup_{t-a\le s\le t} I_H(s). \quad 21$$

Hence we have,

$$I_H(t) = I_{H(0)}(t) + \int_0^t B_H Q_H(a)(1-e^{-\int_{t-a}^t \beta_H I_V(s)ds})da$$

$$\le I_{H(0)}(t) + \int_0^t B_H Q_H(a)(a\beta_H \sup_{t-a\le s\le t} I_V(s))da \quad 22$$

$$I_V(t) = I_{V(0)}(t) + \int_0^t B_V Q_V(a)(1-e^{-\int_{t-a}^t \beta_V I_H(s)ds})da$$

$$\le I_{V(0)}(t) + \int_0^t B_V Q_V(a)(a\beta_V \sup_{t-a\le s\le t} I_H(s))da \quad 23$$

If further we assume that $\int_0^\infty aB_H\beta_H Q_H(a)da<1$ and $\int_0^\infty aB_V\beta_V Q_V(a)da<1,$ then using Lemma 2 we conclude that $\lim_{t\to\infty} I_H(t)=0$ and $\lim_{t\to\infty} I_V(t)=0.$ This shows that the *disease-free equilibrium* $(I_H^*, I_V^*)=(0,0)$ is globally stable.

4.2 The stability of the endemic equilibrium

The *endemic equilibrium* (I_H^*, I_V^*) appears only if $R_0>1.$ Let us see the perturbations of I_H^* and I_V^*, respectively, by $v(t)$ and $u(t).$ Define

$I_H(t) = I_H^* + v(t)$ and $I_V(t) = I_V^* + u(t)$, and substitute these quantities into equation (4) to obtain the following calculations:

$$I_H^* + v(t) = I_{H(0)}(t) + \int_0^t B_H Q_H(a)(1 - e^{-\int_{t-a}^t \beta_H [I_V^* + u(s)]ds})da$$

$$v(t) = -I_H^* + I_{H(0)}(t) + \int_0^t B_H Q_H(a)\left(1 - e^{-\int_{t-a}^t \beta_H I_V^* ds} e^{-\int_{t-a}^t \beta_H u(s)ds}\right)da$$

$$= -\int_0^\infty B_H Q_H(a)(1 - e^{-\beta_H I_V^* a})da + I_{H(0)}(t)$$

$$+ \int_0^t B_H Q_H(a)\left(1 - e^{-\beta_H I_V^* a} e^{-\int_{t-a}^t \beta_H u(s)ds}\right)da$$

$$v(t) = -\int_t^\infty B_H Q_H(a)(1 - e^{-\beta_H I_V^* a})da + I_{H(0)}(t)$$

$$- \int_0^t B_H Q_H(a)(1 - e^{-\beta_H I_V^* a})da + \int_0^t B_H Q_H(a)\left(1 - e^{-\beta_H I_V^* a} e^{-\int_{t-a}^t \beta_H u(s)ds}\right)da$$

$$v(t) = -\int_t^\infty B_H Q_H(a)(1 - e^{-\beta_H I_V^* a})da + I_{H(0)}(t)$$

$$+ \int_0^t B_H Q_H(a) e^{-\beta_H I_V^* a}\left(1 - e^{-\int_{t-a}^t \beta_H u(s)ds}\right)da$$

$$\le -\int_t^\infty B_H Q_H(a)(1 - e^{-\beta_H I_V^* a})da + I_{H(0)}(t)$$

$$+ \int_0^t B_H Q_H(a) e^{-\beta_H I_V^* a} \beta_H a \sup_{t-a \le s \le t} u(s)da$$

Hence, we have

$$|v(t)| \le \left| -\int_t^\infty B_H Q_H(a)(1 - e^{-\beta_H I_V^* a})da + I_{H(0)}(t) \right| + \int_0^t \sup_{t-a \le s \le t} |u(s)| B_H Q_H(a) e^{-\beta_H I_V^* a} \beta_H a da$$

Next define $f(t) = |v(t)|$, $f_0(t) = \left| -\int_t^\infty B_H Q_H(a)(1 - e^{-\beta_H I_V^* a})da + I_{H(0)}(t) \right|$, and $R(a) = B_H Q_H(a) e^{-\beta_H I_V^* a} \beta_H a$. It can be shown that $\int_0^\infty R(a)da < 1$. If $v(t) = u(t)$, that is, the perturbation is symmetrical, then by Lemma 1 we conclude that $\lim_{t\to\infty} v(t) = 0$. This shows that $\lim_{t\to\infty} I_H(t) = I_H^*$. The fact that $\lim_{t\to\infty} I_V(t) = I_V^*$ can be shown analogously. Hence, we conclude that the *endemic equilibrium* $(I_H^*, I_V^*) \neq (0,0)$ is globally stable if $R_0 > 1$.

5 Concluding Remarks

We found a threshold value determining the appearance of the *endemic equilibrium*, in which this equilibrium is occurring only if this threshold value is greater than one. The global stability of this equilibrium is confirmed as long as the perturbation of the equilibrium is symmetrical.

References

[1] Brauer, F. (2002). A Model for an SI Disease in an Age-Structured Population. *Discrete and Continuous Dynamical Systems – Series B.* **2**, 257-264.

[2] Diekmann, O. & J.A.P. Heesterbeek (2000). *Mathematical Epidemiology of Infectious Diseases*. John Wiley & Son. New York.

[3] Esteva, L. & C. Vargas (1998). Analysis of a Dengue Disease Transmission Model, *Math. Biosci.* **150**, 131-151.

[4] Supriatna, A.K. & E. Soewono (2003). Critical Vaccination Level for Dengue Fever Disease Transmission. *SEAMS-GMU Proceedings of International Conference 2003 on Mathematics and Its Applications*, pages 208-217.

[5] Soewono, E. & A.K. Supriatna (2001). A Two-dimensional Model for Transmission of Dengue Fever Disease. *Bull. Malay. Math. Sci. Soc.* **24,** 49-57.

[6] Soewono, E. & A.K. Supriatna. A Paradox of Vaccination Predicted by a Simple Host-Vector Epidemic Model (to appear in an Indian Journal of Mathematics).

ASEP K. SUPRIATNA: Department of Mathematics, Universitas Padjadjaran, Jl. Raya Bandung-Sumedang km 21, Sumedang 45363, Indonesia. Phone/Fax: +62 +22 7794696

EDY SOEWONO: Department of Mathematics & Center of Mathematics, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia. Phone/Fax: +62 +22 250 8126

Dynamics of Semelparous Populations

S.A. van Gils

Department of Applied Mathematics, University of Twente, The Netherlands

Abstract: A semelparous species reproduces only once in its life and dies thereafter. If there is only one opportunity for reproduction per year, and all individuals born in a certain year reproduce k years later, then the population can be divided into year-classes according to the year of birth modulo k. The dynamics is described by a, discrete time, nonlinear Leslie matrix model, where the nonlinearity enters through the density dependent fertility rate. Parameters in the model are, apart from the basic reproduction ratio, the age dependent impact on and sensitivity to the environment.

It is our ultimate goal to be able to classify, in parameter space, depending on the life cycle length k, the possible attractors with emphasis on the Single Year Class state (all but one year class are not present), Multiple Year Class patterns (with some year classes present), heteroclinic cycles and invariant tori.

When the reproductive rate is close to one, the full life-cycle-map can be approximated by a differential equation, which is of Lotka-Volterra type inheriting the cyclic symmetry that is present in the full life-cycle-map. We study the dynamics of this Lotka-Volterra system and give the complete classification of all possible attractors when k=3.

Rotifer Production in a Closed Recirculation System: Experiment, Modeling and Simulation

E. Soewono[1], G. Suantika[2], M. Malvinas[1], A.Y. Gunawan[1]

[1)] Department of Mathematics, Institut Teknologi Bandung, Indonesia
[2)] Departemen of Biology, Institut Teknologi Bandung, Indonesia

Abstract: Rotifers are microscopic organism which live in both fresh-water and sea water habitat. They are commonly used by farmers as the first feeding for fish and shrimp larvae. Rotifers are usually cultivated in a hatchery by using tanks filled with water and equipped with an air-pipe for aeration. However, it is not efficient to use such system since the production cost is high while the quality and the yield production are low. Therefore, a new process, so-called a closed recirculation system, is designed and developed. The basic principle of this system is to maintain the quality of water, i.e. the acid level of the water, for which rotifers are growing so that the period of rotifer culture becomes more longer than that in a conventional one.

In a closed recirculation system, a culture tank of 1000 liter in capacity is used and filled in by sea water of 800 liter. Initially, the population density of rotifers is around 500 individuals/ml. Twice or three times in a day, Green algae or artificial diet based on yeast is added as a nutrient of rotifers. Water-flow in the system is continuously maintained for 24 hours. The first harvest is carried out in the fifth day by taking out an amount of water from the tank and maintaining the density of the remaining rotifers in the tank around 3000 individuals/ml. Water replacement is then done in the tank immediately after the harvest. This replacement is maintained so that the volume of the tank is approximately 800 liter, as same as the initial volume of the tank. The left over waste in the tank after the harvest gives the initial start of the waste in the next one-day period before the next harvest. The replacement process is believed to give significant effect to the population growth. The next harvest is carried out once in a day for 27 days. The initial population of rotifers in every harvest period is always kept at the density of 3000 individuals/ml.

In this talk, we present a mathematical model of the rotifer production mentioned above. We shall find an optimal total production. This optimal production may depend on some parameters that control the level of production after the harvest. We shall also validate our results via the data from the experiment.

An Application of Lambert W Function on a Within-Host Dynamics of Plasmodium Falciparum Model

Hengki Tasman

Department of Mathematics, Institut Teknologi Bandung, Indonesia

Abstract: In this paper we discuss a within-host dynamics of Plasmodium falciparum proposed by Hoshen et al. using delay differential equations. We analyze this model and obtain analytical solution using Lambert W function. We also give some numerical simulations for the model.

On the Vaccination Model for Dengue Disease Transmission

Nuning Nuraini, Kuntjoro Adji S., Edy Soewono

Department of Mathematics, Institut Teknologi Bandung, Indonesia

Abstract: A SIR model for dengue disease transmission is discussed here. It is assumed that two viruses namely strain 1 and strain 2 cause the disease and long lasting immunity from infection caused by one virus may not valid with respect to a secondary infection by the other virus. We introduce an implementation of vaccine with three scenarios which prevents vaccinated healthy persons from catching the disease caused by any type of virus. The vaccine is planned to be administered to a portion of susceptible host and to a newborn baby. In this paper we present a method to estimate the proportion of vaccination to eliminate an infection disease of dengue transmission. The numerical simulation of this model indicates that the scenarios of vaccination delays the outbreaks as well as decreasing the number of first and second infection host for a short period of time.

Keywords: vaccination, susceptible host, threshold parameter

What Happens if a Rolling Disk Nearly Falls Flat?

J.J. Duistermaat

University of Utrecht, The Netherlands

Abstract: In this talk, a rolling disk is a body of revolution which rolls on a horizontal plane under the influence of a constant vertical gravitational force field. The body is supposed to roll without slipping on a sharp rim on the body. The sharp rim is a circle. The center of mass of the body is at the centre of the circle. The body is symmetric about the axis through the center of mass which is perpendicular to the plane of the circular rim. Because no friction is assumed, the total energy of the body is constant as a function of time, and the motion does not converge to a standstill with the disk lying flat on the plane.

This looks like an extremely simple problem of classical mechanics. Actually, it is not difficult to write down the equations of motion, which form a system of ordinary differential equations in an 8-dimensional phase space. It is equally simple to make numerical simulations. However, it turns out to require quite a deep analysis to give a complete qualitative description of what the solutions are doing. For me this is the attraction of the problem.

For a general body of revolution rolling on a horizontal plane, equations discovered in 1897 by Chaplygin lead to the conclusion that for most solutions the angle of the symmetry axis of the body with the horizontal plane is a periodic function of time. Here “most'' means: with the exception of the initial data (i.e., position and velocity coordinates) on a smooth co dimension one sub manifold.

This implies that for most initial data the disk does not fall flat, a fact observed in 1985 by Kolesnikov. A few years ago Richard Cushman proved the surprising converse, that for the initial data on a smooth co dimension one sub manifold F the disk falls flat in a finite time. At this limit time, the point of contact of the disk with the horizontal plane has a limit position on the rim, which acts as a hinge point for the disk turning to the flat position. If the initial data are not on F, but approach a point of F from one of the sides of F, then the disk nearly falls flat, but will rise up again. The point of contact first approaches the point of the rim as when in the case when the disk falls flat, and then races very quickly along the rim to a new limit point. This new limit point acts as the hinge point for the rising disk. The increase of the angle of the point of contact along the rim is equal to

$$\pm\pi\sqrt{1+\frac{mr^2}{I_1}},$$

where the sign depends on the side of F from which F is approached. Here m is the total mass of the disk, r is the radius of the rim, and I_1 is the moment of inertia of the disk about any axis through the centre of mass which lies in the plane of the rim.

Teaching Linear Algebra at University: An Experience

Sri Wahyuni

Department of Mathematics, Gadjah Mada University, Indonesia

Abstract: Linear algebra is a main mathematical subjects taught in science universities. However this teaching has always been difficult and it became an active area for research works in mathematics education in several countries. In recent years there has been much lively debate and creative discussion about improving the teaching of linear algebra.

The goal is to give a synthetic overview of the main results of these works focusing on the most recent developments. The main issues we will address concern to the epistemological specificity of linear algebra and the cognitive flexibility at stake in learning linear algebra. We will also discuss some of the experiences we had and insights we gained there which we found most new and interesting.

Some Experience in Teaching Multivariable Calculus for Sophomores at ITB

Hendra Gunawan

Department of Mathematics, Institut Teknologi Bandung, Indonesia

Abstract: In this talk I will present my experience in teaching multivariable calculus course for second year students at the Department of Mathematics ITB Bandung. In particular, I will explain how I conducted the course and how the students responded. Although I was not fully happy with the students' performance at the end of the course, there were some positive aspects that I would like to share with the audience.

The Ramsey Numbers for Copies Some Tree Versus Wheels and Complete Graph

Hasmawati[1,2], Edy Tri Baskoro[1], Hilda Assiyatun[1]

[1)] Department of Mathematics, Institut Teknologi Bandung, Indonesia
[2)] Department of Mathematics, Hasanuddin University, Makassar, Indonesia

Abstract: For given graphs G and H, the Ramsey number $R(G,H)$ is the smallest natural number n such that for every graph F of order n: either F contains G or the complement of F contains H. This paper investigates the Ramsey number $R(\cup G,H)$, where G contains tree and H are wheels W_m and complete graph K_m. We show that if n is even and $n \geq 3$, then $R(2S_n,W_4) = 3n$. Furthermore, if $n \geq 3$ and m is odd, $m \leq 2n\text{-}1$, then
$R(kS_n,W_m) = 3n\text{-}2+(k\text{-}1)\,n$,
and for arbitrary n and m, then
$$R(\bigcup_{i=1}^{k} T_{n_i},K_m) = R(T_{n_k},K_m)+\sum_{i=1}^{k-1} n_i .$$

Keywords: Ramsey numbers, wheels, tree, complete graph

EXACT SEQUENCE OF PARTIAL ISOMETRIC CROSSED PRODUCT

D. Suratman

ITB, Bandung, Indonesia

Abstract. Let G^+ be the positive cone of a totally ordered abelian group G and let α be an action of G^+ by endomorfism of C^*-algebra A and we refer to a triple (A, G^+, α) as C^*-dynamical system. We study the partial isometric crossed product of C^*-dynamical system (B_{G^+}, G^+, τ), as recently studied by Lindiarni and Raeburn. In this paper we discuss some short exact sequences of partial isometric crossed product $B_{G^+} \times_\tau G^+$.
Key-words: Group, isometric partial, exact sequence, crossed product.

1 Introduction

Suppose A, B and C are C^*-algebras. The sequence

$$0 \longrightarrow A \xrightarrow{\alpha} B \xrightarrow{\beta} C \longrightarrow 0.$$

is a short exact sequence if α is injective, β is surjective and $\alpha(A) = \ker \beta$. Equivalently, C^*-algebra C is the homomorphic image of C^*-algebra C under β and $\alpha(A) = \ker \beta$. In this exact sequence, a homomorphism β induces an isomorphism of $B/\alpha(A)$ onto C.

Suppose G is a totally ordered abelian group with positive cone G^+. Murphy has shown in [3] that C^*-algebra of continous functions of dual group $\widehat{G}$, $C(\widehat{G})$, is the homomorpic image of the Toeplitz algebra $\mathcal{T}(G)$. Here we show that C^*-algebra $C(\widehat{G})$ is the homomorpic image of the partial isometric crossed product $B_{G^+} \times_{\boldsymbol{\tau}} G^+$.

We begin with our discussion of Toeplitz algebra of a totally ordered abelian group G with positive cone G^+ and review the previous results ([3], [4], [1]). In section 3, we focus on partial isometric crossed product $B_{G^+} \times_{\boldsymbol{\tau}} G^+$ and use Theorem 5.6 [2] to prove our main result.

2 Toeplitz Algebras

Suppose G is a totally ordered abelian group with positive cone G^+. Let $\{e_r : r \in G^+\}$ be the usual basis of $\ell^2(G^+)$. The Toeplitz operator T_t on $\ell^2(G^+)$ characterised by $T_t(e_r) = e_{r+t}$ are isometries and satisfy $T_t T_s = T_{t+s}$ for every $t, s \in G^+$. The Toeplitz algebra $\mathcal{T}(G)$ is a C^*-subalgebra of $B(\ell^2(G^+))$ generated by $\{T_t : t \in G^+\}$.

Murphy showed in Theorem 3.14 [4] that $\mathcal{T}(G)$ is the universal C^*-algebra generated by semigroup of isometries, in sense that if B is a C^*-algebra generated by such a semigroup $\{V_t : t \in G^+\}$, then there is unique homomorphism $\psi : \mathcal{T}(G) \longrightarrow B$ such that $\psi(T_t) = V_t$.

The commutator ideal $\mathcal{C}_G$ of Toeplitz algebra $\mathcal{T}(G)$ is an ideal of $\mathcal{T}(G)$ spanned by $\{T_r(1 - T_sT_s^*)T_t^* : r, s, t \in G^+\}$ ([2], Lemma 2.4). The next theorem was proved by Murphy ([3], Theorem 3.7).

Theorem 1 *There is exist a short exact sequence*

$$0 \longrightarrow \mathcal{C}_G \longrightarrow \mathcal{T}(G) \xrightarrow{\boldsymbol{\psi}_{\mathbf{T}}} C(\widehat{G}) \longrightarrow 0.$$

where $\boldsymbol{\psi}_{\mathbf{T}}(T_s) = \epsilon_s, \forall s \in G^+$.

The above theorem say that algebra of continu function of dual group $C(\widehat{G})$ is a homomorphic image of Toeplitz algebra $\mathcal{T}(G)$ under $\boldsymbol{\psi}_{\mathbf{T}}$ and $\ker \boldsymbol{\psi}_{\mathbf{T}}$ is the commutator ideal of $\mathcal{T}(G)$.

Next for $\boldsymbol{\beta} \in$ Aut $C(\widehat{G})$ with $\boldsymbol{\beta}(\epsilon_t) = \epsilon_{-t}$, then $\boldsymbol{\psi}_{\mathbf{T}^*} := \boldsymbol{\beta} \circ \boldsymbol{\psi}_{\mathbf{T}}$ is a homomorphism from $\mathcal{T}(G)$ into $C(\widehat{G})$ and we have corollary as follows.

Corollary 2 *There is a short exact sequence*

$$0 \longrightarrow \mathcal{C}_G \longrightarrow \mathcal{T}(G) \xrightarrow{\boldsymbol{\psi}_{\mathbf{T}^*}} C(\widehat{G}) \longrightarrow 0.$$

where $\boldsymbol{\psi}_{\mathbf{T}^*}(T_s) = \epsilon_{-s}, \forall s \in G^+$.

Proof. Since $C(\widehat{G})$ generated by $\{\epsilon_{-s} : s \in G\}$ then homomorphism $\boldsymbol{\psi}_{\mathbf{T}^*}$ is surjective. ■

Now, we discuss a C^*-algebra related with Toeplitz algebra.
Suppose G^+ be a positive cone of a totally ordered abelian group G. A C^*-dynamical system (A, G^+, α) consists of a C^*-algebra A, a positive cone G^+ and an action α of G^+ to endomorphism of A. A *covariant isometric representation* of (A, G^+, α) on Hilbert space H is a pair (π, V) consisting of a nondegenerate representation π of A on H and a partial isometric representation V of G^+ on H such that

$$\pi(\alpha_t(a)) = V_t\pi(a)V_t, \text{ for every } t \in G^+ \text{ and } a \in A.$$

The *isometric crossed product* of (A, G^+, α) is the C^*-algebra $A \times_\alpha^{\text{iso}} G^+$ generated by a universal covariant partial isometric representation (i_A, i_{G^+}) where for every covariant isometric representation (π, V) on H, there is a nondegenerate representation $\pi \times V$ of $A \times_\alpha^{\text{iso}} G^+$ on H such that $(\pi \times V) \circ i_A = \pi$ and $(\pi \times V) \circ i_{G^+} = W$.

Stacey showed in ([5], Proposition 3.2) that the isometric crossed product $A\times_{\alpha}^{\rm iso}G^+$ exist if (A, G^+, α) has a non-trivial covariant isometric representation.

Adji et al [1] studied isometric crossed product of special C^*-dynamical system $(B_{G^+}, G^+, \boldsymbol{\tau})$. Here, B_{G^+} is C^*-subalgebra of $\ell^\infty(G^+)$ spanned by the characteristic functions $\{1_t : t \in G^+\}$, where

$$\mathbf{1}_t(r) = \begin{cases} 1 \text{ if } r \geq t \\ 0 \text{ else.} \end{cases}$$

for every $r \in G^+$. The action τ defined by

$$\begin{array}{rccc} \tau & : & G^+ & \longrightarrow \text{ End } B_{G^+} \\ & & t & \longmapsto \tau_t \end{array}$$

where

$$\tau_t(f(r)) = \begin{cases} f(r-t) \text{ if } r \geq t \\ 0 \text{ else.} \end{cases}$$

For $t, s, r \in G^+$,

$$\begin{aligned} \tau_t(\mathbf{1}_s(r)) &= \begin{cases} \mathbf{1}_s(r-t) \text{ if } r \geq t \\ 0 \text{ else} \end{cases} \\ &= \begin{cases} 1 \text{ if } r \geq t \text{ dan } r - t \geq s \\ 0 \text{ else} \end{cases} \\ &= \begin{cases} 1 \text{ if } r \geq s + t \\ 0 \text{ else} \end{cases} \\ &= \mathbf{1}_{s+t}(r), \end{aligned}$$

it follow that $\boldsymbol{\tau}_t(\mathbf{1}_s) = \mathbf{1}_{s+t}$.

In ([1], Theorem 2.4), Adji et al showed that the repesentation of $B_{G^+} \times_{\boldsymbol{\tau}}^{\rm iso} G^+$ is faithful whenever the isometric representation of G^+ is non-uniter . So that the Toeplitz algebra $\mathcal{T}(G)$ is a faithful realization of the isometric crossed product $B_{G^+} \times_{\boldsymbol{\tau}}^{\rm iso} G^+$.

3 Partial Isometric Crossed Product

Suppose G^+ be a positive cone of a totally ordered abelian group G. An *parsial isometric representation* W of G^+ on a Hilbert space H is a map $W : G^+ \longrightarrow B(H)$ such that $W(t) := W_t$ is partial isometry and satisfy $W_tW_s = W_{t+s}$ for every $t, s \in G^+$. A *covariant partial isometric representation* of (A, G^+, α) on H is a pair (π, W) consisting of a nondegenerate representation π of A on H and a partial isometric representation W of G^+ on H such that

$\pi(\alpha_t(a)) = W_t\pi(a)W_t$ and $W_t^*W_t\pi(a) = \pi(a)W_t^*W_t$ for every $t \in G^+$ and $a \in A$.

The *partial isometric crossed product* of (A, G^+, α) is the C^*-algebra $A \times_\alpha G^+$ generated by a universal covariant partial isometric representation (i_A, i_{G^+}) where for every covariant partial isometric representation (π, W) on H, there is a non-degenerate representation $\pi \times W$ of $A \times_\alpha G^+$ on H such that $(\pi \times W) \circ i_A = \pi$ and $(\pi \times W) \circ i_{G^+} = W$.

The partial isometric crossed product $A \times_\alpha G^+$ exist and unique up to isomorphism ([2], Proposition 4.7).

Now, consider C^*-dynamical system (B_{G^+}, G^+, τ) stated in §2. Lindiarni and Raeburn [2] showed that partial isometric crossed product $B_{G^+} \times_\tau G^+$ is the universal C^*-algebra generated by a partial isometric representation of G^+ on a Hilbert space H. The Theorem 5.6 in [2] give the following proposition.

Proposition 3 *There is a short exact sequence*

$$0 \longrightarrow \ker \boldsymbol{\theta}_T \longrightarrow B_{G^+} \times_{\boldsymbol{\tau}} G^+ \xrightarrow{\boldsymbol{\theta}_T} \mathcal{T}(G) \longrightarrow 0.$$

Proof. Since the generators T_t of Toeplitz algebra $\mathcal{T}(G)$ are isometries and satisfy $T_t T_s = T_{t+s}$ for every $t, s \in G^+$. Proposition 5.1 in [2] give a homomorphism $\boldsymbol{\theta}_T : B_{G^+} \times_{\boldsymbol{\tau}} G^+ \longrightarrow \mathcal{T}(G)$ such that $\boldsymbol{\theta}_T(i_{G^+}(t)) = T_t$ for every $t \in G^+$. Note that Toeplitz algebra $\mathcal{T}(G)$ generated by $\{T_t : t \in G^+\}$. So $\boldsymbol{\theta}_T$ is surjective. ■

Corollary 4 *There is a short exact sequence*

$$0 \longrightarrow \boldsymbol{\theta}_{T^*} \longrightarrow B_{G^+} \times_{\boldsymbol{\tau}} G^+ \xrightarrow{\boldsymbol{\theta}_{T^*}} \mathcal{T}(G) \longrightarrow 0.$$

Proof. Since the Toeplitz algebra generated by $\{T_t^* : t \in G^+\}$ also. Thus we conclude, there is a surjective homomorphism $\boldsymbol{\theta}_{T^*}$ from $B_{G^+} \times_{\boldsymbol{\tau}} G^+$ onto $\mathcal{T}(G)$ such that $\boldsymbol{\theta}_T(i_{G^+}(t)) = T_t^*$. ■

Proposition 5 *Let $\mathcal{I} = \boldsymbol{\theta}_T \cap \boldsymbol{\theta}_{T^*}$. The map $\boldsymbol{\theta} := \boldsymbol{\psi}_{\mathbf{T}^*} \circ \boldsymbol{\theta}_{\mathbf{T}}$ give a short exact sequence*

$$0 \longrightarrow \mathcal{I} \longrightarrow B_{G^+} \times_{\boldsymbol{\tau}} G^+ \xrightarrow{\boldsymbol{\theta}} C(\widehat{G}) \longrightarrow 0.$$

Proof. The homomorphism θ is surjective because both $\boldsymbol{\psi}_{T^*}$ and $\boldsymbol{\theta}_{\mathbf{T}}$ are surjektive. Now we will show that $\ker \boldsymbol{\theta} = \mathcal{I}$, let $a \in \mathcal{I}$. Since

$$\boldsymbol{\theta}(a) = \boldsymbol{\psi}_{T^*} \circ \boldsymbol{\theta}_{\mathbf{T}}(a) = \boldsymbol{\psi}_{T^*}(0) = 0,$$

then $a \in \ker \boldsymbol{\theta}$, it follws that $\mathcal{I} \subseteq \ker \boldsymbol{\theta}$. On the other hand, let $b \in \ker \boldsymbol{\theta}$. Then

$$0 = \boldsymbol{\theta}(b) = (\boldsymbol{\psi}_{T^*} \circ \boldsymbol{\theta}_{\mathbf{T}})b = \boldsymbol{\psi}_{T^*}(\boldsymbol{\theta}_{\mathbf{T}}(b)).$$

Its means $\boldsymbol{\theta}_{\mathbf{T}}(b) \in \ker \boldsymbol{\psi}_{T^*} = \mathcal{C}_G$. Note that $T_r(1 - T_u T_u^*)T_t^*$ have form $\boldsymbol{\theta}_T(\mathbf{i}_{G^+}(r)(1 - \mathbf{i}_{G^+}(u) i G^+(u)^*)\mathbf{i}_{G^+}(t)^*)$ in $\boldsymbol{\theta}_T(B_{G^+} \times_{\boldsymbol{\tau}} G^+)$. So

$$\boldsymbol{\theta}_{T^*}(\mathbf{i}_{G^+}(r)(1 - \mathbf{i}_{G^+}(u)\mathbf{i}_{G^+}(u)^*)\mathbf{i}_{G^+}(t)^*) = T_r^*(1 - T_u^* T_u)T_t = 0.$$

From there we conclude $\mathbf{i}_{G^+}(r)(1-\mathbf{i}_{G^+}(u)iG^+(u)^*)\mathbf{i}_{G^+}(t)^* \in \ker \boldsymbol{\theta}_{T^*}$. Since $T_r(1-T_uT_u^*)T_t^*$ span $\mathcal{C}_G$, then we have $b \in \ker \boldsymbol{\theta}_{T^*}$. Furthermore, since $\boldsymbol{\psi}_{T^*} \circ \boldsymbol{\theta}_{\mathbf{T}} = \boldsymbol{\psi}_{\mathbf{T}} \circ \boldsymbol{\theta}_{T^*}$, and $(\boldsymbol{\psi}_{\mathbf{T}} \circ \boldsymbol{\theta}_{T^*})b = \boldsymbol{\psi}_{\mathbf{T}}(\boldsymbol{\theta}_{T^*}(b)) = 0$ implies $\boldsymbol{\theta}_{T^*}(b) \in \ker \boldsymbol{\psi}_{\mathbf{T}} = \mathcal{C}_G$. With the same argumen, the fact that $b \in \ker \boldsymbol{\theta}_{\mathbf{T}}$. Thus $b \in \ker \boldsymbol{\theta}_{T^*} \cap \ker \boldsymbol{\theta}_{T^*} = \mathcal{I}$, end the proof. ■

References

[1] S. Adji, M. Laca, M. Nilsen, I. Raeburn, Crossed product by semigroup of endomorphism and the Toeplitz algebra of ordered group, Proc. Amer. Math.Soc. 122(1994), 1133–1141.

[2] J. Lindiarni dan I. Raeburn, *Partial-Isometric Crossed Products by Semigroup of Endomorphisms*, J. Operator Theory 52 (2004), 61 -87.

[3] G.J. Murphy, Ordered group and Toeplitz algebras, J. Operator Theory 18(1987), 303–326.

[4] G.J. Murphy, Toeplitz Operator and algebras, Math. Zeit, **208** (1991), 355–362.

[5] P.J. Stacey, Crossed product of C*-algebra by *-endomorphism, J. Austral. Math. Soc. Ser. A 45(1993), 204–212.

D. Suratman: Ph D student at Department of Mathematics, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia.
Department of Mathematics, Universitas Tanjungpura,
Jl. Imam Bonjol, Pontianak 78124, Indonesia.
E-mail: dede@dns.math.itb.ac.id d_suratman@yahoo.com

NEAR 2:1 RESONANCE IN CONSERVATIVE SYSTEM WITH SINGULAR PERTURBATION (Dynamics of the Energy-Preserving Part)

Fajar Adi Kusumo[a], Johan M. Tuwankotta[a],
Hendra Gunawan[a], Wono Setya Budhi[a]

[a] ITB, Bandung, Indonesia

Abstract. We consider near 2:1 resonance in four-dimensional dynamical system. The vector field of this system can be written as the sum of conservative vector field and dissipative vector field. We assume that order of magnitude of dissipative vector field is much smaller than the conservative one and we use perturbation parameter $0 < \varepsilon \ll 1$ to determine this case. For $\varepsilon = 0$, the system is harmonic oscillator. As we know, all solutions of this one except the origin is periodic solution. For $\varepsilon > 0$ this harmonic oscillator is perturbed by conservative vector field in nonlinear term and dissipative vector field in linear way. This is the reason that we say this system has singular perturbation. In this paper, we use normal form theory for studying behavior of the energy-preserving part of this system, especially behavior of solution near the nontrivial equilibrium, dynamics related to the change of detuning parameter, and bifurcation related to the change of the radius of the sphere.
Key-words: resonance, energy-preserving, conservative.

1 Introduction

Think the system :

$$\begin{aligned} \ddot{x} + \omega_1^2 x &= \varepsilon f(x, \dot{x}, y, \dot{y}) \\ \ddot{y} + \omega_2^2 y &= \varepsilon g(x, \dot{x}, y, \dot{y}) \end{aligned} \tag{1}$$

with $\omega_i > 0, i = 1, 2,\ 0 < \varepsilon \ll 1$. Function f, g are smooth function and there are coupling term in those function. If the function f and g is zero, system (1) is called decoupled. The decoupled system of system (1) is two independent harmonic oscillator with frequencies ω_1 and ω_2. If there are $k_1, k_2 \in \mathbb{N}$ with $k_1\omega_1 - k_2\omega_2 = 0$ and k_1, k_2 relative priem, this situation called $k_1 : k_2$ resonance. If $k_1 + k_2 < 5$ this resonance is called lower order resonance or strong resonance.

The interesting phenomenon in coupled oscillator system is the transfer of energy between oscillator. In strong resonance, the energy transfer is happened more dramatically then higher order ones. This phenomenon can be seen in Fatimah and Ruijgrok [5], Tondl et al. [12], and Arnold [1]. In higher order resonance, the energy exchange between oscillator is small and happened in long time scale. This phenomenon can be seen in Tuwankotta and Verhulst [13]. Because of this, higher-order resonance in a system of coupled oscillators tend to get less attention rather

then lower-order ones. In fact, as noticed in Haller [6], tradition in engineering is to neglect the effect of high order resonance in a system.

However, the result of Broer et al. [2][3] and Tuwankotta and Verhulst [14] in Hamiltonian system, Nayfeh, S.A and Nayfeh A.H. [9], Langford and Zhan [7][8], Nayfeh, AH and Malatkar [10], and Tuwankotta [15] in non-Hamiltonian system, showed that in the case of widely separated frequencies, which can be seen as an extreme type of high-order resonance, the behavior of the system is different from usual high order resonance.

The paper of Tuwankotta [15] studies a system of coupled oscillator with widely separated frequencies. System in [15] is conservative with energy preserving quadratic nonlinearity and there is singular perturbation in the system. In that paper Tuwankotta explain the existence of nontrivial equilibrium and its bifurcation.

In this paper, we consider the same equation but with different assumption. We assume that the frequencies is not widely separated and we consider strong resonance of the system, that is 2:1 resonance. The result of this paper is completing the result of Tuwankotta [15] in other side.

Another motivation to study this system comes from the applications in atmospheric research. In Crommelin [4], a model for ultra low frequency variability in the atmosphere is studied which represents a novel approach to the long time behavior of the atmosphere. In such model, Crommelin studies a ten dimensional system, that the linearized system near an equilibrium has two (among five) pairs of eigenvalues which are $\lambda_1 = 0.00272154 \pm 0.438839i$ and $\lambda_2 = 0.00168416 \pm 0.198707i$. One can see that $Im(\lambda_1)/Im(\lambda_2) = 2.20847 \approx 2$, that is strong resonance. For this reason, in this paper we choose the resonance is 2:1 resonance. In higher order resonance, Crommelin shows that the solutions collapse into nontrivial equilibrium and Tuwankotta in [15] shows the existence of this nontrivial equilibrium and studies the bifurcation on it. In this paper, we are interested to investigate the existence of nontrivial equilibrium in 2:1 resonance and dynamics of the energy-preserving part of the system.

2 Formulation of System

Consider a system of ordinary differential equation in $\mathbb{R}^4$ with $\boldsymbol{z} = (z_1, z_2, z_3, z_4)$, defined by:

$$\dot{\boldsymbol{z}} = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix} \boldsymbol{z} + \varepsilon \boldsymbol{F}(\boldsymbol{z}) \ , 0 < \varepsilon \ll 1 \tag{2}$$

where $D_j, j = 1, 2$ are 2×2 matrices with eigenvalues $\varepsilon\mu_1 \pm i$ and $\varepsilon\mu_2 \pm i(2+\delta)$, μ_1, μ_2, and δ are real number, ε is small parameter with $0 < \varepsilon \ll 1$. Parameter δ is detuning parameter with $\delta = \mathcal{O}(\varepsilon)$. We assume that μ_1 and μ_2 are bounded. Nonlinearity of this system is quadratic, homogeneous polynomial in z satisfying $z \cdot \boldsymbol{F}(\boldsymbol{z}) = 0$. Thus, the flow of the system $\dot{\boldsymbol{z}} = \boldsymbol{F}(\boldsymbol{z})$ is tangent to the sphere $z_1^2 + z_2^2 + z_3^2 + z_4^2 = R^2$, where R is the radius. If $\boldsymbol{F}(\boldsymbol{z}) = 0$, then system (2) is equivalent to the system of two oscillator with dissipations.

In this paper, we consider the strong resonance of system (2), that is 2:1 resonance. This system is similar to the system in Cromelin [4] and Tuwankotta [15], but in those papers, they consider the higher order resonance of the system. So the result of this paper is completing the result of those paper in other side.

To analyze this system, we use the averaging method to find the normal form of system (2). This can be done by applying the transformation into polar coordinate :

$$z_1 \mapsto r_1 \cos(t+\varphi_1), \quad z_2 \mapsto -r_1 \sin(t+\varphi_1),$$
$$z_3 \mapsto r_2 \cos(2t+\varphi_2), \quad z_4 \mapsto -r_2 \sin(2t+\varphi_2)$$

to (2) and then average the resulting equation of motion with respect to t over 2π. See Sanders and Verhulst [11] and Verhulst [16] for detail of the averaging method.

The averaged equation are of the form

$$\dot{r}_1 = \varepsilon G_1(r_1, r_2, 2\varphi_1 - \varphi_2), \quad \dot{r}_2 = \varepsilon G_2(r_1, r_2, 2\varphi_1 - \varphi_2)$$
$$\dot{\varphi}_1 = \varepsilon G_3(r_1, r_2, 2\varphi_1 - \varphi_2), \quad \dot{\varphi}_2 = \varepsilon G_4(r_1, r_2, 2\varphi_1 - \varphi_2)$$

where $G_j, j = 1..4$ at most quadratic function. Thus, we can reduce the dimension of this system into three dimensional system by taking $\varphi = 2\varphi_1 - \varphi_2$ and change the coordinate back into cartesius coordinate by transformation $r = r_1, x = r_2 \cos\varphi$ and $y = r_2 \sin\varphi$. We note that, the averaged equation of system (2) preserves the energy-preserving nature of the nonlinearity. Furthermore, by rotation we can choose the coordinate system such that the equation for r is of the form $\dot{r} = \varepsilon G(r, x)$.

We omit the details of the computation and just write down the reduced averaged equation (or normal form) after re-scaling by time $t \mapsto \varepsilon t$, i.e.

$$\begin{pmatrix} \dot{r} \\ \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} \mu_1 & 0 & 0 \\ 0 & \mu_2 & 0 \\ 0 & 0 & \mu_2 \end{pmatrix} \begin{pmatrix} r \\ x \\ y \end{pmatrix} + \begin{pmatrix} xr \\ \Omega(y)y - r^2 \\ -\Omega(y)x \end{pmatrix} \tag{3}$$

with $\Omega(y) = \delta + 2y$. In this normal form, small parameter ε is no longer present by time parameterization. For developing the analysis, we introduce some definition. Let $G : \mathbb{R}^3 \mapsto \mathbb{R}^3$ is a function defined by

$$G(\xi) = \begin{pmatrix} xr \\ \Omega(y)y - r^2 \\ -\Omega(y)x \end{pmatrix}$$

with $\xi = (r, x, y)^T$ and $\Omega(y) = \delta + 2y$. We also define the function $\mathcal{S} : \mathbb{R}^3 \to \mathbb{R}$ with $\mathcal{S}(\xi) = \frac{1}{2}(r^2 + x^2 + y^2)$, so $\frac{d\mathcal{S}}{dt} = 0$ along the solution of $\dot{\xi} = G(\xi)$. Furthermore, we define $S(R) = \{\xi | r^2 + x^2 + y^2 = R^2, R \geq 0\}$. $S(R)$ is the level set of $\mathcal{S} = R^2$.

3 General Invariant Structures

The normal form (3) has general invariant structures, i.e. the existence of these structures do not depend on the value of its parameters. The invariant

structures are the trivial equilibrium and the invariant manifold $r = 0$. Linearized system near the trivial equilibrium has eigenvalues $\lambda_1 = \mu_1$ dan $\lambda_{2,3} = \mu_2 \pm i\delta$. We have three cases : $\mu_1\mu_2 > 0, \mu_1\mu_2 < 0$ and $\mu_1\mu_2 = 0$.

If $\mu_1\mu_2 > 0$, along the solution of system (3), we have $\dot{\mathcal{S}} = \mu_1 r^2 + \mu_2(x^2 + y^2)$. The function $\dot{\mathcal{S}}$ is positive semi-definite if $\mu_1 > 0$ or negative semi-definite for $mu_1 < 0$. The function $\mathcal{S}$ is a globally defined Lyapunov function. As a consequence, all solutions of system (3) collapse into trivial equilibrium in positive or negative time. In this case, there is no other invariant structure apart from the trivial equilibrium and the invariant manifold $r = 0$.

For $\mu_1\mu_2 < 0$, the trivial equilibrium is unstable. If $\mu_1 > 0$, around the trivial equilibrium, system (3) has one dimensional unstable manifold and two dimensional stable manifold. This stable manifold is the invariant manifold $r = 0$. The situation is reversed in the case $\mu_1 < 0$. In this case, the dynamics of the system is not clear at the moment.

For $\mu_1\mu_2 = 0$, there are three different possibilities, that is for $\mu_1 = 0$, or $\mu_2 = 0$, or $\mu_1 = \mu_2 = 0$. For $\mu_1 = 0$ or $\mu_2 = 0$, the function $\dot{\mathcal{S}}$ is semi-definite. Moreover, all solutions of system (3) collapse into the trivial equilibrium for positive or negative time. In this paper we consider the most degenerate case, that is $\mu_1 = \mu_2 = 0$. In this case, $\dot{\mathcal{S}} = 0$ which means $S(R)$ is invariant under the flow of system (3). The trivial equilibrium is neutrally stable and the phase space is fibered by invariant sphere $S(R)$. Thus, the flow of system (3) can be reduced into a two-dimensional flow on these spheres.

The second invariant structure is the invariant manifold $r = 0$. Tuwankotta in [15] show the existence of this manifold in more general circumstances.

Remark. (Symmetry of the system) We defined two types of transformations i.e. transformation in the phase space $\Phi_i : \mathbb{R}^3 \to \mathbb{R}^3, i = 1, 2$ and in the parameter space $\Psi : \mathbb{R}^3 \to \mathbb{R}^3$. The transformation $\Phi_1(r, x, y) = (-r, x, y)$ keeps the system (3) invariant, and we can reduced the phase space to $\mathcal{D} = \{r \geq 0 | r \in \mathbb{R}\} \times \mathbb{R}^2$. Another important transformation is the combination between two types of transformation, that is $\Phi_2(r, x, y) = (r, x, -y)$ and $\Psi(\delta, \mu_1, \mu_2) = (-\delta, \mu_1, \mu_2)$. This combination keeps the system (3) invariant. Dynamics of the system for $\delta > 0$ is similar with the dynamics for $\delta < 0$ in the opposite direction of y. For this reason, in this paper we assume $\delta > 0$. Analysis in degenerate case, $\delta = 0$, is more complicated, so it is not investigated in this paper.

4 The re-scaled system

For $\mu_1 = \mu_2 = 0$, system (3) has an integral, i.e. $\mathcal{S}(\xi)$. Let ε is a small parameter with $0 < \varepsilon \ll 1$. We re-scale $\mu_1 = \varepsilon\kappa_1$ and $\mu_2 = -\varepsilon\kappa_2, \kappa_1\kappa_2 > 0$. System (3) become :

$$\begin{aligned} \dot{r} &= xr + \varepsilon\kappa_1 r \\ \dot{x} &= \Omega(y)y - r^2 - \varepsilon\kappa_2 x \\ \dot{y} &= -\Omega(y)x - \varepsilon\kappa_2 y \end{aligned} \tag{4}$$

with $\Omega(y) = \delta + 2y$. For $\kappa_1 < 0$ and $\kappa_2 < 0$, in the invariant manifold $r = 0$, all solutions run off to infinity except for the origin. This is motivate us to restrict ourselves to the case where $\kappa_1 > 0$ and $\kappa_2 > 0$. System (4) can be seen as perturbation of system (3) for $\mu_1 = \mu_2 = 0$. Solution of system (3) with $\mu_1 = \mu_2 = 0$ or *unperturb* system, life on the invariant sphere in $\mathbb{R}^3$. It is interesting to see the behavior of the unperturb system which is preserved in the perturbed system. In this paper, we will explain behavior of the unperturb system.

5 Manifold of Equilibria

Recall that we assume $\delta > 0$. For $\varepsilon = 0$ the system (4) becomes :

$$\dot{r} = xr \qquad \dot{x} = \Omega(y)y - r^2 \qquad \dot{y} = -\Omega(y)x \tag{5}$$

There are two manifold of equilibria in this system which lies in the plane $r = 0$ and in the plane $x = 0$.

5.1 A manifold of equilibria in the plane $r = 0$

At the invariant manifold $r = 0$, trivial equilibrium is a stable equilibrium. Non trivial equilibria in this plane is the line $y = -\frac{\delta}{2}$ so it is called manifold of equilibria. If we parameterize $x = x_o$, this manifold can be written as

$$(r, x, y) = (0, x_o, -\frac{\delta}{2}), x_o \in (-\infty, \infty) \tag{6}$$

The eigenvalues of system (5) linearized around (6) are

$$\lambda_1 = 0, \qquad \lambda_2 = 0, \qquad \lambda_3 = -2x_o$$

The λ_1 is the eigenvalue corresponding to the r direction of the sphere in the intersection of the manifold of equilibria (6). For $x_o > 0$, this manifold of equilibria is a stable manifold and for $x_o < 0$ it is unstable. If $x_o = 0$, all eigenvalues are zero.

5.2 A manifold of equilibria in the plane $x = 0$

The other manifold of equilibria of system (5) lies in the plane $x = 0$. The manifold is a curve defined by $r^2 - 2y^2 = \delta y$, that is a hyperbola. We parameterize $y = y_o$, then the manifold of equilibria is

$$(r, x, y) = (\sqrt{(\delta + 2y_o)y_o}, 0, y_o) \tag{7}$$

The eigenvalues of system (5) linearized around (7) are

$$\lambda_1 = 0, \qquad \lambda_{2,3} = \pm\sqrt{-12y_o^2 - 8y_o\delta - \delta^2}$$

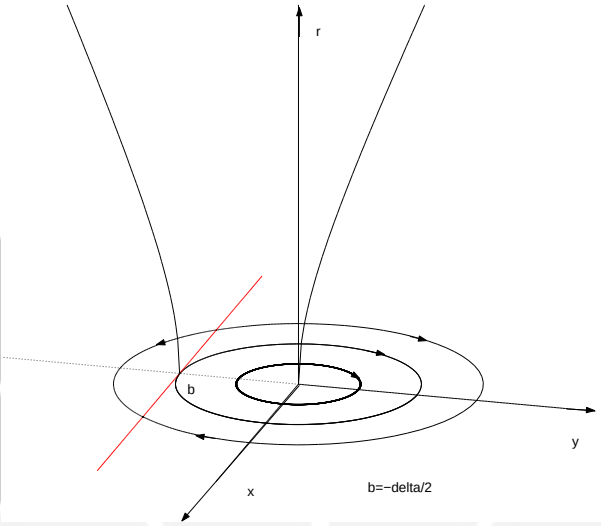


Figure 1: The limit set of the unperturb equation (5), that is the line $y = -\delta/2$ and the hyperbola.

The manifold of equilibria (7) exists for $y_o \leq -\frac{\delta}{2}$ or $y_o > 0$. For these value of y_o, two of the eigenvalues are purely imaginary and the other is zero. In Figure (1), we show the limit set of system (5). Beside that, we also show the dynamics of this system in the invariant manifold $r = 0$. In positive value of x, the manifold of equilibria $y = -\frac{\delta}{2}$ is the stable manifold, but in negative value of x, it becomes unstable.

6 Bifurcation Analysis

Since $S(R)$ is invariant under the flow of system (5), we reduce it into a two-dimensional flow of system (5). We define a bijection map that maps the orbits of system (5) to the orbits of two dimensional system defined in a disc $D(\mathbf{0}, R) = \{(x, y)|x^2 + y^2 \leq \mathbb{R}^2\}$. This map is a projection from the upper half of the sphere to the horizontal plane. The transformed system is

$$\dot{x} = \Omega(y)y - (R^2 - (x^2 + y^2)), \qquad \dot{y} = -\Omega(y)x \tag{8}$$

where $\Omega(y) = \delta + 2y$. Boundary of the disc that invariant under the flow of system (8) is called boundary *the equator*.

The bifurcation point of system (8) as we vary R is $R_o = -\frac{\delta}{2}$ and the critical points of system (8) are $(x_1, y_1) = \left(\pm\sqrt{R^2 - \frac{\delta^2}{4}}, -\frac{\delta}{2}\right)$ and $(x_2, y_2) = \left(0, \frac{-\delta \pm \sqrt{\delta^2 + 12R^2}}{6}\right)$. For $R < |\frac{\delta}{2}|$, boundary the equator is a periodic solution. It

has period

$$T = 4\int_0^R \frac{dy}{(\delta + 2y)\sqrt{R^2 - y^2}}$$

The first critical points exist only for $R \geq R_o$ but the second one exist for all value of R. For $R \geq R_o$, the second critical points lie inside the disc $D(\mathbf{0}, R)$ on the y-axes, but for $R < R_o$, one of the second critical point lies outside that disc. The complete pictures of this situation can be seen in Figure (2). In this figure we show the dynamics of system (5) for $R < R_o$, $R = R_o$, and $R > R_o$.

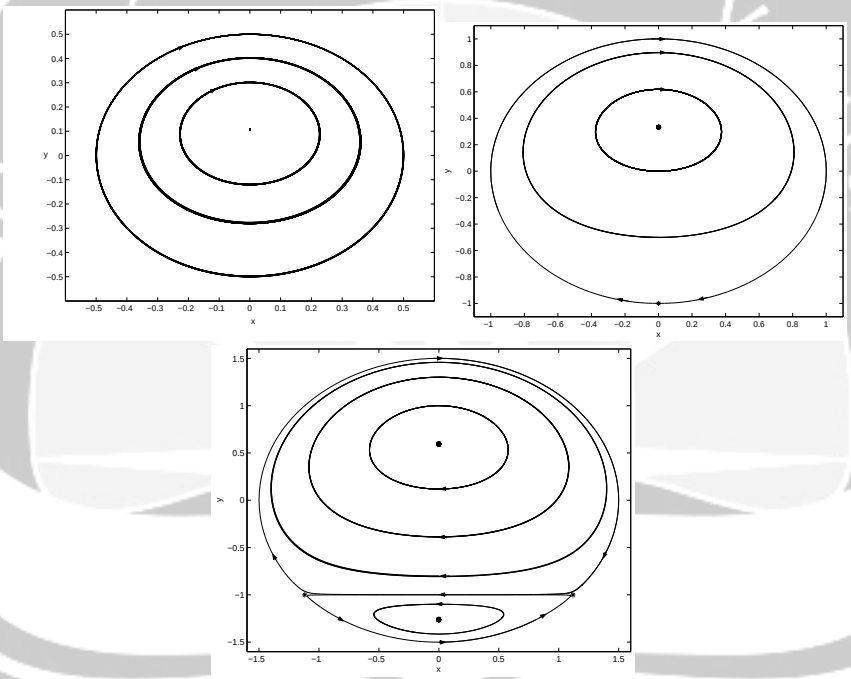


Figure 2: Phase portrait of system (5) as $R \to \infty$. The first picture (upper left) picture is the phase portrait for $R < R_o$, the second one (upper right) is for $R = R_o$, and the third one is for $R > R_o$. In this figure, we use $\delta = 2$.

7 Concluding Remarks

In this paper, we discuss the dynamics of a four dimensional system of the coupled oscillators in 2:1 resonance. In combination with energy-preserving nonlinearity, it makes the solution of the system with $\mu_i = 0, i = 1, 2$ lies in the sphere in $\mathbb{R}^4$. We use the normal form theory to analyze the system and we have completed analysis for the energy preserving part of the normal form.

References

[1] Arnol'd,V.I., *Mathematical Methods of Classical Mechanics*, Springer-Verlag, York etc., 1978

[2] Broer,H.W., Chow,S.N., Kim,Y.,Vegter,G., *A Normally Elliptic Hamiltonian Bifurcation*, ZAMP 44 (1993) 389-432

[3] Broer,H.W., Chow,S.N., Kim,Y.,Vegter,G., *The Hamiltonian System double-zero eigenvalue*, Field Inst. Commun. 4 (1995) 1-19

[4] Crommelin,D.T., *Homoclinic Dynamics : A Scenario for Atmospheric Ultralow-Frequency Variability*, J.Atmos. Sci. 59(9)(2002) 1533-1549

[5] Fatimah,S. and Ruijgrok,M., *Bifurcation in Autoparametric System in 1:1 Internal Resonance with Parametric Excitation*, Int. J. Non-Linear Mechanics, 37(2)(2002) 297-308

[6] Haller,G., *Chaos Near Resonance*, in: Applied Mathematical Sciences, vol 138, Springer, New York, 1999

[7] Langford, W.F., Zhan,K., *Interaction of Andronov-Hopf and Bogdanov-Tekens Bifurcation*, Field Inst. Commun. 24(1999) 365-383

[8] Langford, W.F., Zhan,K., *Hopf Bifurcation Near 0:1 Resonance*, in : C.Chen, Li (Eds), Proceedings of BTNA'98, Springer, New York, 1999, pp. 1-18

[9] Nayfeh,S.A., Nayfeh,A.H.,*Nonlinear Interaction between Two Widely Spaced Modes-external Excitation*, Int. J. Bifurcat. Chaos 3 (1993) 417-427

[10] Nayfeh,A.H., Malatkar,P.,*On the Transfer of Energy between Widely Spaced Modes in Structures*, Nonlinear Dynamics 31 : 225-242, 2003

[11] Sanders,J.A., Verhulst,F., *Averaging Methods in Nonlinear Dynamical System*, Appl. Math. Sciences 59, Springer-Verlag Inc., New York, 1985

[12] Tondl,A., Ruijgrok,M., Verhulst,F., and Nabergoj,R., *Autoparametric Resonance in Mechanical Systems*, Cambridge University Press, New York, 2000

[13] Tuwankotta,J.M., Verhulst,F., *Symmetry and Resonance in Hamiltonian Systems*, SIAM Journal on Appl. Math., vol. 61, number 4, 1369-1385 (2000)

[14] Tuwankotta,J.M., Verhulst,F.,*Hamiltonian System with Widely Separated Frequencies*, Nonlinearity 16 (2) (2003), p.689-706

[15] Tuwankotta,J.M., *Widely Separated Frequencies in Coupled Oscillators with Energy-preserving Quadratic Nonlinearity*, Physica D 182, 2003, p.125-149

[16] Verhulst,F., *Nonlinear Differential Equation and Dynamical System*, Second Edition, Springer-Verlag Inc., Berlin, 1996

Fajar Adi Kusumo: Ph.D student at Department of Mathematics, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia.
Phone: +62 +22 250 2545 room. 303
On leave from Department Mathematics, Gadjah Mada University, Indonesia, Sekip Utara Yogyakarta 55281, Indonesia.
E-mail: fajar_ak@math.itb.ac.id, f_adikusumo@ugm.ac.id

Johan M. Tuwankotta: Department of Mathematics, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia.
Phone: +62 +22 250 2545
E-mail: theo@dns.math.itb.ac.id

Hendra Gunawan: Department of Mathematics, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia.
Phone: +62 +22 250 2545
E-mail: hgunawan@dns.math.itb.ac.id

Wono Setya Budhi: Department of Mathematics, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia.
Phone: +62 +22 250 2545
E-mail: wono@dns.math.itb.ac.id

A REPRESENTATION THEOREM FOR THE SPACE OF MCSHANE INTEGRABLE FUNCTIONS DEFINED ON THE EUCLIDEAN SPACE $\mathcal{R}^n$

Riyadi[a], Soeparna Darmawijaya[b], Sri Daru Unoningsih[b], Widodo[b]

[a] Universitas Sebelas Maret, Surakarta, Indonesia
[b] Universitas Gadjah Mada, Yogyakarta, Indonesia

Abstract. This paper contains a discussion on a representation theorem for an orthogonally additive functional in the space of all McShane integrable functions defined on the Euclidean space $\mathcal{R}^n$. It can be considered as a generalization of the result of Chew Tuan Seng in the real line case.
Key-words: Orthogonally additive functional, McShane integrable function

1 Introduction

A functional $\mathcal{T}$ on a function space X is said to be orthogonally additive if $\mathcal{T}(f+g) = \mathcal{T}(f) + \mathcal{T}(g)$ whenever $f, g \in X$, and $f \perp g$, i.e., f and g have almost disjoint supports. Support of $f \in X$, written by $supp(f)$ is defined by $supp(f) = \{x \in dom(f) : f(x) \neq 0\}$. Based on this definition, it is clear that two functions f and g have disjoint supports if $f(x)g(x) = 0$ almost everywhere in the domain.

Let $\mathcal{R}^n$ denote the $n-$dimensional Euclidean space, $\bar{a}, \bar{b} \in \mathcal{R}^n$ and $\bar{a} \leq \bar{b}$ where $\bar{a} = (a_1, a_2, \cdots, a_n)$ and $\bar{b} = (b_1, b_2, \cdots, b_n)$. An interval $E = [\bar{a}, \bar{b}] \subset \mathcal{R}^n$, is the set of points $\{\bar{x} = (x_1, x_2, \cdots, x_n) \in \mathcal{R}^n : a_i \leq x_i \leq b_i, i = 1, 2, \cdots, n\}$. An interval $E = [\bar{a}, \bar{b}] \subset \mathcal{R}^n$ is called non degenerate whenever $\bar{a} < \bar{b}$, otherwise it is called degenerate interval. A non degenerate interval $E = [\bar{a}, \bar{b}]$ is called a cell. The non negative number:

$$\|\bar{b}\| = \max\{|\bar{b_i}| : 1 \leq i \leq n\}$$

is called the norm of $\bar{b} \in \mathcal{R}^n$.

Let $\mathcal{I}(E)$ denote the collection of all subintervals in a cell E. A set function $\mathcal{F} : \mathcal{I}(E) \to \mathcal{R}$ is said to be *additive* if $\mathcal{F}(E_1 \cup E_2) = \mathcal{F}(E_1) + \mathcal{F}(E_2)$ whenever $E_1, E_2 \in \mathcal{I}(E)$ and $E_1^o \cap E_2^o = \emptyset$. In addition, we shall always assume that $\mathcal{F}(I) = 0$ whenever I is a degenerate interval. A nonnegative additive set function α on $\mathcal{I}(\mathcal{R}^n)$ is called *a volume*. If $A \in \mathcal{I}(\mathcal{R}^n)$, then the number $\alpha(A)$ is called the $\alpha - volume$ of A.

Let $\mathcal{O}$ denote the collection of all open intervals in $\mathcal{R}^n$, then *outer measure* of an arbitrarily set $E \subset \mathcal{R}^n$ is a non negative-extended real number:

$$\mu_\alpha^*(E) = \inf\{\sum_{i=1}^{\infty} \alpha(I_i) : I_i \in \mathcal{O} \text{ for each } i \text{ such that } E \subset \bigcup_{i=1}^{\infty} I_i\}$$

whenever the infimum exist. Based on the outer measure, then we define a measurable set as follows. A set $E \subset \mathcal{R}^n$ is said to be $\mu_\alpha^* - measurable$ if for every $A \subset \mathcal{R}^n$, we have:

$$\mu_\alpha^*(A) = \mu_\alpha^*(A \cap E) + \mu_\alpha^*(A \cap E^c)$$

As usual, the collection of all $\mu_\alpha^* -$ measurable sets in $\mathcal{R}^n$ is a $\sigma -$ algebra of sets on $\mathcal{R}^n$, it is denoted by m and hence $(\mathcal{R}^n, m)$ is a measurable space. Further, the function $\mu_\alpha : m \to \bar{\mathcal{R}}$ defined by $\mu_\alpha(E) = \mu_\alpha^*(E)$ for every $E \in m$ is a measure and then $(\mathcal{R}^n, m, \mu_\alpha)$ is a measurable space.

Let $E = [\bar{a}, \bar{b}] \subset \mathcal{R}^n$ be a cell and $\mathcal{R}$ denote the real number system. A function $s : E \to \mathcal{R}$ is called *a simple function* if there exist $c_1, c_2, \cdots, c_n \in \mathcal{R}$ and $A_1, A_2, \cdots, A_n$ measurable subsets of E with $A_i \cap A_j = \emptyset$ whenever $i \neq j$ such that

$$s = \sum_{i=1}^{n} c_i \chi_{A_i}$$

where χ_{A_i} is a characteristic function on A_i for every i.

Let $\delta : E \to \mathcal{R}$ be a positive function. A finite collection of point-intervals

$$\mathcal{P} = \{(\bar{x}, I)\} = \{(\bar{x}_1, I_1), (\bar{x}_2, I_2), \cdots, (\bar{x}_n, I_n)\}$$

with $I_i \subset N_{\delta(\bar{x}_i)}(\bar{x}_i)$ for every i and $E = \bigcup_{i=1}^n I_i$ is called *a Lebesgue* or *McShane* $\delta - fine\ partition$ on E.

Let α be a volume on E. A function $f : E \to \mathcal{R}$ is said to be *McShane integrable* on E if there is a number A such that for any number $\varepsilon > 0$ there is a positive function δ on E such that for any McShane $\delta -$ fine partition $\mathcal{P} = \{(\bar{x}, I)\} = \{(\bar{x}_1, I_1), (\bar{x}_2, I_2), \cdots, (\bar{x}_n, I_n)\}$ on E we have:

$$\left| A - \mathcal{P} \sum f(\bar{x})\alpha(I) \right| = \left| A - \sum_{i=1}^{n} f(\bar{x}_i)\alpha(I_i) \right| < \varepsilon$$

Because the number A is unique, we shall write:

$$A = (\mathcal{M}) \int_E f \, d\alpha$$

In this paper, $\mathcal{P}(E)$ denotes a collection of Lebesgue partitions on a cell $E = [\bar{a}, \bar{b}] \subset \mathcal{R}^n$ and $\mathcal{M}(E)$ denotes the collection of McShane integrable functions on a cell E. It is known that $\mathcal{M}(E)$ is a linear and an absolutely integral space.

Theorem 1.1 [4] *Let $E = [\bar{a}, \bar{b}] \subset \mathcal{R}^n$. If $f \in \mathcal{M}(E)$ then there exists a sequence of simple functions $\{s_n\}$ on E such that $s_n(\bar{x}) \to f(\bar{x})$ almost everywhere on E as $n \to \infty$ and we have:*

$$\lim_{n \to \infty} \int_E s_n \, d\alpha = \int_E f \, d\alpha$$

Theorem 1.2 [4] *Let $E = [\bar{a}, \bar{b}] \subset \mathcal{R}^n$ and $f_n, f : E \to \bar{\mathcal{R}}$ be functions for every n. If :*

(i) $f_n \to f$ almost everywhere on E as $n \to \infty$ and $f_n \in \mathcal{M}(E)$ for every n,

(ii) $|f_n(\bar{x})| \leq M$ almost everywhere on E for every n and a number $M \geq 0$,

then $f \in \mathcal{M}(E)$ and

$$\lim_{n\to\infty} \int_E f_n \, d\alpha = \int_E f \, d\alpha$$

2 A Banach Space $W_0(E)$

Let $\bar{b} = (b_1, b_2, \cdots, b_n) \in \mathcal{R}^n$, we shall write $\bar{b} \to \infty$, whenever $min_{1\leq i\leq n} b_i \to \infty$. Furthermore, if $E = [\bar{1}, \bar{b}] \subset \mathcal{R}^n$ with $\bar{b} > \bar{1}$, then we mean that $\alpha(E) \to \infty$ whenever $\bar{b} \to \infty$.

Let $f \in \mathcal{M}[\bar{1}, \bar{b}]$ for every $\bar{b} > \bar{1}$, that is $(\mathcal{M}) \int_{\bar{1}}^{\bar{b}} (f) d\alpha$ exists for every $\bar{b} > \bar{1}$. If $\lim_{\bar{b}\to\infty} (\mathcal{M}) \int_{\bar{1}}^{\bar{b}} (f) d\alpha$ exists, we define:

$$(\mathcal{M}) \int_{\bar{1}}^{\bar{\infty}} (f) d\alpha = \lim_{\bar{b}\to\infty} (\mathcal{M}) \int_{\bar{1}}^{\bar{b}} (f)$$

We write $f \in \mathcal{M}[\bar{1}, \bar{\infty})$ if $f \in \mathcal{M}[\bar{1}, \bar{b}]$ for every $\bar{b} > \bar{1}$ and $\lim_{\bar{b}\to\infty} (\mathcal{M}) \int_{\bar{1}}^{\bar{b}} (f) d\alpha$ exists.

Definition 2.1 *Let $E = [\bar{1}, \bar{b}] \subset \mathcal{R}^n$ with $\bar{b} > \bar{1}$ and $f : E \to \mathcal{R}$ be a McShane integrable function on E. We define a subcollection $W_0(E)$ in $\mathcal{M}[\bar{1}, \bar{\infty})$ as follows:*

$$W_0(E) = \{f \in \mathcal{M}(E) : \lim_{\alpha(E)\to\infty} \frac{1}{\alpha(E)} \int_E |f| \, d\alpha = 0\}$$

Theorem 2.2 *$W_0(E)$ is a Banach space with respect to the norm :*

$$\|f\| = sup\{\frac{1}{\alpha(E)} \int_E |f| d\alpha : E = [\bar{1}, \bar{b}] \text{ with } \bar{b} > \bar{1}\}$$

for every $f \in W_0(E)$.

Proof. It is easy to show that $W_0(E)$ is a linear space and $\|\cdot\|$ is a norm on $W_0(E)$. So, we shall only prove that $W_0(E)$ is complete. Let $\{f_n\} \subset W_0(E)$ be an arbitrary Cauchy sequence, that is for any number $\varepsilon > 0$ there exists a natural number N_1 such that if $n, m \geq N_1$ we have:

$$\begin{aligned} \|f_n - f_m\| < \frac{\varepsilon}{\alpha(E)} &\Leftrightarrow \sup\{\frac{1}{\alpha(E)} \int_E |f_m - f_n| d\alpha : E = [\bar{1}, \bar{b}] \text{ with } \bar{b} > \bar{1}\} < \frac{\varepsilon}{\alpha(E)} \\ &\Leftrightarrow \frac{1}{\alpha(E)} \int_E |f_m - f_n| d\alpha < \frac{\varepsilon}{\alpha(E)} \Leftrightarrow \int_E |f_m - f_n| d\alpha < \varepsilon \end{aligned}$$

This means that for every natural numbers m and n, there exists a number $M_{mn} \geq 0$ such that $|(f_m - f_n)(\bar{x})| \leq M_{mn}$ a.e. on E. Therefore, we can choose a natural number N_2 such that if $m, n \geq N_2$ then $M_{mn} < \varepsilon$. Consequently, if $m, n \geq N_2$ we have:

$$|(f_m - f_n)(\bar{x})| < \varepsilon$$

for any $\varepsilon > 0$. This means that the sequence $\{f_n(\bar{x})\}$ is a Cauchy sequence in $\mathcal{R}$, therefore there exists a function f which is McShane integrable on $E = [\bar{1}, \bar{b}]$ with $\bar{b} > \bar{1}$ such that $f_n \to f$ almost everywhere on E, that is for any number $\varepsilon > 0$ there exists a natural number N_3 such that if $n \geq N_3$ we have $|f_n - f| < \varepsilon$. From this inequality, we have:

$$\begin{aligned} & \int_E |f_n - f| d\alpha < \int_E \varepsilon d\alpha = \varepsilon \alpha(E) \\ \Leftrightarrow \quad & \frac{1}{\alpha(E)} \int_E |f_n - f| d\alpha < \varepsilon \\ \Leftrightarrow \quad & \sup\{\frac{1}{\alpha(E)} \int_E |f_n - f| d\alpha\} < \varepsilon \\ \Leftrightarrow \quad & \|f_n - f\| < \varepsilon \end{aligned}$$

In other word, the sequence $\{f_n\}$ is norm convergent to a function f. Furthermore, let $N_0 = \max\{N_1, N_2, N_3\}$ so if $n \geq N_0$ we have:

$$\lim_{\alpha(E)\to\infty} \frac{1}{\alpha(E)} \int_E |f| d\alpha \leq \lim_{\alpha(E)\to\infty} \frac{1}{\alpha(E)} \int_E |f - f_n| d\alpha + \lim_{\alpha(E)\to\infty} \frac{1}{\alpha(E)} \int_E |f_n| d\alpha < \varepsilon$$

for any number $\varepsilon > 0$, this means that $f \in W_0(E)$. Thus $W_0(E)$ is complete, therefore $W_0(E)$ is a Banach space. ■

3 A Representation Theorem For An Orthogonally Additive Functional on $W_0(E)$ in $\mathcal{R}^n$

Let $E = [\bar{1}, \bar{b}] \subset \mathcal{R}^n$ with $\bar{b} > \bar{1}$ and $\mathcal{M}(E)$ denote the collection of McShane integrable functions on a cell E. A sequence $\{f_n\} \subset \mathcal{M}(E)$ is said to be boundedly convergent to a function $f \in \mathcal{M}(E)$, if $\{f_n\}$ converges to f pointwise almost everywhere on E and $\{f_n\}$ is uniformly bounded almost everywhere on E. A functional $\mathcal{F}$ defined on $\mathcal{M}(E)$ is said to be boundedly continuous if $\mathcal{F}(f_n) \to \mathcal{F}(f)$ as $n \to \infty$ whenever $\{f_n\}$ is boundedly convergent to f.

A function $k(\cdot, \cdot) : E \times \mathcal{R} \to \mathcal{R}$ is called a Caratheodory function if $k(\bar{x}, \cdot)$ is continuous for almost all $\bar{x} \in E$ and $k(\cdot, t)$ is measurable for every $t \in \mathcal{R}$.

Lemma 3.1 *Let $E = [\bar{1}, \bar{b}] \subset \mathcal{R}^n$ with $\bar{b} > \bar{1}$, $\mathcal{M}(E)$ denote the collection of McShane integrable functions on E and $\mathcal{F}$ be a functional defined on $\mathcal{M}(E)$. If*

$\mathcal{F}$ is boundedly continuous on $\mathcal{M}(E)$ then $\mathcal{F}(f\chi_E) \to 0$ whenever $\mu(E) \to 0$ for every $f \in \mathcal{M}(E)$.

Proof. Let any $f \in \mathcal{M}(E)$ and any sequence of subsets $\{E_n\} \subset E$ with $\mu(E_n) \to 0$ as $n \to \infty$, i.e. for any number $\varepsilon > 0$ there exist a natural number N_0 such that if $n \geq N_0$ we have $\mu(E_n) < \varepsilon$. This means that, if $n \geq N_0$, we have :

$$\mu\{\bar{x} \in E_n : |f\chi_{E_n}(\bar{x})| > \varepsilon\} \leq \mu(E_n) < \varepsilon$$

In other word, the sequence $\{f\chi_{E_n}\} \to \theta$ in measure, whenever $\mu(E_n) \to 0$. Therefore, there exists a subsequence $\{E_{n(i)}\} \subset \{E_n\}$ such that $f\chi_{E_{n(i)}} \to \theta$ a.e. on E whenever $n(i) \to \infty$. Since F is boundedly continuous, therefore we have:

$$F(f\chi_{E_{n(i)}}) \to F(\theta) = 0$$

whenever $n(i) \to \infty$. Consequently, $\mathcal{F}(f\chi_E) \to 0$ whenever $\mu(E) \to 0$ for every $f \in \mathcal{M}(E)$.

Lemma 3.2 *Let $E = [\bar{1}, \bar{b}] \subset \mathcal{R}^n$ with $\bar{b} > \bar{1}$. If $\mathcal{F}$ is an orthogonally additive and boundedly continuous Functional on $W_0(E)$, then there exists a function $k(\bar{x}, t) : E \times \mathcal{R} \to \mathcal{R}$ such that $k(\bar{x}, t)$ is McShane integrable with respect to $\bar{x}$ on E for every $t \in \mathcal{R}$ with $k(\bar{x}, 0) = \theta$ for almost all $\bar{x} \in E$ and we have:*

$$\mathcal{F}(s) = \int_E k(\cdot, s(\cdot))d\alpha$$

for every simple function s on E.

Proof. Since $\mathcal{F}$ is boundedly continuous, then for any number $\varepsilon > 0$ and $t \in \mathcal{R}$ there exists a number $\delta(\varepsilon, t) > 0$ such that if $A \subset E$ with $\mu(A) < \delta(\varepsilon, t)$ then $|\mathcal{F}(t\chi_A)| < \varepsilon$. In other word, $\mathcal{F}$ as a set function is absolutely continuous with respect to μ. Furthermore, if $E = \bigcup_{i=1}^{\infty} E_i$ with $E_i^o \cap E_j^o = \phi$ for $i \neq j$, then we have :

$$\mathcal{F}(t\chi_E) = \lim_{n\to\infty} \mathcal{F}(t\chi_{\bigcup_{i=1}^n E_i}) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathcal{F}(t\chi_{E_i}) = \sum_{i=1}^{\infty} \mathcal{F}(t\chi_{E_i})$$

This means that $\mathcal{F}$ is an additive and countable set function. According to Radon-Nikodym Theorem there exists a function $k_t^*(\cdot)$ on E such that

$$\mathcal{F}(t\chi_E) = \int_E k_t^*(\cdot)d\alpha$$

for every E. Next, we define a function $k(\bar{x}, t) = k_t^*(\bar{x})$ for every $\bar{x} \in E$ and $t \in \mathcal{R}$. If $t = 0$ then $\mathcal{F}(t\chi_E) = \mathcal{F}(\theta) = 0$, therefore $\int_E k_0^* = 0$ for every E. Thus $k(\bar{x}, 0) = \theta$ for almost all $\bar{x} \in E$. Furthermore, let s be any simple function on E with $s(\bar{x}) = \sum_{i=1}^n t_i\chi_{E_i}(\bar{x})$ where E_i is a pairwise disjoint and measurable sets,

and $E = \bigcup_{i=1}^{\infty} E_i$, then we have:

$$\begin{aligned}\mathcal{F}(s) &= \mathcal{F}(\sum_{i=1}^{n} t_i\chi_{E_i}) = \sum_{i=1}^{n}\int_{E_i} k(\cdot, t_i\chi_{E_i})d\alpha \\ &= \sum_{i=1}^{n}\int_{E_i} k(\cdot, s(\cdot))d\alpha = \int_{E} k(\cdot, s(\cdot))d\alpha\end{aligned}$$

Lemma 3.3 *Let $E = [\bar{1}, \bar{b}] \subset \mathcal{R}^n$ with $\bar{b} > \bar{1}$, $\mathcal{F}$ be an orthogonally additive and boundedly continuous functional on $W_0(E)$, and $k(\bar{x}, t) : E \times \mathcal{R} \to \mathcal{R}$ be a function obtained in Lemma 3.2. If for any number $\delta > 0$ and every bounded closed interval $P = [-a, a]$ with $a > 0$ we define numbers :*

$$W(\delta; P; E) = sup\{\int_E |k(\cdot, t_1) - k(\cdot, t_2)|d\alpha : t_1, t_2 \in P \text{ and } |t_1 - t_2| < \delta\}$$

and

$$W(\delta; P) = sup\{\sum_{i=1}^{n} W(\delta; P; E) : \bigcup_{i=1}^{n} E_i = E,\ E_i^o \cap E_j^o = \phi \text{ for } i \neq j\}$$

then for every interval P, we have $lim_{\delta \to 0^+} W(\delta; P) = 0$.

Proof. If $0 < \delta_1 \leq \delta_2$ and $t_1, t_2 \in P$ with $|t_1 - t_2| < \delta_1$ then $|t_1 - t_2| < \delta_2$. This means that if $0 < \delta_1 \leq \delta_2$ then $W(\delta_1; P) \leq W(\delta_2; P)$. Thus net $\{W(\delta; P) : \delta \downarrow 0\}$ is monotonically decreasing and bounded below, and one of its lower bounds is 0. Therefore $\lim_{\delta \to 0^+} W(E; P)$ exists.
Suppose $\lim_{\delta \to 0^+} W(E; P) = \varepsilon$ for a number $\varepsilon > 0$. Based on its definition, there exists disjoint sets $E_1, E_2, \cdots, E_n$ with $\bigcup_{i=1}^{n} = E$ and numbers $t_1^i, t_2^i,\ i = 1, 2, \cdots, n$ such that $|t_1^i - t_2^i| \leq \delta,\ |t_1^i| \leq a,\ |t_2^i| \leq a$ and

$$\sum_{i=1}^{n}\int_{E_i} |k(\cdot, t_1^i) - k(\cdot, t_2^i)|d\alpha > \varepsilon$$

Next, for each $i = 1, 2, \cdots, n$, we define sets:

$$E_i^+ = \{\bar{x} \in E_i : k(\bar{x}, t_1^i) - k(\bar{x}, t_2^i \geq 0\} \text{ and } E_i^- = E_i - E_i^+$$

and also we define functions:

$$f = \sum_{i=1}^{n}(t_1^i\chi_{E_i^+} + t_2^i\chi_{E_i^-}) \text{ and } g = \sum_{i=1}^{n}(t_2^i\chi_{E_i^+} + t_1^i\chi_{E_i^-})$$

It is clear that $f, g \in W_0(E)$ and we have:

$$\begin{aligned}
\|f\| &= \sup\{\frac{1}{\alpha(E)}\int_E |f|d\alpha : E = [\bar{1}, \bar{b}] \text{ with } \bar{b} > \bar{1}\} \\
&= \sup\{\frac{1}{\alpha(E)}\int_E |\sum_{i=1}^{n}(t_1^i\chi_{E_i^+} - t_2^i\chi_{E_i^-})|d\alpha : E = [\bar{1}, \bar{b}] \text{ with } \bar{b} > \bar{1}\} \\
&\leq \sup\{\frac{1}{\alpha(E)}\int_E \sum_{i=1}^{n}(|t_1^i\chi_{E_i^+}| + |t_2^i\chi_{E_i^-}|)d\alpha : E = [\bar{1}, \bar{b}] \text{ with } \bar{b} > \bar{1}\} \\
&\leq \sup\{\frac{1}{\alpha(E)}\sum_{i=1}^{n}\int_{E_i^+} a\chi_{E_i^+}d\alpha + \frac{1}{\alpha(E)}\sum_{i=1}^{n}\int_{E_i^-} a\chi_{E_i^-}d\alpha : E = [\bar{1}, \bar{b}] \text{ with } \bar{b} > \bar{1}\} \\
&= \sup\{\frac{1}{\alpha(E)}a\sum_{i=1}^{n}\alpha(E) : E = [\bar{1}, \bar{b}] \text{ with } \bar{b} > \bar{1}\} = a
\end{aligned}$$

Similarly, we can show that $\|g\| < a$. Furthermore, we have:

$$\begin{aligned}
f - g &= \sum_{i=1}^{n}(t_1^i\chi_{E_i^+} + t_2^i\chi_{E_i^-}) - \sum_{i=1}^{n}(t_2^i\chi_{E_i^+} + t_1^i\chi_{E_i^-}) \\
&= \sum_{i=1}^{n}((t_1^i - t_2^i)\chi_{E_i^+} + (t_2^i - t_1^i)\chi_{E_i^-})
\end{aligned}$$

From the above result, we can also obtain that: $\|f - g\| \leq a$. Since $\mathcal{F}$ is boundedly continuous then $|\mathcal{F}(f) - \mathcal{F}(g)| < \varepsilon$. On the other hand, we have:

$$\begin{aligned}
&| \mathcal{F}(f) - \mathcal{F}(g) | \\
&= |\int_E k(\cdot, \sum_{i=1}^{n}(t_1^i\chi_{E_i^+} + t_2^i\chi_{E_i^-}))d\alpha - \int_E k(\cdot, \sum_{i=1}^{n}(t_2^i\chi_{E_i^+} + t_1^i\chi_{E_i^-}))d\alpha| \\
&= |\sum_{i=1}^{n}\int_{E_i^+}(k(\cdot, t_1^i\chi_{E_i^+}) - (k(\cdot, t_2^i\chi_{E_i^+}))d\alpha + \sum_{i=1}^{n}\int_{E_i^-}(k(\cdot, t_2^i\chi_{E_i^-}) - (k(\cdot, t_1^i\chi_{E_i^-}))d\alpha| \\
&= |\sum_{i=1}^{n}\int_{E_i^+}(k(\cdot, t_1^i\chi_{E_i^+}) - (k(\cdot, t_2^i\chi_{E_i^+}))d\alpha + \sum_{i=1}^{n}\int_{E_i^-} -(k(\cdot, t_1^i\chi_{E_i^-}) - (k(\cdot, t_2^i\chi_{E_i^-}))d\alpha| \\
&= |\sum_{i=1}^{n}\int_{E_i^+}|k(\cdot, t_1^i\chi_{E_i^+}) - (k(\cdot, t_2^i\chi_{E_i^+})|d\alpha + \sum_{i=1}^{n}\int_{E_i^-}|k(\cdot, t_1^i\chi_{E_i^-}) - (k(\cdot, t_2^i\chi_{E_i^-})|d\alpha \ | \\
&= \sum_{i=1}^{n}\int_{E_i^+}|k(\cdot, t_1^i) - (k(\cdot, t_2^i)|d\alpha + \sum_{i=1}^{n}\int_{E_i^-}|k(\cdot, t_1^i) - (k(\cdot, t_2^i)|d\alpha \\
&= \sum_{i=1}^{n}\int_{E_i}|k(\cdot, t_1^i) - (k(\cdot, t_2^i)|d\alpha > \varepsilon
\end{aligned}$$

Contradiction, thus the assumption is wrong, therefore the right statement is $\lim_{\delta \to 0^+} W(\delta; P) = 0$.

Lemma 3.4 *Let $E = [\bar{1}, \bar{b}] \subset \mathcal{R}^n$ with $\bar{b} > \bar{1}$. If $\mathcal{F}$ is an orthogonally additive and boundedly continuous functional on $W_0(E)$, then the function $k(\bar{x}, t) : E \times \mathcal{R} \to \mathcal{R}$, obtained in Lemma 3.2. is uniformly continuous on every bounded closed interval $P \subset \mathcal{R}$ and for every $\bar{x} \in E$.*

Proof. Let $\bar{x} \in E$ and $P \subset \mathcal{R}$ be any bounded closed interval. Then, according to Lemma 3.3. we have: $\lim_{\delta \to 0^+} W(\delta; P) = 0$, that is:

$$\begin{aligned}
0 &= \lim_{\delta \to 0^+} \sup\{\sum_{i=1}^{n} W(\delta; P; E_i) : \bigcup_{i=1}^{n} E_i, E_i \cap E_j = \phi, i \neq j\} \\
0 &= \lim_{\delta \to 0^+} \sup\{\sum_{i=1}^{n} \sup\{\int_{E_i} |k(\cdot, t_1) - k(\cdot, t_2)| d\alpha : t_1, t_2 \in P \text{ and} \\
&\quad |t_1 - t_2| < \delta\} : \bigcup_{i=1}^{n} E_i, E_i \cap E_j = \phi, i \neq j\} \\
&\geq \lim_{\delta \to 0^+} \{\sup \int_{E} |k(\cdot, t_1) - k(\cdot, t_2)| d\alpha : t_1, t_2 \in P \text{ and } |t_1 - t_2| < \delta\} \\
&\geq \lim_{\delta \to 0^+} \{\int_{E} |k(\cdot, t_1) - k(\cdot, t_2)| d\alpha : t_1, t_2 \in P \text{ and } |t_1 - t_2| < \delta\}
\end{aligned}$$

This means that, if $t_1, t_2 \in P$ with $|t_1 - t_2| < \delta$ we have :

$$\lim_{\delta \to 0^+} \int_{E} |k(\cdot, t_1) - k(\cdot, t_2)| d\alpha = 0$$

That is, if $t_1, t_2 \in P$ with $|t_1 - t_2| < \delta$, there exists a number $M_{(t_1 t_2)} \geq 0$ such that $|k(\bar{x}, t_1) - k(\bar{x}, t_2)| \leq M_{(t_1 t_2)}$ a.e. on E. Therefore, we can choose a number $\delta_0 > 0$ such that if $|t_1 - t_2| < \delta_0$ then $M_{(t_1 t_2)} < \varepsilon$. Consequently, if $t_1, t_2 \in P$ with $|t_1 - t_2| < \delta_0$ we have:

$$|k(\bar{x}, t_1) - k(\bar{x}, t_2)| < \varepsilon$$

for any $\varepsilon > 0$.
In other word, the function $k(\bar{x}, \cdot)$ is uniformly continuous on every bounded closed interval $P \subset \mathcal{R}$ and for every $\bar{x} \in E$.

Theorem 3.5 *Let $E = [\bar{1}, \bar{b}] \subset \mathcal{R}^n$ with $\bar{b} > \bar{1}$. A functional $\mathcal{F}$ is orthogonally additive and boundedly continuous on the space $W_0(E)$ if only if there exists a Caratheodory function $k(\bar{x}, t) : E \times \mathcal{R} \to \mathcal{R}$ with $k(\bar{x}, 0) = 0$ for almost all $\bar{x} \in E$ such that $k(\bar{x}, t)$ is McShane integrable with respect to $\bar{x}$ on E for every $t \in \mathcal{R}$ and we have $\mathcal{F}(f) = \int_E k(\bar{x}, f(\bar{x})) d\alpha$ for every function $f \in W_0(E)$.*

Proof. **Sufficient condition** If $f \in W_0(E)$, then according to Theorem 1.1 there exists a sequence of simple functions $\{s_n\}$ on E, such that $s_n(\bar{x}) \to f(\bar{x})$ almost everywhere on E as $n \to \infty$ and without loss generality, we can assume that $|s_n(\bar{x})| \leq |f(\bar{x})|$ for all n. According to Lemma 3.4, the function $k(\bar{x}, \cdot)$ is uniformly continuous on every bounded closed interval $P \subset \mathcal{R}$ and for each $\bar{x} \in E$, therefore :

$k(\bar{x}, s_n(\bar{x})) \to k(\bar{x}, f(\bar{x}))$ almost everywhere on E and there exists a number $M \geq 0$ such that $|k(\bar{x}, s_n(\bar{x}))| \leq M$ almost everywhere on E. Therefore, according to Dominated Convergence Theorem(Theorem 1.2) $\lim_{n\to\infty} k(\bar{x}, s_n(\bar{x}))$ is McShane integrable and we have:

$$\begin{aligned}\int_E k(\cdot, f(\cdot))d\alpha &= \lim_{n\to\infty} \int_E k(\cdot, s_n(\cdot))d\alpha \\ &= \lim_{n\to\infty} \mathcal{F}(s_n) \\ &= \mathcal{F}(\lim_{n\to\infty} s_n) \\ &= \mathcal{F}(f)\end{aligned}$$

Necessary condition Let $f, g \in W_0(E)$ such that $f \perp g$, that is $f(\bar{x})g(\bar{x}) = 0$ almost everywhere on E. Furthermore, we form sets :

$$A = \{\bar{x} \in E : g(\bar{x}) = 0\},\ B = \{\bar{x} \in E : f(\bar{x}) = 0\} \text{ and}$$

$$C = \{\bar{x} \in E : f(\bar{x}) \neq 0 \text{ and } g(\bar{x} \neq 0\}$$

then we obtain :

$$\begin{aligned}\mathcal{F}(f+g) &= \int_E k(\cdot, (f+g)(\cdot))d\alpha \\ &= \int_A k(\cdot, f(\cdot))d\alpha + \int_B k(\cdot, f(\cdot))d\alpha + \int_C k(\cdot, g(\cdot))d\alpha \\ &\quad + \int_A k(\cdot, g(\cdot))d\alpha + \int_B k(\cdot, g(\cdot))d\alpha + \int_C k(\cdot, f(\cdot))d\cdot \\ &= \int_E k(\cdot, f(\cdot))d\alpha + \int_E k(\cdot, g(\cdot))d\alpha \\ &= \mathcal{F}(f) + \mathcal{F}(g)\end{aligned}$$

Thus $\mathcal{F}$ is orthogonally additive on $W_0(E)$.

Next, let $f \in W_0(E)$ and any sequence $\{f_n\} \subset W_0(E)$ such that $\{f_n\}$ is boundedly convergent to the function f almost everywhere on E, that is there exists a number $M \geq 0$ such that $|f_n(\bar{x})| \leq M$ for every n and $f_n(\bar{x}) \to f(\bar{x})$ almost everywhere on E. Since the function $k(\bar{x}, \cdot)$ is uniformly continuous on every bounded closed interval $P \subset \mathcal{R}$ and for each $\bar{x} \in E$, then there exists a number $N \geq 0$ such that $|k(\bar{x}, f_n(\bar{x}))| \leq N$ for all n and almost everywhere on E. Therefore, according to Dominated Convergence Theorem(Theorem 1.2), we have:

$$\lim_{n\to\infty} \int_E k(\cdot, f_n(\cdot)d\alpha = \int_E k(\cdot, f(\cdot)d\alpha \ \Leftrightarrow\ \lim_{n\to\infty} \mathcal{F}(f_n) = \mathcal{F}(f)$$

In other word, the functional $\mathcal{F}$ is boundedly continuous on $W_0(E)$.

4 Summary

The main result of this paper is a representation theorem for an orthogonally additive functional in the space of all McShane integrable functions defined on the Euclidean space $\mathcal{R}^n$. It can be considered as a generalization of the result of Chew Tuan Seng in the real line case. This result may also be extended in other function spaces, such as in the space of all McShane-Pettis integrable functions defined on the Euclidean space $\mathcal{R}^n$, etc.

References

[1] T.S. Chew(1985), *Orthogonally Additive Functional*, Singapore National University, Singapore.

[2] P.Y. Lee(1989), *Lanzhou Lectures on Henstock Integration*, Wold Scientific, Singapore.

[3] L.I. Paredes(1993), *Orthogonally Additive Functionals and Superposition Operators on $W_0(\phi)$*, University of The Philippines, The Philippines.

[4] W.F. Pfeffer(1993), *The Riemann Approach to Integration*, Cambridge University Press, New York.

Riyadi: Ph D student at Department of Mathematics, Gadjah Mada University, Yogyakarta, Indonesia.
Sebelas Maret University, Surakarta, Indonesia.
E-mail: yadi_laras@yahoo.com

Soeparna Darmawijaya: Department of Mathematics, Gadjah Mada University, Yogyakarta, Indonesia.

Sri Daru Unoningsih: Department of Mathematics, Gadjah Mada University, Yogyakarta, Indonesia.

Widodo: Department of Mathematics, Gadjah Mada University, Yogyakarta, Indonesia.

GENERALIZED DIFFERENTIAL RICCATI EQUATION FOR TWO-PLAYER LINEAR QUADRATIC DYNAMIC GAME DESCRIPTOR SYSTEM

Salmah[a], S.M. Nababan[b], Bambang.S[c], S.Wahyuni[d]

[a] Universitas Gadjah Mada, Yogyakarta, Indonesia
[b] ITB, Bandung, Indonesia
[c] Universitas Gadjah Mada, Yogyakarta, Indonesia
[d] Universitas Gadjah Mada, Yogyakarta, Indonesia

Abstract. Problem of two player linear quadratic dynamic game is considered. Optimal Nash equilibrium for the problem is derived with Hamilton method. For finite time problem, finding optimal control solution is brought to finding solution of two generalized differential Riccati equation. Then the relationship of the existence of the pair of Nash equilibrium solution is studied. For infinite time problem two generalized Riccati algebra equations is considered.
Key-words: Riccati equation, linear quadratic, dynamic game, descriptor system.

1 Introduction

Descriptor systems are believed have a great application for the system modeling because they can preserve the structure of physical systems and can include nondynamic modes and impulsive modes [3].

Non-cooperative open-loop non-zero-sum continuous linear quadratic dynamic game has been studied in [2] and [2]. On non-zero-sum linear quadratic dynamic game, the two players satisfy one differential equation in the state space form, and the players minimize two objective functions in quadratic form. Using Hamilton method the existences and uniqueness of optimal Nash equilibrium are studied. The relationship between the existence of generalized algebra Riccati equations and optimal Nash solution of dynamic games are studied in [1] and [2].

Zero-sum linear quadratic dynamic game of two players is studied in [1], with the players satisfying one differential equation in state space form, and minimizing one objective function in quadratic form.

Linear quadratic dynamic games of two players in descriptor systems, which is zero-sum dynamic games are studied in [7]. The players satisfy one state space differential equation of descriptor system and minimize one objective function in quadratic form.

Using Hamilton method, [4] has studied linear quadratic optimal control descriptor systems. Due to [4] solution of generalized differential Riccati equation can be not exist. With mild assumption of generalized differential Riccati equations [3] that the system will have solutions. For infinite time problem [3] also studied algebra Riccati equations. The numerical non recursive formula is derived to obtain optimal solution of linear quadratic optimal control infinite time. This paper is studying non-zero-sum linear quadratic dynamic games in descriptor system of two players.

Two players linear quadratic dynamic game define as an open-loop game with the players giving control to the system

$$E\dot{x} = Ax + B_1u_1 + B_2u_2 \ , \tag{1.1}$$

$$Ex(0) = Ex_0 , \tag{1.2}$$

with $E \in \Re^{nxn}, A \in \Re^{nxn}, B_1 \in \Re^{nxm_1}, B_2 \in \Re^{nxm_2}$, $x(t)$ descriptor vector of n dimension. While $u_1(t)$, is control vector m_1 dimension which is done by the first player, and $u_2(t)$ is control vector m_2 dimension which is done by the second player. Matrix E is singular with rank $E = r < n$. In the case $rank\ E = n$, the system will be classical differential equation.

The two players minimizing objective function in the Nash sense in the form

$$J_1(u_1,u_2) = \frac{1}{2}x(T)^T E^T K_{1T} Ex(T) + \frac{1}{2}\int_0^T \left(x^T Q_1 x + u_1^T R_{11} u_1 + u_2^T R_{12} u_2\right)dt \ , \tag{1.3}$$

$$J_2(u_1,u_2) = \frac{1}{2}x(T)^T E^T K_{2T} Ex(T) + \frac{1}{2}\int_0^T \left(x^T Q_2 x + u_1^T R_{21} u_1 + u_2^T R_{22} u_2\right)dt \ , \tag{1.4}$$

with all matrices symmetric, further more Q_1 , Q_2 and K_{1T} ,K_{2T} semi positive definite and R_{11} ,R_{12} ,R_{21} ,R_{22} positive definite.

Below is definition for Nash equilibrium of two player game.

Definition 2.1: *The pair* (u_{1*}, u_{2*}) *is called Nash equilibrium for non-cooperative two players dynamic game if the following two equations are satisfied.*

$$J_{1*} = J_1(u_{1*}, u_{2*}) \le J_1(u_1, u_{2*}), \qquad \forall u_1 \in \Re^{m_1},$$

$$J_{2*} = J_2(u_{1*}, u_{2*}) \le J_2(u_{1*}, u_2), \qquad \forall u_2 \in \Re^{m_2}.$$

2 Linear Quadratic regulator Problem

Linear quadratic regulator for finite time ($T < \infty$) and infinite time ($T \to \infty$) of two players and *2* players are studied. Hamilton function is given to prove existence of optimal solution of linear quadratic dynamic game for descriptor system of finite time. Then the relationship between existence of differential Riccati equation solution and existence of optimal Nash solution for linear quadratic dynamic game of descriptor system is studied.

For two players of dynamic game with finite time, necessary conditions for the existence of optimal solution with objective function (1.3), (1.4) and satisfy system (1.1), (1.2) will be derived. First assumption will be given below.

Assumption 2.1: *Descriptor system (1.1), (1.2) is regular, impulse controllable and finite dynamic stabilizable that is*

$\det(sE - A) \neq 0, \quad \forall s \neq 0$, *except for a finite number of* $s \in \Re$,

$\operatorname{Im} E + \operatorname{Im} A(\ker E) + \operatorname{Im} B_i = \Re^n, \quad i = 1,2$,

$rank[sE - A \quad B_i] = n \quad i = 1,2 \quad \forall s, \operatorname{Re}[s] \geq 0$.

Consider Hamilton functions below

$$H_1(t) = \frac{1}{2}\left(x^T Q_1 x + u_1^T R_{11} u_1 + u_2^T R_{12} u_2\right) + \gamma_1^T \left(Ax + B_1 u_1 + B_2 u_2\right), \quad (2.1)$$

$$H_2(t) = \frac{1}{2}\left(x^T Q_2 x + u_1^T R_{21} u_1 + u_2^T R_{22} u_2\right) + \gamma_2^T \left(Ax + B_1 u_1 + B_2 u_2\right), \quad (2.2)$$

with γ_1 and γ_2 are function that will be derived further.

From the Hamilton function definition, theorem for the existence of optimal solution for two players can be derived.

Theorem 2.1: *Assume that x(t) and $u_1(t)$, $u_2(t)$ satisfy equations (1.1), (1.2). Then necessary conditions to minimizing J_1, J_2 are*

$$E^T \frac{\partial \gamma_i}{\partial t} = -\frac{\partial H_i}{\partial x}, \qquad \frac{\partial H_i}{\partial u} = 0, \qquad E\dot{x} = \frac{\partial H_i}{\partial \gamma_i}, \qquad i=1,2.$$

with $E^T \gamma_1(T) = E^T K_{1T} Ex(T)$ *and* $E^T \gamma_2(T) = E^T K_{2T} Ex(T)$. *While $H_1(t)$, $H_2(t)$, are Hamilton functions which is defined in (2.1), (2.2).*

Optimal solutions for two players dynamic game are

$u_i = -R_{ii}^{-1} B_i^T \gamma_i(t)$, $i=1,2$, where $\gamma_i(t)$ satisfy $E^T \dot{\gamma}_i(t) = -Q_i x(t) - A^T \gamma_i(t)$, with boundary conditions

$$E^T \gamma_i(T) = E^T K_{iT} Ex(T) \ . \quad (2.3)$$

The system can be written in the following descriptor system

$$\tilde{E}\dot{\tilde{x}} = \tilde{A}\tilde{x} \ , \quad (2.4)$$

with

$$\tilde{E} = \begin{pmatrix} E & 0 & 0 \\ 0 & E^T & 0 \\ 0 & 0 & E^T \end{pmatrix}, \tilde{x} = \begin{pmatrix} x(t) \\ \gamma_1(t) \\ \gamma_2(t) \end{pmatrix} \text{ and } \tilde{A} = \begin{pmatrix} A & -B_1 R_{11}^{-1} B_1^T & -B_2 R_{22}^{-1} B_2^T \\ -Q_1 & -A^T & 0 \\ -Q_2 & 0 & -A^T \end{pmatrix}.$$

If $\left(s\tilde{E} - \tilde{A}\right)$ regular, then the system will have solution.

Assumption 2.2: *Descriptor system (2.4), (2.3) is regular and impulse free i.e*

$$\operatorname{Im}\tilde{E} + \operatorname{Im}\tilde{A}(\ker\tilde{E}) = \Re^{2n+m} \text{ , and } \det(sE - A) \neq 0, \quad \forall s \neq 0,$$

except for a finite number of $s \in \Re$.

With Assumption (2.2) solution of descriptor system can be derived.

Lemma 2.1: *Define matrices* $X_1(t), X_2(t)$ *with*

$$X_1(t) = \left[\Lambda_{21}(t) + \Lambda_{22}(t)Y_1(0) + \Lambda_{23}(t)Y_2(0)\right]\left[\Lambda_{11}(t) + \Lambda_{12}(t)Y_1(0) + \Lambda_{13}(t)Y_2(0)\right]^{-1}$$

$$X_2(t) = \left[\Lambda_{31}(t) + \Lambda_{32}(t)Y_1(0) + \Lambda_{33}(t)Y_2(0)\right]\left[\Lambda_{11}(t) + \Lambda_{12}(t)Y_1(0) + \Lambda_{13}(t)Y_2(0)\right]^{-1}$$

then $\gamma_1(t) = X_1(t)x(t)$; $\gamma_2(t) = X_2(t)x(t)$, with $(x(t), \gamma_1(t))$, *are solutions of (4.4) and (4.3), with* Λ_{ij} *are found from*

$$\begin{pmatrix} \Lambda_{11}(t) & \Lambda_{12}(t) & \Lambda_{13}(t) \\ \Lambda_{21}(t) & \Lambda_{22}(t) & \Lambda_{23}(t) \\ \Lambda_{31}(t) & \Lambda_{32}(t) & \Lambda_{33}(t) \end{pmatrix} = \tilde{N}\begin{pmatrix} e^{\tilde{F}_t} & 0 \\ 0 & I \end{pmatrix}\tilde{M},$$

while $Y_1(0), Y_2(0)$ *are satisfy* $E^T\gamma_i(0) = Y_i(0)Ex(0), \quad i = 1,2$.

3 Generalized differential Riccati equation

Two generalized differential Riccati equations are given as follow

$$\begin{cases} E^T\dot{K}_1 + A^T K_1 + L_1 A + Q_1 - L_1 B_1 R_{11}^{-1} B_1^T K_1 - L_1 B_2 R_{22}^{-1} B_2^T K_2 = 0 \\ E^T\dot{K}_2 + A^T K_2 + L_2 A + Q_2 - L_2 B_1 R_{11}^{-1} B_1^T K_1 - L_2 B_2 R_{22}^{-1} B_2^T K_2 = 0 \end{cases} \tag{3.1}$$

where

$$L_i E = E^T K_i \quad , \qquad i=1,2. \tag{3.2}$$

Theorem below will study relationship between the existence of descriptor solution and existence of generalized differential equation solution (3.1), (3.2).

Theorem 4.2: *If generalized differential Riccati equation* (3.1), (3.2) *have solutions, then descriptor system will have solutions.*

Proof. Let the players 1, 2 play strategies $u_i(t) = -R_{ii}^{-1} B_i^T K_i(t) x(t), \quad i = 1,2,$ for controlling systems (3.1), (3.2) with $K_i(t), \quad i = 1,2$ solution for (3.1), (3.2).

Define

$$\gamma_i(t) = K_i(t)x(t), \quad i = 1,2.$$

Then the derivations give

$$E^T \dot{\gamma}_i(t) = E^T \dot{K}_i(t)x(t) + E^T K_i(t)\dot{x}(t), \quad i = 1,2.$$

From (3.1), (3.2) then

$$E\dot{x} = Ax(t) - B_1 R_{11}^{-1} B_1^T K_1(t)x(t) - B_2 R_{22}^{-1} B_2^T K_2(t)x(t),$$

or

$$E\dot{x} = (A - B_1 R_{11}^{-1} B_1^T K_1(t) - B_2 R_{22}^{-1} B_2^T K_2(t))x(t).$$

From (3.1), (3.2) we get

$$E^T \dot{K}_i = -A^T K_i - L_i A - Q_i + L_i B_1 R_{11}^{-1} B_1^T K_1 + L_i B_2 R_{22}^{-1} B_2^T K_2.$$

Therefore

$$E^T \dot{\gamma}_i(t) = -A^T K_i x - L_i Ax - Q_i x + L_i B_1 R_{11}^{-1} B_1^T K_1 x + L_i B_2 R_{22}^{-1} B_2^T K_2 x + E^T K_i(t)\dot{x}$$

$$= -A^T K_i x - L_i Ax - Q_i x + L_i B_1 R_{11}^{-1} B_1^T K_1 x + L_i B_2 R_{22}^{-1} B_2^T K_2 x + L_i E\dot{x}$$

$$= -A^T K_i x - L_i Ax - Q_i x + L_i B_1 R_{11}^{-1} B_1^T K_1 x + L_i B_2 R_{22}^{-1} B_2^T K_2 x$$

$$+ L_i Ax - L_i B_1 R_{11}^{-1} B_1^T K_1 x - L_i B_2 R_{22}^{-1} B_2^T K_2 x$$

$$= -A^T K_i x(t) - Q_i x(t),$$

Then

$$\begin{cases} E^T \dot{\gamma}_1(t) = -A^T K_1 x(t) - Q_1 x(t), \\ E^T \dot{\gamma}_2(t) = -A^T K_2 x(t) - Q_2 x(t), \end{cases}$$

or

$$\begin{cases} E^T \dot{\gamma}_1(t) = -A^T \gamma_1(t) - Q_1 x(t), \\ E^T \dot{\gamma}_2(t) = -A^T \gamma_2(t) - Q_2 x(t). \end{cases}$$

This system will has solution.

4 Infinite time problem

Suppose the players satisfy system equation (1.1), (1.2). Control vector that minimizing the following objective function will be found,

$$J_1(u_1, u_2) = \frac{1}{2}\int_0^\infty \left(x^T Q_1 x + u_1^T R_{11} u_1 + u_2^T R_{12} u_2\right) dt\,, \tag{4.1}$$

$$J_2(u_1, u_2) = \frac{1}{2}\int_0^\infty \left(x^T Q_2 x + u_1^T R_{21} u_1 + u_2^T R_{22} u_2\right) dt\,, \tag{4.2}$$

with all matrices symmetric, Q_1 , Q_2 and K_{1T} ,K_{1T} semi definite positive and R_{11} ,R_{12} , R_{21} , R_{22} positive definite.

Given generalized algebra Riccati equation

$$A^T K_i + L_i A + Q_i - L_i B_1 R_{11}^{-1} B_1^T K_1 - L_i B_2 R_{22}^{-1} B_2^T K_2 = 0; \quad L_i E = E^t K_i\,, \quad i=1,2. \tag{4.3}$$

Optimal control for infinite time problem are in the form $u_i = -R_{ii}^{-1} B_i^T K_i x(t)$, $i=1,2$, with K_1 and K_2 are constant matrices which are solutions of (4.3).

5 Generalized eigenvalue problem and existence of differential Riccati solution

Here the existence of two differential Riccati (3.1) and (3.2) solutions are studied. The discussion follows the application of result in [3] for linear quadratic optimal control descriptor system to linear quadratic game problem. Two linear quadratic optimal controls are considered. The first problem is minimizing

$$J_1(u_1, u_2) = \frac{1}{2} x(T)^T E^T K_{1T} E x(T) + \frac{1}{2}\int_0^T \left(x^T Q_1 x + u_1^T R_{11} u_1 + u_2^T R_{22} u_2\right) dt\,, \tag{5.1}$$

with equations (1.1), (1.2) are satisfied. The second problem is minimizing

$$J_2(u_1,u_2)=\frac{1}{2}x(T)^T E^T K_{2T} Ex(T)+\frac{1}{2}\int_0^T\left(x^T Q_2 x+u_1^T R_{11}u_1+u_2^T R_{22}u_2\right)dt\,, \quad (5.2)$$

Necessary conditions for existence of optimal solutions for (5.1), (5.2) satisfy

$$E^T\dot{\gamma}_1(t)=-Q_i x(t)-A^T\gamma_i(t)\,, \qquad i=1,2 \quad (5.3)$$

and

$$u_i(t)=-R_{ii}^{-1}B_i^T\gamma_i(t)\,, \qquad i=1,2 \quad (5.4)$$

From equations (1.1), (5.3) and (5.4) we get systems

$$\begin{pmatrix} E & 0 & 0 & 0\\ 0 & E^T & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} \dot{x}\\ \dot{\gamma}_i\\ \dot{u}_1\\ \dot{u}_2 \end{pmatrix}=\begin{pmatrix} A & 0 & B_1 & B_2\\ -Q_i & -A^T & 0 & 0\\ 0 & B_1^T & R_{11} & 0\\ 0 & B_2^T & 0 & R_{22} \end{pmatrix}\begin{pmatrix} x\\ \gamma_i\\ u_1\\ u_2 \end{pmatrix},\ i=1,2. \quad (5.5)$$

Consider generalized eigenvalue problems

$$\tilde{A}_i z_i=\lambda_i\tilde{E}z_i\,, \qquad i=1,2 \quad (5.6)$$

with

$$\tilde{A}_i=\begin{pmatrix} A & 0 & B_1 & B_2\\ -Q_i & -A^T & 0 & 0\\ 0 & B_1^T & R_{11} & 0\\ 0 & B_2^T & 0 & R_{22} \end{pmatrix}\text{and}\ \tilde{E}=\begin{pmatrix} E & 0 & 0 & 0\\ 0 & E^T & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

By applying the result [3] result it can be proved the existence of differential equations as follows

$$H_{0i}^T A_0+A_0^T H_{0i}+Q_{0i}-H_{0i}^T B_{0i}R_{11}^{-1}B_{0i}^T H_{0i}-H_{0i}^T B_{02}R_{22}^{-1}B_{02}^T H_{0i}=0\,,$$

with

$$A_0=V_2^T AV_1,\quad B_{01}=V_2^T B_1,\quad B_{02}=V_2^T B_2,\quad Q_{01}=V_1^T Q_1 V_1,\quad Q_{02}=V_1^T Q_2 V_1.$$

Also it can be proved the existence of differential equation solution as follows

$$X_{0i}^T A+A^T X_{0i}+Q_i-X_{0i}^T B_1 R_{11}^{-1}B_1^T X_{0i}-X_{0i}^T B_2 R_{22}^{-1}B_2^T X_{0i}=0\,, \qquad i=1,2$$

$$X_{0i}^T E=E^T X_{0i}\,,$$

$$(M^{-1})^T X_{01}M^{-1}=\begin{bmatrix} K_{111} & K_{112}\\ K_{121} & K_{122} \end{bmatrix}\qquad (M^{-1})^T X_{02}M^{-1}=\begin{bmatrix} K_{211} & K_{212}\\ K_{221} & K_{222} \end{bmatrix}$$

$$MB_1=\begin{bmatrix} B_{11}\\ B_{12} \end{bmatrix},\qquad MB_2=\begin{bmatrix} B_{21}\\ B_{22} \end{bmatrix}.$$

Then consider differential Riccati equation

$$\dot{Z}_{i11} + Z_{i11}F_i + F_i^T Z_{i11} - Z_{i11}B_{i1}R_{ii}^{-1}B_{i1}^T Z_{i11} = 0. \qquad i=1,2$$

As in ordinary optimal control theory this equation will have solution, say

$$Z_1(t) = M^T \begin{pmatrix} Z_{111} & 0 \\ 0 & 0 \end{pmatrix} N^{-1},\ Z_2(t) = M^T \begin{pmatrix} Z_{211} & 0 \\ 0 & 0 \end{pmatrix} N^{-1}.$$

By the result of [3] the existence of solutions for differential equation can be derived as follows

$$E^T \dot{Z}_{0i}(t) + Z_{0i}^T(t)A + A^T Z_{0i}(t) + Q_i - Z_{0i}^T(t)B_1 R_{11}^{-1} B_1^T Z_{0i}(t) - Z_{0i}^T(t)B_2 R_{22}^{-1} B_2^T Z_{0i}(t) = 0,$$

$$Z_{0i}^T(t)E = E^T Z_{0i}(t), \quad i=1,2,$$

with final conditions $E^T Z_{0i}(T) = E^T K_{iT} E$, and $Z_{0i}(T) = X_{0i} + Z_i(T)$. Then with the assumptions (2.1), (2,2) the existence of the solutions for differential Riccati equation (3.1), (3.2) can be proved.

References

[1] Basar, Tamer and Olsder, Geert Jan (1995), *Dynamic noncooperative Game Theory*, Second edition, Academic Press, London, San Diego, New York, Boston, Sydney, Tokyo and Toronto, pp.317-345,.

[2] Engwerda, Jacob J, (1998), On the open-loop Nash Equilibrium in the LQ-Games, *Journal on Economic Theory*,.

[3] Katayama, Tohru and Minamino, Katsuki, (1992), Linear Quadratic Regulator and Spectral Factorization for Continuous-Time Descriptor System, *Proceeding of the 31-th Conference on Decision and Control,* Tucson Arizona, 967-972,.

[4] Lewis, FL, (1986), A survey of Linear Singular Systems, *Circuits System Signal Process*, vol.5, no.1, 1986.

[5] Salmah, Bambang S., S.M. Nababan, and S.Wahyuni, (2002), Non-Zero-Sum Linear Quadratic Dynamic Game with Descriptor Systems, *Proceeding of Asian Control Conference, Singapore.*

[6] Salmah, S.M. Nababan, Bambang.S, S.Wahyuni (2004), Masalah Nilai Eigen Diperumum, *Seminar Nasional Matematika* ,Bali, Indonesia.

[7] Salmah, SM Nababan, Bambang S, S Wahyuni (2004), Existence Of Nash Solution For Non-Zero-Sum Linear Quadratic Game With Descriptor System, *Proceeding SEAM Conference,* UGM, Yogyakarta.

[8] Xu, Hua, dan Mizukami, Koichi (1994), Linear Quadratic Zero-sum Differential Games for Generalized State Space Systems, *IEEE Transactions on Automatic Control*, Vol. 39, No.1.

Salmah: Ph D student at Department of Mathematics, Universitas Gadjah Mada, Sekip Utara, Bulaksumur, Yogyakarta 55281, Indonesia.
Department of Mathematics, Universitas Gadjah Mada, Sekip Utara, Bulaksumur, Yogyakarta 55281, Indonesia. Phone/Fax +62 +22 522443
E-mail: syalmah@yahoo.com

S.M. Nababan: Department of Mathematics & Center of Mathematics, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia. Phone/Fax: +62 +22 250 8126

Bambang.S: Department of Mathematics, Universitas Gadjah Mada, Sekip Utara, Bulaksumur, Yogyakarta 55281, Indonesia. Phone/Fax +62 +22 522443

S.Wahyuni: Department of Mathematics, Universitas Gadjah Mada, Sekip Utara, Bulaksumur, Yogyakarta 55281, Indonesia. Phone/Fax +62 +22 522443

Bonds and Options Valuation using a Conditioning Factor

Sankarshan Basu

Indian Institute of Management, Bangalore, India

Abstract: In this talk, I shall first look at methods to calculate bounds for the prices of zero coupon bonds and contingent payments on the interest rates where the interest rates follow a log-normal distribution using two different ways. In the first method a conditioning variable is employed - this is similar to the approach of Basu (1999) and Rogers and Shi (1995). The second method is via a direct expansion. The lower bounds obtained are so accurate that they are essentially the true prices. Then I shall use the approximation technique discussed for the zero coupon bond case to approximate the price of the bond (in fact the lower bound to the price of the bond) for the case of coupon-carrying bonds - both non-defaultable as well as defaultable ones. The second part of the talk deals with pricing of options on assets with stochastic volatility – this is a very interesting problem in mathematical finance and it has widespread uses in the financial industry.

Keyword: contingent claim pricing, stochastic volatility models

Continuous-time Parameter Estimation of Exponential-Affine Term Structure Models

A. Wibowo

University of Twente, the Netherlands

Abstract. The exponential-affine term structure model is a class of models in which the yields to maturity are affine functions of some state vector $x(t)$. This model has been first proposed by Duffie and Kan (1994), and subsumes Vasicek, Cox-Ingersol-Ross and other commonly used interest rate models as special cases. Since the interest rate factors $x(t)$ are not directly observed, unknown parameters in these models need to be estimated on the basis of observing the bond prices of different maturities. Although financial models are commonly formulated in continuous time, all existing parameter estimation techniques discretize the observation equation in time in order to use known statistical or filtering methods. We resolve this incongruity in the present paper by working throughout with the original continuous-time formulation.
Key-words: nonlinear filtering, term structure models

1 Introduction

Short rate modeling of the term structure has been developed as early as 1973 by Merton [17] where he assumes instantaneous short rate to be a Brownian diffusion process with constant coefficients. Later Vasicek [21] extended the model incorporating the mean reversal of the interest rate. These initial models describe the instantaneous short rate as Gaussian processes, which allow negative interest rates. A general equilibrium approach to short rate modeling developed by Cox et al. [6] leads to the modification of the mean reverting diffusion model of Vasicek. The model is known as the square root model and does not allow negative interest rates while maintaining the mean reverting feature. Further generalizations and modifications can be found in Longstaff and Schwartz [15] and Hull and White [13], among others. An even more general model is proposed by Duffie and Kan [9]. Known as the exponential-affine model, it is a popular model for least three reasons. First of all, as the name suggests, bond yields are expressed as linear function with respect to the interest rates factors. Secondly, this class of models reduced the more complicated bond pricing partial differential equation (PDE) into a set of ordinary differential equations (ODE). Thirdly, it spans many of the popular one factor and multi-factor models.

As the development of the interest rate models continues and the models allows greater flexibility, for the purpose of analysis of the yield curve and for pricing of derivatives one needs to estimate the parameters of the models. This too has evolved along with the development of the interest rate model itself. One popular method used to estimate interest rates models is the Generalized Method of Moments (GMM). This method compares the moments of the sample with their

theoretical values. Parameters are chosen such that the values of the theoretical moments are close to those obtained from samples. It has been used by Heston [12], Gibbons and Ramaswamy [11], and Longstaff and Schwartz [15] to estimate parameters of one and two factors Cox-Ingersol-Ross models. A particular case of GMM, which is known both as the Efficient Method of Moments (EMM) and the Simulated Method of Moments (SMM), has been used by Duffie and Singleton [10] in asset pricing and by Dai and Singleton [7] to estimate parameters in the three-factor affine models.

A completely different approach, based on the stochastic filtering theory prevalent in control and communication engineering, also has been used for parameter estimation of term structure models. The work of Ball and Torous [5] is one of the first efforts in this direction, where they fitted spot rates of different maturities to the two-factor Cox-Ingersol-Ross model. Pennachi [19], Babbs and Newman [1] and Lund [16] investigated the two factor generalized Vasicek model. The application of more general exponential affine models is discussed in Duan and Simonatto [8] where they include the estimation of the two-factor Cox-Ingersol-Ross model as an example. In all these studies, the method of maximum likelihood is used. The likelihood function formula contains the (linear) Kalman filter, thereby connecting the solution to that of the filtering theory.

One characteristic of the estimation methods existing in the literature is that they invariably discretize the observation equation in time in order to apply known statistical/filtering techniques. This phenomenon is rather strange, considering the fact that the rest of the analysis in interest rate mathematics is almost always done in continuous time. At this point, all existing literature discretize the observation process and assume that the discretized data is disturbed by a Gaussian white-noise sequence. In the present paper, we resolve this incongruity by working throughout in the original continuous-time formulation.

For the remaining part of this paper we will introduce the mathematical formulation of the exponential affine model, followed by the proposed continuous time maximum likelihood parameter estimation method where we introduce a robust likelihood functional. An example of parameter estimation of the Cox-Ingersol-Ross model using simulated data is given to compare the performance of the robust and the non-robust likelihoods. Finally, some concluding remarks are given at the end.

2 Exponential-Affine Term Structure Models

2.1 Basic Assumptions

Following Duffie and Kan (1994), we assume that the stochastic process

$$\left\{X_t = \left[x_t^1, \ldots, x_t^n\right]^* \in \mathbb{R}^n,\, t \geq 0\right\}$$

to be the short rate factors, and that X_t satisfies the following stochastic differential equation (SDE):

$$dX_t = (AX_t + B)\ dt + \ \mathrm{diag}\left(\mathcal{A}X_t + \mathcal{B}\right)^{1/2}\ dW_t \tag{1}$$

where $A, \mathcal{A} \in \mathbb{R}^{n\times n}$, $B, \mathcal{B} \in \mathbb{R}^{n\times 1}$ are constant matrices. Assume that the instantaneous short rate process is given by the following affine relation:

$$r_t = \psi_0 + \psi_1^* X_t \tag{2}$$

where $\psi_0 \in \mathbb{R}$ and $\psi_1^* \in \mathbb{R}^{n\times 1}$. The vector $W_t = \left[W_t^1, \ldots, W_t^n\right]^* \in \mathbb{R}^{n\times 1}$ is composed of independent Brownian motions under the risk neutral measure $\mathbb{Q}$.

2.2 Bond Pricing

With the setup given above, we denote $V(t,T;x)$, $0 \le t \le T$ as the value of a derivatives instrument that pays off $h(X_T) \in \mathbb{R}$ at a delivery time T, given that $X_t = x$. Using the risk neutral valuation method, $V(t,T;x)$ is given by:

$$V(t,T;x) = E^{\mathbb{Q}}\left(\exp\left(-\int_t^T r_s\, ds\right) h(x_T)\,|\mathcal{F}_t\right)$$

where $\mathcal{F}_t$ is the σ-algebra generated by $\{X_s, 0 \le s \le t\}$. Equivalently, by Feynman-Kac theorem (see e.g. Oksendal [18]), $V(t,T;x)$ satisfies the following partial differential equation (PDE):

$$(\psi_0 + \psi_1^* x)\, V = \frac{\partial V}{\partial t} + \frac{\partial V}{\partial x^*}(Ax + B) + \frac{1}{2}\mathrm{Tr}\left(\frac{\partial^2 V}{\partial x \partial x^*}\mathrm{diag}\left(\mathcal{A}x + \mathcal{B}\right)\right)$$

where

$$\frac{\partial V}{\partial x^*} = \left[\frac{\partial V}{\partial x^1}, \ldots, \frac{\partial V}{\partial x^n}\right],$$

$$\frac{\partial^2 V}{\partial x \partial x^*} = \begin{bmatrix} \frac{\partial^2 V}{\partial x^1 \partial x^1} & & \frac{\partial^2 V}{\partial x^1 \partial x^n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 V}{\partial x^n \partial x^1} & & \frac{\partial^2 V}{\partial x^n \partial x^n} \end{bmatrix}$$

subject to the boundary condition $V(T,T;x) = h(x)$.
Thus, the price of a zero coupon bond that pays a unit amount at the delivery time T is given by:

$$P(t,T;x) = E^{\mathbb{Q}}\left(\exp\left(-\int_t^T r_s\, ds\right)|\mathcal{F}_t\right)$$

Theorem 1 (Duffie and Kan (1994)) *Given the short rate model (2), $P(t,T;x)$ satisfies:*

$$P(t,T;x) = \exp\left(C(t,T)\ x + D(t,T)\right)$$

where

$$\frac{\partial D(t,T)}{\partial t} = \psi_0 - C(t,T)^* B - \frac{1}{2}C^2(t,T)^* \mathcal{B} \tag{3}$$

$$\frac{\partial C(t,T)^*}{\partial t} = \psi_1^* - C(t,T)^* A - \frac{1}{2}C^2(t,T)^* \mathcal{A} \tag{4}$$

$$C^2(t,T) = \begin{bmatrix} C^2(t,T)_1 \\ \vdots \\ C^2(t,T)_n \end{bmatrix}$$

subject to boundary conditions:

$$C(T,T) = 0, \text{ and}$$
$$D(T,T) = 0$$

3 Maximum Likelihood Parameter Estimation

3.1 State-space Representation

Given a set of maturities $\mathcal{T} = \{T_1, T_2, \ldots, T_m\}$, the yield of a zero coupon bond with maturity T_i can be written as:

$$y_P(t,T_i) = \frac{1}{T_i - t}\left\{C^*(t,T_i)\ X_t + D(t,T_i)\right\}, \quad i = 1,\ldots,m \tag{5}$$

where $C^*(t,T_i)$ and $D(t,T_i)$ are solutions of (3) and (4). The linear relation is very convenient and motivates us to consider zero coupon bonds as the basis to estimate the short rate model from.
We first assume that bonds of fixed time to maturities $\tau_i = T_i - t$, $i = 1,\ldots,m$ are available for all $t \geq 0$. In this case, we can write $C^*(t, t+\tau_i) = C^*(0,\tau_i)$ and $D(t,t+\tau_i) = D(0,\tau_i)$. Therefore, (5) becomes:

$$y_P(t,t+\tau_i) = \frac{1}{\tau_i}\left\{C^*(0,\tau_i)\ X_t + D(0,\tau_i)\right\}, \quad i = 1,\ldots,m$$

where $C^*(0,\tau_i)$ and $D(0,\tau_i)$ are time-invariant matrices.
We further assume that the yield process is corrupted by a small noise. It is realistic, considering that in practice bid and ask prices are not equal. In order to apply the model, a unique price have to be determined based on bid and ask prices. For our purpose we set y_P to be equal to the mid-price

$$y_P := y_M(t,T) = \frac{1}{2}\left(y^B(t,T) + y^A(t,T)\right)$$

where $y^B(t,T)$ and $y^A(t,T)$ are bid and ask prices respectively.
Let us now denote

$$y_M(t) = \begin{bmatrix} y_M(t,t+\tau_1) \\ \vdots \\ y_M(t,t+\tau_m) \end{bmatrix} \in \mathbb{R}^m,\ C^* = \begin{bmatrix} \frac{1}{\tau_1}C^*(0,\tau_1) \\ \vdots \\ \frac{1}{\tau_m}C^*(0,\tau_m) \end{bmatrix} \in \mathbb{R}^{m\times n},$$

$$D = \begin{bmatrix} \frac{1}{\tau_1}D(0,\tau_1) \\ \vdots \\ \frac{1}{\tau_m}D(0,\tau_m) \end{bmatrix} \in \mathbb{R}^m$$

and to write the actual observation $y(t)$ as:

$$\begin{aligned} y(t) &= y_M(t) + \Sigma_0 \varepsilon(t) \\ &= C^* X_t + D + \Sigma_0 \varepsilon(t) \end{aligned}$$

where we have assumed that $y_M(t)$ is corrupted by Gaussian *white noise* $\varepsilon(t) \in \mathbb{R}^m$. The matrix $\Sigma_0 \in \mathbb{R}^{m \times m}$ is a constant, allowing possible correlations between the noises.
Although it is possible to model noisy observations with the white noise process $\varepsilon(t)$ (see e.g. [4]), we will follow the standard approach by defining an integrated observation $Y(t)$ of $y(t)$ as:

$$dY_t = (C^* X_t + D)\ dt + \Sigma_0 d\tilde{W}_t \tag{6}$$

where each components of the vector $\tilde{W}(t)$ are independent to each components of $W(t)$.

3.2 Robust Likelihood Functional

Let θ denotes the vector of unknown parameters and $\mathcal{Y}_t = \{Y_s, 0 \leq s \leq t\}$ be the observations up to time t. Given the partially observable system (1) and (6), the log-likelihood function is given by:

$$\begin{aligned} \mathbf{L}^{\theta}_{affine}(\mathcal{Y}_T) = &\int_0^T (\Sigma_0 \Sigma_0^*)^{-1} \left[C^*(\theta) \hat{X}_s + D(\theta), dY_s \right] - \\ &\frac{1}{2} \int_0^T \left\| (\Sigma_0 \Sigma_0^*)^{-1/2} \left(C^*(\theta) \hat{X}_s + D(\theta) \right) \right\|^2 ds \end{aligned} \tag{7}$$

where the filtered state $\hat{X}_t$, defined as the conditional expectation of X_t given the observation $\mathcal{Y}_t$, is given by:

$$\hat{X}_t = \frac{\int_{R^n} x\ q(t,x)\ dx}{\int_{R^n} q(t,x)\ dx}$$

where $q(t,x)$ satisfies the Zakai equation (see e.g. Poor [20]) :

$$dq(t,x) = \mathcal{L}^* q(t,x)\ dt + q(t,x) \left\langle (\Sigma_0 \Sigma_0^*)^{-1} (C^* x + D), dY(t) \right\rangle$$

and

$$\mathcal{L}\xi = \frac{\partial \xi}{\partial x^*} (Ax + B)\ dt + \frac{1}{2} \mathrm{Tr} \left(\mathrm{diag}(\mathcal{A}x + \mathcal{B}) \frac{\partial^2 \xi}{\partial x \partial x^*} \right)$$

To evaluate the log-likelihood function, we have to calculate the Itô integral appearing in the second term of $\mathbf{L}^{\theta}_{affine}(\mathcal{Y}_T)$. Note, however, that the observation equation (6) is only an idealization. We can never observe Y_t precisely, since the Brownian motion term is nowhere differentiable almost surely. Observed data is always a band-limited approximation, albeit of a large enough bandwidth to justify our model, of the ideal observation process $\dot{Y}_t$ coming from (6). A little more

precise, let $M^k(s)$ be an ideal low-pass filter with cut-off frequency $2\pi k$, actual observation y^k is a band-limited approximation:

$$y^k(t,\theta,\omega) = \int_{R^n} M^k(t-s)\ dY(s,\theta,\omega)$$

It is shown by Balakrishnan [3] that as $k \longrightarrow \infty$, the the log-likelihood function converges to:

$$\begin{aligned}\tilde{\mathbf{L}}^{\theta}_{affine}(\mathcal{Y}_T) = &\int_0^T \left\langle (\mathbf{\Sigma}_0\mathbf{\Sigma}_0^*)^{-1}\left(C^*(\theta)\,\hat{X}_s + D(\theta)\right),\, y(s)\right\rangle ds \\ &-\frac{1}{2}\int_0^T \left|\left|(\mathbf{\Sigma}_0\mathbf{\Sigma}_0^*)^{-1/2}\left(C^*(\theta)\ \hat{X}_s + D(\theta)\right)\right|\right|^2 ds \\ &-\frac{1}{2}\mathrm{Tr}\left(\int_0^T (\mathbf{\Sigma}_0\mathbf{\Sigma}_0^*)^{-1} C^*(\theta)\ \hat{P}(s)\ C\ (\theta)\right) ds \end{aligned} \tag{8}$$

where $\hat{P}(s)$ is the conditional covariance of $X(t)$ given the observation $\mathcal{Y}_t$. $y(t)$ can be approximated with the actual bond yields.
The log-likelihood (8) is valid for the general exponential-affine model with yield observations. Parameter estimates are given by parameters that maximize the log-likelihood, i.e.

$$\hat{\theta}_{MLE} = \underset{\theta}{\mathrm{argmax}}\ \tilde{\mathbf{L}}^{\theta}_{affine}(\mathcal{Y}_T)$$

In the following section, we will compare the performance of the both the non-robust likelihood (7) and its robust counterpart (8) by taking the one-factor non-linear Cox-Ingersol-Ross model as a special case of the exponential-affine model.

4 Parameter Estimation of the Cox-Ingersol-Ross Model

4.1 Model

The Cox-Ingersol-Ross short rate model is given by:

$$dr(t) = \kappa(\theta - r(t))\ dt + \sigma\sqrt{r(t)}\ dW(t)$$

for which, the corresponding bond yield of time to maturity τ at time t is:

$$y(t,\tau) = C(\tau)\ r(t) + D(\tau)$$

where

$$\begin{aligned} C(\tau) &= \frac{1}{\tau}\frac{2(e^{\gamma\tau}-1)}{2\gamma + (\kappa+\lambda+\gamma)(e^{\gamma\tau}-1)} \\ D(\tau) &- -\frac{q+1}{\tau}\log\left(\frac{2\gamma\exp(\tau(\kappa+\lambda+\gamma)/2)}{2\gamma+(\kappa+\lambda+\gamma)(e^{\gamma\tau}-1)}\right) \\ \gamma &= \sqrt{(\kappa+\lambda)^2+2\sigma^2} \\ q &= \frac{2\kappa\theta}{\sigma^2} - 1 \end{aligned}$$

κ	θ	σ	λ
0.1862	0.0654	0.0481	-0.0741

Table 1: Cox-Ingersol-Ross parameters reported in Jong and Santa-Clara (1999).

4.2 Data

We simulate 100 paths of 250 weekly data based on parameters reported in de Jong and Santa-Clara [14]. These parameters are presented on Table (1). We include bond yields with time to maturities 3 months, 6 months, 1, 2, 3, 5, 7, 10, 20 years.

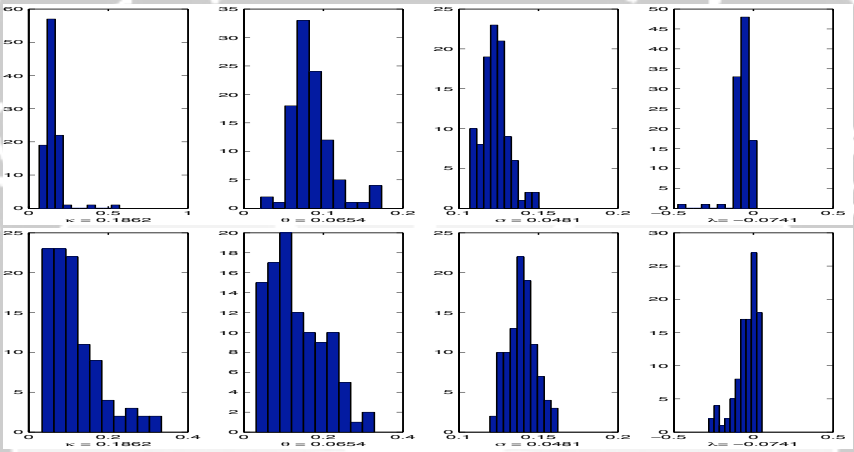


Figure 1: Parameter estimates (100 paths) σ_0=10 basis points. Figures in the first and second rows show histograms of the parameter estimates using the robust and non-robust likelihoods respectively.

4.3 Results

In this study we employ the Nelder-Mead's simplex method (MATLAB's fminsearch) which is a local optimization procedure. In all estimations, true parameter values were taken as initial guesses to the optimization procedure. In order to apply the proposed continuous-time method, interpolated data is used in the computation of the log-likelihood.

The observation error covariance term, Σ_0 is assumed to be $\sigma_0 \mathbf{I}_9$, where σ_0 = 10 and 0.1 basispoints. The resulting parameter estimates from 100 paths are presented as histograms in Figure (1) and (2) given σ_0 of 10 basis points and 0.1 basis points respectively. From both figures, we can conclude that resulting parameter estimates are generally better when using the robust likelihood. With the robust likelihood, parameter estimates of κ, θ and λ were reasonably close to their true values. However, the estimates of the volatility σ is biased. It

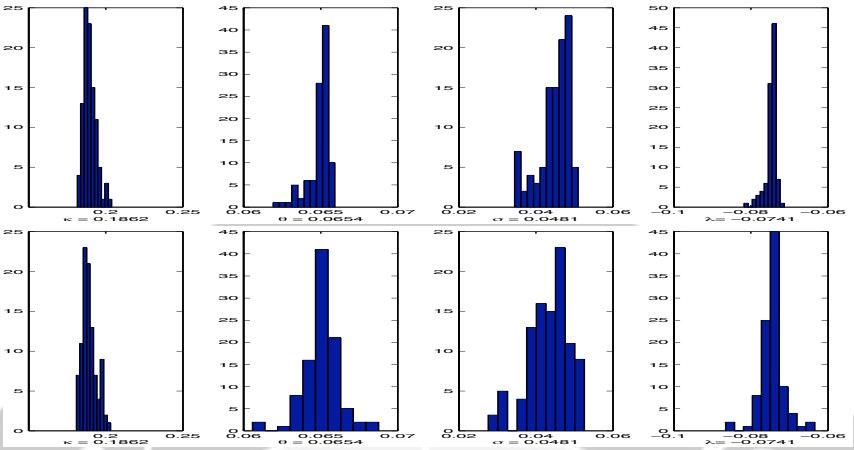


Figure 2: Parameter estimates (100 paths) σ_0=0.1 basis points. Figures in the first and second rows show histograms of the parameter estimates using the robust and non-robust likelihoods respectively.

is interesting to note that the performance both methods were comparable in estimating the volatility σ. When the observation errors is small (0.1 basis points), we see in Figure (2) that the performance of both likelihoods, were comparable across all parameters with a hint of better estimates resulted from the robust likelihood. We note that estimates of the volatility σ do not bias as much as in the case when the observation errors is larger.

5 Conclusions

In this paper we have proposed a continuous-time MLE of the exponential-affine model. The main feature of the method is that it retains the continuous-time formulation of both the interest rate factors and the corresponding bond prices. For this purpose, we have introduced a robust formulation of the likelihood functional to use the actual bond yields observations. Taking the Cox-Ingersol-Ross model as a special case, we have shown through simulated weekly data the robust likelihood provides better parameter estimates compared to those given by the non-robust counterpart.

References

[1] S. H. Babbs & K. B. Nowman (1999), Kalman Filtering of Generalized Vasicek Term Structure Models, *Journal of Financial and Quantitative Analysis*, 115–130

[2] A. Bagchi (1975), Continuous Time Systems Identification with Unknown Noise Covariance, *Automatica*, **11**, 533–536

[3] A.V. Balakhrisnan (1973), *Stochastic Differential Systems*, Springer-Verlag

[4] A.V. Balakhrisnan (1977), Likelihood Ratios for Signals in Additive White Noise, *Applied Mathematics and Optimization*, **3**, 341–356

[5] C. A. Ball & W. N. Torous (1996), Unit Roots and the Estimation of Interest Rate Dynamics, *Journal of Empirical Finance*, **3**, 215–238

[6] J. C. Cox, J. E. Ingersol, S. A. Ross (1985), A Theory of the Term Structure of Interest Rates, *Econometrica*, **53**, 385–407

[7] Q. Dai & K. J. Singleton (2000), Specification Analysis of Affine Term Structure Models, *Journal of Finance*, **LV**, 1943–1978

[8] J. C. Duan & J. G. Simonato (1995), *Estimating and Testing Exponential-Affine Term Structure Models by Kalman Filter*, *Review of Quantitative Finance and Accounting*, **13**, 102–127

[9] D. Duffie & R. Kan (1994), Multi-factor Term Structure Models, *Philosophical Transactions of the Royal Society of London A*, **347**, 577–586

[10] D. Duffie & K. J. Singleton (1993), Simulated Moments Estimation of Markov Models of Asset Prices, *Econometrica*, **61**, 929–952

[11] M. R. Gibbons & K. Ramaswamy (1993), A Test of the Cox, Ingersoll, and Ross Model of the Term Structure, *Review of Financial Studies* **6**, 619-658

[12] S. L. Heston (1988), Testing Continuous Time Models of the Term Structure of Interest Rates, *Working Paper*, *Carnegie Mellon University*, 1–27

[13] J. C. Hull & A. D. White, Pricing Interest Rate Derivative Securities (1990), *Review of Financial Studies*, **3-4**, 573–592

[14] F. de Jong & P. Santa-Clara (1999), The Dynamics of the Forward Interest Rate Curve: A Formulation With State Variables, *Journal of Financial and Quantitative Analysis*, **34**, 131–157

[15] F. A. Longstaff & E. S. Schwartz (1992), Interest Rate Volatility and the Term Structure: A Two-Factor General Equilibrium Model, *Journal of Finance*, **47**, 1259–1282

[16] J. Lund (1997), Non Linear Kalman Filtering Techniques for Term Structure Models, *Working Paper*, *University of Aarhus*, 1–34

[17] R. C. Merton (1973), The Theory of Rational Option Pricing, *Bell Journal of Economics and Management Science*, **4**, 141–183

[18] B. Oksendal (2003), *Stochastic Differential Equations: An Introduction with Applications*, Springer-Verlag

[19] G. G. Pennachi (1991), Identifying the Dynamics of Real Interest Rates and Inflation: Evidence Using Survey Data, *Review of Financial Studies*, **4**, 53–86

[20] V. Poor (1994), *An Introduction to Signal Detection and Estimation*, Springer-Verlag

[21] O. Vasicek (1977), An Equilibrium Characterization of the Term Structure, *Journal of Financial Economics*, **5**, 177–188

A. Wibowo: Department of Applied Mathematics, University of Twente, Postbus 217, 7500 AE, Enschede, the Netherlands.
E-mail: a.wibowo@math.utwente.nl

DEFAULT CORRELATION

Michael Rampisela

Ernst & Young Advisory Services, Jakarta, Indonesia

Abstract. A problem in a credit portfolio is the difficulty of modeling default correlations. Default correlation measures the strength of the default relationship between two borrowers. The historically observed joint probability of default between two firms is usually zero. This lack of data makes it difficult to estimate any type of credit correlation directly from history. This paper describes a model to capture default correlation that has more readily estimated parameter namely asset correlation. In addition, some of the most important properties of default correlation are also described. By knowing the individual firms' probability of default (PD or EDF) and the correlation of their assets values, the likelihood of both defaulting at the same time can be calculated.
Key-words: Credit risk modeling, correlation

1 Introduction

Credit risk is risk that a borrower (obligor or counterparts) will fail to repay money owed to the bank. Credit risk is conventionally defined using the concepts of expected credit loss and unexpected credit loss. Obviously credit losses are not constant across the economic cycle. The credit portfolio models are designed to quantify this volatility.

Modeling portfolio credit risk in credit portfolio is neither analytically nor practically easy. The idea is to aggregate the credit risk of all individual borrowers in a portfolio. The output is not the sum of risks of individual borrowers. Hence, one of the portfolio credit risk problem is the difficulty of modeling correlations.

For equities, the correlations can be directly estimated by observing high-frequency liquid market prices. For credit quality, the lack of data makes it difficult to estimate any type of credit correlation directly from history. We may assume that credit correlations are uniform across the portfolio, or proposing a model to capture credit correlations that has more readily estimated parameters. Two important elements of credit correlation are probability of default and loss given default.

1.1 Probability of Default

Default risk is the uncertainty regarding a borrower's ability to pay its debts. This risk can be quantified by the so-called probability of default (i.e. PD or EDF expected default frequencies), which describes the probability that the borrower will fail to meet its obligation. In addition to the above definition, an obligor can be categorized as defaulted obligor if one of the following conditions holds:

1 Any views expressed represent those of the author only and **not** necessarily those of Ernst & Young Advisory Services.

- the bank considers that the obligor is unlikely to pay its entirely credit obligation without any alternative for the bank to liquidates collateral(s);
- the obligor is past due more than 90 days on any credit obligation (including overdraft).

1.2 Loss Given Default (LGD)

Loss given default (LGD) is an estimate of the loss a creditor will incur if the borrower of a loan defaults. In other words it is the fraction of the debt the bank is not likely to recover from the borrower once it has defaulted. LGD is typically stated as a percent of the total debt value, it is one minus the recovery rate. Almost all recent credit risk models assume probability of default and loss given default to be statistically independent.

2 Default Correlation

Default correlation measures the strength of the default relationship between two borrowers (i.e. firm). If there is no relationship, then the default is independent and the correlation is zero. When two borrowers are correlated, this means that the probability of both defaulting at the same time is heightened, i.e. it is larger than it would be if they were completely independent.

The borrower will default when its market asset value falls below the face value of obligations (the default point). This means that the joint probability of default is the likelihood of both borrowers' market asset values being below their respective default points in the same time period. This probability can be determined quite readily from knowing:

- the borrowers' current market asset values A_0;
- their asset volatilities σ_A;
- the correlation between the two borrowers' market asset values ρ_{asset}.

The historically observed joint frequency of default between two companies is usually zero. Two borrowers have some chance of jointly defaulting, but nothing in their default history enables us to estimate the probability since neither has ever defaulted. We can measure the default correlation between two borrowers, using their asset correlation and their individual probabilities of default. The correlation between two borrowers' asset values can be empirically measured from their equity values.

Figure 1 illustrates the ranges of possible future asset values for two different borrowers. The two intersecting lines represent the default points for the two borrowers. For instance, if borrower X's asset value ends up below DD_x (the point represented by the vertical line), then borrower X will default.

The intersecting lines divide the range of possibilities into four regions. The upper right region represents those asset values for which neither borrower X nor borrower Y will default. The lower left region represents those asset values for which both will default. The probabilities of all these regions taken together must equal one.

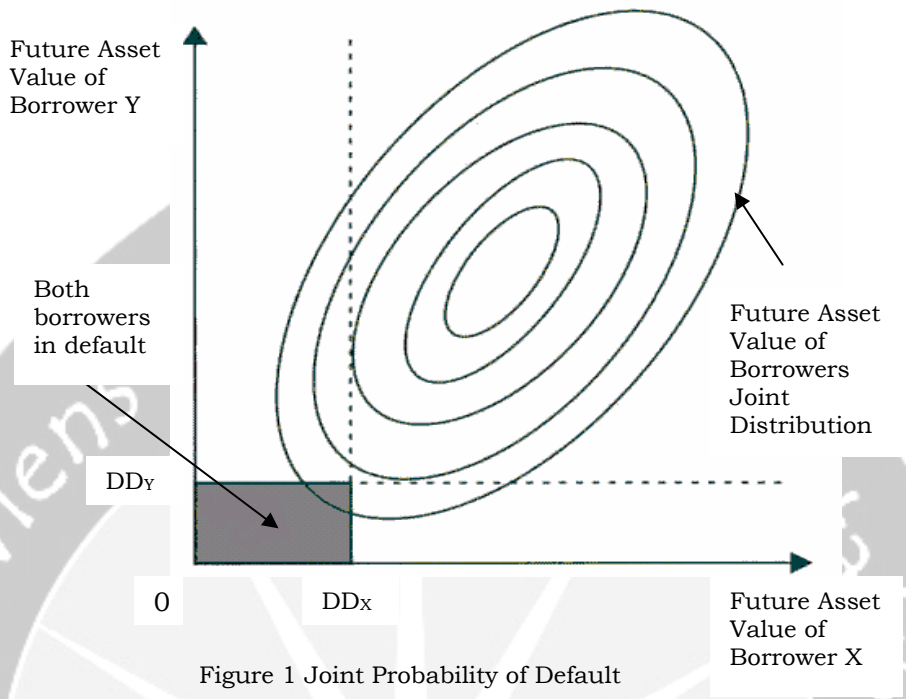


Figure 1 Joint Probability of Default

If the asset values of the two borrowers were independent, then the probabilities of the regions could be determined simply by multiplying the individual probabilities of default and non-default for the two borrowers. If the two borrowers' assets are positively correlated, then the probability of both asset values being high or low at the same time is higher than if they were independent and the probability of one being high and other being low is lower. By knowing the individual borrowers' probability of default, and knowing the correlation of their asset values, the likelihood of both defaulting at the same time can be calculated. The time series of a borrower's asset values can be determined from its equity value. The correlation between two borrowers' asset values can be calculated from their respective time series.

2.1 Model of Default Correlation

We assume the asset value of a borrower is lognormally distributed. It translates into normal distribution of asset return. We denote the distribution of the standardized asset returns of borrower 1 and borrower 2 by $X_1 \sim \phi(0,1)$ and $X_2 \sim \phi(0,1)$. The asset (return) correlation coefficient is denoted by ρ_{asset}. Respectively, we denote the value of asset return triggering default (default threshold) for borrower 1 and borrower 2 by $(-DD_1)$ and $(-DD_2)$. Furthermore we assume that in the case of default the defaulted borrower will pay nothing to the bank (the loss given default equals to one). Hence, we can derive the following binomial distribution describing default or non-default:

$$\tilde{X}_1 = \begin{cases} 1(\text{D}) & \text{if } x_1 \le -DD_1, \text{ i.e. default wi th probabilit y} \quad \Phi^{-1}(-DD_1) = \text{EDF}_1 \\ 0(\text{ND}) & \text{if } x_1 > -DD_1, \text{i.e. no default wi th probabilit y} \quad 1-\Phi^{-1}(-DD_1) = 1-\text{EDF}_1 \end{cases}$$

$$\tilde{X}_2 = \begin{cases} 1(\text{D}) & \text{if } x_2 \le -DD_2, \text{ i.e. default with probability} \quad \Phi^{-1}(-DD_2) = \text{EDF}_2 \\ 0(ND) & \text{if } x_2 > -DD_2, \text{i.e. no default with probability} \quad 1-\Phi^{-1}(-DD_2) = 1-\text{EDF}_2 \end{cases}$$

$\tilde{X}_1 = D$ and $\tilde{X}_2 = D$ respectively denote the events that borrower 1 and borrower 2 default at the horizon.

From the definition of covariance, the Pearson (linear) coefficient of correlation between the default events and is given by

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = Corr\left(1\{. \le x_1\}, 1\{. \le x_2\}\right)$$

where: σ_{12} covariance between the default event 1 and 2
σ_1 standard deviation of event 1
σ_2 standard deviation of event 2

We assume that the default process is a two-state event, then the default events $\tilde{X}_1 = D$ and $\tilde{X}_2 = D$ are binomial. We obtain the (coefficient of) correlation between the two default events:

$$\rho_{12}^{def} = \frac{P\left(\tilde{X}_1 = D, \tilde{X}_2 = D\right) - EDF_1 . EDF_2}{\sqrt{EDF_1\left(1 - EDF_1\right)}\sqrt{EDF_2\left(1 - EDF_2\right)}} \qquad (1)$$

The numerator of equation (1) represents the difference of the actual probability of both borrowers defaulting and the probability of both defaulting if they were independent. If the asset values are independent, then the default correlation is zero. The denominator reflects the standard deviation of default rates under the binomial distribution of each borrower.

The joint probability of default is

$$P\left(\tilde{X}_1 = D, \tilde{X}_2 = D\right) = \int_{-\infty}^{\Phi^{-1}(EDF_1)} \int_{-\infty}^{\Phi^{-1}(EDF_2)} N_2\left(x_1, x_2, \rho_{asset}\right) dx_2 dx_1$$

Where $N_2(x_1, x_2, \rho_{asset})$ is a bivariate normal distribution function. Hence, the asset correlation ρ_{asset} influences the corresponding default correlation by entering the joint probability of default term. The derivation of the correlation between those two default events (i.e. equation 1) can be found in the appendix.

Now we consider that the loss given default does not equal to one and it is variable. The average of the loss given default is $\overline{LGD}$. LGDs are assumed to be independent

across borrower. We obtain the (coefficient of) correlation between the two default events:

$$\rho_{12}^{def} = \frac{P\left(\tilde{X}_1 = D, \tilde{X}_2 = D\right) - EDF_1.EDF_2}{\sqrt{EDF_1\left(1-EDF_1\right)+\frac{(1-\mathrm{LGD}_1)}{4\mathrm{LGD}_1}EDF_1}\sqrt{EDF_2\left(1-EDF_2\right)+\frac{(1-\mathrm{LGD}_2)}{4\mathrm{LGD}_2}EDF_2}} \quad (2)$$

This equation is almost the same with equation 1 except the standard deviation terms contain the variable loss given default. The derivation of the default correlation that the loss given default does not equal to one (i.e. equation 2) can be found in the appendix.

2.2 Characteristics of Default Correlation

Some of the most important properties of default correlation are described in this section. If expected probability of default change, default point will also change accordingly. We consider the default-correlations having loss given default equal to one.

Property 1
When both borrowers have a probability of default (EDF) of 50%, its default correlation reaches the upper bound:

$$\rho^{def}{}_{\max} = \frac{2}{\pi}\arcsin\left(\rho_{asset}\right) \quad (3)$$

The proof of this property can be found in [4].

Property 2
If a borrower has a very small probability of default (EDF), its default correlation (ρ^{def}) with other borrowers can be approximately zero.

$$\lim_{EDF_1 \to 0} \rho^{def}\left(EDF_1, EDF_2, \rho_{asset}\right) = 0$$

The same result holds for EDF2 $\to$0.

$$\lim_{EDF_2 \to 0} \rho^{def}\left(EDF_1, EDF_2, \rho_{asset}\right) = 0$$

Both property 1 and property 2 limits the range value of default correlation. Since ρ^{def} is continuous in EDF_1, EDF_2 and ρ_{asset}, it takes on values between the upper bound in equation (3) and zero for a fixed ρ_{asset}. The higher the asset correlation between two borrowers the higher the default correlation between them.

Figure 2 describes the default correlations vs. the asset correlations for both borrowers having the same probability of default. The probability of default varies from 0.01% to 50% (value that the upper bound occurs). Figure 3 describes the default correlations vs. the asset correlations for a borrower having probability of default 1% and the other having probability of default from 0.01% to 50%.

We observe from figure 2 and figure 3 that there are two different default correlation characteristics: two equal EDF borrowers and two different EDF borrowers. The default correlation converges to one when the asset (return) correlation between two equal EDF borrowers is close to one. The closer both default-probabilities to 50% the closer the default correlation to the upper bound. On the contrary, the default correlation does not converge to one for two different EDF borrowers when the asset correlation approaches one. We can also notice that low default-probabilities lead to low default correlation.

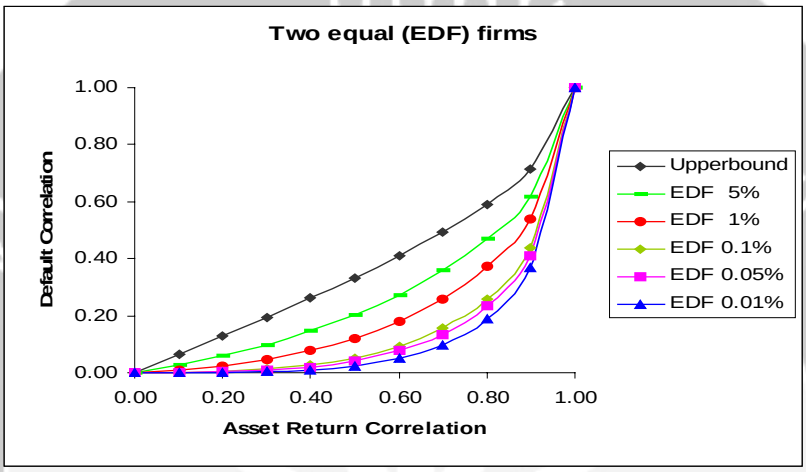


Figure 2 Default Correlation of Two Equal EDFs Borrowers

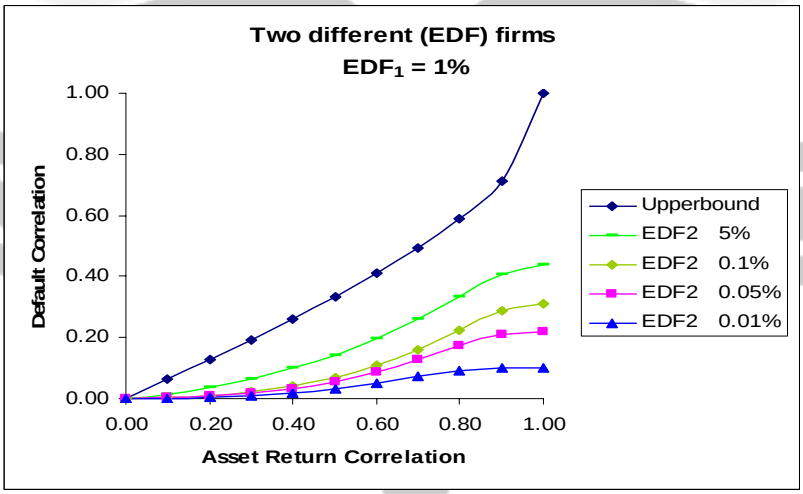


Figure 3 Default Correlation of Two Different EDFs Borrowers

From figure 4 we notice that if one borrower has a very low EDF (≤ 0.05%), its default correlation (ρ^{def}) with other borrowers can be approximately zero (property 2). The previous characteristics still hold; the default correlation converges to one when the asset correlation is close to one only if both probability of default are the same.

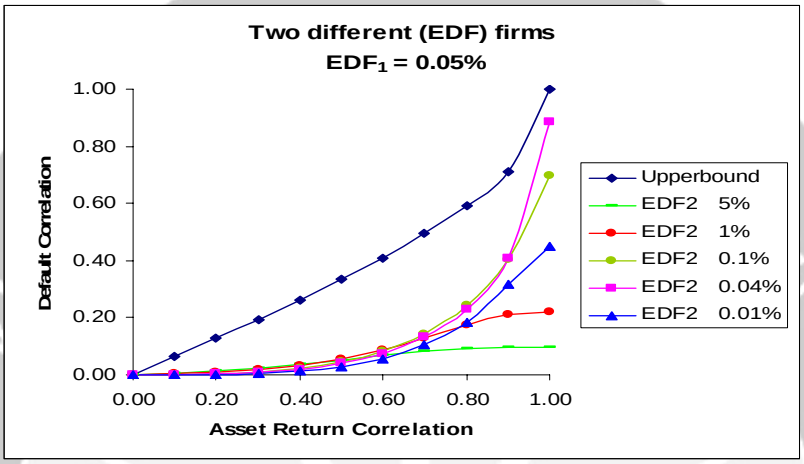


Figure 4 Default Correlation of a Very Low EDF Borrower

In most practical applications, probability of default is lower than 50%. This implies that deterioration in the credit qualities of both borrowers lead to an increase of default correlation. Since rating agencies take time to downgrade companies whose credit quality worsens, using an inappropriate credit rating for EDF estimation leads to miscalculation of the portfolio's unexpected loss due to inappropriate default correlation.

Default correlation is much lower than the corresponding asset correlation unless both probabilities of default are the same and the asset correlation is very large (close to one). If asset correlation is used instead of default correlation, this yields a misleading result since the unexpected losses would be higher than it should be.

In figure 5, a borrower has a very low EDF (0.05%). Its default correlations with other borrowers (for example 1%) are very low especially for low asset correlation (below 50%). Loss given default is another important parameter determining default correlation. The lower the loss given default the lower the default correlation.

ρ_{asset}	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
ρ^{def} LGD=1	0.31%	0.90%	1.89%	3.44%	5.73%	8.89%	12.94%	17.61%	21.51%	22.26%
ρ^{def} LGD=0.9	0.30%	0.87%	1.84%	3.35%	5.58%	8.64%	12.59%	17.13%	20.93%	21.65%
ρ^{def} LGD=0.4	0.23%	0.65%	1.37%	2.50%	4.16%	6.45%	9.40%	12.79%	15.62%	16.16%
ρ^{def} LGD=0.1	0.10%	0.27%	0.58%	1.06%	1.76%	2.72%	3.97%	5.40%	6.59%	6.82%
ρ^{def}_{max}	6.38%	12.82%	19.40%	26.20%	33.33%	40.97%	49.36%	59.03%	71.29%	100%

Figure 5 Relationship between Asset Correlation and Default Correlation
EDF_1= 1%; EDF_2= 0.05%

In figure 6 both probabilities of default (EDF) are set to 1%. We observe that for various values of loss given defaults the default correlation is one when the asset correlation is one. In this case the values of default correlation are always higher than those shown in figure 5. This condition does not hold for borrowers having different probability of default.

We consider borrower 1 having EDF_1= 1% and borrower 2 having EDF_2= 5% and then 20% (see figure 7 and 8). When the asset correlation is low (below 30%), the higher the difference between both probability-of-defaults the higher the value of default correlation. On the other hand, when the asset correlation is high (above 70%), the lower the difference between both probabilities of default the higher the value of default correlation. In all cases the default correlation is far less than the upper bound $\rho^{def}{}_{max}$.

ρ_{asset}	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
ρ^{def} LGD=1	0.94%	2.42%	4.63%	7.76%	12.09%	17.99%	26.00%	37.14%	53.84%	100%
ρ^{def} LGD=0.9	0.92%	2.36%	4.50%	7.55%	11.76%	17.50%	25.29%	36.13%	52.37%	100%
ρ^{def} LGD=0.4	0.68%	1.76%	3.35%	5.63%	8.77%	13.05%	18.86%	26.94%	39.05%	100%
ρ^{def} LGD=0.1	0.29%	0.74%	1.41%	2.37%	3.69%	5.50%	7.95%	11.35%	16.45%	100%
ρ^{def}_{max}	6.38%	12.82%	19.40%	26.20%	33.33%	40.97%	49.36%	59.03%	71.29%	100%

Figure 6 Relationship between Asset Correlation and Default Correlation
$EDF_1 = EDF_2$ =1%

ρ_{asset}	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
ρ^{def} LGD=1	1.52%	3.63%	6.40%	9.94%	14.33%	19.67%	26.02%	33.33%	40.80%	43.97%
ρ^{def} LGD=0.9	1.48%	3.53%	6.22%	9.66%	13.93%	19.12%	25.30%	32.40%	39.66%	42.74%
ρ^{def} LGD=0.4	1.10%	2.62%	4.62%	7.17%	10.33%	14.18%	18.77%	24.03%	29.42%	31.71%
ρ^{def} LGD=0.1	0.46%	1.09%	1.93%	2.99%	4.32%	5.92%	7.84%	10.04%	12.29%	13.24%
ρ^{def}_{max}	6.38%	12.82%	19.40%	26.20%	33.33%	40.97%	49.36%	59.03%	71.29%	100%

Figure 7 Relationship between Asset Correlation and Default Correlation
EDF_1= 1%; EDF_2= 5%

ρ_{asset}	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
ρ^{def} LGD=1	2.05%	4.44%	7.10%	9.94%	12.85%	15.62%	17.99%	19.56%	20.09%	20.10%
ρ^{def} LGD=0.9	1.99%	4.30%	6.88%	9.64%	12.46%	15.15%	17.44%	18.97%	19.48%	19.49%
ρ^{def} LGD=0.4	1.44%	3.12%	4.99%	6.99%	9.03%	10.98%	12.64%	13.75%	14.12%	14.13%
ρ^{def} LGD=0.1	0.58%	1.26%	2.01%	2.81%	3.64%	4.42%	5.09%	5.54%	5.69%	5.69%
$\rho^{def}{}_{max}$	6.38%	12.82%	19.40%	26.20%	33.33%	40.97%	49.36%	59.03%	71.29%	100%

Figure 8 Relationship between Asset Correlation and Default Correlation
EDF_1= 1%; EDF_2= 20%

3 Conclusion

Correlation plays an important role in credit risk portfolio modeling. Due to correlation between borrowers, the credit risk is not simply a summation of its individual risk. In this report we describe the work done to calculate the default correlation between two borrowers with more readily estimated parameters (i.e. probability of default, loss given default and asset correlation).

Some of the most important properties of default correlation are also described in the report. Default correlation is much lower than the corresponding asset correlation unless both probabilities of default are the same and the asset correlation is very large (close to one). If asset correlation is used instead of default correlation, this yields a misleading result since the credit risk portfolio would be higher than it should be.

References

[1] Altman, E.I. & Sauders, A. (1998), Credit Risk Measurement: Development over the Last 20 Years, *Journal of Banking and Finance*, 21, 1721-1742.

[2] Crosbie, P.J. & Bohn, J.R. (2001), *Modeling Default Risk*, Working Paper, KMV LLC, San Francisco.

[3] Embrechts, P. & McNeil, A. (1999), Strauman, D., *Correlation and Dependence in Risk Management: Properties and Pitfalls*, Eidgenössiche Technische Hochschule, Zürich.

[4] Erlenmaier, U. (2001), *Models of Joint Defaults in Credit Risk Management: An Assessment*, Alfred-Weber-Institut, Universität Heidelberg, Heidelberg, Germany.

[5] Frey, R., McNeil, A.J. & Nyfeler, M.A. (2001), *Modelling Dependent Defaults; Asset Correlations Are Not Enough*, Eidgenössiche Technische Hochschule, Zürich.

[6] Kealhofer, S. & Bohn, J.R. (2001), *Portfolio Management of Default Risk*, Working Paper, KMV LLC, San Francisco.

[7] Ong, M.K. (1999), *Internal Credit Risk Models: Capital Allocation and Performance Measurement*, Risk Books, London.

[8] Rice, J.A. (1995), *Mathematical Statistics and Data Analysis*, Duxbury Press, Belmont, California, USA.

[9] Saunders, A. (1999), *Credit Risk Measurement New Approaches to Value at Risk and Other Paradigms*, John Wiley & Sons, New York.

APPENDIX

From the definition of covariance, the Pearson (linear) coefficient of correlation between the default events $\tilde{X}_1 = D$ and $\tilde{X}_2 = D$ is given by

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = Corr\left(1_{\{\cdot \le x_1\}}, 1_{\{\cdot \le x_2\}}\right) \tag{A-1}$$

where: σ_{12} covariance between the default event 1 and 2
σ_1 standard deviation of event 1
σ_2 standard deviation of event 2

We assume that the default process is a two-state event, then the default events $\tilde{X}_1 = D$ and $\tilde{X}_2 = D$ are binomial. Furthermore we assume that in the case of default the defaulted borrower will pay nothing to the bank (the loss given default equals to one). Their corresponding standard deviations are

$$\begin{aligned} \sigma_1 &= L\overline{G}D_1\sqrt{P\left(\tilde{X}_1 = D\right)\left[1 - P\left(\tilde{X}_1 = D\right)\right]} \\ &= \sqrt{P\left(\tilde{X}_1 = D\right)\left[1 - P\left(\tilde{X}_1 = D\right)\right]} \\ &= \sqrt{EDF_1\left(1 - EDF_1\right)} \end{aligned} \tag{A-2}$$

and

$$\sigma_2 = L\overline{G}D_2\sqrt{P\left(\tilde{X}_2 = D\right)\left[1 - P\left(\tilde{X}_2 = D\right)\right]}$$

$$= \sqrt{P\left(\tilde{X}_2 = D\right)\left[1 - P\left(\tilde{X}_2 = D\right)\right]}$$
$$= \sqrt{EDF_2\left(1 - EDF_2\right)} \quad \text{(A-3)}$$

Both equation (A-2) and (A-3) describe the variability of loss around its expected value $EDF_i * LGD_i = EDF_i$ (i=1,2).
Now, the covariance is

$$\sigma_{12} = P\left[\tilde{X}_1 = D, \tilde{X}_2 = D\right] - P\left[\tilde{X}_1 = D\right]P\left[\tilde{X}_2 = D\right]$$
$$= P\left(\tilde{X}_1 = D, \tilde{X}_2 = D\right) - EDF_1.EDF_2 \quad \text{(A-4)}$$

Hence using equation (A-1), (A-2), (A-3) and (A-4) we obtain the (coefficient of) correlation between the two default events

$$\rho_{12}^{def} = \frac{P\left(\tilde{X}_1 = D, \tilde{X}_2 = D\right) - EDF_1.EDF_2}{\sqrt{EDF_1\left(1 - EDF_1\right)}\sqrt{EDF_2\left(1 - EDF_2\right)}} \quad \text{(A-5)}$$

The numerator of equation (A-5) represents the difference of the actual probability of both borrowers defaulting and the probability of both defaulting if they were independent. If the asset values are independent, then the default correlation is zero. The denominator reflects the standard deviation of default rates under the binomial distribution of each borrower.

The joint probability of default is

$$P\left(\tilde{X}_1 = D, \tilde{X}_2 = D\right) = \int_{-\infty}^{\Phi^{-1}(EDF_1)} \int_{-\infty}^{\Phi^{-1}(EDF_2)} N_2\left(x_1, x_2, \rho_{asset}\right) dx_2 dx_1 \quad \text{(A-6)}$$

Where: $N_2(x_1, x_2, \rho_{asset})$ is a bivariate normal distribution function.

$$P\left(\tilde{X}_1\tilde{X}_2\right) = \int_{-\infty}^{\Phi^{-1}(EDF_1)} \int_{-\infty}^{\Phi^{-1}(EDF_2)} \frac{dx_2 dx_1}{2\pi\sqrt{1-\rho_{asset}^2}} \exp\left\{-\frac{1}{2\left(1-\rho_{asset}^2\right)}\left[x_1^2 - 2x_1x_2\rho_{asset} + x_2^2\right]\right\} \quad \text{(A-7)}$$

The equation (A-7) can be simplified by using transformation

$$x_1 = u$$
$$x_2 = v\sqrt{1-\rho_{asset}^2} + u\rho_{asset} \quad \text{(A-8)}$$

We obtain

$$P\left(\tilde{X}_1 = D, \tilde{X}_2 = D\right) = \int_{-\infty}^{\Phi^{-1}(EDF_1)} \int_{-\infty}^{\frac{\Phi^{-1}(EDF_2)-\rho_{asset}u}{\sqrt{1-\rho_{asset}^2}}} \frac{dvdu}{2\pi}$$

$$\exp\left\{-\frac{1}{2\left(1-\rho_{asset}^2\right)}\left[u^2 - 2u\left(v\sqrt{1-\rho_{asset}^2}+u\rho_{asset}\right)\rho_{asset} + \left(v\sqrt{1-\rho_{asset}^2}+u\rho_{asset}\right)^2\right]\right\} \quad \text{(A-9)}$$

After arranging some terms, we get

$$P\left(\tilde{X}_1 = D, \tilde{X}_2 = D\right) = \int_{-\infty}^{\Phi^{-1}(EDF_1)} \int_{-\infty}^{\frac{\Phi^{-1}(EDF_2)-\rho_{asset}u}{\sqrt{1-\rho_{asset}^2}}} \frac{dvdu}{2\pi} \exp\left\{-\frac{\left(u^2+v^2\right)}{2}\right\} \quad \text{(A-10)}$$

The variables u and v can be separated as follows

$$P\left(\tilde{X}_1\tilde{X}_2\right) = \int_{-\infty}^{\Phi^{-1}(EDF_1)} \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{u^2}{2}\right\} \int_{-\infty}^{\frac{\Phi^{-1}(EDF_2)-\rho_{asset}u}{\sqrt{1-\rho_{asset}^2}}} \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{v^2}{2}\right\}dvdu \quad \text{(A-11)}$$

Since the inner integral is a cumulative standard normal function, the correlation of two facility-values at the horizon becomes

$$P\left(\tilde{X}_1 = D, \tilde{X}_2 = D\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\Phi^{-1}(EDF_1)} \exp\left\{-\frac{u^2}{2}\right\}\Phi\left[\frac{\Phi^{-1}\left(EDF_2\right)-\rho_{asset}u}{\sqrt{1-\rho_{asset}^2}}\right]du$$

$$P\left(\tilde{X}_1 = D, \tilde{X}_2 = D\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\Phi^{-1}(EDF_1)} \exp\left\{-\frac{x_1^2}{2}\right\}\Phi\left[\frac{\Phi^{-1}\left(EDF_2\right)-\rho_{asset}x_1}{\sqrt{1-\rho_{asset}^2}}\right]dx_1 \quad \text{(A-12)}$$

Hence, the asset correlation ρ_{asset} influences the corresponding default correlation by entering the joint probability of default term. Finally we obtain the (coefficient of) default correlation between borrower 1 and borrower 2 as follows:

$$\rho_{12}^{def} = \frac{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\Phi^{-1}(EDF_1)} \exp\left\{-\frac{x_1^2}{2}\right\} \Phi\left[\frac{\Phi^{-1}(EDF_2) - \rho_{asset}\, x_1}{\sqrt{1-\rho_{asset}^2}}\right] dx_1 - EDF_1.EDF_2}{\sqrt{EDF_1(1-EDF_1)}\sqrt{EDF_2(1-EDF_2)}} \quad \text{(A-13)}$$

Now we consider that the loss given default does not equal to one and it is variable. LGDs are assumed to be independent across borrower. The standard deviations become

$$\begin{aligned}
\sigma_1 &= \sqrt{P\left(\tilde{X}_1 = D\right)\left[1 - P\left(\tilde{X}_1 = D\right)\right]L\bar{G}D_1^2 + \sigma_{LGD}^2\, P\left(\tilde{X}_1 = D\right)} \\
&= \sqrt{EDF_1(1 - EDF_1)L\bar{G}D_1^2 + \frac{(1 - L\bar{G}D_1)L\bar{G}D_1}{4} EDF_1} \\
&= L\bar{G}D_1\sqrt{EDF_1(1 - EDF_1) + \frac{(1 - L\bar{G}D_1)}{4L\bar{G}D_1} EDF_1}
\end{aligned} \quad \text{(A-14)}$$

And

$$\begin{aligned}
\sigma_2 &= \sqrt{P\left(\tilde{X}_2 = D\right)\left[1 - P\left(\tilde{X}_2 = D\right)\right]L\bar{G}D_2^2 + \sigma_{LGD}^2\, P\left(\tilde{X}_2 = D\right)} \\
&= \sqrt{EDF_2(1 - EDF_2)L\bar{G}D_2^2 + \frac{(1 - L\bar{G}D_2)L\bar{G}D_2}{4} EDF_2} \\
&= L\bar{G}D_2\sqrt{EDF_2(1 - EDF_2) + \frac{(1 - L\bar{G}D_2)}{4L\bar{G}D_2} EDF_2}
\end{aligned} \quad \text{(A-15)}$$

Now, the covariance is

$$\begin{aligned}
\sigma_{12} &= P\left(\tilde{X}_1 = D, \tilde{X}_2 = D\right)L\bar{G}D_1.L\bar{G}D_2 - P\left(\tilde{X}_1 = D\right)L\bar{G}D_1.P\left(\tilde{X}_2 = D\right)L\bar{G}D_2 \\
&= P\left(\tilde{X}_1 = D, \tilde{X}_2 = D\right).L\bar{G}D_1.L\bar{G}D_2 - EDF_1.L\bar{G}D_1.EDF_2.L\bar{G}D_2
\end{aligned} \quad \text{(A-16)}$$

Hence using equation (A-1), (A-14), (A-15) and (A-16) we obtain the (coefficient of) correlation between the two default events

$$\rho_{12}^{def} = \frac{\int_{-\infty}^{\Phi^{-1}(EDF_1)} \int_{-\infty}^{\Phi^{-1}(EDF_2)} N_2(x_1, x_2, \rho_{asset})\, dx_2 dx_1 - EDF_1.EDF_2}{\sqrt{EDF_1(1 - EDF_1) + \frac{(1 - L\bar{G}D_1)}{4L\bar{G}D_1} EDF_1}\sqrt{EDF_2(1 - EDF_2) + \frac{(1 - L\bar{G}D_2)}{4L\bar{G}D_2} EDF_2}} \quad \text{(A-17)}$$

This equation is almost the same with equation A-5 except the standard deviation terms contain the variable loss given default.

MICHAEL RAMPISELA [2]: Business Risk Services (BRS), Ernst & Young Advisory Services, Gedung Bursa Efek Jakarta, Jl. Jend. Sudirman Kav. 52-53, Jakarta 12190. Phone: +62-21-52895521.
Email: Michael.Rampisela@id.ey.com.

[2] Any views expressed represent those of the author only and **not** necessarily those of Ernst & Young Advisory Services.

CONSISTENT MODELING OF DIVIDENDS AND FUTURES

M.H. Vellekoop[a,†] & J.W. Nieuwenhuis[b]

[a] University of Twente, The Netherlands
[b] Rijksuniversiteit Groningen, The Netherlands
[†] Corresponding author

Abstract. We propose a framework for the modeling of futures and dividends, based on the concept of tradeable securities. We show how dividends can be incorporated in a consistent manner, both for the case of continuous dividend payments, and in the case of cash dividends. We then show that future contracts may be treated as a special case of these dividend models.

Key-words: Dividends, Futures, Financial modeling.

1 Introduction

Nowadays it is perfectly understood, at least at a conceptual level, how to price European options with non-dividend paying stocks as the underlying asset [4, 7]. In this note we will clarify stock-price models *with* dividends. By introducing the economic concept of a tradeable we are able to conceptually reduce models with dividends to models without dividends.
In the standard Black-Scholes model with one stock and one bank account with fixed interest rate $r > 0$ the stock and bank account are considered to be basic tradeables. For the stock it is clear that one can trade it and as our bank account is equivalent to a zero coupon bond on a fixed time interval $[0, T]$, it is also clear that we may view our bank account as a tradeable. It is assumed that there are no transaction costs, that shortselling is allowed and that the products are perfectly divisible. For the market participants we assume that they possess a perfect memory, an assumption that is reflected in the use of the concept of filtration. Given the assumptions above, every product that can be made from these two basic tradeables by a reasonable self-financing strategy (to be defined later on) is a tradeable as well.
Introducing dividends in such models means that the ex-dividend price process of the stock can no longer be thought of as a tradeable. Indeed it is clear that nobody would like to invest in such an asset without receiving the dividend stream. It is therefore useful to investigate how dividends can be incorporated in models for markets of tradeables.
This paper will be split into three parts. In the first part we will briefly discuss continuous dividends and in the second part the general case of both continuous and discrete dividends. We will discuss futures in the last part, where after a formal definition of future price processes we can show that a future can be considered to be a special case of the dividend-carrying assets treated before.

2 Continuous Dividends

We assume given a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,T]}, P)$ where the filtration $(\mathcal{F}_t)_{t\in[0,T]}$ is the usual one associated with a given standard Brownian Motion $W : \Omega \times [0,T] \to \mathbb{R}$ where $T > 0$ is a given fixed time-horizon. Our bank account is given as

$$dB_t \;=\; r_t\, B_t\, dt$$

where $r : \Omega \times [0,T] \to \mathbb{R}_+$ is[1] a given strictly positive adapted continuous process so r is not necessarily a function of time alone (let alone a constant).
We also assume that the adapted càdlàg stochastic process $S : \Omega \times [0,T] \to \mathbb{R}_{++}$ describes the price of one unit of stock ex-dividend. We further assume the existence of an adapted stochastic process $\delta : \Omega \times [0,T] \to \mathbb{R}_+$ satisfying

$$\mathbb{E}\int_0^T \delta_u du \;<\; \infty \qquad (P-\text{a.s.})$$

The interpretation of this process is as follows. Assume you own x shares of stock during the time interval $]t_a, t_b]$ with $0 \le t_a \le t_b \le T$, then you will receive at time t_b the amount of money $x\int_{t_a}^{t_b}\delta_u du$ as dividend. It goes without saying that in our context x may be negative or noninteger as well.
We will not assume that S (the ex-dividend price process of the stock) is a tradeable and this is economically perfectly clear: possession of one unit of stock during the time interval $]t_a, t_b]$ or $]\bar{t}_a, t_b]$ where $\bar{t}_a \neq t_a$ leads to different results at time t_b. In fact, we will need to construct the tradeable $\tilde{S}$ from S.
Informally we would reason as follows. Suppose we start at time zero with $x_0 = 1$ unit of stock. When we are at time $t \in [0,T]$ we have x_t stocks. Let $\epsilon > 0$ be small. At time $t+\epsilon$ we receive $\epsilon x_t \delta_t$ money units which we immediately invest in stock so

$$x_{t+\epsilon} \;=\; x_t + \frac{\epsilon x_t \delta_t}{S_{t+\epsilon}}$$

Now let us define $\tilde{S}_t = x_t S_t$. We are then inclined to view $\tilde{S}$ as the price process of a tradeable. Now we have

$$\begin{aligned}
\tilde{S}_{t+\epsilon} &= x_{t+\epsilon}S_{t+\epsilon} \\
&= x_t S_{t+\epsilon} + x_t\delta_t\epsilon \\
&= x_t(S_t + S_{t+\epsilon} - S_t) + x_t\delta_t\epsilon \\
&= x_t S_t + x_t(\Delta\tilde{\tilde{S}}_t + \delta_t\epsilon) \\
&= \tilde{S}_t + x_t(\Delta S_t + \delta_t\epsilon)
\end{aligned}$$

or

$$\Delta\tilde{S}_t \;=\; x_t\Delta S_t + x_t\delta_t\epsilon$$

Formally we therefore proceed as follows. We are looking for a predictable adapted stochastic process $x : \Omega \times [0,T] \to \mathbb{R}$ with $x_0 = 1$ almost surely and an adapted

[1] We use the notation $\mathbb{R}_+ = \{x \in \mathbb{R} \mid x \ge 0\}$ and $\mathbb{R}_{++} = \{x \in \mathbb{R} \mid x > 0\}$

stochastic process $S : \Omega \times [0,T] \to \mathbb{R}$ such that the following equations are satisfied simultaneously:

$$\tilde{S}_t = \tilde{S}_0 + \int_0^t x_u d(S_u + D_u) \tag{2.1}$$

$$\tilde{S}_t = x_t S_t \tag{2.2}$$

where

$$D_t = \int_0^t \delta_u du$$

defines the cumulative dividend process, and where we assume the above equations to be well-defined, i.e. $S + D$ is a semi-martingale.
Our economically motivated intuition says that in an arbitrage-free market model there is precisely one predictable adapted process x such that the equations above are satisfied, and we will now show that this intuition is correct. We will use the well-known fact [4, 7] that in an arbitrage-free market containing $\tilde{S}$ and B, there exists at least one measure $\mathbb{Q}$ such that $\tilde{S}/B$ is a martingale under $\mathbb{Q}$.

Theorem 2.1 *Assume that $S + D$ is a semi-martingale. Then there exists a unique process x such that $(\tilde{S}, B)$ is arbitrage-free while the equations (2.1)-(2.2) above are satisfied.*

Proof. We assumed that $S+D$ is a semi-martingale and as D is a semi-martingale as well it follows that S is also a semi-martingale. From the general theory of stochastic integration it follows that $\tilde{S}$ is a semi-martingale too. As S is assumed to be strictly positive and cadlag it follows from Ito's lemma and a localization argument (see for example Protter [8]) that $1/S$ is a semi-martingale and hence x must be a semi-martingale too.
The pair $(\tilde{S}, B)$ is assumed to be arbitrage-free and we therefore know that $\tilde{S}/B$ is a martingale under $\mathbb{Q}$. But this implies, as we work in a Brownian setting, that $\tilde{S}$ is a *continuous* semi-martingale. But this means that $S + D$ is continuous (see for example [8]) and since D is continuous, S has to be continuous too. And since

$$\begin{aligned} [S,x]_t &= S_t x_t - \int_0^t S_u dx_u - \int_0^t x_u dS_u \\ &= S_0 x_0 - \int_0^t S_u dx_u + \int_0^t x_u dD_u \end{aligned}$$

by equation (2.1)-(2.2) so this shows that $\int_0^t S_u dx_u$, and hence x, has finite variation. Applying Ito's rule to $S = Sx$ and using the fact that x has finite variation we find

$$dS_t = S_t dx_t + x_t dS_t$$

and combining this with

$$dS_t = x_t dS_t + x_t dD_t$$

we find that

$$dx_t = \frac{x_t}{S_t} dD_t$$

As D is non-decreasing, so is x. From the general theory of stochastic differential equations it follows that

$$\begin{aligned} dx_t &= \frac{x_t}{S_t} dD_t \\ x_0 &= 1 \end{aligned}$$

has a unique strong predictable solution on $[0, T]$ (see for example Protter [8]). In fact

$$x_t = e^{\int_0^t \frac{\delta_u}{S_u} du}$$

which completes the proof. ■

Note that a proof of existence for the process x_t could easily be settled using the last few equations, but it is the *uniqueness* which is of particular interest here.

3 Continuous and Discrete Dividends

We would now like to be able to define on the same probability space with Brownian filtration an asset process which may pay a discrete (i.e. cash) dividend equal to $\tilde{D}$ on time $t_D \in]0, T[$ where $\tilde{D} \in \mathcal{F}_{t_D}$ and such that

$$S_{t_D -} - \tilde{D} > 0 \qquad (P - \text{a.s.})$$

where S again describes the ex-dividend process. In fact, we would even like to consider cases where an asset pays both continuous and discrete dividends.
At this point we note that consistent treatment of dividends may still lead to different models under different assumptions. Quite often, nice properties of models without dividends (such as the lognormal distributions in the standard Black-Scholes model) disappear when the ex-dividend processes are adjusted with cash dividends. As a result, certain approximations have been proposed in the literature to find option prices when the underlying asset pays cash dividends in the future.(see for example the overview in [5]). Possibilities include the assumption that the asset price minus the present value of all dividends to be paid until the maturity of the option follows a Geometric Brownian Motion (the so called *Escrowed Model*), or that the asset price plus the forward value of all dividends (from past dividend dates to today) follows a Geometric Brownian Motion (the *Forward Model*). Simple computations show that in the Escrowed and the Forward Model the prices for European options on dividend-paying assets reduce to the equivalent European prices on assets without dividend payments, but with an adjusted value of the current stock price or strike, respectively. Moreover, for American options it is quite easy to adjust computation methods for assets without dividends to those which include dividends. This seems to be the reason why these approaches have

been popular among practitioners. There are, however, serious modeling problems for these approaches. The Escrowed Model formulated above admits arbitrage opportunities for American options [1]. The reason for this is obvious: different asset price process dynamics are assumed for products up until the first dividend date. One can fix this by changing the definition to an assumption that the asset price minus the present value of all dividends to be paid in the future follows a Geometric Brownian Motion (an *Adjusted Escrowed Model*) but this would mean that the prices of options will depend on the dividends which are being paid after the options have expired which may be unsatisfactory as well, since this means that a trader would have to adjust the price of a two-year option once his view on the five-year dividend prediction changes [9]. All this exemplifies the need for a consistent framework to model cash dividends.

Let V_t, S_t and B_t be cadlag ex-dividend price processes for assets V, S and B which are strictly positive and let D_t^V, D_t^S and D_t^B denote the corresponding right-continuous cumulative dividend processes which are increasing (and hence of finite variation) and such that $S_t + D_t^S$ and $B_t + D_t^B$ are both **continuous** semi-martingales. We will assume that $D_0^S = D_0^B = D_0^V = 0$ throughout the paper.

We would like to define the notion of *replicability* i.e. the idea that the price process of a certain asset V can be mimicked by trading in other assets.

Definition 3.1 *We say that an asset V can be replicated using assets S and B iff there exist adapted and predictable processes ϕ^S and ϕ^B such that for all $t \in [0, T]$*

$$V_{t-} = \phi_t^S S_{t-} + \phi_t^B B_{t-} \tag{3.1}$$

$$d(V_t + D_t^V) = \phi_t^S d(S_t + D_t^S) + \phi_t^B d(B_t + D_t^B) \tag{3.2}$$

where the first equation for $t = 0$ should be read as $V_0 = \phi_0^S S_0 + \phi_0^B B_0$ (i.e. without taking left-hand side limits).

In this paper we will always use the above definition with $D^V \equiv 0$ in (3.2) so one could argue that it is more natural to leave this term out, but we keep the possibility that V has its own dividend process open here, to emphasize the symmetry in the replication formula.
Note that for continuous processes without dividends we find the classical definition of replication back, but the lefthand side limits in the first equations are an important difference compared to the case without dividends. Indeed we can no longer say that

$$V_t = \phi_t^S S_t + \phi_t^B B_t$$

as in the usual formulations in the absence of dividends, but instead

$$V_t + \Delta D_t^V = \phi_t^S (S_t + \Delta D_t^S) + \phi_t^B (B_t + \Delta D_t^S)$$

which is of course a reformulation of (3.1) since $X_t + \Delta D_t^X = X_{t-}$. This continuity property needs to be *assumed* (for obvious economic reasons) for our assets S and B but can then be *derived* for V, using results from stochastic calculus.

If we define

$$\phi_t^S = \psi_t^S V_{t-}/S_{t-} \tag{3.3}$$
$$\phi_t^B = \psi_t^B V_{t-}/S_{t-} \tag{3.4}$$

then

$$\frac{dV_t + dD_t^V}{V_{t-}} = \psi_t^S \frac{dS_t + dD_t^S}{S_{t-}} + \psi_t^B \frac{dB_t + dD_t^B}{B_{t-}}$$

for certain predictable adapted processes ψ^S and ψ^B such that

$$\psi_t^S + \psi_t^B = 1$$

The interpretation is that the rate of return of V (which equals the difference in value based on changes in *both* the ex-dividend price *and* the dividends, divided by the price *before* any dividends have been paid out) is based on percentages invested in assets S and B. Working with percentages guarantees in an intuitive manner that we only consider strategies which do not necessitate cash withdrawal or injection, i.e. it is a convenient way to define self-financing strategies. However, our definition above is slightly more general in the sense that it allows the price processes becoming zero for certain times as well.
Throughout the paper we will assume $D^B = 0$ i.e. our bank account does not pay dividends (or coupons), only interest.

Theorem 3.1 *Let $D_t^S = \tilde{D}\mathbf{1}_{[t_D,T]}(t)$. Then there exists an asset price process V with $D_t^V = 0$ such that (1) V is a continuous semi-martingale, and (2) V can be replicated using S only, i.e. such that $\phi^B \equiv 0$.*

Proof. We show that the process $V = \tilde{S}$ with

$$\tilde{S}_t = S_t + \mathbf{1}_{[t_d,T]}(t)\frac{\tilde{D}}{S(t_D)}S_t$$

satisfies the requirements. Since $S + D^S$ and D^S are semi-martingales, so is S and hence $\tilde{S}$. First notice that

$$t < t_D : \quad \tilde{S}_t = S_t \qquad \tilde{S}_{t-} = S_{t-} \tag{3.5}$$
$$t = t_D : \quad \tilde{S}_{t_D} = S_{t_D} + \tilde{D} \qquad \tilde{S}_{t_D-} = S_{t_D-} \tag{3.6}$$
$$t > t_D : \quad \tilde{S}_t = S_t(1 + \tfrac{\tilde{D}}{S_{t_D}}) \qquad \tilde{S}_{t-} = S_{t-}(1 + \tfrac{\tilde{D}}{S_{t_D}}) \tag{3.7}$$

which shows that $V = \tilde{S}$ is indeed left-continuous (and hence continuous), because $S_t + \Delta D_t^S = S_{t-}$ implies that $S_{t_D} + \tilde{D} = S_{t_D-}$ and also that $S_t = S_{t-}$ when $t \neq t_D$. Since we would like to have $\psi^B = \phi^B \equiv 0$ we must have $\psi^S = 1$ and we see from (3.3) that it is thus left to prove that

$$\frac{d\tilde{S}_t}{\tilde{S}_{t-}} = 1\frac{d(S_t + D_t^S)}{S_{t-}} \tag{3.8}$$

for all $t \in [0,T]$. But we have for $t < t_D$ that

$$d\tilde{S}_t \;=\; dS_t \;=\; dS_t + dD_t^S \;=\; \frac{\tilde{S}_{t-}}{S_{t-}} d(S_t + D_t^S)$$

as required, where we have used (3.5) and the definition of D_t^S. For $t = t_D$ we find, using (3.6)

$$d\tilde{S}_t \;=\; dS_t + dD_t^S \;=\; \frac{\tilde{S}_{t-}}{S_{t-}} d(S_t + D_t^S)$$

and finally, for $t > t_D$

$$\begin{aligned} d\tilde{S}_t &= dS_t + \frac{\tilde{D}}{S_{t_D}} dS_t \;=\; (1 + \frac{\tilde{D}}{S_{t_D}}) d(S_t + D_t^S) \\ &= \frac{\tilde{S}_{t-}}{S_{t-}} d(S_t + D_t^S) \end{aligned}$$

so we are done. ■

The approach taken in the previous theorem formalizes the idea that we would like to reinvest dividend payouts in the asset which pays the dividends. If we put the dividend proceeds in the bank account instead we would expect to create a value process

$$S_t^B \;=\; S_t + \mathbf{1}_{[t_d,T]}(t) \tilde{D} \frac{B_t}{B_{t_D}}$$

and this process can indeed also be replicated from S and B.

Corollary 3.1 *The asset price process S^B is a continuous semi-martingale which can be replicated from S and B.*

Proof. We have

$$\begin{aligned} S_{t_D}^B &= S_{t_D} + \tilde{D} \\ S_{t_D-}^B &= S_{t_D-} \\ S_{t_D+}^B &= S_{t_D+} + \tilde{D} \end{aligned}$$

so we have continuity for $t = t_D$ and in all other points S and B are both continuous, so there S^B is continuous too. It is a semi-martingale since $S + D$ was assumed to be a semi-martingale while $S + D - S^B$ is clearly of bounded variation. We therefore need to find ψ^S and ψ^B such that

$$\frac{dS_t^B}{S_t^B} \;=\; \psi_t^S \frac{d(S_t + D_t^S)}{S_{t-}} + \psi_t^B \frac{dB_t}{B_t}$$

which by (3.8) is equivalent to

$$\frac{dS_t^B}{S_t^B} \;=\; \psi_t^S \frac{d\tilde{S}_t}{\tilde{S}_t} + \psi_t^B \frac{dB_t}{B_t} \tag{3.9}$$

If we choose

$$\psi_t^S = \frac{\tilde{S}_t}{S_t^B} \tag{3.10}$$

$$\psi_t^B = \frac{B_t}{B_{t_D} S_t^B} \tilde{D} \mathbf{1}_{]t_d,T]}(t) \tag{3.11}$$

the analysis goes through just like in the previous proof. ■

Notice that

$$\lim_{t \downarrow t_D} \psi_t^B S_t^B = \tilde{D}$$

which means that we indeed invest the dividend payout in the bank account in this case. Also note that

$$\psi_{t_D}^B S_{t_D}^B \neq \tilde{D}$$

since ψ^B is left-continuous but not right-continuous in t_D.

Corollary 3.2
If an asset can be replicated using the assets S and B, then it can be replicated using the assets $\tilde{S}$ and B.
If an asset can be replicated using the assets S and B, then it can be replicated using the assets S^B and B.

Proof. The first statement can be concluded immediately from (3.8). For the second one notice that we are looking for adapted and predictable ζ^S and ζ^B such that

$$\begin{aligned} dV_t &= V_t \zeta_t^S \frac{dS_t^B}{\tilde{S}_t^B} + V_t \zeta_t^B \frac{d\tilde{B}_t}{B_t} \\ 1 &= \zeta_t^S + \zeta_t^B \end{aligned}$$

while from the first part of the corollary we know that there exist adapted and predictable ξ_t^S and ξ_t^B such that

$$\begin{aligned} dV_t &= V_t \xi_t^S \frac{d\tilde{S}_t}{\tilde{S}_t} + V_t \xi_t^B \frac{dB_t}{B_t} \\ 1 &= \xi_t^S + \xi_t^B \end{aligned}$$

But (3.9) shows that

$$\frac{d\tilde{S}_t}{\tilde{S}_t} = \frac{1}{\psi_t^S} \frac{dS_t^B}{S_t^B} - \frac{\psi_t^B}{\psi_t^S} \frac{dB_t}{B_t} \tag{3.12}$$

so the result follows by taking

$$\begin{aligned} \zeta_t^S &= \tfrac{\xi_t^S}{\psi_t^S} \\ \zeta_t^B &= \xi_t^B - \xi_t^S \tfrac{\psi_t^B}{\psi_t^S} \end{aligned}$$

since these sum to one, as required. ■

We now consider an arbitrage-free market with the assets $(\tilde{S}, B)$ in it. We know that there exists a measure $\mathbb{Q}$, equivalent to our original measure P, such that $\tilde{S}/B$ is a martingale under $\mathbb{Q}$.

Definition 3.2 *We say that V is the price process of a tradeable asset iff*

1. *It can be replicated using $\tilde{S}$ and B, and*
2. *V/B is a martingale under $\mathbb{Q}$.*

The second part of the definition is needed since the first part only implies that V/B is a *local* martingale, as the following theorem shows.

Theorem 3.2 *If a continuous asset price process V can be replicated using $\tilde{S}$ and B then there exists a unique adapted predictable process ϕ such that*

$$d\frac{V_t}{B_t} = \phi_t d\frac{\tilde{S}_t}{B_t}$$

Proof. Since V, $\tilde{S}$ and B are all continuous processes we may apply Ito's rule to find

$$d\frac{\tilde{V}_t}{B_t} = \frac{d\tilde{V}_t}{B_t} - \frac{\tilde{V}_t}{B_t}\frac{dB_t}{B_t}$$

and using the fact that V can be replicated we can rewrite this as

$$\begin{aligned} d\frac{\tilde{V}_t}{B_t} &= \frac{\tilde{V}_t}{B_t}\left(\psi_t^{\tilde{S}}\frac{d\tilde{S}_t}{\tilde{S}_t} + \psi_t^B\frac{dB_t}{B_t}\right) - \frac{\tilde{V}_t}{B_t}\frac{dB_t}{B_t} \\ &= \frac{\tilde{V}_t}{B_t}\left(\psi_t^{\tilde{S}}\frac{d\tilde{S}_t}{\tilde{S}_t} + (\psi_t^B - 1)\frac{dB_t}{B_t}\right) \end{aligned}$$

but since $\psi^{\tilde{S}} + \psi^B = 1$ this gives

$$\begin{aligned} d\frac{\tilde{V}_t}{B_t} &= \frac{\tilde{V}_t}{\tilde{S}_t}\psi_t^{\tilde{S}}\left(\frac{d\tilde{S}_t}{\tilde{S}_t} - \frac{\tilde{S}_t}{B_t}\frac{dB_t}{B_t}\right) \\ &= \frac{\tilde{V}_t}{\tilde{S}_t}\psi_t^{\tilde{S}} d\frac{\tilde{S}_t}{B_t} \end{aligned}$$

which proves the existence result if we take

$$\phi_t = \frac{\tilde{V}_t}{\tilde{S}_t}\psi_t^{\tilde{S}}$$

Uniqueness now follows because if both ϕ^1 and ϕ^2 satisfy the requirements, then $\int_0^t(\phi_u^1 - \phi_u^2)d\frac{\tilde{S}_u}{B_u}$ equals zero for all t which implies that $\phi_t^1 - \phi_t^2$ equals zero for almost all t. ■

We will know show how the framework developed so far can be used to price futures.

4 Modeling Futures

We will use the following definition.

Definition 4.1 *We call $m : \Omega \times [0,T] \to \mathbb{R}$ the futures price process associated with delivery of asset S at time T if the following holds:*

- *m is a semi-martingale and $m_T = S_T$*
- *For* **all** *previsible processes ψ the following process M is a tradeable:*

$$\begin{cases} dM_t &= M_t \frac{dB_t}{B_t} + \psi_t dm_t \\ M_0 &= 0 \end{cases} \tag{4.1}$$

Notice that delivery involves the ex-dividend price, and not the price of the tradeable $\tilde{S}$.
We will use the notation M^ψ for the process M to remind ourselves that it depends on the process ψ. Note that the process ψ in the definition above has the interpretation of a futures trading strategy: ψ_t represents the number of futures contracts in our position at time t. Our definition reflects the fact that we may enter the futures market at any time at zero costs. What we do is to 'invest' the proceeds of the futures strategy ψ into the so-called *margin-account* M which earns the riskfree rate.
This approach is different from the usual one (see for example [2]) where margin accounts are never taken into account explicitly. Our treatment here is inspired by the paper by Pozdnyakov and Steele on the martingale framework for futures pricing [6], but our definition differs from theirs. We only impose that m is such that M^ψ/B is a $\mathbb{Q}$-martingale on $[0,T]$ (i.e. that M^ψ is a tradeable in economic parlance) and we do *not* need to impose any regularity conditions on m from the start. Another difference with the approach in [6] is that we introduce a whole *collection* of tradeables from the very beginning and this is completely in line with the fact that one may enter a futures contract at any time in real life.

The following is then immediate:

Theorem 4.1 *The margin account process can be replicated using a zero ex-dividend process with pays continuous dividends equal to the futures price.*

Proof. Taking $\phi^S_t = \psi_t,\ \phi^B_t = M_t/B_t$ and $S_t = 0,\ D^S_t = m_t$ replicates $V_t = M_t$ with $D^V_t = 0$, see equations (3.1)-(3.2). ■

We now set out to explore the ramifications of our definition.

Theorem 4.2 *We have for all $t \in [0,T]$,*

$$m_t \ = \ \mathbb{E}^{\mathbb{Q}}[S_T \mid \mathcal{F}_t].$$

Proof. Take ψ equal to the previsible process B and define

$$\tilde{M}_t \;=\; \frac{M_t^\psi}{B_t}.$$

Since M^ψ is a tradeable we must have that $\tilde{M}$ is a $\mathbb{Q}$-martingale on the Brownian filtration, and since B is of finite variation we find that

$$d\tilde{M}_t \;=\; \frac{dM_t^\psi}{B_t} \;-\; \frac{M_t^\psi\, dB_t}{B_t^2} \;=\; \psi_t \frac{dm_t}{B_t} \;=\; dm_t$$

Hence m is a martingale under $\mathbb{Q}$ and since $m_T = S_T$ we then automatically find the result. ■

Notice that the last equation shows that if the number of futures in our possession equals the bank account process (i.e. $\psi_t = B_t$ for all $t \in [0,T]$) then

$$M_T^\psi / B_T - M_0^\psi \;=\; S_T - m_0.$$

In fact, (4.1) allows the explicit solution

$$M_t^\psi \;=\; M_0^\psi + \int_0^t \psi_u e^{\int_u^t r_v\, dv} dm_u$$

for all $0 \le t \le T$. This shows why futures can be used to hedge contingent claims based on future values of the stock price process S.

Example: Hedging using futures.
If we want to hedge a contingent claim $X \in \mathcal{F}_T$ then the price process C_t of this claim should satisfy

$$\begin{aligned} C_t &= \phi_t^S S_t \;+\; \phi_t^B B_t \\ dC_t &= \phi_t^S dS_t \;+\; \phi_t^B dB_t \end{aligned}$$

for certain predictable processes ϕ^S and ϕ^B. If we want to use the future on S with delivery time T instead of the asset S itself, we can use the tradeable M^ψ so we would like to have

$$\begin{aligned} C_t &= M_t^\psi \;+\; \xi_t^B B_t \\ dC_t &= dM_t^\psi \;+\; \xi_t^B dB_t \end{aligned}$$

for a certain predictable processes ξ^B. But using our definition of a marigin account this gives

$$\begin{aligned} dC_t &= M_t \tfrac{dB_t}{B_t} \;+\; \psi_t dm_t \;+\; \xi_t^B dB_t \\ &= C_t \tfrac{dB_t}{B_t} \;+\; \psi_t dm_t \end{aligned}$$

and we recognize in this equation the usual formulation of the dynamics of a hedging strategy involving futures, without marigin accounts. Ito calculus then gives that

$$d\tfrac{C_t}{B_t} \quad = \quad \tfrac{\psi_t}{B_t} dm_t \tag{4.2}$$

while

$$\xi^B_t \quad = \quad \tfrac{C_t}{B_t} \; - \; \tilde{M}^\psi_t$$

from which the strategies ψ and ξ^B can be derived using martingale representation theorems. For example, in the Black-Scholes case where under $\mathbb{Q}$

$$d\tfrac{S_t}{B_t} \quad = \quad \tfrac{S_t}{B_t} \sigma \, dW_t$$

we find using the previous theorem that

$$\begin{aligned} m_t \quad &= \quad \mathbb{E}^{\mathbb{Q}}[S_T \mid \mathcal{F}_t] \; = \; S_t B_T / B_t \\ dm_t \quad &= \quad B_T d\tfrac{S_t}{B_t} \end{aligned}$$

but in the Black-Scholes case

$$d\tfrac{C_t}{B_t} \quad = \quad f(S_t, t) d\tfrac{S_t}{B_t}$$

for a certain function f which can be calculated explicitly in the Black-Scholes Model (the so-called contingent claim *Delta value*). So (4.2) shows that the correct hedging strategy using futures is given by

$$\psi_t \quad = \quad \tfrac{B_t}{B_T} f(S_t, t)$$

5 Conclusions

We have shown how dividends and futures can be modeled consistently within the framework of tradeable securities. At some points in the derivations we made explicit use of the strong properties (for example, continuity) of martingales on a Brownian filtrations. An obvious interesting research direction would be a generalization to other filtrations, for example those generated by jump-diffusions, and we hope to address such questions in future work.

References

[1] Beneder, R. and Vorst, T. (2002), Options on dividend paying stocks. In *Recent developments in Mathematical Finance*, Shanghai, 2001, World Science Publishing, River Edge, NJ, 204 – 217.

[2] Björk, T. (2004), *Arbitrage Theory in Continuous Time* (2nd edition), Oxford University Press, Oxford.

[3] Black, F. and Scholes, M. (1973) The pricing of options and corporate liabilities. *Journal of Political Economy*, **81**, 637 - 654.

[4] Duffie, D. (2001), *Dynamic Asset Pricing Theory* (3rd edition), Princeton University Press, London.

[5] Frishling, V. (2002), A Discrete Question. *RISK magazine*, **15**.

[6] Pozdnyakov, V. and Steele, M. (2004), On the Martingale Framework for Futures Prices, *Stochastic Processes and Their Applications*, **1**09 , 69–77.

[7] Musiela, M. & Rutkowski, M. (1997), *Martingale Methods in Financial Modelling*, Springer, New York.

[8] Protter, P. (1997), *Stochastic Integration and Differential Equations* (2nd edition), Springer, New York.

[9] Vellekoop, M.H. and Nieuwenhuis, J.W. (2004), Efficient pricing of derivatives on assets with discrete dividends. Working Paper, submitted for publication.

Michel Vellekoop: Department of Mathematics, University of Twente, P.O. Box 217, 7500 AE, Enschede, the Netherlands.
E-mail: m.h.vellekoop@ewi.utwente.nl

Hans Nieuwenhuis: Faculty of Economics, University of Groningen, P.O. Box 800, 9700 AV, Groningen, The Netherlands.
E-mail: j.w.nieuwenhuis@eco.rug.nl

Modelling Attitude at University of Western Australia

Neville Fowkes

University of Western Australia

Abstract: Modelling is as much an art as a science, and, whilst there are good modelling books around, nothing beats personal experience from the point of view of communicating the skills and attitudes required. Thus, if at all possible, staff as well as students should be actively involved in the process of modelling.

I'll briefly describe my thoughts on these issues. Then I shall discuss a range of problems that I have worked on in recent years that have come out of the food industry that involve modelling the relevant physical and chemical processes.

Breakfast cereal (Uncle Toby's): Is it possible to improve the efficiency of the cooking process by first soaking the grain before cooking?

Beer production (BrewTech): Sometimes there is inefficient or incomplete conversion of starch in grain to alcohol.

Ice cream (Brownes): Temperature variations can cause ice cream to become 'icey'. Why?

Labelling (Southcorp): Sometimes bubbles appear in wine labels. Can one improve the operation of the labelling machine?

Teaching Mathematical Modelling at UTM

Norsarahaida S. Amin, Zainal Abdul Aziz,

Univ Techn Malaysia, 81310 UTM Skudai, Johor

Abstract: This presentation provides an insight into the teaching of mathematical modelling aimed at stimulating interest in mathematics and its applications. Mathematical modelling requires strong interaction between the abstract concepts of a mathematical system and their physical aspects that are being modelled. Thus it is imperative that teaching methods enlighten and motivate students to appreciate the interrelation between mathematics and the real problems of physical phenomena. An innovative and resourceful approach to teaching mathematical modelling will be suggested to encompass the need to inspire and enrich the mathematical awareness and acumen among students.

Mathematical Modeling Course; *bringing real world problems in class room activity*

Edy Soewono, Kuntjoro A. Sidarto

Department of Mathematics, Institut Teknologi Bandung

Abstract: University staff in engineering departments as well as in science or mathematics departments nowadays is being challenged with a difficult responsibility: *how to train students to deal with complex problems which they may face in the working environment*. Many staffs acknowledge that they are only able to give standard text book examples, which are far from simulating real world situations, in their routine courses.

Strong demand in the job market requires graduates to comprehend several real life aspects in problem solving such as ability to work in a team, collaborate with non-math people, identify and formulate problems and select proper tools to attack the problems. Mathematical Modelling course not only offers all those aspect, but also motivate students to struggle and show all their effort which may improve their appreciation toward mathematics. This type of course should be put as one of key courses in the mathematics curriculum in line with recommendation of CUPM 2004.

This course has been offered as a compulsory course in the fourth year at the mathematics department ITB. Six staffs volunteer to develop the course which requires several months of preparation for selecting problems from surrounding ITB, industries and MCM/ICM. The experience shows that students enjoy the challenge and show no fear to deal with real and complex problems which they never know and never being trained how to solve. This positive attitude of students which has nothing to do with their breadth of mathematical background is unfortunately not easy to be shared among the staff. Long training (only) in mathematical rigor and formal approach turns out to be a potential barrier in appreciating a non-formal working process like mathematical modelling. This is the more difficult challenge for mathematical community in overcoming the big problem in teaching such as decrease of qualified students in mathematics departments and lacking of motivation in mathematics.

On the Initial Sea Surface Fields that Led to the Indian Ocean Tsunami using GPS measurements in Southeast Asia

Julie Pietrzak[1], Anne Socquet, Wim Simons, David Ham, Christophe Vigny, Robert Jan Labeur, Ejo Schrama, Jurjen Battjes, Guus Stelling and Deepak Vatvani

Environmental Fluid Mechanics Section, CiTG, TUDelft Netherlands

Abstract: Data collected at ~60 Global Positioning System (GPS) sites in southeast Asia are used to determine the bed displacements (uplift) that occurred in the Indian Ocean during the 26 December 2004 Sumatra-Andaman earthquake. This data are then used as the initial surface displacement fields in a numerical model of the tsunami. The GPS data are first inverted in order to determine the slip and then vertical bed displacements are derived from these results. However, the inversion process requires knowledge of the underlying fault geometry. Two realisations of the slip and corresponding uplift are therefore presented. Modelled tsunami arrival times were compared to satellite and coastal tide gauge data.

A first inversion of the GPS data based on local fault maps could be available within 30 minutes. Here we present a novel use of GPS data and demonstrate how that data could be an important component of a future tsunami warning systems. The effects of surface wave dispersion are discussed.

1883 Krakatau Volcano Tsunami; Facts and Modelling

Efim Pelinovsky[1], Irina Didenkulova[1] and Byung Ho Choi[2]

[1)] Laboratory of Hydrophysics, Institute of Applied Physics, Nizhny Novgorod, Russia
[2)] Department of Civil and Environmental Engineering, Sungkyunkwan University, Suwon, Korea

Abstract: The 1883 Krakatau volcanic eruption has generated a destructive tsunami higher than 40 m on the Indonesian coast where more than 36,000 lives were lost. Sea level oscillations related with this event have been reported on significant distances from the source in the Indian, Atlantic and Pacific Oceans. Tide-gauge records of the Krakatau event are digitized and analyzed. The problem of tsunami source of volcanic origin is discussed. The worldwide propagation of the tsunami waves generated by the Krakatau volcanic eruption is studied numerically using two conventional models: ray tracing method and two-dimensional linear shallow-water model. The results of the numerical simulations are compared with available data of the tsunami registration.

Non Hydrostatic Computation of Wave Run-Up on Slopes

G.S. Stelling[1], M. Zijlema

[1)] TU Delft, The Netherlands

Abstract: Stelling and Duinmeyer [1], propose a numerical technique that in essence is based upon the classical staggered grids and implicit numerical integration schemes, but that can be applied to problems that include rapidly varied flows as well. Rapidly varied flows occur, for instance, in hydraulic jumps and bores. Also inundation of dry land implies sudden flow transitions. Here conservation properties become crucial. The numerical method combines the efficiency of staggered grids with conservation properties so as to ensure accurate results for rapidly varied flows, as well as in expansions as in contractions. The resulting method is very efficient for the simulation of large-scale inundations.

In Stelling and Zijlema [2] a numerical technique is presented for the approximation of vertical gradient of the non-hydrostatic pressure arising in the Reynolds-averaged Navier-Stokes equations for simulating non-hydrostatic free-surface flows. It is based on the Keller-box method that takes into account the effect of non-hydrostatic pressure with a very small number of vertical grid points. As a result, the proposed technique is capable of simulating relatively short wave propagation, where both frequency dispersion and non-linear effects play an important role, in an accurate and efficient manner. Numerical examples show that accurate wave characteristics are already achieved with only two layers.

The present contribution combines the approaches described in [1] and [2]. The resulting method is very efficient to describe wave run up on slopes. The major energy loss follows from the momentum conservation of the applied method. Some additional energy loss is follows from a simple turbulence model. Examples, consisting of comparisons with measurements, will be given.

References:

[1] Stelling, GS, Duinmeijer, S.P.A. A staggered conservative scheme for every Froude number in rapidly varied shallow water flows. *International journal for numerical methods in fluids 43*, 1329-1354, 2003

[2] Stelling G, Zijlema M. An accurate and efficient finite-difference algorithm for non-hydrostatic free-surface flow with application to wave propagation, *International journal for numerical methods in fluids* 43 (1): 1-23 SEP 10 2003

Robust optimization models in investment science

Shuzhong Zhang

Dept. of Systems Engineering & Engineering Management, The Chinese University of Hong Kong

Abstract: In this talk we shall discuss several robust optimization formulations arising from investment problems. Investment, by its very definition, is a science dealing with uncertainties. One part of the uncertainty is related to the knowledge about the distributions of the random variables. We shall discuss how robust optimization can help in dealing with this type of uncertainties.

Optimal Codes of Length n and Maximal Distance m

A.J van Zanten

Delft University of Technology, The Netherlands

Abstract

A short induction proof is presented of a well-known expression for the maximal size of a binary anticode of length n and maximum distance m. This expression was first proved by Kleitman, in the context of extremal set theory, while Katona had already derived an equivalent result in an earlier paper.

1 Introduction

In algebraic coding theory one usually studies binary (n, M, d)-codes. An (n, M, d)-code is a set of M binary words of length n and minimal Hamming distance d. Optimality questions arise when two of the three parameters are fixed. More generally one can define such optimality problems for q-ary codes over an alphabet $\{0, 1, , q-1\}$, where the Hamming distance between two words is defined as the number of positions in which they differ. Occasionally, one studies binary codes with a maximum for the distance between two codewords, and one speaks of *anticodes* (cf. e.g. [3, 6, 7]). In particular one can ask for the maximal size $N(n, m)$ of an anticode with codeword length n and maximum distance m. In [7] it was conjectured that this maximal size is equal to

$$\begin{aligned} N(n,m) &= \sum_{i=0}^{k} \binom{n}{i}, \text{for } m \text{ even}, \\ N(n,m) &= \sum_{i=0}^{k} \binom{n}{i} + \binom{n-1}{k}, \text{for } m \text{ odd}, \end{aligned}$$

where $k = \lfloor \frac{m}{2} \rfloor$, and $1 \leq m < n$. The r.h.s. of these equalities are equal to the contents of a ball with radius $\frac{m}{2}$, with an additional term for odd values of m. Although this result seems quite obvious, an immediate proof was not at hand. Actually, it was already Erdös who posed this question (for even m), and which was proved later by Kleitman in [5], and also by Katona in [4], in a slightly different setting, in terms of extremal set theory. We shall present a relatively short induction proof in terms of binary words, based on the well-known addition theorem for binomial coefficients. In Section 3 we briefly indicate how our method possibly can be used to obtain a similar result for q-ary codes, as derived by Ahlswede, Cai and Zhang in [1] and by Ahlswede and Khachiatran in [2].

2 Proof of The Theorem

Let C be a set of $N := |C|$ binary words of length n. We shall call C a binary code of length n and of size N. As usual, the *Hamming distance* $d(\mathbf{v}, \mathbf{w})$ between two words $\mathbf{v}$ and $\mathbf{w}$ of C is defined as the number of positions j, $1 \leq j \leq n$, where $\mathbf{v}$ and $\mathbf{w}$ differ, i.e. where $v_j \neq w_j$. The maximal distance in C is called the diameter $d(C)$ of the code, so

$$d(C) := max\{d(\mathbf{v}, \mathbf{w}) | \mathbf{v}, \mathbf{w} \in C\}. \tag{1}$$

Let $N(n,m)$ be the maximal possible size of a binary code of length n and diameter m, $0 \leq m \leq n$. We define $\mathcal{F}$ as the family of binary codes of length n, diameter m and with size equal to $N(n,m)$. For any code $C \in \mathcal{F}$ we introduce the subcodes

$$C_i := \{\mathbf{v} | \mathbf{v} \in C, v_n = i\} \tag{2}$$

for $i \in \{0,1\}$.

Proposition 2.1 *For any n and m, $0 < m < n$, there exists a code $C \in \mathcal{F}$, such that $d(\mathbf{v},\mathbf{w}) \leq m-2$ for all $\mathbf{v}$, $\mathbf{w} \in C_1$.*

Proof. Take an arbitrary code $B \in \mathcal{F}$. If B satisfies the above inequality we define $C := B$. If B does not satisfy the inequality of the Proposition, we shall transform B into another code of $\mathcal{F}$ which does satisfy this inequality.

A. First we transform B into a code $D \in \mathcal{F}$, such that D_1 does not contain a pair of words at mutual distance m. Let $S \subseteq B_1$ be defined as

$$S = \{\mathbf{x} | \mathbf{x} \in B_1, \exists \mathbf{y} \in B_1 \text{ with } d(\mathbf{x},\mathbf{y}) = m\}. \tag{3}$$

It will be clear that either $S = \emptyset$ or $|S| \geq 2$. If $S \neq \emptyset$ we define for each $\mathbf{x} \in S$ a word $\mathbf{x}'$ with $x'_j = x_j, 1 \leq j < n$, and $x'_n = 0$. None of these words $\mathbf{x}'$ is in B, since the condition $d(\mathbf{x},\mathbf{y}) = m$ in (3) implies $d(\mathbf{x},\mathbf{y}) = m+1$. Furthermore, we have for any $\mathbf{x}'$

$$\begin{aligned} d(\mathbf{x}',\mathbf{v}) &= d(\mathbf{x},\mathbf{v}) - 1 \leq m-1, \text{ if } \mathbf{v} \in B_0, & (4)\\ d(\mathbf{x}',\mathbf{v}) &= d(\mathbf{x},\mathbf{v}) + 1 \leq (m-1)+1 = m, \text{ if } \mathbf{v} \in B_1 \backslash S, & (5)\\ d(\mathbf{x}',\mathbf{v}') &= d(\mathbf{x},\mathbf{v}) \leq m, \text{ if } \mathbf{v} \in S. & (6) \end{aligned}$$

Hence, by replacing all $\mathbf{x} \in S$ by their counterpart $\mathbf{x}'$, we obtain a code D which has the property stated in the beginning of this part of the proof.

B. Next, we shall transform D into a code $C \in \mathcal{F}$, such that C_1 does not contain a pair of codewords at mutual distance $> m-2$.
Let $T \subseteq D_1$ be defined as

$$T = \{\mathbf{x} | \mathbf{x} \in D_1, \exists \mathbf{y} \in D_1 \text{ with } d(\mathbf{x},\mathbf{y}) = m-1, x_1 = y_1\}. \tag{7}$$

It will be obvious that T is the union of two disjoint sets T_0 and T_1, such that in T_i one has $x_1 = y_1 = i$, and $|T_i| = \emptyset$ or $|T_i| \geq 2$, for $i \in \{0,1\}$. If $T \neq \emptyset$, we define for each $\mathbf{x} \in T$ a word $\mathbf{x}'$ with $x'_1 = x_1 + 1 \pmod 2$, $x'_j = x_j$ for $1 < j < n$ and $x'_n = 0$. None of these words $\mathbf{x}'$ is in D, since the conditions $d(\mathbf{x},\mathbf{y}) = m-1$ and $x_1 = y_1$ in (7) imply $d(\mathbf{x}',\mathbf{y}) = (m-1)+2 = m+1$. Furthermore, we have for any $\mathbf{x}'$

$$\begin{aligned} d(\mathbf{x}',\mathbf{v}) &\leq d(\mathbf{x},\mathbf{v}) + 1 - 1 \leq m, \text{ if } \mathbf{v} \in D_0, & (8)\\ d(\mathbf{x}',\mathbf{v}) &\leq d(\mathbf{x},\mathbf{v}) + 1 \leq (m-1)+1 = m, \text{ if } \mathbf{v} \in D_1 \backslash T, & (9)\\ d(\mathbf{x}',\mathbf{v}') &\leq d(\mathbf{x},\mathbf{v}) \leq m-1, \text{ if } \mathbf{v} \in T. & (10) \end{aligned}$$

Hence, by replacing all $\mathbf{x} \in T$ by their counterparts $\mathbf{x}'$, we obtain a code $E^{(1)} \in \mathcal{F}$, such that $d(E_1^{(1)}) \leq m-1$, and if $E_1^{(1)}$ contains a pair $(\mathbf{x}, \mathbf{y})$ with $d(\mathbf{x}, \mathbf{y}) = m-1$, then $x_1 \neq y_1$. In a similar way we eliminate pairs $(\mathbf{x}, \mathbf{y})$ from $E_1^{(1)}$ with $d(\mathbf{x}, \mathbf{y}) = m-1$ and $x_2 = y_2$, yielding a code $E^{(2)} \in \mathcal{F}$, such that $d(E_1^{(2)}) \leq m-1$, and if $E_1^{(2)}$ contains a pair $(\mathbf{x}, \mathbf{y})$ with $d(\mathbf{x}, \mathbf{y}) = m-1$, then $x_1 \neq y_1$ and $x_2 \neq y_2$. Proceeding like this, we finally end up with a code $E^{(n-1)} \in \mathcal{F}$ with $d(E^{(n-1)}) \leq m-1$, and with the property that if $E_1^{(n-1)}$ contains a pair of words $\mathbf{x}$ and $\mathbf{y}$ with $d(\mathbf{x}, \mathbf{y}) = m-1$, then $x_1 \neq y_1, x_2 \neq y_2, \ldots, x_{n-1} \neq y_{n-1}$. Since $x_n = y_n = 1$, it would follow in such a case that $d(\mathbf{x}, \mathbf{y}) = n-1$, and hence $m = n$. This last equality contradicts the condition $m < n$, and so, $E_1^{(n-1)}$ does not contain a pair of words $\mathbf{x}$ and $\mathbf{y}$ with $d(\mathbf{x}, \mathbf{y}) > m-2$. Therefore, we can define the code $C := E^{(n-1)}$ which satisfies the statement in the Proposition.■

Remark The method used in the above proof seems to be similar to the "pushing-techniques" applied in [1, 2] and in related papers.

The following property is an immediate consequence. Its proof is accomplished by omitting the last coordinate of all words in the code C as mentioned in Proposition 2.1.

Corollary 2.2 *The maximal possible size $N(n, m)$ of a binary code of length n and diameter m, $1 < m < n$, satisfies the recurrence relation*

$$N(n, m) \leq N(n-1, m) + N(n-1, m-2).$$

We are ready now to prove the well-known result proved by Kleitman in [5] and by Katona in [4].

Theorem 2.3 *Let C be a binary code of length $n \geq 1$ with diameter $m, 1 \leq m \leq n$, which has maximal possible size $N(n, m)$. Then $N(n, n) = 2^n$, whereas for $m < n$ one has*

$$N(n, m) = \begin{cases} \sum_{i=0}^{k} \binom{n}{i} & m \text{ even}, \\ \sum_{i=0}^{k} \binom{n}{i} + \binom{n-1}{k} & m \text{ odd}, \end{cases}$$

where $k := \lfloor \frac{m}{2} \rfloor$.

Proof. The equality $N(n, n) = 2^n$ is trivial. For $m < n$, we first prove the weaker relations

$$N(n, m) \leq \begin{cases} \sum_{i=0}^{k} \binom{n}{i} & m \text{ even}, \\ \sum_{i=0}^{k} \binom{n}{i} + \binom{n-1}{k} & m \text{ odd}, \end{cases} \tag{11}$$

Assume that the inequalities (11) hold for all values less than some fixed $n > 1$. Let m be even. By Corollary 2.2 and by the induction assumption, we have for $m < n-1$,

$$N(n, m) \leq \sum_{i=0}^{k} \binom{n-1}{i} + \sum_{i=0}^{k-1} \binom{n-1}{i} = 1 + \sum_{i=1}^{k} \binom{n-1}{i} + \sum_{i=1}^{k} \binom{n-1}{i-1} = \sum_{i=0}^{k} \binom{n}{i}.$$

Since the code consisting of all words in a ball of radius k with some fixed word -say $\mathbf{0}$ -as its center has precisely this many words, the right hand side of the above inequality is sharp. So, the Theorem has been proved now for n, in case that m is even and $0 < m < n-1$. To complete the case of m is even, we have to prove the Theorem for $m = n-1$, when n is odd. The only restriction on C now is that if a word is in C its complement is not in C. According to the Pigeon Hole Principle, we obtain

$$N(n, n-1) = 2^{n-1} = \sum_{i=0}^{k} \binom{n}{i}, \text{ with } k = \frac{n-1}{2}.$$

The proof for odd values of m is completely similar. So, the Theorem also holds for the value n. Since for $n = 1$ the Theorem is trivial, we have proved it for all $n \geq 1$, according to the principle of mathematical induction.■

3 Generalization for q-ary Anticodes

Our method can easily be generalized to derive an upper bound for the maximal size of a q-ary anticode of length n and maximal distance m. If we denote this maximal size by $N_q(n, m)$, we obtain, completely similar to the derivation of (11)

$$N_q(n,m) \leq \begin{cases} \sum_{i=0}^{k} \binom{n}{i}(q-1)^i & m \text{ even,} \\ \sum_{i=0}^{k} \binom{n}{i}(q-1)^i + \binom{n-1}{k}(q-1)^{k+1} & m \text{ odd,} \end{cases} \tag{12}$$

However, the upper bound in (12) is not sharp for all integers m in the interval $[1, n-1]$. In order to obtain an equality sign in (12) for $1 \leq m \leq m_0$, where m_0 is an integer as described in refs. [1] and [2], we have to generalize the arguments used in the last part of the proof of Theorem 2.3 for such an integer m_0. This work is still under investigation.

References

[1] Ahlswede, R., Cai, N. Zhang, Z., *Diametric theorem in sequence spaces, Combinatorica,* vol. 12 (1992), pp. 1-17.

[2] Ahlswede, R. and Khachiatran, L.H., *Diametric theorem in Hamming spaces-optimal anticodes, Adv. Appl. Math.,* vol. 20 (1998), pp. 429-449.

[3] Farrell, P.G., *Linear binary anticodes, Electronic Letter,* vol. 6 (1970), pp. 419-421.

[4] Katona, G.O.H., *Intersection theorem for systems of finites sets, Acta Math. Acad. Sci. Hungar,* vol. 15 (1964), pp. 329-337.

[5] Kleitman, D.J., *On a combinatorial conjecture of Erdös, J. Comb. Theory,* vol. 1 (1966), pp. 209-214.

[6] Maki, G.K., *Maximum-distance Linear codes, IEEE, Trans. Inform. Theory,* vol. 17 (1971), p. 632.

[7] Reddy, S.M., *On block codes with specified maximum distance, IEEE, Trans. Inform. Theory,* vol. 18 (1973), pp. 823-824.

A.J. van Zanten: Department of Mathematics
Faculty of Electrical Engineering, Mathematics and Computer Sciences
Delft University of Technology
P.O. BOX 5031, 2600 GA Delft, The Netherlands

On Scheduling and Optimization

A. Aman

Department of Mathematics, Institut Pertanian Bogor, Indonesia

Abstract: Scheduling is a class of problems in operations research which is concerned by questions that arise in production planning, in computer control, and generally in all situations in which scarce resources have to be allocated to activities over time. In this presentation we will address some classification of the scheduling problems, and present some results involving a single machine.

On Balanced Uniform Counting Sequences

I.N Suparta [a,b], A.J van Zanten[a]

[a] Delft University of Technology, The Netherlands
[b] IKIP Singaraja, Bali - Indonesia

Abstract

The Hamming distance between two n-bit strings (codewords) is the number of bit positions where they differ. A uniform counting sequence of length n is a list of all 2^n binary n-bit codewords such that any two successive codewords in the list have the same Hamming distance, including the last and the first codeword. This notion generalizes a cyclic Gray code where any two successive codewords differ in precisely one bit position. A balanced uniform counting sequence is a uniform counting sequence such that the total number of changes in bit position i differs at most 2 from the total number of changes in position j for any pair $i, j \in \{1, 2, ..., n\}$. Robinson and Cohn (*IEEE Transaction on Computers*, C-30, No. 1, (1981) 17-23) conjectured that balanced uniform counting sequences exist for any length n. In this paper we introduce a construction for uniform counting sequences, and we prove that in some cases the produced uniform counting sequences are balanced.

Key-words: Counting sequences, uniform counting sequences, balanced uniform counting sequences, Gray codes.

1 Introduction

A counting sequence $\mathcal{O}(n)$ of length n is a list of all 2^n binary n-tuples (codewords of length n) such that each codeword appears exactly once (See [4]). We shall index codewords in $\mathcal{O}(n)$ from 0 until $2^n - 1$ and denote the jth codeword in $\mathcal{O}(n)$ by $\mathbf{x}_j$, $0 \leq j < 2^n$. Let s_i, $i \in [2^n] := \{1, 2, ..., 2^n\}$, be the set of bit positions in which the codewords $\mathbf{x}_{i-1}$ and $\mathbf{x}_i$ differ in the list of $\mathcal{O}(n)$. The ordered sequence of s_i, for all $i \in [2^n]$, is called the transition sequence of $\mathcal{O}(n)$. The transition sequence of $\mathcal{O}(3)$ in Fig.1 (bit positions are counted from right to left) is equal to
$\{1, 2, 3\}, \{2, 3, 4\}, \{1, 2, 3, \}, \{1, 3, 4\}, \{1, 2, 4\}, \{1, 3, 4\}, \{1, 2, 3\}, \{2, 3, 4\}, \{1, 3, 4\}, \{2, 3, 4\}, \{1, 2, 4\}$,
$\{2, 3, 4\}, \{1, 2, 3\}, \{1, 2, 4\}, \{1, 3, 4\}, \{1, 2, 4\}$.

0000	1100
0111	0001
1001	1111
1110	0100
0011	1010
1000	1101
0101	0110
0010	1011.

Figure 1: A uniform counting sequence of length 4

When we consider the sequence $\mathcal{O}(n)$ as a closed sequence (as we mostly do), $\mathbf{x}_{2^n}$ will be identified with $\mathbf{x}_0$. As usual, the *Hamming distance* between two codewords of the same length is defined to be the number of bit positions where they differ. If for every $i \in [2^n]$ the cardinality of s_i is equal to t, then the counting sequence $\mathcal{O}(n)$ is called an (n, t)-*uniform counting sequence* or, shortly, an (n, t)-*sequence*. It is well known that n-bit *cyclic Gray codes* are $(n, 1)$-sequences. Uniform counting sequences have applications in many areas such as testing and fault diagnosis in combinational logic circuits (cf. [3, 4]).
It is easy to see that an (n, t)-sequence does not exist whenever t is even, since successive codewords then always have the same parity. Therefore, from now on we assume that t is odd.

The transition sequence of (n,t)-sequences will be denoted by $\bar{S}(n,t)$. In the special case of $(n,1)$-sequences (Gray codes) we shall also use the notation $\bar{S}(n)$. The number of changes in position i in a counting sequence $\mathcal{O}(n)$ of length n, $i \in [n] := \{1,2,...,n\}$, is called the *transition count* of bit position i in $\mathcal{O}(n)$. In an (n,t)-sequence, the transition count of bit position i is denoted by $TC_{(n,t)}(i)$. For the simplicity $TC_{(n,1)}(i)$ will be written as $TC_n(i)$. The distribution $TC_{(n,t)} = (TC_{(n,t)}(1), TC_{(n,t)}(2), ..., TC_{(n,t)}(n))$ is called the *transition count spectrum* of an (n,t)-sequence. It is obvious that in an (n,t)-sequence we have

$$\sum_{i=1}^{n} TC_{(n,t)}(i) = t \cdot 2^n. \tag{1}$$

An (n,t)-sequence which has the property that for all $i,j \in [n]$, $|TC_{(n,t)}(i) - TC_{(n,t)}(j)| \leq 2$ is called *balanced*, and it is called *totally balanced* if for all $i,j \in [n]$, $TC_{(n,t)}(i) = TC_{(n,t)}(j)$. We can verify that the $(4,3)$-sequence in Fig.1 is (totally) balanced with transition count spectrum $(12,12,12,12)$. Robinson and Cohn in [4] gave a method for the construction of uniform counting sequences based on linear codes. They arrange the codewords of a linear code $[n,k,t]$-code such that each pair of successive codewords have minimum distance t, by making use of a minimum weight basis and by a Gray code of length k for the enumeration of all linear combinations of the basis codewords. To obtain all words of length n, this arrangement is followed by its cosets which are also built up in such a way that the uniformity with respect to the Hamming distance between any two successive codewords is maintained(cf. also [8, 9, 10], where the same principle is applied). This construction however, does not guarantee that the resulting sequences are balanced. Based on that technique, with some adjustments, Robinson and Cohn in [4] could produce balanced $(7,3)$ and $(8,5)$-sequences. But the existence of balanced (n,t)-sequences for any odd t and $n > t$ is a long standing conjecture of Robinson and Cohn in [4].

In this paper we introduce a simple technique for the construction of uniform sequences where in some cases the resulting sequences are balanced.

2 Uniform distribution

Let n be a fixed positive integer and let $2^n = q \cdot n + r$, $0 \leq r < n$. We distinguish between two cases.

Case I. q is even. This implies r is even. We define a set $Q \subseteq [n-2]$ such that the cardinality of Q is equal to $\frac{r}{2}$, and we define the integers p_i, $1 \leq i \leq n$, according to

$$p_i := \begin{cases} q, & i \in [n] \backslash Q, \\ q+2, & i \in Q. \end{cases} \tag{2}$$

Remark If n is a power of two, then the value of r is zero. This implies that the value of p_i is equal to q for all $i \in [n]$.

Case II. q is odd. This implies that $n+r$ is even. Here, we define $Q \subseteq [n-2]$ as a set with cardinality $\frac{n+r}{2}$, and we define

$$p_i := \begin{cases} q-1, & i \in [n] \backslash Q, \\ q+1, & i \in Q. \end{cases} \tag{3}$$

It has been proven that balanced Gray codes exist for any length n(See e.g. [1, 2, 4, 6, 7]). Furthermore, in [7] Suparta and van Zanten showed that any balanced Gray code of length n will have a

transition count spectrum $(p_1, p_2, ..., p_n)$ where p_i is determined using (2) or (3).

Now, let $\mathcal{D}$=$(\underbrace{a, ..., a}_{k_a}, \underbrace{b, ..., b}_{k_b})$ be a distribution of integers a and b with $1 \leq k_a \leq k_b$, and let $\rho = k_b - \lfloor \frac{k_b}{k_a} \rfloor \cdot k_a$. It is clear that $0 \leq \rho < k_a$. Furthermore, let also $\mathcal{B}_1^i = \underbrace{b, ..., b}_{\lfloor \frac{k_b}{k_a} \rfloor}$, $1 \leq i \leq k_a - \rho$, and $\mathcal{B}_2^j = \underbrace{b, ..., b}_{\lfloor \frac{k_b}{k_a} \rfloor + 1}$, with $1 \leq j \leq \rho$.

Rearrange the distribution $\mathcal{D}$ using the following Construction A (we assume each $\mathcal{B}_1^i$ and $\mathcal{B}_2^j$ as one component when determining the cyclic distance in $\mathcal{D}$).

Construction A

1. Distribute $\mathcal{B}_1^i$ and $\mathcal{B}_2^j$ in $\mathcal{D}$, ordered according to their superscript, such that for all i, $i \in [k_a - \rho]$, the (cyclic) distances from $\mathcal{B}_1^i$ to $\mathcal{B}_1^{i+j}$ and from $\mathcal{B}_1^i$ to $\mathcal{B}_1^{i-j}$, for all $j \in [\lfloor \frac{k_a - \rho}{2} \rfloor]$, differ at most 1. Here we use that the addition of superscripts takes place in the cycle $(1\ \ 2\ \ \cdots\ \ k_a - \rho)$.

2. Insert a at the front of $\mathcal{B}_1^i$ and of $\mathcal{B}_2^j$.

A distribution of $\mathcal{D}$ which is arranged according to Construction A will be called *uniform*, and it is called *totally uniform* if the distance from $\mathcal{B}_1^i$ to $\mathcal{B}_1^{i-j}$ is equal to the one from $\mathcal{B}_1^i$ to $\mathcal{B}_1^{i+j}$, with $\mathcal{B}_1^i \neq \mathcal{B}_1^{i-j}$ for some $j \geq 1$. For example, consider the distribution $\mathcal{D}$=$(2, 2, 2, 2, 2, 4, 4, 4, 4, 4, 4, 4)$. Here we have $k_2 = 5$, $k_4 = 7$, $\lfloor \frac{7}{5} \rfloor = 1$, and $\rho = 2$. Therefore, we have 3 $\mathcal{B}_1$'s and 2 $\mathcal{B}_2$'s, with $\mathcal{B}_1 = 4$ and $\mathcal{B}_2 = 4, 4$. A uniform distribution of $\mathcal{D}$ is $(2, \mathcal{B}_1^1, 2, \mathcal{B}_1^2, 2, \mathcal{B}_2^1, 2, \mathcal{B}_1^3, 2, \mathcal{B}_2^2,)$= $(2, 4, 2, 4, 2, 4, 4, 2, 4, 2, 4, 4)$. Whereas $(2, \mathcal{B}_2^1, 2, \mathcal{B}_2^2, 2, \mathcal{B}_1^1, 2, \mathcal{B}_1^2, 2, \mathcal{B}_1^3)$=$(2, 4, 4, 2, 4, 4, 2, 4, 2, 4, 2, 4)$ is not uniform, since the difference of distances between $\mathcal{B}_1^1$ and $\mathcal{B}_1^2$, and between $\mathcal{B}_1^1$ and $\mathcal{B}_1^{1-1} = \mathcal{B}_1^3$, not counting the integers 2, is equal to $2 > 1$. We see that $\mathcal{D}$ can not be arranged as a *totally* uniform distribution. But it will be clear that any distribution of two integers a and b can be transformed into a uniform one.

For an arbitrary distribution $(a_1, a_2, ..., a_n)$, we shall say that a_n and a_1 are consecutive and that a_1 is the successor of a_n. Groups consisting of t consecutive elements of a distribution will be called t-blocks.

Lemma 2.1 *Let $\mathcal{D}$ be a uniform distribution of the integers a and b. We shall write t_α and t_β for the t-blocks which contain the largest and the smallest number of b's, respectively. Then we have that the number of b's in t_α is at most one more than the number of b's in t_β.*

Proof: Assume that the longest sequence of consecutive $\mathcal{B}_2$'s in $\mathcal{D}$ contains m $\mathcal{B}_2$'s, $m > 0$. Rule 1 of Construction A implies that the shortest sequence of consecutive $\mathcal{B}_2$'s in $\mathcal{D}$ contains $m - 1$ $\mathcal{B}_2$'s. We shall write $\mathbb{B}_m$ and $\mathbb{B}_{m-1}$ for the longest and the shortest of these sequences, respectively. Apart from the distribution of a's in $\mathcal{D}$, each $\mathcal{B}_1$ will be sandwiched between two $\mathbb{B}_m$'s or between $\mathbb{B}_m$ and $\mathbb{B}_{m-1}$. Moreover, the number of $\mathbb{B}_m$'s between $\mathcal{B}_1^i$ and $\mathcal{B}_1^{i+j}$ and the number of $\mathbb{B}_m$'s between $\mathcal{B}_1^i$ and $\mathcal{B}_1^{i-j}$ differ at most 1 due to Rule 1. This fact proves the implication of the Lemma.■

An immediate consequence of Lemma 2.1 is the following.

Corollary 2.2 *If $|b - a| = 2$ in Lemma 2.1, then the absolute value of the difference between the sums of the components of t_α and of t_β respectively, is at most 2.*

3 A construction of uniform counting sequences based on transition sequences of Gray codes

For a fixed positive integer n, let σ be the cycle $(1\ 2\ ...\ n)$. We notice here that for every element a of this cycle σ, and for every integer $i \equiv j(mod\ n)$ we have

$$a+i := \begin{cases} a+j, & 1 \le a+j \le n, \\ a+j-n, & otherwise. \end{cases}$$

Next we define for every element a of σ the integer $a^{(i)} = a+i$, for every integer i.
For instance, with respect to the cycle $\sigma = (1\ 2\ 3\ 4\ 5)$, one has $3^{(1)} = 3+1 = 4$, $4^{(3)} = 4+3-5 = 2$, $3^{(-1)} = 3+4-5 = 2$, and $2^{(-3)} = 2+2 = 4$. For each m, t, with $0 < m \le t < n$, we define a mapping $\Phi_{m|t}$ from the set $[n]$ into the power set $\mathbf{2}^n$ of n, which maps every integer $a \in [n]$ to $\Phi_{m|t}(a) := \{a, a^{(m)}, a^{(m+1)}, ..., a^{(m+t-2)}\}$.
Let us consider the above cycle σ. We have for examples $\Phi_{1|3}(2) = \{2,3,4\}$, $\Phi_{3|3}(2) = \{2,5,1\}$, etc.
Below we introduce a construction of uniform sequences which based on transition sequence of Gray codes.

Construction B
Let n and odd t, $1 \le t \le n-1$, be fixed integers which have a greatest common divisor $gcd(n,t)$ equal to m. Let $\bar{S}(n) = s_1, s_2, ..., s_{2^n}$ be the transition sequence of a Gray code $G(n)$. Start with an n-bit codeword (usually with the zero codeword). Apply the sequence $\bar{S}(n,t) = \Phi_{m|t}(s_1), \Phi_{m|t}(s_2), ..., \Phi_{m|t}(s_{2^n})$ to generate the next $2^n - 1$ codewords.

In [5], we proved that Construction B can always produce an (n,t)-sequence if n and t are relatively prime. Furthermore we proved that for any relevant values of t, balanced (n,t)-sequences exist if n is prime or if n has the property that $2^n = n \cdot p + r$, with $r = 0, 2, n-2$, for some integer p.
Here, we extend the results of [5] w.r.t. balanced uniform counting sequences based on Corollary 2.2.
Let $m = 1$ in Construction B, or equivalently, let n and t be relatively prime. Consider a Gray code of length n with transition count spectrum TC_n=$(TC_n(1), TC_n(2), ..., TC_n(n))$. Furthermore, consider the uniform counting sequence produced by applying Construction B for $m = 1$ with transition count spectrum $TC_{(n,t)}$=$(TC_{(n,t)}(1), TC_{(n,t)}(2), ..., TC_{(n,t)}(n))$. Based on the mapping $\Phi_{1|t}$, we have that

$$TC_{(n|t)}(i) = TC_n(i) + TC_n(i^{(-1)}) + TC_n(i^{(-2)}) + \cdots + TC_n(i^{(-(t-1))}). \tag{4}$$

Since every distribution of type $\mathcal{D}$ can be transformed into a uniform distribution, due to Corollary 2.2 we have the following main theorem of this Section.

Theorem 3.1 *For every integer n and odd t, $1 \le t \le n-1$, with $gcd(n,t) = 1$, there exists a balanced (n,t)-sequence.*

4 Computer results

By computer calculations we checked that for certain n-values, Construction B produces (n,t)-sequences for all t, $t \le n-1$, with $gcd(n,t) > 1$. For instance, we can construct $(6,3)$, $(9,3)$, $(12,3)$, $(15,3)$, $(20,3)$, $(10,5)$, $(15,5)$, $(20,5)$, $(14,7)$ and $(21,7)$-sequences. Moreover, by adjusting the distribution of the transition count spectra, we can construct balanced $(9,3)$, $(15,3)$, $(15,5)$, and $(21,7)$-sequences.
We firmly conjecture that Construction B will produce an (n,t)-sequence for every pair of n and t, $1 \le t \le n-1$.

Lemma 4.1 *For every positive integer k, one has*

$$2^{3^k} = 3^k \cdot q + (3^k - 1),$$

for some integer q.

Proof: We prove this Lemma by induction to k. It is obvious that the Lemma is true for $k = 1, 2, 3$. Assume now that the Lemma is true for $k \geq 3$. So, we have that $2^{3^k} = 3^k \cdot s + (3^k - 1)$, for some integer s. Then we have

$$\begin{aligned} 2^{3^{k+1}} &= 2^{3^k \cdot 3} \\ &= (3^k \cdot s + (3^k - 1))^3 \\ &= (3^k \cdot s)^3 + 3(3^k \cdot s)^2(3^k - 1) + 3(3^k \cdot s)(3^k - 1)^2 + (3^k - 1)^3 \\ &= (3^k \cdot s)^3 + 3(3^k \cdot s)^2(3^k - 1) + 3(3^k \cdot s)(3^k - 1)^2 + 3^{3k} - 3(3^{2k}) + (3 \cdot 3^k - 1). \end{aligned}$$

We see that the first four terms on the right hand side are divisible by 3^{k+1}, for all $k \geq 1$. Hence, we have $2^{3^{k+1}} = 3^{k+1}.q + (3^{k+1} - 1)$, for some q. Because of the principle of mathematical induction we have proved the Lemma now.■

By considering eq. (2), Lemma 4.1 implies that a transition count spectrum of a balanced Gray code of length 3^k, for some non-negative integer k, can be arranged as $(a, b, a, b, ..., a, b, a)$, with $b = a + 2$. According to Corollary 2.2, if Construction B succeeds to produce an $(3^k, t)$-sequence, then the resulting sequence will be balanced when starting from a balanced Gray code with transition count spectrum ordered as $(a, b, a, b, ..., a, b, a)$.

Remark By using Construction B for the construction of uniform counting sequences, we can determine the transition count spectrum of the resulting counting sequence by using the distribution of the transition count spectrum of the Gray codes which we use as the basis of the construction. Since for any distribution $(p_1, p_2, ..., p_n)$, with $\sum_{i=1}^{n} p_i = 2^n$, p_i is even for every $i \in [n]$, and $|p_i - p_j| \leq 2$, $i, j \in [n]$, a balanced Gray code exists having transition count spectrum $(p_1, p_2, ..., p_n)$, we expect to be able to produce "approximately balanced" or balanced uniform counting sequences for every n and odd $t < n$. A proof of the validity of Construction B for any pair of n and odd $t < n$ with $gcd(n, t) > 1$, is under investigation.

References

[1] Ádám A. (1968), *Truth Functions and the Problem of their Realization by Two-Terminal Graphs,* Akadémiai Kiadó, Budapest.

[2] Bhat, G.S., and C. D. Savage (1996), Balanced Gray codes, *The Electronic Journal of Combinatorics,* 3, paper 25.

[3] Hayes, J.P. (1978), Generation of optimal transition count tests, *IEEE Trans. Computers,* C-27, 36-41.

[4] Robinson, J.P. and M. Cohn (1981), Counting sequences, *IEEE Trans. Computers,* C-30, 17-23.

[5] Suparta, I N. and A.J. van Zanten (2005), On balanced maximal counting sequences and balanced uniform counting sequences, *Tech. Report CS 05-01, Institute for Knowledge and Agent Technology, University Maastricht,* Maastricht, The Netherlands.

[6] Suparta, I N. and A.J. van Zanten (2003), Balanced Gray codes, *Tech. Report CS 03-03, Institute for Knowledge and Agent Technology, University Maastricht,* Maastricht, The Netherlands.

[7] Suparta, I N. and A.J. van Zanten, A note on balanced Gray codes, *submitted for publication.*

[8] van Zanten, A.J. (2001), Cyclic distance-preserving codes on constant-weight basis, *Discrete Applied Mathematics,* 114, 289-294.

[9] van Zanten, A.J. (1993), Minimal-change order and separability in linear codes, *IEEE Trans. Inform. Theory,* IT-39, 1988-1989.

[10] van Zanten, A.J., and A. Lukito (1999), Construction of certain cyclic Distance- preserving codes having linear-algebraic characteristics, *Designs, Codes and Cryptography,* 16, 185-199.

I Nengah Suparta: Department of Applied Mathematics, TU Delft
PO.BOX. 5031, 2600 GA Delft, the Netherlands
Phone/Fax:+31-15-2783613/+31-15-2787295.
E-mail: N.Suparta@ewi.tudelft.nl

A.J van Zanten: Department of Applied Mathematics, TU Delft
PO.BOX. 5031, 2600 GA Delft, the Netherlands
Phone/Fax: +31-15-2785821/+31-15-2787295.
E-mail: A.J.vanZanten@twi.tudelft.nl

Long-Distance Wave-Group Propagation using a Variational Boussinesq Model

Gert Klopman, Brenny van Groesen

Applied Analysis and Math. Physics, University of Twente, Enschede, The Netherlands

Abstract: Water waves at sea are always grouped: sequences of higher waves are alternated by sequences of lower waves. Due to non-linearity, the wave groups are accompanied by bound sub- and super-harmonic wave components, as well as by amplitude dispersion besides the linear frequency dispersion. When modelling the wave motion, one has to distinguish the horizontal propagation space of the waves and a vertical cross-space. Starting from a variational principle we integrate out the vertical cross-space by using a Boussinesq-type polynomial approximation for the vertical flow structure. This results in a model maintaining the positive-definiteness of the Hamiltonian, which leads to good dynamic behaviour of the approximate equations. At the conference, we will show how the model can be used to propagate wave groups above slowly-varying sea bed over large distances, predicting the changes in the wave group shape and the generation of free long waves (which are important for moored ship dynamics and coastal morphology).

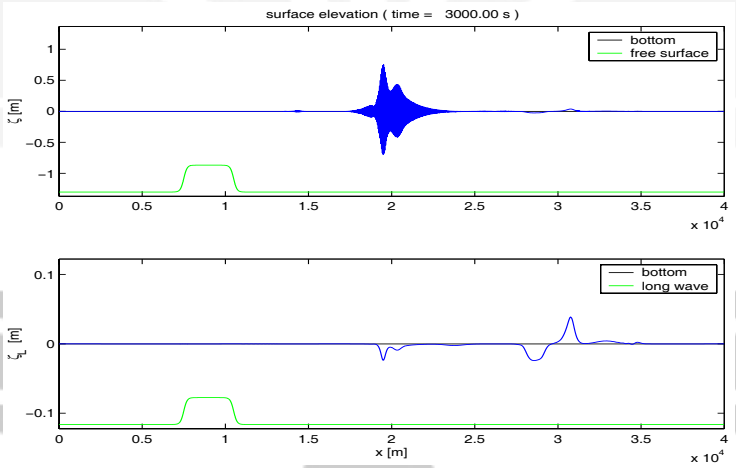


Figure caption: Upper graph: Sea bed shape (green line, not on scale) and free-surface elevation (blue line) after an initial soliton has propagated for 600 wave periods. Lower graph: Low-pass filtered long-wave part of the free-surface elevation (blue line) in the upper graph.

Keywords: variational method, non-linear water waves, inhomogeneous media

Kinematics of Shot-Geophone Migration

Chris Stolk

Applied Analysis and Math. Physics, University of Twente, Enschede, The Netherlands

Abstract: Imaging of reflection seismic data requires the estimation of a velocity model (reference medium). Prestack migration is used to assess whether a choice of the velocity model is consistent with the data, by using the semblance principle. We compare shot-geophone migration with migration methods based on data binning, such as prestack Kirchhoff migration. Image artifacts, present in binning based methods due to multiple raypaths connecting source and receiver points with subsurface points, are absent in shot-geophone migration, when the migration velocity is kinematically correct and when events to be migrated arrive in the data along non-turning rays. Common image gathers produced by shot-geophone migration exhibit the appropriate semblance property in either offset domain (focussing at zero offset) or angle domain (focussing at zero slope). Thus shot-geophone migration may be a particularly appropriate tool for migration velocity analysis of data exhibiting structural complexity.

THE BLOCK STORAGE PROBLEM AT PT PAL

Pritanto E.
PT PAL Indonesia, Surabaya

Abstract. In order to maximize utilization of the expensive docks in the shipyard, blocks which build up a ship have to made ready as many as possible. This means a sufficient number of block storage areas should be available. Even though similar blocks can be stacked, the storage areas are also used for joining blocks together into grand-blocks and they are loaded according to a fixed sequence and a rigid schedule. In practice, there is uncertainty with respect to the arrival and departure moment of blocks to/from the storage areas. A need of extra storage areas therefore might be inevitable and the decision as to where to place blocks in the areas should be made in some optimal fashion. The major difficulty stems from a situation that blocks visit the storage areas several times. Simple decision strategies are discussed and evaluated in a dynamic simulation model of the shipyard. Experiments with the model show that the current strategies may need an optimization technology in order to get a better result.
Key-words: block storage, stacking, simulation, optimization

1 Introduction

Improvements in shipbuilding technology have been successful in increasing productivity, particularly with support from a growingly effective use of information technology. But, problems still remain in practice such as late developing engineering changes, unexpected delays in material availability or labor absenteeism that may limit the improvement program. Some of the problem sources are not under the direct control of the shipyard. The problems influence the production process which relates directly to product delivery to the customer.

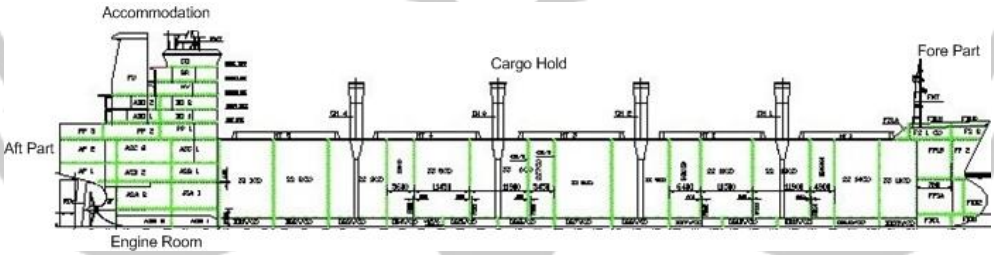


Figure 1: DSBC ship type which is built in series by the shipyard.

Problems above become complicated by the current circumstances faced by the shipyard, for example:

1. Schedule is tight:
    - Strict erection period and limited resources (time schedule). Erection means the last assembly stage in which blocks are joined together at the building dock. Full orders of new ships ask for strict erection periods as

any extra time will accumulate to the next ships. Figure 1 shows a ship type called the 50,000 DWT Double Skin Bulk Carrier (DSBC) which is built in series by the shipyard and exported to sail across the international oceans. Another ship type is the 18,500 DWT Dry Cargo Vessel (DCV) which is built also in series in the side launching area, not in the end launching area or dock like the DSBC.

- Limited block storage areas (space schedule). The lower block building rate than the block erection rate demands for more blocks ready at the storage areas waiting for erection. Storage areas in most cases are used for maintaining a continue production flow between workshops. What typical is for shipyards with limited storage areas is that blocks entering the storage areas for more than once, a re-entrant case. The storage areas are also leased for constructing offshore jacket-type platforms & pipelines, repairing and yet loading large & heavy structures with the help of the goliath crane onto a barge which is moored near to the dock way out. So, both block building and storage areas at present are critical resources and they should be planned accurately in the course of time.

2. More out-of-dock time is needed for accuracy control of blocks joined and on-block outfitting e.g. installing pipe systems, ladders, ducting systems before erection. Inaccurate blocks could not fit together in dock and it will be laborious and time consuming to repair. The expected higher erection rate will improve utilization of the expensive building dock and finally increase the number of ships produced per year.

Little research so far has been reported on the shipbuilding block storage problem. In early work, [3] developed a spatial scheduling algorithm using partial enumeration and decomposition. [1] reported a successful completion of simulation-based space utilization at Aker Ostsee, Germany. Blocks were not stacked and sufficient storage areas were divided for each workshop so that blocks entered the related storage areas only once.

Solution made by the shipyard management is that blocks have to be built earlier and as many as possible but the finished blocks needs storage areas. They have to wait for the loading time into dock. This manner allows the preceding stages to produce continuously without having to slow down or speed up production to match the strict erection period. The earlier block building provides also slack time which is still required to absorb potential delays during the preceding fabrication, assembly, block blasting and painting processes. Certain blocks can be joined together into grand-blocks because dock is an area where work is carried out less efficiently than at the other production stages. Practical experiences cannot help because the tight schedule did not happen in the past. A balanced mix of large and small ships was built in the past but now mostly big ships are built at the same time. Decisions made for the block storage should be optimal in order to guarantee the contracted delivery time of ships. Otherwise, the shipyard has to pay the daily penalty cost for late delivery and will lose the customer satisfaction.

Discrete event simulation software SiMPLE++ (now Tecnomatix's eM-Plant) has been used to build a shipyard model. Simulation offers an excellent method of observing shipbuilding processes, identifying problems like bottlenecks and evaluating various alternatives to increase the production. Strategies based on logical thoughts for solving the problems are discussed and tested with the help of

simulation. The result seems to be sufficient but the author is eager to know another method that may provide a maximum result. In mathematical world, there is an operational research/management science (OR/MS) method known as optimization that may improve the existing result.

The paper is constructed as follows. Section 2 briefly describes the block storage problem in the shipyard. The problem is modeled and simulated in SiMPLE++. Experiments are carried out with the model and the results are discussed in Section 3. Limitation of the current results and a plan to embed an optimization procedure into the model are given in Section 4. The conclusions are presented in Section 5.

2 Problem description

The shipbuilding process can be simply described as follows, see Figure 2. The Steel Stock House (SSH) supplies painted plate and profiles in the form of work packages to the fabrication shop. The cutting process is relatively fast. The forming process requires a high level of craftsmanship and the actual duration is usually longer than the estimation. Interim products from the cutting machines are moved to the Line A and to the Line B, while from the forming process to Line B. The Line A and Main Panel Line (MPL) and Grand Assembly Indoor (GA) are grouped into the panel line that builds blocks of the middle section, cargo hold parts of ships. The Line B and Curved Block Line (CBL) are grouped into the curved line that provides blocks for the aft, fore and engine room parts of ships.

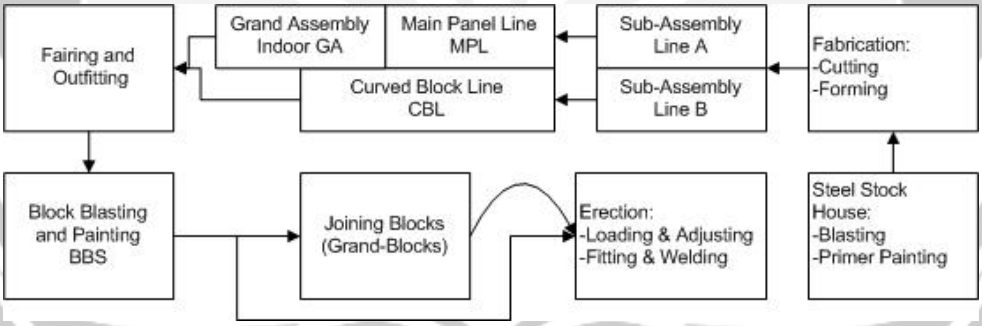


Figure 2: Shipbuilding processes.

Subsequently, blocks are stored in 32 slots waiting for entering the Block Basting Shop (BBS). These 32 slots are grouped under the BBS storage. Blocks are faired and outfitted first. Fairing is repairing any distortion of components caused by improper welding. Outfits e.g. supporting structures, pipes, ladders are installed in steps starting from the sub-assembly process until on-board of ship. Even though demanding a high 3D-CAD technology, the earlier the installation of outfits is the easier and the cheaper way to do it. In practice, the schedule of outfit installation goes along the schedule of building blocks. As a result, the installation of outfits

may occur behind schedule i.e. on-board of ship because of the strict erection period.

The BBS have four shelters. One shelter is for blasting and painting one block except BBS number 1 which can handle two blocks simultaneously. There are three coats of painting after blasting. When blocks are queuing, the finished 1st-coat blocks are moved out the shelter and continued at open-air. If blocks are joined together at outside of dock, the blasting and painting processes are started after finishing the join. Joining blocks after the BBS means rework of paint. Blocks on the 32 slots basically are not stacked. They are stacked if the slots are full and blocks have to satisfy block structural constraints. Both rectangular and arbitrary-shaped blocks are placed within the same two-dimensional rectangular slots; see the next Figure 3.

From the BBS, a block is placed in one of the grand-block storage areas which have 40 slots before loaded and joined together within the dock. In these 40 slots, blocks can be stacked up to three levels. The 2nd and 3rd coats of painting can be completed on the 40 slots or if free on the previous 32 slots. Installing outfits of the engine room blocks is allocated in the special unit outfitting area with movable roof.

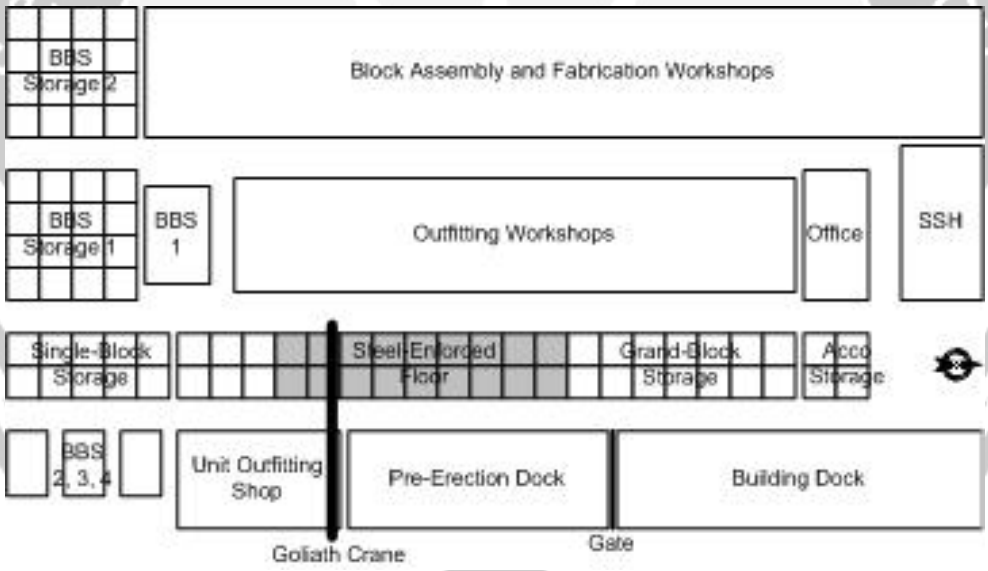


Figure 3: Shipyard layout.

The arrival time of blocks either to the BBS or to the 32 slots is determined by the master schedule. Due to disturbances during the preceding production stages, blocks may arrive not as scheduled. The departure time of blocks from the 32 slots should be decided in order to obtain the best possible way for the blocks to enter the BBS and so to enter the next 40 slots. There are several things to be considered

i.e. the number of free slots in the 32 slots, the estimated block's processing time at the BBS, the number of free slots in the 40 slots and the expected loading time of blocks into dock. The difficulty is that the sequence of loading blocks is in opposition to the staking of blocks i.e. first-processed-first-loaded against first-stacked-last-loaded. In particular when the access to storage areas is limited e.g. areas at the corner position, some blocks have to be moved out before the transporter can reach a block to be loaded.

There are three types of blocks to be considered for the 40 slots:

1. The blocks-to-be-joined on storage areas with lattice floor (steel-enforced-concrete-floor) - 1st priority
2. The blocks-to-be-joined anywhere on storage areas – 2nd priority
3. The single blocks - 3rd priority

where all blocks-to-joined named as grand-block needs 2 parallel slots. The 40 slots are put together under the grand-block-storage. Only single blocks that will be loaded shortly into dock are placed at the nearby single-block storage which has 10 slots. Two slots are usually made free for a 150/300 ton transporter to go into so that the goliath crane can pick up a block on from the transporter.

A grand-block may consist of 2 blocks (1 layer) or 6 blocks (3 layers). So, the 40 slots have 18 slots with steel-enforced-concrete-floor and 22 slots with only concrete-floor. The scarcity of these particular slots determines the priority. When these slots are full, blocks have to be allocated at extra storage areas. Blocks that have to be joined together must to be moved first into the grand-block storage. They need a goliath crane and a stable ground, slots under the crane, for accurate positioning. Single blocks from the extra storage areas need also the goliath crane and the single-block storage in order to be loaded into dock. The transportation to/from the extra storage areas needs also a 150/300 ton transporter and eventually a barge across the sea. The transporter needs a sufficient access for maneuvering, in particular when moving a large block. Only single blocks are planned at the extra storage areas or blocks that have to be joined together but must wait for their partners and the particular slots. The accommodation storage is designed for joining the thin-shell accommodation blocks together. The storage should be close to the aft part of ships so that the loading of accommodation blocks follows a short route and their thin-shell structures will not deform much.

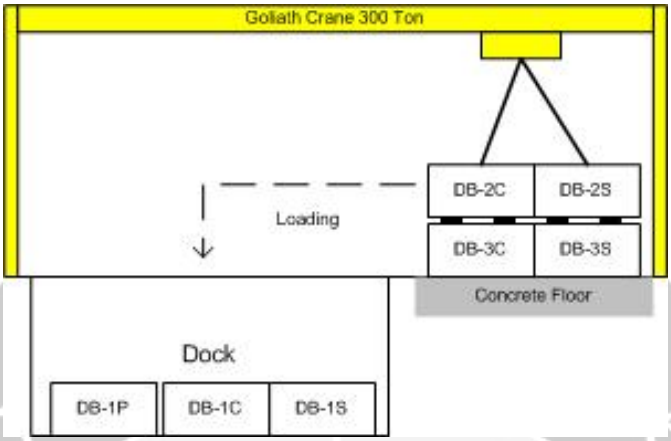


Figure 4: Loading a stacked grand-block into dock.

To imagine, a ship -DSBC type- may have 152 blocks which consist of

- 30 grand-blocks with 2 blocks per grand-block. 6 grand-blocks are of type 1 and 24 are of type 2
- 2 grand-blocks with 6 blocks per grand-block
- 1 grand-block with 7 blocks per grand-block
- 73 single blocks

The blocks-to-be-joined are pairs of blocks that will be loaded & adjusted and then fitted & welded at the grand-block storage. Before joining, these blocks have to be placed by the goliath crane onto the allocated slots and during joining, there must be no blocks placed on top of the pair. The joining process needs two weeks, after which the pair can be shifted on top of another joined pair.

From 24 grand blocks of type 2 of a ship, they are

- 13 grand-blocks are coded with CS
- 5 grand-blocks are coded with PP
- remaining blocks with different codes

Only grand-blocks with the same codes can be stacked. Their longitudinal girders are lined up with each other. 6 grand blocks are of type 1 cannot be stacked. From 73 single blocks of a ship, only 34 blocks can be stacked. Note that a single block occupies one slot. Because of the serial production of ships, blocks of different ships but the same codes can be stacked, see Figure 4.

3 Simulation of the problem

In re-entrant flow shops as mentioned before, it is difficult to model analytically but relatively easy to deal with heuristics [2]. The following simple decision strategies or heuristics are expected to decrease the cost for extra storage areas, transportation, handling, traffic and lateness. They are:

- First-in-last-out (FILO) for reducing unnecessary restacking. This rule relates to reducing the operational and maintenance cost of the goliath crane.

- First priority is given to blocks which need joining together into grand-blocks and those that can be stacked until 2-3 levels. If the 32 slots before the BBS are empty, they can also be used for finished grand-blocks and single blocks waiting for loading times into dock. The unstable asphalt of these slots cannot be used for joining blocks. If a must, stacking until 2 levels is still allowable.
- Because the required duration for joining blocks together is about two weeks. One has to make reservations. Otherwise, free slots at the grand-block storage might not be available and it could disturb the loading schedule. If blocks come out from the BBS at the same time and free slots are available, they can be moved to those slots and start the joining process. In case of no free slot with e.g. steel-enforced-concrete-floor, blocks have to wait at extra storage areas. The blocks are added into a list of queuing blocks and sorted in decreasing priority. When slots are free, blocks with high priority has the right to get first.
- When one block is behind another, the finished BBS block could go to a free slot and reserve a parallel slot for its partner. If not, the finished BBS block could wait somewhere, even at extra storage areas, and order two parallel slots for the moment its partner finishing the BBS process.
- The joined blocks basically can be loaded into dock as soon as they are finished. This way makes two slots free. However, the loading has to follow a fix loading sequence i.e. after their predecessors are loaded and then adjusted. Only blocks at the bottom position of a ship can be loaded without waiting for the finished adjustment of their predecessors.
- Suppose LoadTime is the loading time according to the given master schedule and StartGrandTime is the estimated finish BBS time of those two blocks, the earliest time. One can delay the joining process until (LoadTime - two weeks), the latest time and one can use those free slots for others. If (SlackTime = LatestTime - EarliestTime), one can create a heuristic that if (SlackTime > x days), then better to wait somewhere.
- Repainting is needed when blocks are blasted and painted at the BBS first then joined together. Ideally, blocks are joined together and completely outfitted first and then going to the BBS. The latter means at least supports for outfitting items that require mounting with heat have to be entirely installed.
- Joining blocks together before painting is allowable if many free slots are available or grand-blocks can be stacked. Moving out the grand-blocks is not feasible because of the limited maneuvering ability of transporter and the narrow access. Decision of whether earlier, later or just-in-time joining is not straight-forward when many ships are built at the same time and work progress of blocks is not following the given master schedule. The interrelated blocks of a ship would be practically easier to be managed if their work progress is close by each other. Inappropriate decision may cause jam.

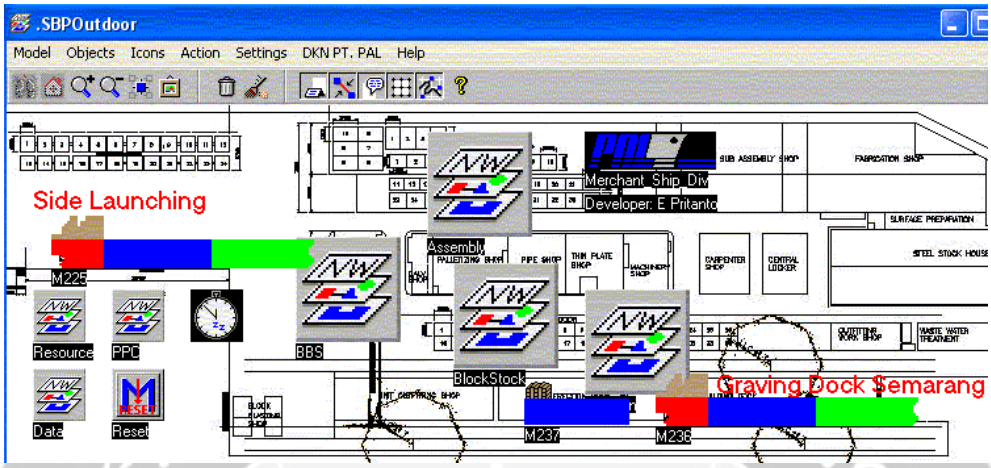


Figure 5: Model of the shipyard in SiMPLE++.

A simulation model of the shipyard has been developed in SiMPLE++ (now known as eM-Plant). SiMPLE++ stands for Simulation in Production, Logistics and Engineering. It is the standard software for object-oriented, graphical, and integrated modeling, simulation, and visualization of systems and business processes. For the strategic level, simulation can demonstrate the construction sequence of a new ship to a promising owner on computer in the offer phase. It can be shown whether or not a specific ship can be built and if so, how. For the tactical level, alternative factory layouts and operations can be designed and analyzed by taking various factors and resources into account (e.g. workers, equipment and transport infrastructure). For the operational level, if some components are delivered late or if some problem is encountered during the fabrication or assembly processes, then the impact can be examined in advance. A common method used is what-if analysis. For example, what if another two welders are added for expediting a late block, what if the goliath crane stops working? Or what if the number of frame carriers is not sufficient? See the first layer of the shipyard model in Figure 5.

181	182	183	184	185	186	187
2005/09/26 07:30:				2005/10/03 07:30:		
Layer1	Layer2	Layer3		Layer1	Layer2	Layer3
M237-DBBS-8C+D				M237-DBBS-8C+D		
M237-ASB-1PS				M237-ASB-1PS		
M237-ASA-2PS				M237-ASA-2PS		
M237-ASA-1PS				M237-ASA-1PS		
M237-DBBS-9C+D				M237-DBBS-9C+D		
M229-ADB-1PS				M229-ADB-1PS		
M237-DBBS-10C+				M237-DBBS-10C+		
M237-ASB-2PS				M237-ASB-2PS		
M237-AP-1PS				M237-AP-1PS		
M236-Hatch-1						
M229-DBBS-7C+D				M229-DBBS-7C+D		
M229-DBBS-5C+D	M229-DBBS-3C+D					
M237-DBBS-7C+D						
M229-DBBS-3S+D						
M237-DBBS-7P+D						
M229-DBBS-1C+D						

Figure 6: List of slots used in the grand-block storage. Right below is the 2D animation of grand-blocks on the assigned slots.

Figure 6 shows a weekly list of slots used in the grand-block storage i.e. storage at the side of dock and under the goliath crane. One can see the date time in the first row. The first four letters in a cell e.g. M237 are the ship project number and the remaining letters e.g. ASB-1PS are the grand-block name. Even though similar grand-blocks can be stacked until 3 layers in these areas, the heuristics mentioned before do not allow them. As a result, a large number of single blocks are located at extra storage areas. It could be a large number of single blocks stacked in the BBS storage waiting for partners or entering the BBS. This situation can be seen in the below Figure 7. The date time shown relates to the week number 88. About 80 slots are used, that is 34 single-blocks (17 grand-blocks) in the grand-block storage and 46 single-blocks in the BBS storage in which some of them are stacked in two layers. Because the BBS storage, see Figure 3, has buildings at the North and West sides, the transporter that move in/out blocks will have restricted access. Accordingly, positioning of a block on a free slot should consider the required repositioning of other blocks i.e. to which areas they have to be moved temporarily.

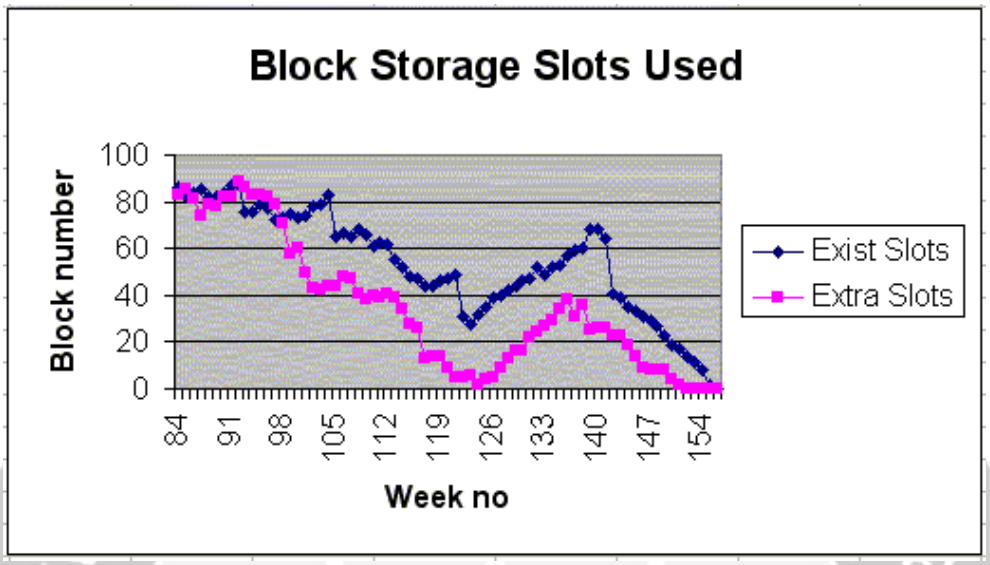


Figure 7: Number of storage slots used. Existing slots consists of those in the BBS storage and the grand-block storage.

Figure 6 shows also two types of direction namely the transversal direction (e.g. the third column in the animation: M237-DBBS-8C+ DBBS-8S and the longitudinal one (the first two columns: M229-DBBS-7P+DBBS-8P). When loading a grand-block into a dock, the goliath crane with two wire ropes cannot rotate it. The direction of a grand-block in the storage has to be the same direction as that in the dock. Only light weight grand-blocks can be rotated by the goliath crane with simply one wire rope. The direction code can be seen in the block name suffix. Different suffixes means e.g. C+S laying out in the transversal direction while the same ones e.g. P+P in the longitudinal direction. In case of building five DSBC in series, the first two of them, M236 and M237, have grand-blocks mostly in transversal direction while the next three, M229, M230 and M231, in the longitudinal one. This discrepancy is caused by the made-to-order characteristic of building ships even in series and same owner. The not easy implication to the block storage is when dividing the available slots to the transversal and longitudinal grand-blocks, in particular when reserving slots for joining blocks together and when the blocks are not in the same work progress.

Another practical problem that can be seen in Figure 6 is the size of grand-blocks. For example, one can see in the animation that grand-blocks M237-ASB-1PS, M237-ASA-2PS, M237-ASA-1PS and M237-ASB-2PS (column numbers 4, 5, 6 and 10) are wider than the available slots. Because there are electrical supply panels and consumable outlets along the grand-block storage, slots without panels and outlets can be occupied by the wider grand-blocks. The implication is that the wider and transversal grand-blocks should get a higher priority when reserving free slots. A longer longitudinal grand-block may occupy a small area of its neighboring slots.

Although single blocks can use the grand-block storage for the short term until the slots are required, transportation of the single blocks should be minimized. The cost of lift-up/down and move those blocks is high and there are risks that those blocks could slip out and their forms could distort. To sum it up, the block storage at extra areas which are far from the shipbuilding plant should be minimized. This means reducing the traffic of block movement and the cost for preparing the extra storage areas. The problem looks for better heuristics or another technology to solve it fast and efficiently, in particular for the dynamic operational level.

4 Plan to embed optimization

Shortcomings of simulation are inability to evaluate a huge range of trial-and-error options and to organize the search to the best solutions. Since simulation can cope with the complexities and uncertainties posed by the real world problems, a combination of simulation and optimization will unify advantages of both.

If the real world is so complex, an optimization technology may not be able to solve the problem completely. It is then necessary to restrict part of the real world in which the optimization technology can contribute much to the problem. In terms of planning, the technology should be embedded within the simulation model and it will look as if a real-time optimization in the real world.

To start with the application of an optimization technology, it is crucial to structure the problem. For example,

1. Scope: Finish BBS-Start Erection
2. Goal:
   - Minimize the number of extra storage areas
   - Minimize the cost of block handling (e.g. stacking) and transportation.
3. Constraints:
   - Fixed BBS finish date
   - Fixed erection loading start date
   - Number of slots in the BBS storage and the grand-block storage.

The extension could be starting from the finish assembly with actual assembly finish date and the fixed erection sequence. It means covering the dynamics in the preceding fabrication & assembly processes and improving the finish erection (launching) date.

Fortunately, the Centre for Mathematical Modelling and Simulation, Bandung Institute of Technology, Indonesia and Prof Jan Bisschop from the Department of Applied Mathematics, University of Twente, Netherlands offered helps to the shipyard for applying an appropriate optimization technology to the problem. The preliminary result will be published in the next paper.

5 Conclusion

This paper provides details of the block storage problem at PT PAL Indonesia and the use of simple heuristics to solve the problem.

Acknowledgment

The author would like to thank to Prof Jan Bisschop from the Department of Applied Mathematics, University of Twente, The Netherlands and Dr Andonowati & Yudith Prasasti from the Centre for Mathematical Modelling and Simulation ITB, Indonesia who help the author to find a better solution of the block storage problem by applying an optimization technology.

References

[1] Krause, M., F. Roland, D. Steinhauer and M. Heinemann (2005), Discrete Event Simulation: An Efficient Tool to Assist Shipyard Investment and Production Planning, *Journal of Ship Production*, **20-3**, 176–182.

[2] Morton, T.E. and D.W. Pentico (1993), *Heuristic Scheduling Systems*, John Wiley & Sons Inc., New York.

[3] Park K., K. Lee, S. Park and S. Kim (1996), Modelling and Solving the Spatial Block Scheduling Problem in Shipbuilding Company, *Computers Industrial Engineering*, **30-3**, 357-364.

ERWIN PRITANTO PHD: Merchant Ship Division, PT PAL Indonesia, Ujung Surabaya PO.BOX.1134, Indonesia. Phone: +62-31-3292275 (Ext.4281)
E-mail: Erwin_Pritanto@pal.co.id

An Optimization Model for Stacking

Johannes Bisschop

Department of Applied Mathematics, University of Twente, The Netherlands

Abstract: During this presentation the mathematical formulation of a basic stacking optimization model with known block arrival and departure moments will be developed. This model is based on the a priori construction of admissible stacks, and minimizes the number of extra storage slots required to place all blocks. The model is a binary-programming model. This type of model is not guaranteed to solve in polynomial time. Nevertheless, optimal solutions have been found for model instances with up to 150 blocks, where the measured solution time was in the order of minutes. Next, some extensions of the problem together with the corresponding formulation are presented in detail. The talk ends with an overview of potential model extensions that have not yet been implemented.

A Simulation Model for Stacking

Yudith Prasasti

Center for Mathematical Modelling and Simulation ITB, Indonesia

Abstract: In an operational setting it is quite natural to design simple decision rules for the placement of blocks in stacks. These rules can then be used not only by the storage manager, but also by the person that develops a simulation model of the overall ship construction process. The design goal behind each rule is to minimize the number of so-called shifts, i.e. any extra block movements after a block has been placed. During this talk, some simple block placement decision rules are explained and illustrated for a few selected small examples. The results are then compared with the results derived from the optimization model.

Nonnegative Solutions of the Hamilton-Jacobi Equations

Yudi Soeharyadi

Department of Mathematics, Institut Teknologi Bandung, Indonesia)

Abstract: Nonnegative solutions to the Hamilton Jacobi equation of the form $u_t + H(\nabla u) + G(\cdot, u) = 0$ are considered, where the Hamiltonian H and the perturbation term G satisfy certain regularity. The theory of nonlinear semigroup of operators necessitates us to work with time dependent equation. Lack of regularity forces the use elliptic regularization. Regularity of solutions in term of Burch classes will also be discussed.

Keywords: Hamilton-Jacobi equations, nonnegative solutions, Burch classes, regularity

Fractional Integral Operators

H. Gunawan[1], Eridani[2]

[1)] Department of Mathematics, Institut Teknologi Bandung, Indonesia
[2)] Department of Mathematics, Campus C, Mulyorejo, Airlangga University, Surabaya 60115, Indonesia

Abstract: In this talk I will present some recent results on generalized fractional integral operators (and their modified versions) on generalized Morrey spaces (and generalized Campanato spaces, respectively). These results were obtained by Nakai (2001), Eridani (2002), Gunawan (2003), and Eridani, Gunawan and Nakai (2004). Some possible applications will also be discussed.

Keywords: Fractional integral operator, Riesz potentials, function spaces, boundedness

Riesz Potentials in BMO and L^{∞}

Eridani

Department of Mathematics, Campus C, Mulyorejo, Airlangga University, Surabaya 60115, Indonesia

Abstract: In this paper we prove the boundedness of Riesz potential I_α from Morrey spaces to Campanato spaces. Actually, in this case, we consider only functions with compact support. We also consider the special case of our result. That is, we consider I_α in BMO and L^{∞}.

A LEVEL SET METHOD FOR MULTI-PHASE STEAM DISPLACING OIL IN A SATURATED POROUS MEDIUM

M. Muksar[1], E. Soewono[2], S. Siregar[3]

[1] Dept. of Mathematics UM, Malang, Indonesia
[2] Dept. of Mathematics ITB, Bandung, Indonesia
[3] Dept. of Petroleum Engineering ITB, Bandung, Indonesia

Abstract. A one-dimensional model of multi-phase steam displacing oil in a saturated porous medium is considered here. We consider the fluids are incompressible displacement and the porous medium is flat and homogeneous. Using the improved interface model, the interface, separating the steam zone from the liquid zone, is considered as a moving internal boundary. The model equations are derived from the mass and energy balances, coupled with the Darcy's law for multi-phase flow. The interface velocity is determined from Rankine-Hugoniot condition of the mass and energy balances. We investigate numerically the model using a level set method. We discretize the model equations by higher order TVD Runge-Kutta scheme in time and conservative finite difference WENO scheme in space. Using varying oil viscosities and steam injection velocities, some numerical results are obtained.

Key-words: Level set method, interface, multi-phase flow, higher order

1 Introduction

A problem of steam displacing oil in a saturated porous medium has been investigated experimentally and numerically [12, 5], and analytically [3, 8, 9]. An important characteristic of their model was the occurring of steam condensation front (interface) as an internal boundary which separates the steam zone (consisting of water, oil, and steam) from the liquid zone (consisting of only water and oil). The temperature was assumed to be constant, which is the boiling temperature in the steam zone and the reservoir temperature in the liquid zone. The interface velocity was given as a constant velocity, calculated from the local heat balance.

In this paper, we improved a one dimensional interface model of the problem. The model equations are derived from the mass and energy (enthalpy) balances, coupled with the Darcy's law for multi-phase flow (water, oil, and steam) without the capillary and gravity effects. The interface velocity, assumed to be equal to the total Darcy's velocity in the liquid zone, is determined from the Rankine-Hugoniout condition of the mass and the energy balances. The temperature not assumed to be constant anymore.

The evolution of the interface is computed implicitly by applying a level set method. The main idea of the method is that the interface is always represented

as a zero level set of a smooth function defined on whole domain. In the implementation of the method, we only need to define the level set function in a neighborhood of the interface. The model equations are discretized by higher order TVD (total-variation-diminishing) Runge-Kutta scheme in time and a conservative finite difference WENO (weighted essentially non-oscillatory) scheme in space [6, 7, 13].

In section 2, we describe the improved one dimensional interface model of the problem, and the governing equation of the model is presented in section 3. In the section 4, we write the level set method and its algorithm according to the problem, and in the section 5 the discretizations is given. Some numerical results are shown in section 6 and in the last section we draw the conclusion and discussion.

2 Improved One Dimensional Interface Model

Here we improve the one dimensional interface model done by Bruining & van Duijn [3]. We consider the porous medium with a length L is flat and homogeneous with constant porosity φ. Therefore we ignore the gravity effect and the model is investigated in one dimension. The fluid is assumed to be incompressible displacement and the oil is a distillable component. We then disregard the capillary effect from the fluid flow. The interface is assumed as a steam condensation front where all steam condenses to become water, and it separates the steam zone (consisting of water, oil, and steam) from the liquid zone (consisting of water and oil) (see Figure 1 (i)).

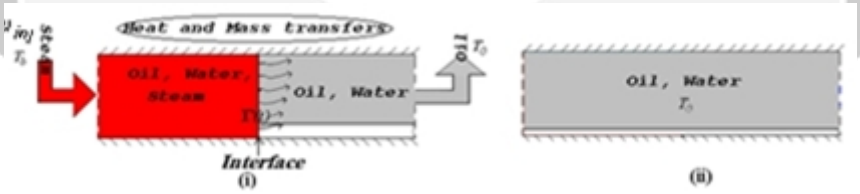


Figure 1: (i) Improved interface model (ii) Initial condition

Initially, the porous medium consist of only oil and connate water at the reservoir temperature T_0, see Figure 1 (ii). The steam is injected into the injection well with a constant velocity u_{inj} at the boiling temperature T_b. It heats and drives the liquid toward the production well and the oil will be produced within the production temperature. Hence the fluids transport and heat transfer occur in the porous medium, governed by the mass and heat balances coupled with the Darcy's law for multi-phase flow. In the steam zone, a process of steam condensation occurs while the heat losses away from the boiling temperature to a prior unknown interface temperature T_{int} at the interface where all steam abruptly condenses to become water. Therefore we assume the fluids saturation and gradient temperature are discontinuous at the interface. In the liquid zone, liquids move

forward with a prior unknown interface velocity V while the heat losses away from the interface temperature to the production temperature at the production well. Here, we assume that the production temperature is equal to the initial reservoir temperature at the production well. The interface velocity V is determined from the Rankine-Hugoniout conditions of the mass and energy balances.

3 Governing Equation

We denote the water, steam, and oil saturations, temperature, and total Darcy's velocity by $S_w(x,t)$, $S_g(x,t)$, $S_o(x,t)$, $T(x,t)$, and $u(x,t)$ respectively. The functions are defined on a domain $\mathcal{D} = [0,L] \times [0, t_{max}]$, where t_{max} is a time spent by the interface to the production well.

The mass and energy balances, coupled with the Darcy's law for multi-phase flow without capillary and gravity effects read as follows.

$$\varphi \partial_t \left(\rho_o S_o\right) + \partial_x \left(\rho_o u f_o\right) = 0, \tag{1}$$
$$\varphi \partial_t \left(\rho_{w,w} S_w\right) + \partial_x \left(\rho_{w,w} u f_w\right) = q, \tag{2}$$
$$\varphi \partial_t \left(\rho_{g,w} S_g\right) + \partial_x \left(\rho_{g,w} u f_g\right) = -q, \tag{3}$$
$$\partial_t \left[(1-\varphi) H_r + \varphi(S_o H_o + S_w H_w + S_g H_g)\right] + \partial_x \left[u \left(f_o H_o + f_w H_w + f_g H_g\right)\right] = \partial_x \left[\kappa(T) \partial_x T\right], \tag{4}$$

where $f_i = \frac{\lambda_i}{\sum_{j=w,g,o} \lambda_j}$, $\lambda_i = -k \frac{k_{ri}}{\mu_i}; i = w,g,o$, $k_{rw} = \left(\frac{S_w - S_{wc}}{1-S_{wc}}\right)^4$, $k_{ri} = \left(\frac{S_i}{1-S_{wc}}\right)^4$; $i = g,o$, $H_r = \rho_r C_P^r \left(T - T_0\right)$, $H_w = \rho_{w,w} C_P^w \left(T - T_0\right)$, $H_g = \rho_{g,w} C_P^{g,w} \left(T - T_0\right) + \rho_{g,w} \Lambda$, $H_o = \rho_o C_P^o \left(T - T_o\right)$, and $q \geq 0$ is a steam condensation rate.

We can show if $S_w = S_{wc}$, $S_g = 0$, or $T = T_b$ then $q = 0$, implying $u = u(t)$. Therefore, we consider $q = Q(S_w - S_{wc}) S_g (T_b - T)$ for a positive constant Q.

By scaling $S_{wD} := \frac{S_w - S_{wc}}{1-S_{wc}}$, $S_{iD} := \frac{S_i}{1-S_{wc}}; i = g,o$, $T_D := \frac{T - T_0}{T_b - T_0}$, $x_D := \frac{x}{L}$, $u_D := \frac{u}{u_{inj}}$, $t_D := \frac{u_{inj} t}{\varphi L (1 - S_{wc})}$, and neglecting the indexes D, we rewrite (1), (2), (3), and (4) as follows.

$$\partial_t \left(\rho_o S_o\right) + \partial_x \left(\rho_o u f_o\right) = 0, \tag{5}$$
$$\partial_t \left(\rho_{w,w} S_w\right) + \partial_x \left(\rho_{w,w} u f_w\right) = Q^* S_w S_g (1 - T), \tag{6}$$
$$\partial_t \left(\rho_{g,w} S_g\right) + \partial_x \left(\rho_{g,w} u f_g\right) = -Q^* S_w S_g (1 - T), \tag{7}$$
$$\partial_t \left[\frac{(1-\varphi) H_r + \varphi S_{wc} H_w}{\varphi (1 - S_{wc})} + S_w H_w + S_g H_g + S_o H_o\right] + \partial_x \left[u \left(f_w H_w + f_g H_g + f_o H_o\right)\right] = \left(\frac{\Delta T}{u_{inj} L}\right) \partial_x \left[\kappa(T) \partial_x T\right], \tag{8}$$

with initial and boundary conditions

$$S_w(x,0) = \begin{cases} 0, \\ 0, \end{cases} \quad S_o(x,0) = \begin{cases} 0, \\ 1, \end{cases} \quad T(x,0) = u(x,0) = \begin{cases} 1, & \text{if } x = 0, \\ 0, & \text{if } 0 < x \leq 1, \end{cases} \tag{9}$$

$$S_w(0,t) = 0, \quad S_w(1,t) = 0, \quad S_o(0,t) = 0, \quad S_o(1,t) = 1,$$
$$T(0,t) = 1, \quad T(1,t) = 0, \quad u(0,t) = 1, \quad u(1,t) = u^+(t), \tag{10}$$

where $Q^* = QL(1 - S_{wc})^2 \Delta T$ and $\Delta T = T_b - T_r$.
Note that we also use $\mathcal{D}$ for the dimensionless domain.

By adding (5), (6), (7), and using $\sum_{j=w,g,o} S_j = 1$ and $\sum_{j=w,g,o} f_j = 1$, we get an equation for the total Darcy's velocity as follow.

$$\frac{\partial u}{\partial x} = -\frac{Q^*}{\rho} S_w(1 - S_w - S_o)(1 - T), \quad \text{where} \quad \rho = \frac{\rho_{w,w}\rho_{g,w}}{\rho_{w,w} - \rho_{g,w}}. \tag{11}$$

Using (11), we can show that (6) and (7) are equivalen. We then simplify and rewrite the governing equtations (5), (6) (8), and (11) in the form

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{G}(\mathbf{U})}{\partial x} = 0, \quad \frac{\partial u}{\partial x} = -\frac{Q^*}{\rho} S_w(1 - S_w - S_o)(1 - T), \tag{12}$$

where $\mathbf{U} = (S_w, S_o, (\alpha_1 + \alpha_3 + (\alpha_2 - \alpha_3)S_w + (\alpha_4 - \alpha_3)S_o)\, T)^T$, $\mathbf{G}(\mathbf{U}) = \left(u(f_w + \frac{\rho}{\rho_{w,w}}), uf_o, u((\alpha_3 + (\alpha_2 - \alpha_3)f_w + (\alpha_4 - \alpha_3)f_o)T + \frac{\rho\Lambda}{\Delta T}) - (\frac{\kappa(T)}{u_{inj}L})\frac{\partial T}{\partial x}\right)^T$, and $\alpha_1 = ((1-\varphi)\rho_r C_P^r + \varphi S_{wc}\rho_{w,w}C_P^w)\,/\,(\varphi(1 - S_{wc}))$, $\alpha_2 = \rho_{w,w}C_P^w$, $\alpha_3 = \rho_{g,w}C_P^{g,w}$, $\alpha_4 = \rho_o C_P^o$ are positive constants.
Here we will investigate a class solution of piecewise smooth functions which are discontinuous at the interface.

Let the interface $x = \Gamma(t)$ move with a velocity $V(t)$ equal to a prior unknown liquid velocity in the liquid zone $u^+(t)$. We denote the steam zone by $\Omega = [0, \Gamma(t))$, and the liquid zone by $\Omega^c = (\Gamma(t), 1]$. We set $\kappa(T) = \kappa_g$, $(x,t) \in \Omega$, and $\kappa(T) = \kappa_l$, $(x,t) \in \Omega^c$. Then we summarize the equation for each zones, and the interface conditions according to the improved one dimensional interface model as follows.

3.1 Steam Zone Problem (SZP)

Using $\kappa(T) = \kappa_g$ and from (12) we get governing equations in the steam zone as follows.

$$\frac{\partial \mathbf{U}^g}{\partial t} + \frac{\partial \mathbf{G}^g(\mathbf{U}^g)}{\partial x} = 0, \quad \frac{\partial u^g}{\partial x} = -\frac{Q^*}{\rho} S_w^g(1 - S_w^g - S_o^g)(1 - T^g), \tag{13}$$

where $\mathbf{U}^g = (S_w^g, S_o^g, (\alpha_1 + \alpha_3 + (\alpha_2 - \alpha_3){S_w}^g + (\alpha_4 - \alpha_3)S_o^g)T^g)^T$, $\mathbf{G}^g(\mathbf{U}^g) = \left(u^g(f_w^g + \frac{\rho}{\rho_{w,w}}), u^g f_o^g, u^g((\alpha_3 + (\alpha_2 - \alpha_3)f_w^g + (\alpha_4 - \alpha_3)f_o^g)T^g + \frac{\rho\Lambda}{\Delta T}) - (\frac{\kappa_g}{u_{inj}L})\frac{\partial T^g}{\partial x}\right)^T$,

with initial and boundary conditions

$$S_w(x,0)=0, \quad S_o(x,0)=0, \quad T(x,0)=1, \quad u(x,0)=1, \tag{14}$$

$$S_w(0,t)=0, \quad S_w(\Gamma(t),t)=S_w^-, \quad S_o(0,t)=0, \quad S_o(\Gamma(t),t)=S_o^-,$$
$$T(0,t)=1, \quad T(\Gamma(t),t)=T_{int}, \quad u(0,t)=1, \quad u(\Gamma(t),t)=u^-. \tag{15}$$

3.2 Liquid Zone Problem (LZP)

Because of $S_g=0$ (there is no steam in the liquid zone), the total Darcy's velocity u is only a prior unknown time independence, i.e. $u=u^+(t)$. Hence we simplify the equations (12) using $\kappa(T)=\kappa_l$ to get

$$\frac{\partial \mathbf{U}^l}{\partial t}+\frac{\partial \mathbf{G}^l(\mathbf{U}^l)}{\partial x}=0, \tag{16}$$

where $\mathbf{U}^l=\left({S_o}^l,\left(\alpha_1+\alpha_2+(\alpha_4-\alpha_2){S_o}^l\right)T^l\right)^T$,
$\mathbf{G}^l(\mathbf{U}^l)=\left(u^+f_o^l, u^+(\alpha_2+(\alpha_4-\alpha_2)f_o^l)T^l-(\frac{\kappa_l}{u_{inj}L})\frac{\partial {T^l}}{\partial x}\right)^T$,
with respect to initial and boundary conditions

$$S_o(x,0)=1, \quad T(x,0)=0, \quad u(x,0)=0, \tag{17}$$

$$S_o(\Gamma(t),t)=S_o^+, \quad S_o(1,t)=1, \quad T(\Gamma(t),t)=T_{int}, \quad T(1,t)=0. \tag{18}$$

3.3 Interface Condition

Using the Rankine-Hugoniout condition for (12) and by assuming that the interface velocity $V(t)$ is equal to a prior unknown $u^+(t)$, from our calculations we get interface conditions as follows. The interface velocity is given by

$$V=u^-+(\frac{\Delta T}{\rho u_{inj}L})\frac{(\kappa_l {T_x}^l-\kappa_g {T_x}^g)}{((C_P^{g,w}-C_P^{w,w})\Delta T T_{int}+\Lambda)}, \tag{19}$$

S_o^- is one of the roots of $(1-S_w^--S_o^-)\,V=\frac{\rho}{\rho_{g,w}}V-u^-\left(f_w^-+f_o^-+\frac{\rho}{\rho_{w,w}}\right)$,
S_w^+ is one of the roots of $(1-S_w^+-S_o^-)\,V=V(1-f_w^+)-u^-f_o^-$, and S_w^- is given. Another conditions u^- and T_{int} are obtained from (11) and the local heat balance in (12).

4 Level Set Method

Here we use an idea of a level set method for tracking a discontinuity in hyperbolic conservation laws done by Aslam [1, 2] to track the interface. The interface is represented by zero level set of a smooth function Φ constructed and defined on $\mathcal{D}$. One of the advantages of the method is that the interface is never explicitly computed from the interface velocity. Hence, we could increase the order of the accuracy depending on the scheme used to approximate the function Φ.

Here, we modify the domain only on a neighborhood of the interface, which moves with respect to time. The modification should reduce the computation time. The method successfully have applied for two-phase steam displacing water in a saturated porous medium as a simplified of the problem by assuming there is no oil in the porous medium, see [10, 11].

4.1 Construction of Level Set Equation

Suppose we have an initial interface $\frac{d\Gamma(t)}{dt}|_0 = V(0)$ and interface $x = \Gamma(0) = \Gamma_0$. We want to construct a smooth level set function $\Phi(x,t)$ in which the evolution of the interface $x = \Gamma(t)$ is computed from a zero level set of $\Phi(x,t)$, i.e. $\Phi(\Gamma(t),t) = 0$. Initially, we extend the initial condition in each zones (14), (17) onto the whole domain $[0,1]$ to get

$$\mathbf{U}_0^g(x) = (S_{w0}{}^g(x), S_{o0}{}^g(x), T_0^g(x))^T, \quad u_0^g(x), \tag{20}$$
$$\mathbf{U}_0^l(x) = (S_{o0}{}^l(x), T_0^l(x))^T, \quad u_0^+(t). \tag{21}$$

By using (13) - (20) and (16) - (21), we then obtain smooth solutions

$$\mathbf{U}^g(x,t) = (S_w^g(x,t), S_o^g(x,t), T^g(x,t))^T, \quad u^g(x,t), \tag{22}$$
$$\mathbf{U}^l(x,t) = (S_o^l, T^l(x,t))^T, \quad u^+(t), \tag{23}$$

defined on $\mathcal{D}$ with $1 \geq T^g(x,t) \geq T^l(x,t) \geq 1$ respectively. Note that the extended functions (20) and (21) are not unique. Therefore we choose the extensions such that the solutions (22) and (23) are obtained.

Next, we define a speed function $F(x,t)$ as continuous extension of the interface velocity $V(t)$ by

$$F(x,t) = u^- + \left(\frac{\Delta T}{\rho u_{inj} L}\right)\left(\frac{\kappa_w {T^l}_x{}^+ - \kappa_s {T^g}_x{}^-}{(C_P^{g,w} - C_P^{w,w})T_{int} + \Lambda}\right). \tag{24}$$

Since T_x^g and T_x^l are only continuous, then $F(x,t)$ is a continuous too. For numerical computations, we smoothen $F(x,t)$ in the neighborhood of the interface to get a smooth extension of V. This will be discussed in subsection4.2.
Now we look for a level set function $\Phi(x,t)$, moving with speed function $F(x,t)$, in which $\Gamma(t)$ is a zero level set of this $\Phi(x,t)$, i.e. $\Phi(\Gamma(t)) = 0$. We consider a level set $\Phi(x,t) = c$ for any constant c. By using the chain rule and the fact that the speed F along this level set is $dx/dt = F(x,t)$, we have

$$\partial_t \Phi + F\partial_x \Phi = 0, \ (x,t) \in \ \mathcal{D}, \ \Phi(x,0) = \Phi_0, \tag{25}$$

where the initial condition Φ_0 is given as a distance function $\Phi_0(x,0) = x - \Gamma(0)$. From the smooth solution $\Phi(x,t)$ of (25), we obtain an interface as a zero level set of $\Phi(x,t)$. Furthermore, the piecewise smooth solution of (12) is defined by

$$(\mathbf{U}(x,t), \ u(x,t)) = \begin{cases} (\mathbf{U}^g(x,t), \ u^g(x,t)) \text{ jika } \Phi(x,t) < 0, \\ (\mathbf{U}^l(x,t), \ u^+(t)), \text{ jika } \Phi(x,t) \geq 0. \end{cases} \tag{26}$$

Furthermore we investigate the solution Φ of the level set equation (25). Because of the speed function $F(x,t)$ is not constant, then the shape of solution Φ in (25) will change and no longer to be a distance function even after some short time. To anticipate a loss of the numerical error, we reinitilize (in the numerical iteration) the initial condition of (25) with a distance function in each timestep using

$$\partial_\tau \Phi_d = sgn(\Phi)(1 - \partial_x \Phi_d), \tag{27}$$

to steady state, where $\Phi_d(x,0) = \Phi(x,t)$ is not a distance function, and sgn is a sign function, suggested by [4].

4.2 Continuous Extension of Speed Function

Suppose for each t, the speed function $F(x,t)$ as an extension speed function of V away from the interface $\Gamma(t)$ onto the domain $\mathcal{D}$, i.e $F(\Gamma(t),t) = V(t)$. Since F is not smooth at this interface, it may contribute to large numerical errors in the computation of Φ in (25). Here we smoothen the function F at the interface by solving the advection equation suggested in [4]

$$\partial_\tau F + sgn(\Phi_d)\partial_x F = 0,\ u(x,0) = u(x,t) \tag{28}$$

where Φ_d a distance function from $\Gamma(t)$. The solution F of (28), will smoothly extend V away from the interface, and the value of V on the interface does not change because Φ_d is zero on the interface, $\Phi_d(\Gamma(t),t) = 0$.

4.3 Algorithm

In our numerical computation, we start the method only at the initial interface Γ_0 with the initial interface velocity $V(0)$. Therefore the method is applied only on the neighborhood of this initial interface, $\mathcal{D}_r$, and we repeat the method for some time. We can now summarize the method in an algorithm as follows.

1. Initialize $S_{w0},\ S_{g0},\ T_0,\ u_0, \Gamma_0$ and Φ_0 as a distance function from Γ_0.
2. Define the extended initial condition $\mathbf{U}_g(x,0) = (S^g_{w0}(x), S^g_{o0}(x), T^g_0(x))^T$, $u^g_0(x)$, and $\mathbf{U}_l(x,0) = (S^l_{o0}, T^l_0(x))^T$ onto whole domain [0,1].
3. Find $\mathbf{U}^g(x,t) = (S^g_w(x,t), S^g_0(x,t), T^g(x,t))^T$, $u^g(x,t)$, from (13) - $\mathbf{U}_g(x,0)$, $u_0(x)$ and $\mathbf{U}^l(x,t) = (S^l_0, T^l(x,t))^T$, from (16) - $\mathbf{U}_l(x,0)$.
4. Compute the speed function $F(x,t)$ using (24) and then be smoothen by (28).
5. Construct $\Phi(x,t)$ using (25) and then reinilize to be exact distance function Φ_d by solving (27).
6. Compute the new interface as a zero of the distance function Φ_d.
7. Define the solution by $(\mathbf{U}(x,t),\ u(x,t)) = \begin{cases} (\mathbf{U}^g(x,t),\ u^g(x,t)) \text{ if } x \le \Gamma(t), \\ (\mathbf{U}^l(x,t),\ u^+(t)),\ \text{if } x > \Gamma(t). \end{cases}$
8. Repeat the steps 2 through 7 for the next timestep.

5 Discretization

In the numerical calculations, we take uniform grid sizes $\Delta x = \frac{1}{N}$ for the domain [0,1], where $N+1$ is the number of gridpoins, and $\Delta x_r = \frac{\Delta x}{N_r+1}$ for the restricted (refining) domain $[x_{ri} - \delta, x_{ri+1} + \delta]$, where $x_{ri} < \Gamma(t) \leq x_{ri+1}$, N_r is the number of points added in an interval $[x_i, x_{i+1}]$, and $\delta = 1.5(N_r+1)\Delta x_r$. The main timestep Δt is depending on the grid size Δx and the CFL number of the TVD Runge-Kutta scheme. The numerical solution of s is denoted by $s_i^n = s(x_i, t_n)$, where $x_i = (i-1)\Delta x,\ i = 1...N+1$ and $t_n = n\Delta t,\ n = 1...n_{maxiter} = t_{max}/\Delta t$.

Recall now that equations (13), (16) (25), (27), and (28) need to be discretized. In general, the partial differential equation in (13) and (16) can be written in form

$$\partial_t s + \partial_x f(s) = \partial_x Q(s, s_x). \tag{29}$$

5.1 TVD Runge-Kutta Time Discretization

To avoid any instabilities arising from the evolutions of (25), (27), and (29) we use a third order TVD Runge-Kutta scheme in time as described in [13]. We consider

$$s_t = L(s), \tag{30}$$

L is the spatial operator of either (25), (27), or (29) . The time discretization of (30) is

$$\begin{aligned} s^{(1)} &= s^n + \Delta t L(s^n), \\ s^{(2)} &= (3/4)s^n + (1/4)s^{(1)} + (1/4)\Delta t L(s^{(1)}), \\ s^{n+1} &= (1/3)s^n + (2/3)s^{(1)} + (2/3)\Delta t L(s^{(2)}), \end{aligned} \tag{31}$$

where $\Delta t \leq c\Delta x/\alpha$, c is a CFL number, $\alpha = \max|f'(s)|$.

5.2 Spatial Discretization

In order to apply the fifth order WENO scheme, we rewrite (29) in the form

$$s_t = -\partial_x f(s) + \partial_x Q(s, s_x) \equiv L1(s). \tag{32}$$

$L1(s)$ is approximated by $L1_i^n = -\frac{\hat{f}^n_{i+1/2} - \hat{f}^n_{i-1/2}}{\Delta x} + \hat{Q}_i^n$. $\hat{Q}_i^n$ is the fourth order central difference approximating to the diffusion term $Q(s, s_x)$, see [10, 11]. Using a conservative global Lax-Friedrichs flux splitting, $\hat{f}^n_{i+1/2}$ is written by

$$\begin{aligned} \hat{f}^n_{i+1/2} &= \hat{f}^{+n}_{\ i} + \hat{f}^{-n}_{\ i+1}, \\ \text{where} \quad \hat{f}^{+n}_{\ i} &= WENO5(f^{+n}_{\ i-2}, f^{+n}_{\ i-1}, f^{+n}_{\ i}, f^{+n}_{\ i+1}, f^{+n}_{\ i+2}), \\ \hat{f}^{-n}_{\ i+1} &= WENO5(f^{-n}_{\ i+3}, f^{-n}_{\ i+2}, f^{-n}_{\ i+1}, f^{-n}_{\ i}, f^{-n}_{\ i-1}), \\ f^{+n}_{\ i} &= (f(s_i^n) + \alpha s_i^n)/2, \quad f^{-n}_{\ i} = (f(s_i^n) - \alpha s_i^n)/2, \\ \alpha &= \max_s(|f'(s)|). \end{aligned} \tag{33}$$

The fifth order WENO scheme, $WENO5$, is defined as in [13].
For the level set equation (25), we rewrite in the form

$$\partial_t \Phi = -F\Phi_x \equiv L2(\Phi). \tag{34}$$

$L2(\Phi)$ is approximated by $L2_i^n = -F_i^n \hat{\Phi}_{x_i}^n$. Then $\hat{\Phi}_x$ is approximated using fifth order WENO scheme for Hamilton-Jacobi equation [6]. Define $D^-{}_i^n = (\Phi_i^n - \Phi_{i-1}^n)/\Delta x$, $D+_i^n = (\Phi_{i+1}^n - \Phi_i^n)/\Delta x$,then the $\hat{\Phi}_x$ is given by

$$\hat{\Phi}_x = \begin{cases} WENO5(D^-{}_{i-2}^n, D^-{}_{i-1}^n, D^-{}_i^n, D^-{}_{i+1}^n, D^-{}_{i+2}^n) & \text{if } F_i^n \geq 0, \\ WENO5(D^+{}_{i+2}^n, D^+{}_{i+1}^n, D^+{}_i^n, D^-{}_{i+1}^n, D^-{}_{i+2}^n) & \text{otherwise.} \end{cases}$$

By applying the process as used in (25) we discretize similarly the equation (27) using the function $sgn(\Phi)$ smoothed by $sgn_\epsilon(\Phi)_i^n = \Phi_i^n / \sqrt{\Phi_i^{n2} + \epsilon^2}$, $\epsilon = \Delta x_r$, and a time step $\Delta\tau = \Delta x_r/5$. As done in [4], we discretize the equation (28) by a first order upwind scheme using $CFL = \Delta\tau/\Delta x$ number 0.5 as follow.

$$\begin{cases} \text{if } sgn_i^n(\Phi) > 0, & \text{then } u^{newn}{}_i = u^{oldn}{}_i - 0.5(u^{oldn}{}_i - u^{oldn}{}_{i-1}), \\ \text{if } sgn_i^n(\Phi) < 0, & \text{then } u^{newn}{}_i = u^{oldn}{}_i + 0.5(u^{oldn}{}_{i+1} - u^{oldn}{}_i). \end{cases}$$

Note that if $x_{j-1} < \Gamma < x_j$ for some j, then $u^{newn}{}_j = u^{oldn}{}_j - 0.5(u^{oldn}{}_i - V)$ and $u^{newn}{}_{j-1} = u^{oldn}{}_{j-1} + 0.5(V - u^{oldn}{}_{j-1})$.

6 Numerical Results

Here, we use the steam injection u_{inj}, the oil viscosity μ_o, and the left shock water saturation S_w^- as parameters. We plot the evolutions of the temperature, fluids saturation, and total Darcy's velocity at $t = 0$, $t = \frac{t_{max}}{4}$, $t = \frac{t_{max}}{3}$, $t = \frac{t_{max}}{2}$ and $t = t_{max}$, and the interface velocity evolution, shown in Figure 2 as follow.

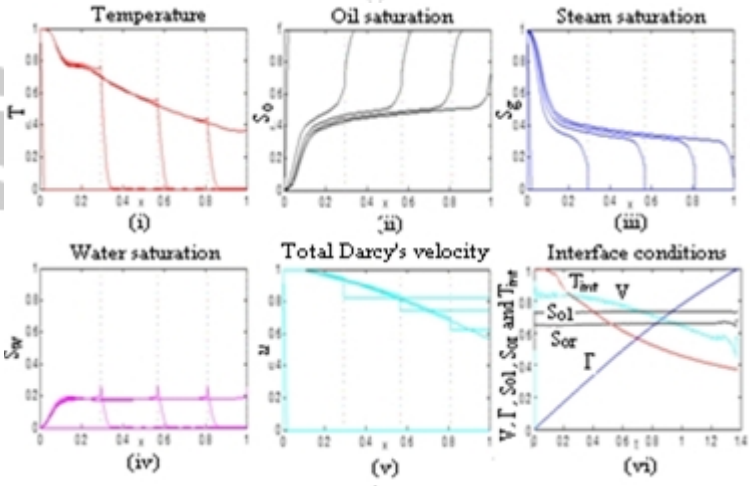


Figure 2: Evolutions of fluids saturation, temperature, total Darcy's velocity, and interface condition with $u_{inj} = 5.00$ e-05, $\mu_o = 2.45$ e-03, $S_w^- = 0.20$.

The Figure indicate that the gradient temperature, fluids saturation, and total Darcy's velocity are discontinuous at the interface (see (i) - (v)). The temperature and velocity of the interface are decreasing with respect to the time and position. The left and right shock oil saturations are constant.

By using some steam injection velocities and left shock water saturations with the fixed oil viscosity, we have the numerical results, given in the Figure 3 and Table 1 as follows. This numerical results show that the interface velocity and sweeping

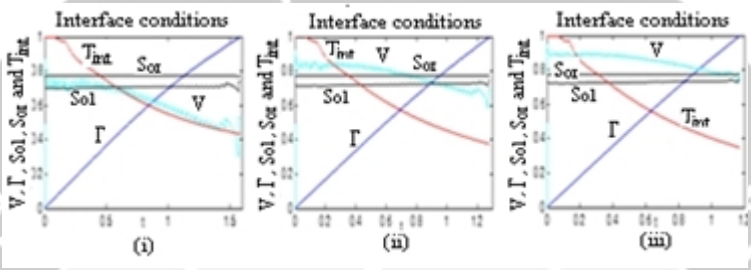


Figure 3: Interface conditions with (i) u_{inj} = 3.02 e-05, (ii) u_{inj} = 5.00 e-05, (iii) u_{inj} = 7.50 e-05, and fixed μ_o = 5 e-03, S^- = 0.15.

Table 1: Total oil recovery and sweeping time with respect to parameter u_{inj} with μ_o = 2.45 e-03 and S_w^- = 0.15 (left), S_w^- = 0.20 (right).

u_{inj}	Total oil recovery (%)		Sweeping time (hour)	
(m/s)	S_w^- = 0.15	S_w^- = 0.20	S_w^- = 0.15	S_w^- = 0.20
3.02 e-05	55.92	58.44	489.907	575.024
5.00 e-05	55.03	57.23	234.642	247.777
7.50 e-05	54.66	56.86	141.885	146.223

time t_{max} depend on the steam injection velocity u_{inj}, but the total oil recovery does not. The left shock water saturation influence to the total oil recovery and time sweeping.

When we use some oil viscosities with fixed the steam injection velocity, we have the numerical results, shown in the Figure 4 and Tabel 2 as follows.

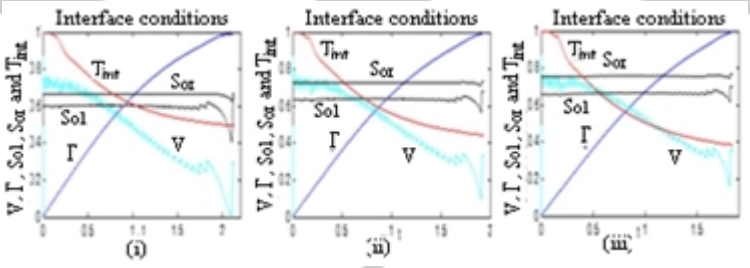


Figure 4: Interface conditions with (i) μ_o = 1 e-03, (ii) μ_o = 2.45 e-03, (iii) μ_o = 5 e-03 , and fixed u_{inj} = 3.02 e-05, S^- = 0.20 tetap.

Table 2: Total oil recovery and sweeping time with respect to parameter μ_o with $u_{inj} =$ 5.00 e-05 and $S_w^- = 0.15$ (left), $S_w^- = 0.20$ (right).

μ_o (Pa s)	Total oil recovery (%)		Sweeping time (Jam)	
	$S_w^- = 0.15$	$S_w^- = 0.20$	$S_w^- = 0.15$	$S_w^- = 0.20$
1 e-03	59.35	61.64	240.115	254.345
2.45 e-03	55.03	57.28	234.642	247.777
5 e-03	51.61	53.86	230.281	243.860

The results show that the total oil recovery depends on the oil viscosity μ_o, but the interface velocity and sweeping time t_{max} do not.

Furthermore, when we assume that the temperature at the steam zone is the constant boiling temperature, we get that the interface velocity is constant and the temperature is decreasing drastically from the boiling temperature at the interface to the reservoir temperature away from the interface as shown within Figure 5. This result is similar to the result of [3] within isothermal condition.

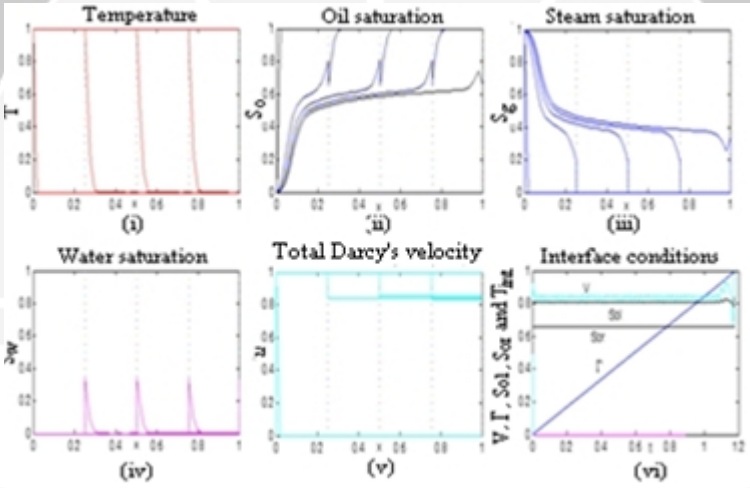


Figure 5: Evolutions of Fluids Saturation, Temperature, Total Darcy's Velocity, and interface condition by assuming the constant boiling temperature at the steam zone with $u_{inj} = 5.00$ e-05, $\mu_o = 2.45$ e-03, $S_w^- = 0.20$.

7 Conclusion and Discussion

From our investigation we have some concluding remarks as follows. (1) The level set method can be applied to multi-phase steam displacing oil in a saturated porous medium problem in which the method is used to track implicitly the evolution of the interface as a zero of a level set function. (2) Accuracy of tracking interface using the level set method is higher order. (3) The steam injection velocity influence to the interface velocity and sweeping time. (4) The oil viscosity influences

to the total oil recovery. (5) The left shock water saturation influence to total oil recovery and time sweeping.

Some future researches should be investigated: (1) including the capillary effect, the problem will be complicated and interested. (2) Furthermore, we can consider the interface as a zone, not as a point anymore. (3) Numerical solution of the problem in two or three dimension is also interesting.

Acknowledgment

Part of the research is sponsored by Research Consortium of Oil and Gas Recovery for Indonesia (OGRINDO) Department of Petroleum Engineering ITB Bandung Indonesia .

References

[1] Aslam, T.D. (2001), A Level Set Algorithm for Tracking Discontinuities in Hyperbolic Conservation Laws I: Scalar Equations, *Journal of Computational Physics*, **167**, 413-438.

[2] Aslam, T.D. (2003), A Level Set Algorithm for Tracking Discontinuities in Hyperbolic Conservation Laws II: Systems of Equations, *Journal of Scientific Computing*, **19**, 37-62.

[3] Bruining, J., and van Duijn, C.J. (2000), Uniqueness Conditions in a Hyperbolic Model for Oil Recovery by Steamdrive, *Computational Geosciences* **4**, 65 – 98.

[4] Chen, S. Merriman, B., Osher, S., Smereka, P. (1997), A Simple Level Set Method for Solving Stefan Problems, *Journal of Computational Physics*, **135**, 8 - 29.

[5] Godderij, P. (1997), A Three Dimensional Interface Model for Steamdrive in Heterogenous Reservoir, *Thesis*, TU Delft, Netherland.

[6] Jiang, G.S., and Peng, D. (2000), Weighted ENO Schemes for Hamilton-Jacobi Equations, *SIAM J. SCI. COMPUT.*, **21**, 2126– 2143.

[7] Liu, X-D., Osher, S., Chan, T. (1994), Weighted Essentially Non-Oscillatory Schemes, *Journal Comput. Phys.*, **115**, 200 – 212.

[8] Muksar, M., Siregar, S., Soewono, E. (2002), Mathematical Model for Steam-Water Flow in a Porous Medium, *Natural Jurnal*, **6**, 116 – 120.

[9] Muksar, M., Siregar, S., Soewono, E. (2002), One Dimensional Model of Steam Displacing Water in a Saturated Porous Medium, *supl. Proc. ITB*, **34**, 197 – 219.

[10] Muksar, M., Siregar, S., Soewono, E. (2002), Level Set Method for Steam-Water Flow in a Porous Medium, *Matematika: Journal of Mathematics and Its Learning* (special edition), 1025 – 1031.

[11] Muksar, M., Soewono, E., Siregar, S. (2003), A Level Set Method for Two-Phase Steam Displacing Water in A Saturated Porous Medium, *Proceeding of the International Conference 2003 On Mathematics and Its Applications*, 309 – 320.

[12] Palmgren, C. (1992), Oil Recovery by Steamdrive; An Interface Approach, *Thesis*, TU Delft, Netherland.

[13] Shu, C. W. (1997), Essentially Non-Oscillatory and Weighted Essentially Non-Oscillatory Schemes for Hyperbolic Consevation Laws., *ICASE Report* 97-65, Langley Reseach Center Hampton, Virginia 23681-2199.

M. Muksar: Ph. D student at Department of Mathematics, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia.
Department of Mathematics Universitas Negeri Malang, Jl. Veteran 2 Malang, Indonesia.
E-mail: mmuksar@yahoo.com

E. Soewono: Department of Mathematics Institut Teknologi Bandung, Jl. Ganesa 10 Bandung 40132, Indonesia
E-mail: esoewono@bdg.centrin.net.id

S. Siregar: Department of Petroleum Engineering Institut Teknologi Bandung, Jl. Ganesa 10 Bandung 40132, Indonesia.
E-mail: ss@tm.itb.ac.id

CONVERGENCE OF FINITE DIFFERENCE APPROXIMATION FOR TWO CHANNELS DISSIPATION MODEL*)

Sumardi[a], Suparna D[a]., Lina Aryati[a]

[a] UGM, Yogyakarta, Indonesia

Abstract. In this paper the classical solution of two channel dissipation model is approximated by finite difference method. By proving the stability and consistency of this method for two channels model dissipation and using the Lax Equivalence theorem we prove that the method is convergent. Furthermore it will be demonstrated that the numerical solutions tend to the exact steady state solution as the time limits infinity.
Key-words: partial differential equation, Chemical engineering

*)This paper is part joint supervision between Gadjah Mada University and F.P.H. Van Beckum, Netherland

1 Introduction

For introducing for two channels dissipation model is think of a gas flowing in a long pipe at velocity *c*. At a certain point $x = 0$ in the pipe, let another gas be injected, creating a concentration level *A* in the flow. Suppose the injection has started at time $t = 0$, and will go on permanently ever since. Now it is of interest to see how this sudden jump of gas concentration will propagate down the pipe. What do we see?

At first sight we would think the concentration $u(x,t)$ would be carried along with the gas, satisfying the translation equation

$$u_t + c\,u_x = 0\,. \tag{1}$$

In the present case u is a step function with two levels: 0 and *A*. According to equation (1), at any time t the jump would exactly be at position $x = ct$, the concentration being still zero for $x > ct$, while for $x < ct$ it would already be *A*.

However, this is not exactly what we see in practice: we will rather see the front loosing its steepness more and more.

Intuitively we blame this to velocity differences in the flow: some particles run faster than others; along the pipe's axis the velocity might be larger than along the pipe wall; maybe there is some turbulence; particles may even change now and then from being a faster one to being a slower one or vice versa.

Think of two separated pipes, each carrying half of the flow. Pipe 1 contains the faster particles, having a speed c_1 say, and we just assume one single value c_1 for all particles in this pipe. Pipe 2 conveys the slower particles, all with speed c_2. Particles being slow first and becoming fast later, or vice versa, are modelled by crossing over from pipe 1 to pipe 2 and back.

Thus, Van Beckum (4) arrived at the following set of equations for the concentration u in pipe 1 and the concentration v in pipe 2:

$$\begin{aligned} u_t + c_1 u_x &= \alpha (v-u) \\ v_t + c_2 v_x &= \alpha (u-v) \end{aligned} \tag{2}$$

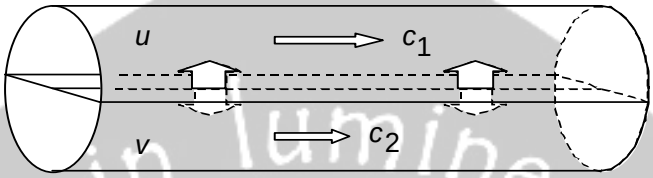


Figure 1: Two channels dissipation model

where α is the percentage of the exchanging concentration.
So instead of one PDE of second order, we now have **two of the first order**. In this paper the classical solution partial differential equations with initial values and boundary conditions is approximated by finite difference method. By proving the stability and consistency of this method and using the Lax Equivalence theorem we prove that the method is convergent. Furthermore it will be demonstrated that the numerical solutions tend to the exact steady state solution as the time limits infinity.

2 Lax equivalence theorem

Let be V Banach space, $V_0 \subseteq V$ a dense subspace V. Let $L : V_0 \subseteq V \to V$ be a linear operator. The operator L usually unbounded and can be thought of as a differential operator. Consider that initial value problem (Cauchy abstract)

$$\begin{aligned} &\frac{du(t)}{dt} = Lu(t), \qquad 0 \le t \le T, \\ &u(0) = u_0 \end{aligned} \tag{3}$$

The next definition gives the meaning of a solution of the problem (3).

Definition 2.1. A function $u : [0,T] \to V$ is a *solution* of the initial value problem (3) if for any $t \in [0,T], \quad u(t) \in V_0$

$$\lim_{\Delta t \to 0} \left\| \frac{1}{\Delta t}\big(u(t+\Delta t) - u(t)\big) - Lu(t) \right\| = 0 \tag{4}$$

and $u(0) = u_0$.

Definition 2.2 The initial value problem (3) is *well-posed* if for $u_0 \in V_0$, there is a unique solution $u = u(t)$ and the solution depends continuously on the initial value : There exists a constant $c_0 > 0$ such that $u(t)$ and $\bar{u}(t)$ are the solutions for the initial values $u_0, \bar{u}_0 \in V_0$, then

$$\sup_{0 \le t \le T} \left\| u(t) - \bar{u}(t) \right\|_V \le c_0 \left\| u_0 - \bar{u}_0 \right\|_V . \tag{5}$$

Proposition 2.1 *Assume the problem (3) is well-posed and the solution is denoted as*

$$u(t) = S(t)u_0, \quad u_0 \in V_0 . \tag{6}$$

Then the solution operator $S(t)$ *is linear.*

Definition 2.3 Assume the operator $S(t) : V_0 \subseteq V \to V$ can be uniquely extended to a linear continuous operator $S(t) : V_0 \subseteq V \to V$ with

$$\sup_{0 \le t \le T} \left\| S(t) \right\|_V \le c_0 .$$

For $u_0 \in V - V_0$, we call $u(t) = S(t)u_0$ the *generalized solution* of the initial value problem (3).

Proposition 2.2 *For any* $u_0 \in V$, *the generalized solution* $u(t) = S(t)u_0$ *of the initial value problem (3) is continuous in t.*

Proposition 2.3 *Assume the problem (3) is well-posed. Then for all* $t_1, t_0 \in [0, T]$ *such that* $t_1 + t_0 \le T$, *we have* $S(t_1 + t_0) = S(t_1)S(t_0)$.

The existence of the well-posed solution is shown more details on theoretical analysis of the semigroups of linear operators by Goldstein [2].

Now we introduce a finite difference method defined by one-parameter family of uniformly bounded linear operators

$$C(\Delta t) : V \to V, \quad 0 < \Delta t \le \Delta t_0 . \tag{7}$$

Here $\Delta t_0 > 0$ is a fixed number. The family $\{C(\Delta t)\}_{0 < \Delta t < \Delta t_0}$ is said to be uniformly bounded if there is a constant c such that

$$\left\| C(\Delta t \right\| \le c \quad \forall \Delta t \in (0, \Delta t_0].$$

The approximate solution is then defined by

$$u_{\Delta t}(m\Delta t) = C(\Delta t)^m u_0, \quad m = 1,2,3,\cdots \tag{8}$$

Definition 2.4 *The difference method* is the approximate solution of the problem (3) that is defined by

$$u_{\Delta t}(m\Delta t) = C(\Delta t)^m u_0, \quad m = 1,2,3,\cdots \tag{8}$$

Definition 2.5 The difference method (8) is *consistency* if there exists a dense subspace V_c of V such that for all $u_0 \in V_c$, for the corresponding solution u of the initial value problem (3) , we have :

$$\lim_{\Delta t \to 0} \left\| \frac{1}{\Delta t} \left(C(\Delta t)u(t) - u(t + \Delta t) \right) \right\| = 0 \text{ isuniformly on } [0, T] \tag{9}$$

Proposition 2.4 *If difference method [8] is consistency, then* $\frac{C(\Delta t)-I}{\Delta t}$ *is a convergent approximation of the operator L.*

Definition 2.6 The difference method (8) is *convergent* if for any fixed $t \in [0,T]$, any $u_0 \in V$, we have

$$\lim_{\Delta t_i \to 0} \left\| C(\Delta t_i)^{m_i} - S(t)u_0 \right\| = 0 \tag{10}$$

where $\{m_i\}$ is sequence of integer and $\{\Delta t_i\}$ a sequence of step sizes such that $\lim_{i \to \infty} m_i \Delta t_i = t$.

Definition 2.7 The difference method (8) *is stable* if the operator

$$\left\{ C(\Delta t)^m \middle| 0 < \Delta t \le \Delta t_0, m\Delta t \le T \right\} \tag{11}$$

are uniformly bounded; i.e., there exists a constant $M_0 > 0$ such that

$$\left\| C(\Delta t)^m \right\|_{V \to V} \le M_0 \quad \forall m : m\Delta t \le T \quad \forall \Delta t \le \Delta t_0 . \tag{12}$$

We now give the main result of this section.

Theorem 2.1 (Lax Equivalence theorem) *Suppose the initial value problem (3) is well-posed. For a consistency difference method, stability is equivalent to convergence.*

3 Difference method Approximation for Two Channel Dissipation Model

Let us consider problem the two channel dissipation model that was derived as in introduction:

$$\begin{aligned} u_t + c_1 u_x &= \alpha (v-u) \\ v_t + c_2 v_x &= \alpha (u-v) \end{aligned} \tag{13}$$

In this paper we choose for case $c_1 = -c_2$ and then by scaling the equation (13) can be written as

$$\begin{aligned} \partial_t u + \partial_x u &= \beta(v-u) \\ \partial_t v - \partial_x v &= \beta(u-v) \end{aligned} \tag{14}$$

The partial differential equations are given the initial value:

$$u(x,0) = u_0(x) \quad v(x,0) = v_0(x) \quad 0 < x < 1, \tag{15}$$

and the boundary conditions:

$$u(0,t) = a \qquad v(1,t) = b \qquad t > 0 . \tag{16}$$

This boundary value problem (14)-(16) has the steady state solution:

$$u(x)=\frac{\beta(b-a)}{\beta+1}x+a,\qquad v(x)=\frac{\beta(b-a)}{\beta+1}x+a+\frac{b-a}{\beta+1}. \tag{17}$$

By using Cauchy-Kowalewski Theorem if u_0 and v_0 is analytic in some neighborhood of the point $x_0\in(0,1)$, then the boundary value problem has classical solution. This solution will be approximated by difference method.
The first stage the problem is written in the form of operator differential equation:

$$\begin{pmatrix}u_t\\ v_t\end{pmatrix}=\begin{pmatrix}-\partial_x-\beta & \beta\\ \beta & -\partial_x-\beta\end{pmatrix}\begin{pmatrix}u(t)\\ v(t)\end{pmatrix}$$

$$\begin{pmatrix}u(0)\\ v(0)\end{pmatrix}=\begin{pmatrix}u_0\\ v_0\end{pmatrix}$$

Let be operator $L=\begin{pmatrix}-\partial_x-\beta & \beta\\ \beta & -\partial_x-\beta\end{pmatrix}$ and $w(x,t)=\begin{pmatrix}u(x,t)\\ v(x,t)\end{pmatrix}$, so the problem :

$$\frac{dw(x,t)}{dt}=Lw(x,t),\qquad 0\le t\le T, \tag{18}$$
$$w(x,0)=w_0$$

It is clear that L is linear operator and

$$V=C_0[0,1]\times C_0[0,1]=\left\{\begin{pmatrix}u\\ v\end{pmatrix}\in C[0,1]\times C[0,1]\middle| u(0)=a, v(1)=b\right\}$$

with maximum norm V is Banach space . For difference method, we have difference scheme:

$$\frac{u_j^{n+1}-u_{j-1}^n}{\Delta t}=\beta(v_j^{n+1}-u_j^{n+1})$$
$$\frac{v_j^{n+1}-v_{j+1}^n}{\Delta t}=\beta(u_j^{n+1}-v_j^{n+1}) \tag{19}$$

This difference equation can be written as:

$$u_j^{n+1}=\frac{(1+\beta\Delta t)u_{j-1}^n+\beta\Delta t v_{j+1}^n}{1+2\beta\Delta t}\qquad j=0,1,2,3,\cdots,N$$

$$v_j^{n+1}=\frac{(1+\beta\Delta t)v_{j+1}^n+\beta\Delta t u_{j-1}^n}{1+2\beta\Delta t}\qquad n=0,1,2,\cdots \tag{20}$$

$$u_0^n=a\ \ v_N^n=b,\ \ u_j^0=u_0(x_j)\ \ v_j^0=v_0(x_j)$$

For the difference scheme, we define the operator $C(\Delta t)$ by the formula

$$C(\Delta t)\begin{pmatrix}u(x)\\ v(x)\end{pmatrix}=\begin{pmatrix}(\frac{1}{2}+r)u(x-\Delta x)+(\frac{1}{2}-r)v(x+\Delta x)\\ (\frac{1}{2}-r)u(x-\Delta x)+(\frac{1}{2}+r)v(x+\Delta x)\end{pmatrix} \quad [19]$$

where $\Delta x=\Delta t$ and $r=\dfrac{1}{2+4\beta\Delta t}<\dfrac{1}{2}$ for all $x\pm\Delta x\in[0,1]$.

Then $C(\Delta t):V\to V$ is a linear operator and it can be shown that

$$\left\|C(\Delta t)\begin{pmatrix}u(x)\\ v(x)\end{pmatrix}\right\|=\left\|\begin{matrix}(\frac{1}{2}+r)u(x)+(\frac{1}{2}-r)v(x)\\ (\frac{1}{2}-r)u(x)+(\frac{1}{2}+r)v(x)\end{matrix}\right\|$$

$$=\left\|\begin{pmatrix}\frac{1}{2}+r & \frac{1}{2}-r\\ \frac{1}{2}-r & \frac{1}{2}+r\end{pmatrix}\begin{pmatrix}u(x)\\ v(x)\end{pmatrix}\right\|$$

$$\le(\tfrac{1}{2}+r)\left\|\begin{pmatrix}u(x)\\ v(x)\end{pmatrix}\right\|$$

So

$$\|C(\Delta t)\|\le\tfrac{1}{2}+r<1 \quad [20]$$

and the family $\{C(\Delta t)\}$ is uniformly bounded. The difference method is

$$w_{\Delta t}(t_m)=C(\Delta t)w_{\Delta t}(t_{m-1})=C(\Delta t)^m w_0,\quad m=1,2,3,\cdots N_t$$

or

$$w_{\Delta t}(\cdot,t_m)=C(\Delta t)^m w_0(\cdot),\quad m=1,2,3,\cdots N_t.$$

The difference method generates an approximate solution that is defined for $x\in[0,1]$ and $t=t_m, m=0,1,2,\cdots,N_t$:

$$w_{\Delta t}(x_j,t_{m+1})=\begin{pmatrix}u(x_j,t_{m+1})\\ v(x_j,t_{m+1})\end{pmatrix}=\begin{pmatrix}(\frac{1}{2}+r)u(x_{j-1},t_m)+(\frac{1}{2}-r)v(x_{j+1},t_m)\\ (\frac{1}{2}-r)u(x_{j-1},t_m)+(\frac{1}{2}+r)v(x_{j+1},t_m)\end{pmatrix}$$

$$1\le j\le N_x-1,\ 0\le m\le N_t-1 \quad [21]$$

$$w(x_j,0)=\begin{pmatrix}u(x_j,0)\\ v(x_j,0)\end{pmatrix}=\begin{pmatrix}u_0(x_j)\\ v_0(x_j)\end{pmatrix}\quad 0\le j\le N_x$$

$$u(0,t_m)=a,\ v(1,t_m)=b\qquad 0\le m\le N_t$$

We see that the relation between the approximate solution $w_{\Delta t}$ and the solution $\begin{pmatrix}u\\ v\end{pmatrix}$ by the scheme difference [20] is

$$w_{\Delta t}(x_j,t_m)=\begin{pmatrix}u(x_j,t_m)\\ v(x_j,t_m)\end{pmatrix}=w_j^m=\begin{pmatrix}u_j^m\\ v_j^m\end{pmatrix}. \qquad [22]$$

As for the consistency, we take $V_c=V_0\subset C^\infty[0,1]\times C^\infty[0,1]$. By using Cauchy-Kowalewski theorem in reference [3] for the initial value functions in V_c , we have the solution which is infinitely smooth. Now using Taylor expansion at (x,t), we have

$$C(\Delta t)\begin{pmatrix}u(x,t)\\ v(x,t)\end{pmatrix}-\begin{pmatrix}u(x,t+\Delta t)\\ v(x,t+\Delta t)\end{pmatrix}$$

$$=\begin{pmatrix}(\frac{1}{2}+r)u(x-\Delta x,t)+(\frac{1}{2}-r)v(x+\Delta x,t)-u(x,t+\Delta t)\\ (\frac{1}{2}-r)u(x-\Delta x,t)+(\frac{1}{2}+r)v(x+\Delta x,t)-v(x,t+\Delta t)\end{pmatrix}$$

$$=\begin{pmatrix}(\frac{1}{2}+2r)u_{xx}(x,t)+2ru_{xt}(x,t)-\frac{1}{2}u_{tt}(x,t)+4\beta ru_t(x,t)\\ (\frac{1}{2}+2r)v_{xx}(x,t)-2rv_{xt}(x,t)-\frac{1}{2}v_{tt}(x,t)-4\beta rv_t(x,t)\end{pmatrix}(\Delta t^2)+o(\Delta t^3)$$

So

$$\left\|\frac{1}{\Delta t}\left(C(\Delta t)\begin{pmatrix}u(x,t)\\ v(x,t)\end{pmatrix}-\begin{pmatrix}u(x,t+\Delta t)\\ v(x,t+\Delta t)\end{pmatrix}\right)\right\|\le c\Delta t\,, \qquad (23)$$

where

$$c=\sup_{0<x<1,0\le t\le T}\left\{u(x,t),v(x,t),D^1u(x,t),D^1v(x,t),D^2u(x,t),D^2v(x,t)\right\}$$

Thus we have the consistency of the scheme.

According to (20); $\|C(\Delta t)\|<1$ and so $\|C(\Delta t)^m\|<1,\quad m=1,2,3,\cdots$. Then the family operators $\left\{C(\Delta t)^m\,\middle|\,0<\Delta t\le\Delta t_0, m\Delta t\le T\right\}$ are uniformly bounded. Then the scheme difference [20] is stable. Because of the stability and consistency of this method and using the Lax Equivalence theorem we prove that the method is convergent.

This method will be demonstrated that the numerical solutions tend to the exact steady state solution as the time limits infinity. We explain the phenomenon in figure with parameter $a=0.5, b=2.5, \beta=1$ and the initials value are:

$$w(x,0)=\begin{pmatrix}u(x,0)\\ v(x,0)\end{pmatrix}=\begin{pmatrix}0.5\\ 2x^2+0.5\end{pmatrix}. \qquad [24]$$

From in the figure we see that the solutions tend to the exact steady solution:

$$w(x)=\begin{pmatrix}u(x)\\ v(x)\end{pmatrix}=\begin{pmatrix}x+0.5\\ x+1.5\end{pmatrix}. \qquad [25]$$

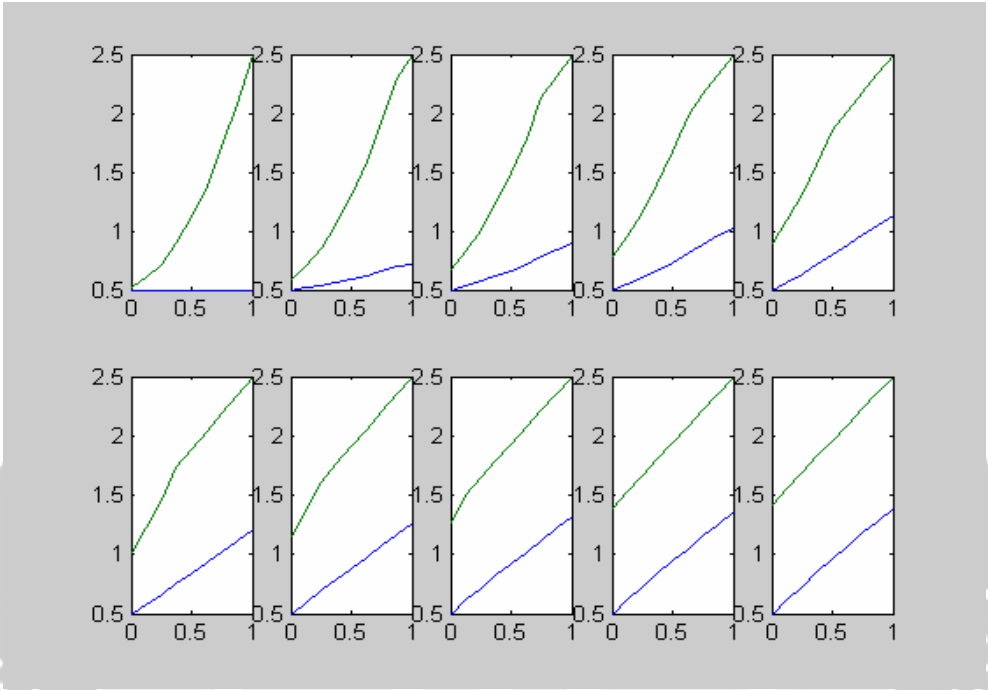


Figure 2: Graphic solution two channel dissipation model

Acknowledgment

For writing this paper, we would like to thank to dr. F.P.H. Frits van Beckum who gave me the problem when he visited at Department of Mathematics, Gadjah Mada University.

References

[1] Atkinson, K. E. & Weimin Han (2001), *Theoretical Numerical Analysis*: *A Functional Analysis Framework,* Springer, New York.

[2] Goldstein, J.A. (1985), *Semigroups of linear operator and applications*, Oxford University Press, New York

[3] Petrovsky, I.G. [1954], *Lecturer on Partial Differential Equation*, Interscience, New York.

[4] Van Beckum F.P.H. (2003), Travelling wave solution of a coastal zone non-Fourier dissipative model, *Proceedings of the Symposium on Coastal Zone Management*, Bandung, Editor: C.B. Vreugdenhil & E. Soewono, 89 – 101.

SUMARDI: Department of Mathematics , Gajah Mada University, Sekip Utara Yogyakarta 55281, Indonesia. Phone/Fax: +62 +27452243, 902126, 513339.
E-mail: mas_mardi@yahoo.com

SOEPARNA DARMAWIJAYA: Department of Mathematics , Gajah Mada University, Sekip Utara Yogyakarta 55281, Indonesia. Phone/Fax: +62 +27452243, 902126, 513339.
E-mail: swahyuni@indosat.net.id

LINA ARYATI: Department of Mathematics , Gajah Mada University, Sekip Utara Yogyakarta 55281, Indonesia. Phone/Fax: +62 +27452243, 902126, 513339.
E-mail: lina@math-ugm.web.id

Analytical Study of the Gas Lift Performance Curve and Optimum Gas Injection Rate in a Gas Lift Technique

Deni Saepudin[1,3], Edy Soewono[1], Kuntjoro Adji Sidarto[1], Agus Yodi Gunawan[1] , Septoratno Siregar[2]

[1)] Department of Mathematics, Institut Teknologi Bandung, Indonesia
[2)] Department of Petroleum Engineering, Institut Teknologi Bandung, Indonesia
[3)] Sekolah Tinggi Teknologi Telkom, Bandung, Indonesia

Abstract: The main objective in oil production system using Gas Lift technique is to obtain the optimum gas injection rate which yields the maximum production rate. Relationship between gas injection rate and production rate is described by gas lift performance curve (GLPC). In case where the GLPC is a unimodal curve and a large enough amount of injection gas available, the optimum gas injection rate is the peak of the GLPC. Therefore, in this case, existence of the optimum solution is guaranteed. Obtaining the optimum gas injection is something important because excessive gas injection will reduce production rate, and also it is expensive due to the high gas prices and compressing costs. In this paper, we discuss the characteristic of the GLPC for a production well, for which one phase flow (liquid) in the reservoir, and two phase flow (liquid and gas) in the tubing. By assuming that gradient pressure satisfies *Hagedorn & Brown* equation, it can be shown that the GLPC is a unimodal curve.

Keywords: Constrained Optimization, Gas Lift Performance Curve, Gas Injection Rate, Liquid Production Rate.

DISTANCE LEARNING WITH WEB-BASED: THE OPPORTUNITIES AND THE CHALLENGES

R. Poppy Yaniawati
Universitas Pasundan, Bandung, Indonesia

Abstract. Dirjen Dikti Depdiknas has strictly warned and instructed not to allow the higher education learning to have the distance classes, no matter it is aimed at servicing to the society that has lack of needed education. One of the alternatives to solve this problem is through the distance learning, this could give the society an even distribution of education. This will be very effective way especially for Indonesia which has many separated islands both for large ones and even smaller ones. E-learning is a kind of web-based distance learning. The advantage to implement of this program things are: 1) E-learning is of much benefit to society; 2) The growth of internet users both in quality and quantity; and 3) By using its facilities, internet can be a great media for e-learning. Besides the advantages, there are some hindrances things are: 1) limited infrastructures; 2) the difficulty of facilities to access; and 3) the lack of awareness of using internet in the society.

Key words: distance learning, web-based, e-learning, internet

1. Introduction

Since on January 2002, there has been a problem about implementation of distance classes by higher education learning, especially S1 and S2 programs. It was aimed at servicing to the society that has lack of needed education, so they can more economize of their cost accommodation for study. However, Dirjen Dikti Depdiknas has strictly warned and instructed to stop it. The reason is the implementation of it is a less scientific atmosphere, and it was not conducted in campus environment.

One of the alternatives to solve this problem is through the distance learning, this could give the society an even distribution of education. This will be very effective way especially for Indonesia which has many separated islands both for large ones and even smaller ones. The distance learning has given by the RI law section 31 number 20 in 2003 (in Kartasasmita, 2003).

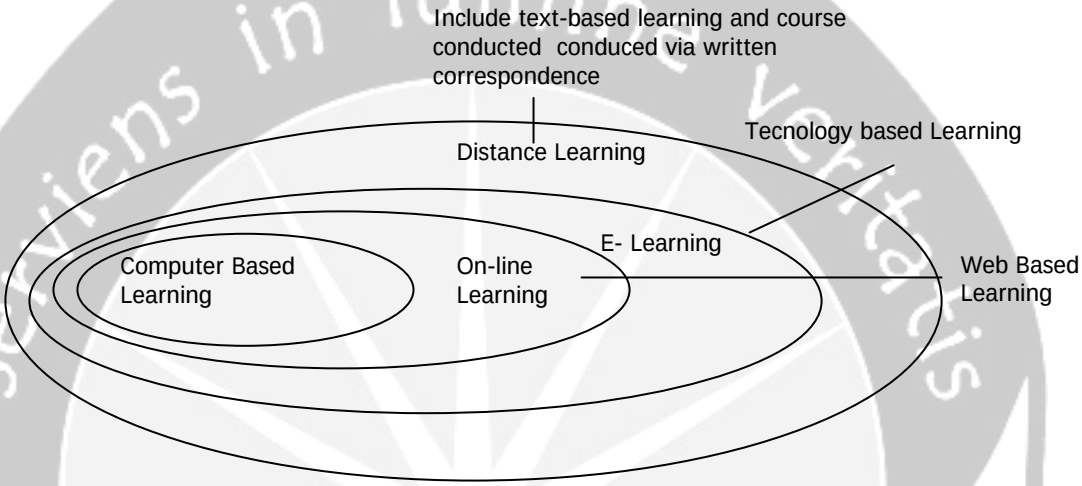
According to Keegan (1986); McLean (2002) (in Soekartawi, 2004), there are advantages of distance learning:

- Flexible; the students can study without limited times and places.
- Less to disturb daily activities students.
- Every one can study through distance learning.
- It makes the students be more autonomous in studying.

E-learning is a kind of web-based distance learning. According to Kartasasmita (2003), e-learning is a kind of distance learning, especially access to internet. Web-based learning system is often known with on-line learning or virtual

learning that part of the e-learning. Furthermore, Savel (Kartasasmita, 2004) explains that e-learning integrate electronic technology and education, by using internet more dominant. Besides according to Linde (2004), e-learning is a formal or informal learning with electronic media, such as internet, intranet, CDROOM, video tape, DVD, TV, mobile-phone, PDA, etc.

Based on the above opinions, the terminology of e-learning is wider than on-line learning as presented on figure-1.



Source: WR Hambrecht + Co, http://www.wrhambracht.com (Simamora, 2003)

Figure 1. Learning Terminology

2. Opportunities

Cisco (in Kamarga, 2002) explained the philosophy of e-learning are:

a. E-learning is the medium of information, communication, education, and training based on online.

b. E-learning provides many tools that could be an added value of conventional study, there fore can face challenges of globalization.

c. E-learning doesn't change conventional model learning in the classes.

d. Various capacity of audience depends on the content type.

It was shown that e-learning is the combination between information and communication in education, so teachers can give many lessons via Internet that could be accessed any time where ever they would. The audiences can develop their learning process by looking for reference and information from other sources.

The numbers of Internet users in the world increase very fast. Internet has a big potential for e-learning, among other: 1) internet could be accessed every moment; 2) the audiences or teachers can give their opinions freely; 3) public society can access, correction, and control subject matter. Besides, the internet can give the opportunities to develop perception from other website.

There are some facilities by Internet, among other: 1) electronic mail (e-mail); 2) mailing list (mills); 3) file transfer protocol (FTP); 4) newsgroup; 5) world wide web (www).

3. Challenges

Many influences and advantages of e-learning in learning culture cannot eliminate problems caused by this system. Criteria to implement learning based on e-learning at school level, especially in learning mathematics, is accessibility, affordability, and reliability technology (Supriadi, 2002).

According to Bullen (Soekartawi, 2003), using the internet for learning can make some problems, among other:

1) Shortcoming of interaction between teachers and students, or among students themselves.
2) Learning process tendency to train then to educate.
3) Change of teacher function from conventional learning technique to ICT learning technique.
4) The students that haven't high learning motivation will tendency fail.
5) Internet only found in special places.
6) A less capability-using computer by user.

Implementation of e-learning, instructor would be an urgent factor that is as motivator for students to learn. According to Purbo (1996), the instructor must be transparent to give information about all aspect activities in learning, there fore the students could learn better to reach a good outcome. The under line information consist of:

1) Time allocation to study of matter learning and do some tasks.
2) Skill technology needed for students to accelerate learning activities.
3) Facilities and equipments that are needed in learning activities.

Besides, the instructors must be active in discussion in e-learning, for example:

1) To response each information explained by students.
2) To prepare and expose matter from variety references.
3) To provide guidance and encourage students to do interaction.
4) To provide feed back to individual and continuous.

E-learning critiques said that besides e-learning activities reach area is limited, quantity of direct contact among students or between instructor and students are

very poor, also the opportunity of student to socialize is limited too (Wildavsky, 2001).

4. Conclusions

There are some opportunities and challenges to implement of distance learning with web-based. The advantage of this program things are: 1) E-learning is of much benefit to society; 2) The growth of internet users both in quality and quantity; and 3) By using its facilities, internet can be a great media for e-learning. Besides the advantages, there are some hindrances things are: 1) limited infrastructures; 2) the difficulty of facilities to access; and 3) the lack of awareness of using internet in the society.

Concerning with this, to implement e-learning should consider five factors as:

1) Student (learner); e-learning system is made suitable with student characteristic and behavior.
2) Subject matter; restructuring is made suitable with technology format, so it would be give value added than conventional classes.
3) Organization; police and commitment from top organization is needed to implement e-learning, so it can be socialize.
4) System process; it is business process implementation of e-learning that consist of administrator, teacher (instructor), technician, design of matter, and those are must be synergy and interactive.
5) Technology; as tools to support in reaching effectiveness of e-learning.

According to Kartasasmita (2003), implementation of e-learning should have in a small group (consist of 5-7 students) and tutor. Tutor give subject matter which is designed by him/herself, from internet, or another sources.

References

[1] Abidin, A. dan Nawi, R. (2002). *E-learning: Penerekoan Media Pembelajaran Terkini*. E-learning Unit. Universiti Malaysia Sarawak.

[2] Kamarga, H. (2002). *Belajar Sejarah melalui E-learning*. Jakarta: Intimedia.

[3] Kartasasmita, B. (2003). *Catatan Pengembangan e-Learning dalam Budaya Belajar Kini.* Makalah Seminar pada tanggal 8 Desember 2003 di ITB Bandung.

[4] Kartasasmita, B. (2004). *Berkenalan dengan e-Learning*. Makalah Seminar pada tanggal 10 Agustus 2004 di UPI Bandung.

[5] Linde, E. (2004). *Online Teaching and Learning*. Makalah Seminar pada tanggal 16 Pebruari 2004 di Unpad Bandung.

[6] Oetomo, B.S.D. (2002). *E-Education: Konsep, Teknologi dan Aplikasi Internet Pendidikan.* Yogyakarta: ANDI.

[7] Purbo, O. (1996). *Internet untuk Dunia Pendidikan*. Makalah. Institut Teknologi Bandung.

[8] Simamora, L. (2003). "E-learning: Konsep dan Perkembangan Teknologi yang Mendukungnya". *Cakrawala Pendidikan: E-learning dalam Pendidikan*. Jakarta: Universitas Terbuka.

[9] Soekartawi, (2003). *E-learning di Indonesia dan Prospeknya di Masa Mendatang*. Makalah Seminar Nasional pada tanggal 3 Pebruari 2003 di Universitas Kristen Petra Surabaya.

[10] Soekartawi. (2004). "Beberapa Kesulitan dalam Pelaksanaan Pembelajaran Berbasis Web pada Sistem Pendidikan Jarak Jauh". *Teknologi Pembelajaran*. Jakarta: PUSTEKOM.

[11] Supriadi, D. (2002). *Internet Masuk Sekolah: Pemberdayaan Guru dan Siswa dalam era Sekolah Berbasis E-learning*. Makalah pada Seminar Implementasi E-learning untuk Sekolah Menengah. PT Telkom Bandung.

[12] Wildavsky, B. (2001). *Want More from High School? Special Report: E-learning*. [online]. Tersedia: http:/www.usnews/edu/elearning/articles.

R. POPPY YANIAWATI: Ph. D student at Department of Mathematics, Universitas Pendidikan Indonesia, Jl. Dr. Setiabudhi 229 Banding 40154, Indonesia. Phone/Fax: +62 +22 2001197; Department of Mathematics, Universitas Pasundan, Jl. Taman Sari 6-8 Bandung, Indonesia. Phone/Fax: +62 +22 4205317

E-mail: opyaniawati@bdg.centrin.net.id

LEARNING MATHEMATICS EASILY BY INTERACTIVE INSTRUCTION SOFTWARE : A New Paradigm in Information and Communication Era

Rahayu Kariadinata

UPI Bandung -Indonesia

Abstract: A progress in global communication and information has an effect on various aspects of life, including a shift in life, work and learning patterns. Now, the information and communication technology is used very much in a learning process. A variety of computer applications in teaching learning are known as *computer-based instruction (CBI), computer-assited instruction/learning (CAI/CAL),* or *computer-managed instruction (CMI).* In the computer-based learning, the computer can be used as a very attractive learning means or medium when being supported by the availability of instruction software, and can help the teacher with their task as " substitute" for role of a teacher in instiling a concept that the teacher at the same time can become a trainer, councelor and instructor. In mathematics instruction, the software can employ Macromedia Flash. Its a modern program to create the application web. The macromedia falash enables the application to be equipped with components, such as animation, picture, text, sound, interactive animation, and so on. The software is also equipped with the video. The components are collectively refered to as multimedia application. Something special abuot the software being equipped with the componens is that it can prrovidespecific services. The film and video present situations supporting *authentic learning* to create the student"s learning motivation and provide a proper *situatedness learning.* The picture and animation can make the presentations more concrete and realistic. The animation can also show more picture combination displayed in sequels. The subject matters and saved data can be displayed again in a quick, exact and simple manner. In mathematics instruction, the software can help the student study certain topics requiring high accuracy. For example, grah and diagrams can be presented very easily. The animation enables the space to be move aroung (rotated) and split by its sides and the student to understand the space concept. Some researches suggest that the learning by animation produces a better result than by a static picture. The animation has some functions that can be used to direct the student's attention to an important aspect of an object

Keywords: computer-based instruction, software, macromedia flash

1. Introduction

It can't be denied that the advance in information and communication technology (ICT), today has been arising the significant changing in the field of life aspect, in the working patterns, learning patterns as well as the patterns of community life. Through in information and communication technology, the user of the relative service will gain several kinds of information and various resources and the place quickly and easily.

By using the advancement of tchnology, Centron (1988) said that learning process to master the science and technology quicker, efficient, and its process also more indiovidually in line with the need of the student moreover as a whole.

Futher, Azra (2004:2) said that the advancement of information and communication technology was enables to smooth the democratization process and equity in the field of learning. Today the teacher was not only one of the humans creative capital resources in the learning process. The information and communication technology will constanly more develop, so the it will make posibly the students to access by themselves the various kinds of sources of learning nearly limities

Today, computer is a media of information and communication technology that most used in the education. One of the aspects that differenciate computer with the other electronic tools among other the high capability interactive as the vechicle in disseminating a lot of kinds of information as well as the vechicle to gain the feedback for the students. Moreover the computer memory capaty make it possibly the students to click again the subbject matter.

2. A new Paradigm of Instruction in Information and Communication Era

Today the information and communication technology is very rapidly, specially the computer technology has been changed the aspect of human life a lot. With the computer aid, anyone can work more rapidly. So in the education, anyone doesn't depend fisically only in the school/campus. Learning media is an alternative of source information and learning resource for who else that wants it. The instruction system has been changed, that is the changing of paradigm from "teaching" into "learning". The instruction paradigm oriented at gaining the aim in the framework of preparing the students to become the human that can learn independently (independent learners)

Today, it has been introduced a variety of computer applications in teaching learning are known as *computer-based instruction (CBI), computer-assited instruction/learning (CAI/CAL),* or *computer-managed instruction (CMI).*In non formal education or "training" its applications is *computer-based training (CBT*).

3. The Interactive Instruction Software Make Learning Mathematics Easily

In the mathematics instruction the potentiality of computer technology as a media so a big. Through software in the kline with the computer can be the effective tool in helping to learn of mathematics (Fletcher, in Kusumah ,2003:18-3). The essential concepts can be explorated by the students under the guidance of the teacher, so that the students can understand the mathematical concepts deeply.

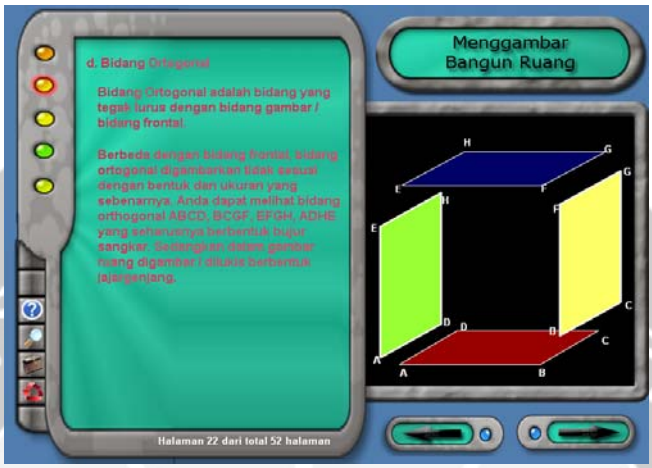
In making the interactive instruction software it can use Macromedia Flash programme, this programme consisting of basic competency, material presentation, the examples of exercises, drill completed with feedback and comprehensive drill.

The statde programme can be used individually by the students easily and help the brain to understand the concept that has been taught without boring.

Macromedia flash is sophisticated programme for making application web, the macromedia flash application with various component such as animation, picture, text, sound, and video.
The advantage of using software completed with the stated component, it can provide the special services. Film and video clip make the situation that can *"authentic learning*" that can arise the learning process of the student and preparing an enough *situatedness learning.*

Picture and animation can make the presentation more concrete and realistic. With animation it also can mixing and matching various picture that presented consecutively. The subject matters and saved data can be displayed again in a quick, exact and simple manner.

In mathematics instruction, the software can help the student study certain topics requiring high accuracy. For example, graph and diagrams can be presented very easily. The animation enables the space to be move aroung (rotated) and split by its sides and the student to understand the space concept.



Some researches suggest that the learning by animation produces a better result than by a static picture. The animation has some functions that can be used to direct the student's attention to an important aspect of an object.

To simplity the subbject matters spesialy that consisting of basic principle of each theory can be made using interactive instruction software, so that the complicated impression for mathematic will fade away from student through. Through interactive instruction software it can present the scope and coverage of konwledge broadly and completely.

The phase of communication in learning with interactive instruction software (*http://202.159.18.43./jsi/2toha.htm*) is : 1) computer present the subject matters, 2) the students learn the subject matters, 3) computer ask question, 4) the students give response, 5) computer examine the stated response, if its true, computer present another subject matters, but if the answer wrong, computer give the true answer and its explanation. At the second grade, it enriched with the more variative interaction, eg, the student click the qustion, and the computer answer, the student want the computer to move the object that can be seen in the screen or on the contrary, the computer wants the student to move the stated object. By doing, so character of the interactive learning, simulation, dialogistic, and pedagogic can be felt by the student.

The followings are some advantages of interactive instruction software according Dubin and clement (Munir,2001:10) is :

a. The existing of interactive learning, and materializing the stimulan relationship and answer, it can arise inspiration and growing the interest
b. The remedial occurrence, where the computer give facilities for repetition if iy is necessary, aside to strengthen the learning process and improve the memory
c. Feedback, the computer help the student to have the feedback of the lessons freely, and can push the student motivation.

4. Conclusion

In this global era, it is time we use technology as a vehicle in the various aspects of life, in education the student use of interactive instruction software is one of the effort to learning mathematics easily.

The using of mixing components such as picture, text, sound, interactive animation, colour, visualisation, video in interactive instruction software can maximize the role of senses in receiving information for memory system. Beside that the visualisation presentation, can be followed easily and the simulation that presented can help the mind in understanding the subbject matters that can be taught without boring.

REFERENCES

[1] Anggoro, Toha, M. *Long Distance Education and Its Application in Indonesia.* [On Line] Avail : *(http://202.159.18.43./jsi/2toha.htm)* (July,20,2004)

[2] Azra,A. (2004). *Instruction Paradigm in Global Era.* Paper Presented in National Seminar of Instruction Technology. Jakarta : Open University, Pustekkom, Depdiknas, and IPTPI

[3] Centron, M.J. (1988). An American Renaissance in the Year 2000. The Futurist July-August

[4] Kusumah, Y. (2003). *Design and Development Mathematical Subject Matters Model Interactive Based Technology.* Paper presented in National Seminar Science and Mathematis Education. Bandung : JICA and UPI

[5] Munir. (2001). *Application of Multimedia Technology in Teaching Learning.* The Journal of Education. Vol.3(21)

RAHAYU KARIADINATA: PhD student at Department of Mathematics Education, UPI BANDUNG, Jl. Setiabudhi Bandung, Indonesia; Department of Mathematics Education, IAIN Bandung, Jl.A.H.Nasution, Ujung Berung – Bandung, Indonesia
E-mail: rahayu61@yahoo.com

STATISTICAL MODELING BY USING NEURAL NETWORKS

Subanar

Universitas Gadjah Mada, Yogyakarta, Indonesia

Abstract. In recent years an impressive array of publications has appeared claiming considerable successes of neural networks in data analysis and engineering applications. In fact, the most commonly used neural networks architecture particularly in data analysis is the multi-layer perceptron (MLP), also known as feed-forward neural networks (FFNN). FFNN in term of statistical modeling can be seen as a wide class of flexible nonlinear regression models, discriminant analysis, and data reduction models. This paper explains what neural networks are, translates neural network jargon into statistical jargon, and shows the relationship between neural networks and statistical models such as nonlinear regression models, discriminant analysis, and data reduction models.

Key-words: neural networks, data analysis, statistical models

1 Introduction

During the last few years, modeling to explain nonlinear relationship between variables and some procedures to detect this nonlinear relationship have grown in a spectacular way and received a great deal of attention. An overview and further discussion on the subject can be found in [10]. This fact also happens in field of statistical modeling, particularly in time series modeling and econometrics. Due to computational advances and increased computational power, nonparametric models that do not make assumptions about the parametric form of the functional relationship between the variables to be modelled have become more easily applicable.

Neural Networks (NN) model is a prominent example of such a flexible functional form. The use of the NN model in applied work is generallly motivated by a mathematical result stating that under mild regularity conditions, a relatively simple NN model is capable of approximating any Borel-measureable function to any given degree of accuracy (see e.g. [6, 9, 12, 13, 33]).

Today's research is largely motivated by the possibility of using NN model as an instrument to solve a wide variety of application problems such as pattern recognition, signal processing, process control, and time series forecasting. Sarle [27] stated that NN are used in three main ways:

- as models of biological nervous systems and "intelegence"
- as real-time adaptive signal processors or controllers implemented in hardware for applications such as robots
- as data analytic methods.

This paper is concerned with NN for data analysis.

Wong, Lai, and Lam [34] surveyed the use of NN model in business application on 1994–1998 periods. This survey identified 302 research articles of NN model are that distributed on several field application, those are on accounting or auditing, finance, human resources, information system, marketing or distribution, and production or operation research. NN models are also growth and applied on medical field. Some examples of the NN application in this area are for myocardinal infarction diagnoses (see e.g. [2, 19]), classification of EEG signals [21], and PET scan [15]. Additionally, Somoza and Somoza [28] also have applied NN model in psychiatry field.

2 Feedforward Neural Networks

Multilayer perceptron (MLP), also known as feedforward neural networks (FFNN), is probably the most commonly used NN architecture in engineering application. Typically, applications of NN for regression, time series modelling and classification (discriminant analysis) are based on the FFNN architecture [27]. Some references that contain general concept and form of FFNN model can be found in [3, 8, 26].

FFNN can be seen as a flexible class of nonlinear functions. They receive a vector of inputs x and compute a response or output $\hat{y}(x)$ by propagating x through the interconnected processing elements. The processing elements are arranged in layers and the data, x, flows from each layer to the successive one. Within each layer, the inputs to the layer are nonlinearly transformed by the processing elements and propagated to the next layer. Finally, at the output-layer $\hat{y}(x)$, which can be scalar or vector valued, is computed.

In a typical FFNN with one hidden-layer, such as illustrated in Figure 1, the response value $\hat{y}(x)$ is computed as

$$\hat{y}_{(k)} = f_{\bullet}^{o}\left[\sum_{j=1}^{q}[w_{\bullet j}^{o} f_j^{h}(\sum_{i=1}^{p} w_{ji}^{h} x_{i(k)} + b_j^{h}) + b_{\bullet}^{o}]\right], \tag{1}$$

where

$x_{i(k)}$	=	input variables, $(i = 1,2,\ldots, p)$
$\hat{y}_{(k)}$	=	response (output variable) value
k	=	number of input-target pairs $(x_{i(k)}, y_{(k)})$, $(k = 1,2,\ldots,K)$
w_{ji}^{h}	=	weights from the ith input to the jth neuron of the hidden layer
b_j^{h}	=	bias at the jth neuron of the hidden layer, $(j = 1,2,\ldots,q)$
f_i^{h}	=	activation function at the jth neuron of the hidden layer
$w_{\bullet j}^{o}$	=	weights from the jth neuron of the hidden layer to neuron at output layer
$b_{\bullet}^{o}$	=	bias at neuron of the output layer
$f_{\bullet}^{o}$	=	activation function at the neuron of the output layer.

The nonlinearity enters into the function $\hat{y}_{(k)}$ through the so called activation function f_i^{h} at hidden layer and $f_{\bullet}^{o}$ at output layer, usually a “smooth” threshold function. Some common activation functions are:

- linear or identity: $f(x) = x$
- logistic: $f(x) = (1+e^{-x})^{-1}$
- hyperbolic tangent: $f(x) = \tanh(x)$
- Gaussian: $f(x) = e^{-x^2/2}$

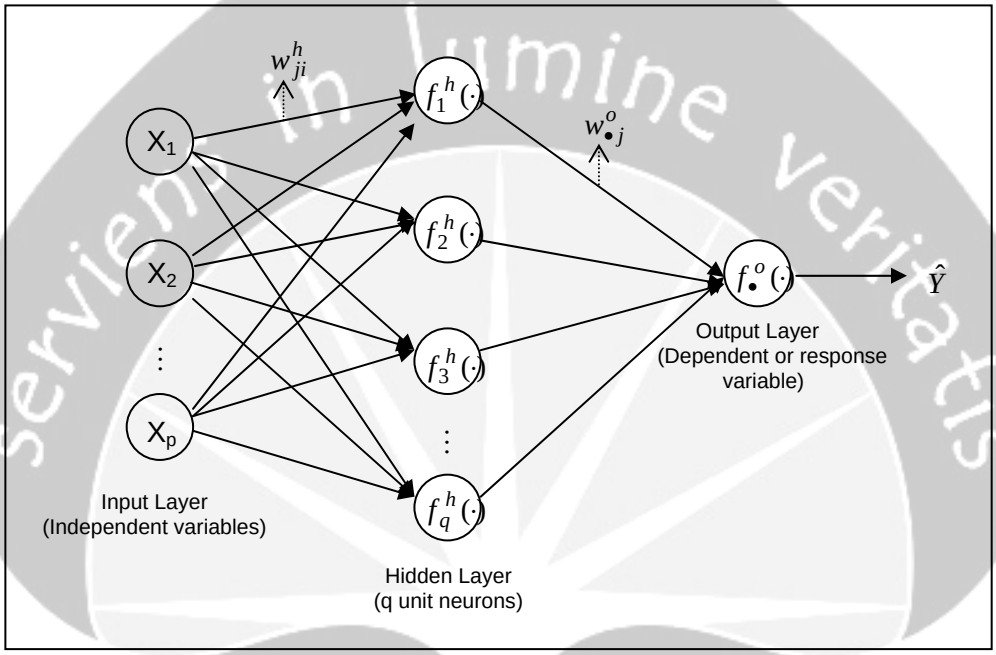


Figure 1. Architecture of FFNN with single hidden layer, p input units, q unit neurons at hidden layer, and one unit neuron at output layer.

2.1 Neural Networks and Statistical Jargon

The terminology in the NN literature is quite different from that in statistics, eventhough many NN models are similar or identical to well-known statistical models. For example, in the NN literature:

- variables are called *features*
- independent variables are called *inputs*
- predicted values are called *outputs*
- dependent variables are called *targets* or *training values*
- residuals are called *errors*
- estimation is called *training, learning, adaptation,* or *self-organization*
- observations are called *patterns* or *training pairs*
- parameter estimates are called *weigths*

- regression and discriminant analysis are called *supervised learning*
- data reduction is called *unsupervised learning*
- cluster analysis is called *adaptive vector quantization*
- interpolation and extrapolation are called *generalization*

The terms *sample* and *population* in statistics do not seem to have NN equivalents. However, the data in NN are often divided into a *training set* and *testing set* for cross-validation.

2.2 Relationship between FFNN and Statistical Models

In general, the relationship between NN and statistical models can be found in [4, 5, 16, 25, 26, 27]. In this section, we will give a brief review of NN for data analysis, particularly the relationship between FFNN to statistical modeling.

FFNN includes estimated weigths between the inputs and the hidden layer, and the hidden layer uses nonlinear activation functions such as the logistic function, the FFNN becomes genuinely nonlinear model, i.e., nonlinear in the parameters. In this case, FFNN can be seen as nonlinear regression. FFNN can have multiple inputs and outputs (Figure 1 is multiple inputs with single output), and this architecture is similar to multivariate multiple nonlinear regression.

FFNN with nonmetric data (dichotomus or polycothomus) in target values is identical to logistic regression and nonlinear discriminant analysis. In this case, FFNN often use a multiple logistic function to estimate the conditional probabilities ef each class. A multiple logistic function is called a *softmax* activation function in the NN literature.

The NN literature distinguishes between supervised and unsupervised learning. In the supervised learning, the goal is to predict one or more target variables from one or more input variables. Supervision consists of the use of target values in training. Supervised learning is usually some form of regression and discriminant analysis. The goal in most forms of unsupervised learning is to construct feature variables from which the observed variables, which are both input and target variables, can be predicted. Unsupervised Hebbian learning constructs quantitative features. In most cases, the dependent variables are predicted by linear regression from the feature variables. Hence, as is well-known from statistical theory, the optimal feature variables are the principal components of dependent variables.

3 Model Selection in Neural Networks

In the application of FFNN, it contains limited number of parameters (weights). How to find the best FFNN model, that is, how to find an accurate combination between number of input variables and unit nodes in hidden layer (imply the optimal number of parameters), is a central topic on the some NN literatures that discussed on many articles and books (see e.g. [3, 11, 23, 26]).

In general, there are two procedures usually used to find the best FFNN model (the optimal architecture), those are “general-to-specific” or “top-down” and “specific-to-

general" or "bottom-up" procedures. "Top-down" procedure is started from complex model and then applies an algorithm to reduce number of parameters (number of input variables and unit nodes in hidden layer) by using some stopping criteria, whereas "bottom-up" procedure works from a simple model. The first procedure in some literatures is also known as "pruning" (see e.g. [22, 23]), or "backward" method in statistical modeling. The second procedure is also known as "contructive learning" and one of the most popular is "cascade correlation" (see e.g. [7, 17, 20]), and it can be seen as "forward" method in statistical modeling.

Kaashoek and Van Dijk [14] introduced a "pruning" procedure by implementing three kinds of methods to find the best FFNN model; those are incremental contribution ($R^2_{incremental}$), principal component analysis, and graphical analysis. Whereas, Swanson and White [29, 30, 31] applied a criteria of model selection, SBIC or Schwarz Bayesian Information Criteria, on "bottom-up" procedure to increase number of unit nodes in hidden layer and input variables until finding the best FFNN model.

In recent development, procedure of inference statistics was also applied to determine the best FFNN model. In this case, the concept of testing hypotesa, parameter distribution and the use of some criteria for model selection are applied to find the optimal FFNN model. Terasvirta and Lin [32] were among the first researchers who applied this procedure to find the optimal number of unit nodes in hidden layer on FFNN model with single hidden layer. Some latest articles about FFNN model building by using inference statistics can be seen in [1, 18].

4 Conclusion

Statistical models and NN are not competing methodologies for data analysis. There is considerable many similarities between the two models. NN include several models, such as FFNN, that are useful for statistical applications. Statistical methodology is directly applicable to NN in a variety of ways, including estimation criteria, optimization algorithm, testing hyphotesis, and diagnostic check.

References

[1] Anders, U., and O. Korn (1999), Model selection in neural network, *Neural Networks*, **12**, 309–323.

[2] Baxt, W.G. (1991), Use of an artificial neural network for the diagnosis of myocardial infarction, *Annals of Internal Medicine*, **115**, 843–848.

[3] Bishop, C. M. (1995), *Neural Network for Pattern Recognition*, Oxford: Clarendon Press.

[4] Cheng, B. and D. M. Titterington (1994), Neural Networks: A Review from a Statistical Perspective, *Statistical Science*, **9**, 2–54.

[5] Cherkassky, V., J.H. Friedman and H. Wechsler eds. (1994), *From Statistics to Neural Networks: Theory and Pattern Recognition Applications*, Berlin: Springer-Verlag.

[6] Cybenko, G. (1989), Approximation by superpositions of a sigmoidal function, *Mathematics of Control, Signals and Systems,* **2**, 304–314.

[7] Fahlman, S. E. and C. Lebiere (1990), The Cascade-Correlation Learning Architecture, in Touretzky, D. S. (ed.), *Advances in Neural Information Processing Systems 2*, Los Altos, CA: Morgan Kaufmann Publishers, pp. 524-532.
[8] Fine, T. L. (1999), *Feedforward Neural Network Methodology*, Springer, New York.
[9] Funahashi, K. (1989), On the approximate realization of continuous mappings by neural networks, *Neural Networks*, **2**, 183–192.
[10] Granger, C. W. J. and T. Terasvirta (1993), *Modeling Nonlinear Economic Relationships*, Oxford: Oxford University Press.
[11] Haykin, H. (1999), *Neural Networks: A Comprehensive Foundation*, second edition, Prentice-Hall, Oxford.
[12] Hornik, K., M. Stinchombe and H. White (1989), Multilayer feedforward networks are universal approximators, *Neural Networks*, **2**, 359–366.
[13] Hornik, K., M. Stinchombe and H. White (1990), Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks, *Neural Networks*, **3**, 551–560.
[14] Kaashoek, J. F. and H.K. Van Dijk (2001), *Neural Networks as Econometric Tool*, Report EI 2001–05, Econometric Institute Erasmus University Rotterdam.
[15] Kippenhan, J.S., W.W. Barker, S. Pascal, J. Nagel and R. Duara (1992), Evaluation of a neural network classifier for PET scans of normal and Alzheimer disease subjects, *Journal of Nuclear Medicine*, **33**, 1459–1467.
[16] Kuan, C.M. and H. White (1994), Artificial Neural Networks: An econometric perspective, *Econometric Reviews*, **13**, 1–91.
[17] Littmann, E. and H. Ritter (1996), Learning and generalization in cascade network architectures, *Neural Computation*, **8**, 1521–1539.
[18] Medeiros, M. C., T. Terasvirta and G. Rech (2002), *Building Neural Network for Time Series: A Statistical Approach*, SSE/EFI Working Paper Series in Economics and Finance No 508.
[19] Pazos, A., V. Maojo, F. Martin and N. Ezquerra (1992), A neural network approach to assess myocardial infarction, In: Lun *et al.* (eds.), *Medinfo*: **92**, 659–663: Amsterdam, Elsevier.
[20] Prechelt, L. (1997), Investigation of the CasCor Family of Learning Algorithms, *Neural Networks*, **10**, 885-896.
[21] Reddy, D.C. and D.R. Korrai (1992), Neural Networks for classification of EEG signals, In: Lun *et al.* (eds.), *Medinfo*: **92**, 653–658: Amsterdam, Elsevier.
[22] Reed, R. (1993), Pruning algorithms - A survey, *IEEE Transactions on Neural Networks*, **4**, 740-747.
[23] Reed, R. D. and R.J. Marks II (1999), *Neural Smithing*, MIT Press, Cambridge, MA.
[24] Ripley, B.D. (1993), *Statistical Aspects of Neural Networks*, in O.E. Barndorff-Nielsen, J.L. Jensen and W.S. Kendall, eds., Networks and Chaos: Statistical and Probabilistic Aspects, Chapman & Hall.
[25] Ripley, B.D. (1994), Neural Networks and Related Methods for Classification, *Journal of the Royal Statistical Society*, Series B, **56**, 409-456.
[26] Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge.
[27] Sarle, W. (1994), Neural network and Statistical Models. *In Proceeding 19th A SAS Users Group Int. Conf.*, pp. 1538–1550. Cary: SAS Institute.

[28] Somoza, E. and J.R. Somoza (1993), A neural network approach to predicting admission in a psychiatric emergency room, *Medical Decision Making*, **13**, 273–280.
[29] Swanson, N. R. and H. White (1995), A model-selection approach to assessing the information in the term structure using linear models and artificial neural networks, *Journal of Business and Economic Statistics*, **13**, 265–275.
[30] Swanson, N. R. and H. White (1997a), Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models, *International Journal of Forecasting*, **13**, 439–461.
[31] Swanson, N. R. and H. White (1997b), A model-selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks, *Review of Economic and Statistics*, **79**, 540–550.
[32] Terasvirta, T. and C.F. Lin (1993), *Determining the number of hidden units in single hidden-layer neural network model,* Research Report 1993/7, Bank of Norway.
[33] White, H. (1990), Connectionist nonparametric regression: Multilayer feed forward networks can learn arbitrary mapping, *Neural Networks*, **3**, 535–550.
[34] Wong, B. K., V.S. Lai and J. Lam (2000), A bibliography of neural network business applications research: 1994–1998, *Computers and Operations Research,* **27**, 1045–1076.

SUBANAR: Department of Mathematics, Universitas Gadjah Mada, Bulaksumur, Yogyakarta 55281, Indonesia.
E-mail: subanar@yahoo.com

RADIAL BASIS FUNCTION AS STATISTICAL MODELING FOR FINANCIAL DATA

Brodjol S.[a], Subanar[b] and S. Guritno[b]

[a] Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[b] Universitas Gadjah Mada, Yogyakarta, Indonesia

Abstract: The radial basis function (RBF) neural network is approximation function methods. The network depends on number of centers and learning is used for training. Regulation parameter is used to get smooth function and to solve a problem inverse that near singular matrix. Financial data especially Indonesia Inflation data will be used as application data. The approximation function will be measured by some criteria statistics and will be compare ARIMA model. Three RBF network will be try on the data and the results are RBF network better than ARIMA model but nothing best model in model network.

Key-words: radial basis function, neural network, regulation, criteria statistics

1. Introduction

Radial basis functions (RBF) were first introduced by Powell to solve the real multivariate interpolation problem. This problem is one of the principal fields of research in numerical analysis. In the field of neural networks, RBF were first used by Broomhead and Lowe [2]. Other major contributions to the theory, design, and applications of RBF can be found in papers by Moody and Darken [6], and Poggio and Girosi [7]. The paper by Poggio and Girosi explains that the use of *regularization theory* applied to this class of neural networks. In the last years, Lazaro et.al [5] has applied EM algorithm to training RBF model, and Rivas et.al [10] introduce Evolving RBF to forecast time series data. Sutijo and Subanar [11] compared linear model and nonlinear model using Neural network. Sutijo, subanar and Guritno [12] has applied Regulation theory on simulation data.

The RBF network is a neural network approached by viewing the design as a curve-fitting (approximation) problem in a high dimensional space. Learning is equivalent to finding a multidimensional function that provides a best fit to the training data, with the criterion for "best fit" being measured in some statistical sense. This methods is motivation behind the RBF network that researcher work on strict interpolations in a multidimensional space. In a neural network the hidden units is a set of "functions" that compose a "basis" for the input patterns (vectors). These functions are called *radial basis functions*.

The basic design of a RBF network consists of three separate layers. The input layer is the set of source nodes. The second layer is a hidden layer of high dimension. The output layer gives the response of the network to the activation

patterns applied to the input layer. The RBF network consists some of parameters (weights) to be estimate. To find best model of RBF network necessary best combination for number of input variables, number of hidden units; centre and width of each hidden units. This is a central topic in some literature RBF network such as Bishop [1], Ripley [9], Fine [3], Haykin [4], Reed and Marks II [8], and Lazaro et al. [5]. This paper will be discuss reconstruc-tion problem and learning RBF for forecasting base on economics data especially Indonesia inflation data. The result of RBF will be compare with time series modeling.

2. Interpolation Problem

The network can be designed to perform a nonlinear mapping from the input space to the hidden space, and a linear mapping from the hidden space to the output space. The network represents a map from *p*-dimensional input space to the single di-mensional output space, expressed as

$$s : \Re^p \rightarrow \Re^1 \qquad (1)$$

The interpolation problem, in its *strict* sense can be stated as follows:

Given a set of N different points $\{x_i \in \Re^N \,|\, i = 1, 2, \ldots, N\}$ and a corresponding set of N real numbers $\{x_i \in \Re^1 \,|\, i = 1, 2, \ldots, N\}$, find a function $F : \Re^N \rightarrow \Re^1$ that satisfies the interpolation condition:

$$F(x_i) = d_i \;;\; i = 1,2,\ldots., N \qquad (2)$$

The interpolating surface has to pass through *all* the training data points. The RBF technique consists of choosing a function that has the following form :

$$F(x) = \sum_{i=1}^{N} w_i \phi(\|x - x_i\|) \qquad (3)$$

where $\{\phi(\|x - x_i\|) \,|\, i = 1, 2, 3, \ldots, N\}$ is a set of N random (usually nonlinear) functions, known as *radial basis functions*, and $\|\ \|$ represents a *norm* that is generally Euclidean. The known data points $\{x_i \in \Re^p \,|\, i = 1, 2, \ldots, N\}$ are the *centers* of radial basis functions. An architecture of RBF neural network is :

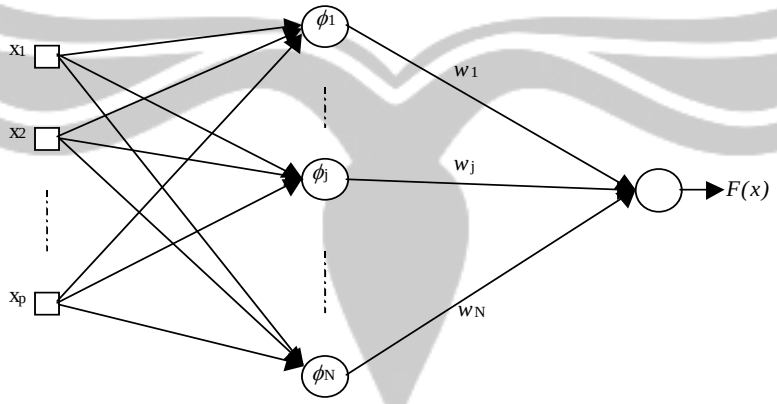


Figure 1. Archtecture the Radial Basis Function Neural Network

If the interpolation conditions equation (2) is inserted in (3), the following set of simultaneous linear equations can be obtained for the unknown coefficients (weights) of the expansion $\{wi\}$:

$$\begin{bmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1N} \\ \phi_{21} & \phi_{22} & \cdots & \phi_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{N1} & \phi_{N2} & \cdots & \phi_{NN} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{bmatrix} \tag{4}$$

where $\phi_{ij} = \phi(\|x_i - x_j\|)$ $i,j = 1, 2, 3, \ldots, N$

let $d = [d_1 \; d_2 \; \cdots \; d_N]'$, $w = [w_1 \; w_2 \; \cdots \; w_N]$ and $\phi = \phi_{ij}$

The matrix ϕ is called interpolation matrix, and equation (4) can be written in the compact form :

$$\phi \; w = d \tag{5}$$

If the data points are all distinct, the interpolation matrix is positive definite, and weight vektor w can be formed as follows :

$$w = \phi^{-1} d \tag{6}$$

In practice equation (5) cannot be solved when matrix ϕ is close to singular. Regularization theory can solve this problem by perturbating the matrix ϕ to $\phi + \lambda \mathbf{I}$.

3. Regularization Theory

The fundamental idea of regularization is to *stabilize* the solution in terms of some auxiliary nonnegative functional that embeds prior information, e.g., smoothness constraints on the input-output mapping. The approximation function is denoted by $F(\mathbf{x})$. The function F is obtained by minimizing a *cost functional* $\xi(F)$ that maps functions to the real line. Haykin expresses the cost func-tional using two terms of regularization as follows:

$$\xi(F) = \xi_s(F) + \lambda \xi_c(F) \tag{7}$$

where $\xi_s(F)$ is *standard error term* that measures the standard error (distance) between the desired response *di* and the actual response *yi* for training samples i = 1, 2, .., *N*. The term $\xi_c(F)$ is *regularization term* that depends on the geome-tric properties of the approximation function $F(\mathbf{x})$. The symbol λ is a positive real number called *regularization parameter*. The main aim of the regularization is to minimize the cost functional $\xi(F)$. The cost functional can be written interms of the desired response *di*, the actual response *yi*, and the regularization parameter λ as follows:

$$\xi(F) = \frac{1}{2} \sum (d_i - F(x_i))^2 + \frac{1}{2} \lambda \|PF\|^2 \tag{8}$$

where **P** is a linear (pseudo) differantial operator that contains the prior information about the form of the solution. The *Frechet differential* can be employed to do the minimization function. The *Frechet differential* has the following form:

$$d\xi(F,h) = \left[\frac{d}{d\beta}\xi(F+\beta h)\right]_{\beta=0} \quad (9)$$

where h(x) is a fixed function of the vector **x,** and β is a muti index. A necessary condition for the function F(x) to be a relative extremum of the functional $\xi(F)$ is that the Frechet differential $d\xi(F,h)$ be zero at F(x) for all $h \in H$, as expressed by

$$d\xi(F,h) = d\xi_s(F,h) + \lambda d\xi_c(F,h) = 0 \quad (10)$$

$$d\xi_s(F,h) = \left[\frac{d}{d\beta}\xi_s(F+\beta h)\right]_{\beta=} = -\sum_{i=1}^{N}[d_i - F(x_i)]h(x_i) \quad (11)$$

$$d\xi_c(F,h) = \left[\frac{d}{d\beta}\xi_c(F+\beta h)\right]_{\beta=0} = (Ph, PF)_H \quad (12)$$

Haykin states that the Frechet differential $d\xi(F,h)$ is zero for every h(**x**) in H space if and only if the following condition is satisfied:

$$P^*PF(x) = \frac{1}{\lambda}\sum_{i=1}^{N}(d_i - F(x_i))\delta(x - x_i) \quad (13)$$

where P* is the adjoin of differential operator P and $\delta(x - x_i)$ is a delta function located at x_i.

Let G(**x**;**x***i*) be a Green's function centered at **x**i. A Green's function G(**x**;**x***i*) is any function that satisfies the partial differential equation ***P****P**G*(**x**;**x***i*) = 0, Haykin gives the solution F(**x**) for the differential equation (13) with the following equation:

$$F(x) = \frac{1}{\lambda}\sum_{i=1}^{N}(d_i - F(x_i))G(x - x_i) \quad (14)$$

Haykin declares that the solution of the regularization problem lies in an N-dimensional subspace of the space of smooth functions, and the set of Green's functions G(**x**;**x***i*) centered at **x***i*, i =1, 2,..., N.

Since we have the solution F(**x**) to the regularization problem, our next step is to determine the unknown coefficients in the equation (14). Let us denote:

$$F(x_j) = \sum_{i=1}^{N} w_i G(x_j, x_i) \quad (15)$$

The definitions needed to be introduced can be given as follows;

$$F = [F(x_1)\ \ F(x_2)\ \ldots\ F(x_N)]' \quad , \quad d = [d_1\ d_2\ \ldots 'd_N]'$$

$$G = \begin{bmatrix} G(x_1;x_1) & G(x_1;x_2) & \cdots & G(x_1;x_N) \\ G(x_2;x_1) & G(x_2;x_2) & \cdots & G(x_2;x_N) \\ \vdots & \vdots & \ddots & \vdots \\ G(x_N;x_1) & G(x_N;x_2) & \cdots & G(x_N;x_N) \end{bmatrix}$$

$$w = [w_1\ \ w_2\ \ \ldots\ \ w_N]',$$

$$w = \frac{1}{\lambda}(d - F) \quad (16)$$

Now, the equation (15) can be rewritten in matrix form as follows,

$$F = Gw \quad (17)$$

When F is eliminated between equations (16) and (17), the following equation can be obtained:

$$(G + \lambda I)w = d \tag{20}$$

the associated Green's function $G(x;xi)$ is a symmetric function. Poggio and Girosi give a unique solution of the linear system of equations (20) as follows:

$$w = (G - I\lambda)^{-1} d \tag{21}$$

and Haykin concludes that the solution the regularization problem is expressed by (15) and he also states that if the Green's functions in equation (15) are radial basis functions, the solution can be rewritten as follows:

$$F(x) = \sum_{i=1}^{N} w_i G\left(\left\|x - x_i\right\|\right) \tag{22}$$

where $$G(x : x_i) = \exp\left(-\frac{1}{2\sigma_i^2}\left\|x - x_1\right\|^2\right) \tag{23}$$

where $G(x; x_i)$ is a *multivariate Gaussian function* characterized by a *mean vector* x_i and common *variance* σ_i^2.

4. Result Study

The data are used in this study is time series data (monthly) from January 1999 until April 2005. The data are Indonesia economic data, especially inflation data. In this study data divided into two part. First, data training is used to modeling, from Januari 1999 until December 2004. Second, data testing is used to model validation result first part. To determined best model in first part used statistics criteria *Mean Square Error* (MSE). The model is applied on second data to measure goodness of fit. Some goodness of fit model are *Mean Percentage Error* (MPE), *Mean Absolute Deviation* (MAD), and *Mean Absolute Percentage Error* (MAPE).

Generally time series analysis is used to get forecast model and value of the forecast. Stationary data and information about observation that influence to respond are needed in time series analysis (ARIMA models). Model identification is based on Time series plot, plot ACF and PACF. Plots data show that Indonesia Inflation data has nonseasonal and seasonal pattern and the observation (t-1), (t-11) and (t-12) have straight influence to respond.

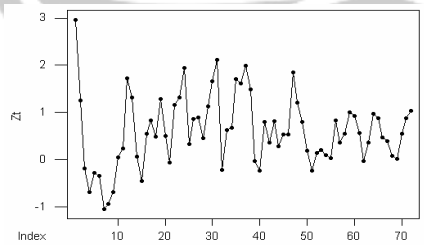


figure 2. Time series Plot data

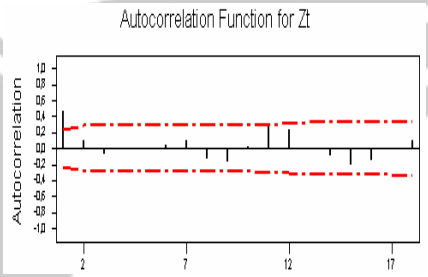


figure 3. Plot ACF data

Indonesia Inflation data has some ARIMA models , the best two of the models are ARIMA$(0,0,1)(0,0,1)^{11}$ and ARIMA$(1,0,1)(1,0,1)^{11}$. These models have MSE 0.2685 and 0.2743.

According the plots we know that respond (t observation) depend on observation (t-1), (t-11) and (t-12), so these observation are input of radial basis function neural network. In this study are used three architecture RBFNN models :

1. Number of unit in hidden layer equal number of points (model 1).
2. Number of unit in hidden layer is 12 (model 2).
3. Number of unit in hidden layer is 12 with regulation parameter (model 3).

Fist model based on standart architecture RBFNN model, that used number of unit in hidden layer equal number of point. Second and third model used 12 unit in hidden layer, it is based on number of month in a year. In third model there is additional regulation parameter. The regulation parameter is used to show smoothing effect in model validation. The result of RBFNN based on training and testing data for three models above is shown on table 1.

Table 1. Statistics of training and testing data

Architecture NN	MSE	MAD	MAPE	MPE
Model 1	0.000	0.000	0.000	0.000
	1.654	0.887	3.645	-1.696
Model 2	0.253	0.396	2.548	-1.058
	0.127	0.792	1.378	-1.198
Model 3	0.281	0.428	2.684	-1.013
	1.243	0.848	1.549	-1.233

☐ : training model ☐ : testing model

Table 1, show that the first model is good model in training data but in testing data is bad. Second and third models show that the second model is better than third model. This condition mean the adding regulation parameter do not give better result on Indonesia inflation data. According MSE, second model smaller than ARIMA model.

5. Conclusion

According the theory unit in hidden layer is equal number of point data, but we are possible to reduce the unit. The adding regulation parameter can be smooth the function but it is not always give better the result. In case Indonesia inflation data show that there is seasonal pattern. According ARIMA model, the best model is ARIMA(0,0,1)(0,0,1)11. Function approximation base on RBFNN for the data, there is not best model, because first model is good in training data but second model is good in testing data.

References

[1] Bishop, C. M., (1995). *Neural Network for Pattern Recognition.* Oxford: Clarendon Press.

[2] Broomhead, D.S. and Lowe, D. (1988). Multivariable functional interpolation and adaptive network. *Complex Systems*, **2**, 321–355.

[3] Fine, T. L. (1999). Feedforward *Neural Network Methodology*. Springer, NY.
[4] Haykin, H. (1999). *Neural Networks: A Comprehensive Foundation*, second edition, Prentice-Hall, Oxford.
[5] Lazarro, M., Santamaria, I., and Pantaleon, C. (2003). A new EM-based training algorithm for RBF networks. *Neural Networks*, **16**, 69–77.
[6] Moody, J. and Darken, C. (1989). Fast learning in networks of locally tuned processing units. *Neural Computation*, **1** (2), 281–294.
[7] Poggio, T. and Girosi, F. (1990). Network for approximation and learning. *Proceedings of IEEE*, 78 (9), 1491–1497.
[8] Reed, R. D. and Marks II, R. J. (1999). *Neural Smithing*. MIT Press, Cambridge, MA.
[10] Rivas, V.,M., Merelo, J.J, (2004). Evolving RBFNN untuk time series forcasting, With EvRBF. *Informatica Science*, 165, 7-20.
[11] Sutijo, B., dan Subanar (2004) *Uji Nonlinearitas yang Diabaikan dalam time series*, ICSM, Unisba, Bandung
[12] Sutijo, B., Subanar dan Guritno, S.(2005) *Efek Regulasi Dalam Estimasi Fungsi dengan Pendekatan Jaringan Fungsi Radial Basis*, Seminar sehari Matematika dan Komputer, UNS, Surakarta.

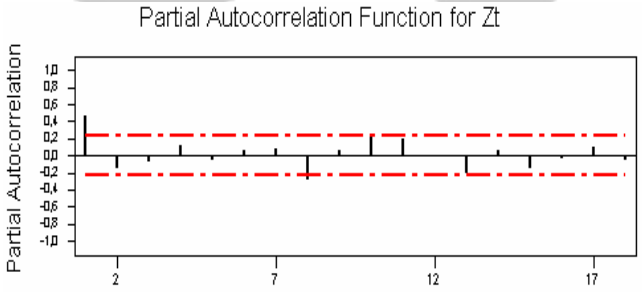
BRODJOL SUTIJO: Ph.D student at Department of Mathematics, Universitas Gadjah Mada, Bulaksumur, Yogyakarta 55281, Indonesia.
Department of Statistics, Institut Teknologi Sepuluh Nopember, Kampus ITS Keputih Surabaya 60111, Indonesia.
E-mail: sutijo_b@yahoo.com

SUBANAR: Department of Mathematics, Universitas Gadjah Mada, Bulaksumur, Yogyakarta 55281, Indonesia.
E-mail: subanar@yahoo.com

SURYO GURITNO : Department of Mathematics, Universitas Gadjah Mada, Bulaksumur, Yogyakarta 55281, Indonesia.
E-mail: suryoguritno@ugm.ac.id

Appendix :

Plot PACF



MODELING OF FINANCIAL DATA BY USING FEEDFORWARD NEURAL NETWORKS

Suhartono[a], Subanar[b]

[a] Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[b] Universitas Gadjah Mada, Yogyakarta, Indonesia

Abstract. The feedforward neural network (FFNN) model has been the most popular form of artificial neural network model used for forecasting, particularly in economics and finance. In this paper, we elucidate the application of FFNN as a means of modeling financial data. We particularly focus on the model building of FFNN as time series model and use inflation rates in Indonesia as a case study. A comparison is drawn between FFNN model and the best existing models based on traditional econometrics time series approach. The best models are selected on forecasting ability by using the MSE and RMSE, particularly on the dynamic forecast. The results show that FFNN models outperform the traditional econometric time series model.

Key-words: feedforward neural networks, inflation, dynamic forecasting

1. Introduction

During the last few years, the use of the neural networks (NN) in economics literature, particularly in the areas of financial statistics and exchange rates, has grown and received a great deal of attention. Some publications about it can be found in [16, 19, 20, 26, 33].

Feedforward Neural Networks (FFNN) model is the most popular form of NN models used for forecasting, particularly in economics and finance. FFNN is a class of flexible nonlinear models that can discover patterns adaptivey from the data. The use of the NN model in applied work is generally motivated by a mathematical result stating that under mild regularity conditions, a relatively simple NN model is capable of approximating any Borel-measureable function to any given degree of accuracy (see e.g. [13, 15, 17, 18, 35]).

The investigation of nonlinearites in time series data is important to macro-economic theory as well as forecasting, as illustrated, in seminal work by Brock and Hommes [10], or Barnett, Medio and Serletis [4]. Recently, many studies have applied NN models to macroeconomic time series, particularly on the modeling and forecasting inflation. Some paper about inflation forecasting by using NN can be found in [11, 21, 23, 25, 27, 28, 32].

This paper discuss and investigate the usefulness of FFNN for forecasting inflation in Indonesia. Two main issues about the effect of increasing fuel price (also known as BBM) and Islamic Calendar effects (price tend to increase during Ramadhan and the Eids holiday) to the inflation fluctuation are also studied. Finally, a comparison

is drawn between FFNN model and the best existing models based on traditional econometrics time series approach.

2. Inflation Forecasting

The investigation about forecasting inflation in a specific country has been received a great attention for many macroeconomics researcher. For most central banks, inflation is at least one monetary policy objective. Given typical time lags, monetary policy needs to be concerned with future inflation. Current inflation levels, which are themselves the result of past policies, may provide only insufficient information. Inflation forecasts that link future inflation to current developments can bridge this gap. This paper attempts to develop an inflation forecasting model for Indonesia that could serve as input for policy setting by the Bank Indonesia (BI).

Moshiri and Cameron [25] did a comparison study between NN and econometrics models for forecasting inflation in Canada. Stock and Watson [28] and Chen, Racine and Swanson [11] have studied NN for forecasting inflation in USA. Kabundi, Marais and Greyling [21] have discussed and compared between NN and econometrics models for forecasting inflation in South Africa. McNelis and McAdam [23] also have studied about forecasting inflation in USA, Japan and some Europe countries by using "*Thick Model*" and NN.

In Indonesia, modeling inflation has been studied by Arief [3] and Anglingkusumo [2]. Arief [3] used econometrics approach by implementing three models; Meiselman model, Anderson-Karnosky model, and Causal model developed by Hsiao. Anglingkusumo [2] implemented P-star model for monetary analysis of inflation.

2.1. Econometrics Time Series Approach

Modeling and forecasting inflation by using econometrics time series approach is usually used by many researchers in many decades, especially compared to the used of NN model. In this section, we will give a brief review of some forecasting models from econometrics time series approach that used in this paper, particularly ARIMA, Intervention Analysis and Calendar Variation Model.

2.1.1. ARIMA Model

The ARIMA model belongs to a family of flexible linear time series models that can be used to model many different types of seasonal as well as nonseasonal time series. The seasonal ARIMA model can be expressed as: (see e.g. [8, 34])

$$\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D y_t = \theta_q(B)\Theta_Q(B^S)\varepsilon_t \ , \qquad (1)$$

where S is the seasonal length, B is the back shift operator and ε_t is a sequence of white noises with zero mean and constant variance.

2.1.2. Intervention Analysis Model

Intervention analysis model is a time series method that usually used to explain the effect of external and internal factors to the time series data. Some papers about the application of intervention analysis model can be found in [6, 9, 14, 22, 24, 30, 31].

The general class of intervention analysis models can be written as: (see e.g. [8] and [34])

$$Y_t = \frac{\omega_s(B)}{\delta_r(B)} B^b I_t + \frac{\theta_q(B)\Theta_Q(B^s)}{\phi_p(B)\Phi_P(B^s)} a_t \tag{2}$$

where b is the time delay for the intervention effect and I_t is intervention variable.

2.1.3. Calendar Variation Model

Calendar variation effects model was originally given by Bell and Hillmer [5]. Suhartono and Sampurno [29] studied the effect of Eids holiday (as Islamic calendar effects) to the increasing number of train and plane passengers at Jakarta-Surabaya route by using calendar variation model. This approach also used by Bokil and Schimmelpfennig [7] for forecasting inflation in Pakistan. In general, the calendar variation model can be written as (see [12])

$$Y_t = \alpha_1 C_t + \frac{\theta_q(B)\Theta_Q(B^s)}{\phi_p(B)\Phi_P(B^s)} a_t \tag{3}$$

where α_1 is the effect magnitude of calendar variation variable and C_t is calendar variation variable.

2.2. Feedforward Neural Network

Neural networks (NN) are a class of flexible nonlinear models that can discover patterns adaptively from the data. Theoretically, it has been shown that given an appropriate number of nonlinear processing units, NN can learn from experience and estimate any complex functional relationship with high accuracy. Empirically, numerous successful applications have established their role for pattern recognition and time series forecasting.

Feedforward Neural Networks (FFNN) is the most popular NN models for time series forecasting applications. Figure 1 shows a typical three-layer FFNN used for forecasting purposes. The input nodes are the previous lagged observations, while the output provides the forecast for the future values. Hidden nodes with appropriate nonlinear transfer functions are used to process the information received by the input nodes.

The model of FFNN in figure 1 can be written as

$$y_t = \beta_0 + \sum_{j=1}^{q} \beta_j f\left(\sum_{i=1}^{p} \gamma_{ij} y_{t-i} + \gamma_{oj} \right) + \varepsilon_t , \quad (4)$$

where p is the number of input nodes, q is the number of hidden nodes, f is a sigmoid transfer function such as the logistic:

$$f(x) = \frac{1}{1+e^{-x}} , \quad (5)$$

$\{\beta_j, j = 0,1,\cdots,q\}$ is a vector of weights from the hidden to output nodes and $\{\gamma_{ij}, i = 0,1,\cdots,p; j = 1,2,\cdots,q\}$ are weights from the input to hidden nodes. Note that equation (4) indicates a linear transfer function is employed in the output node.

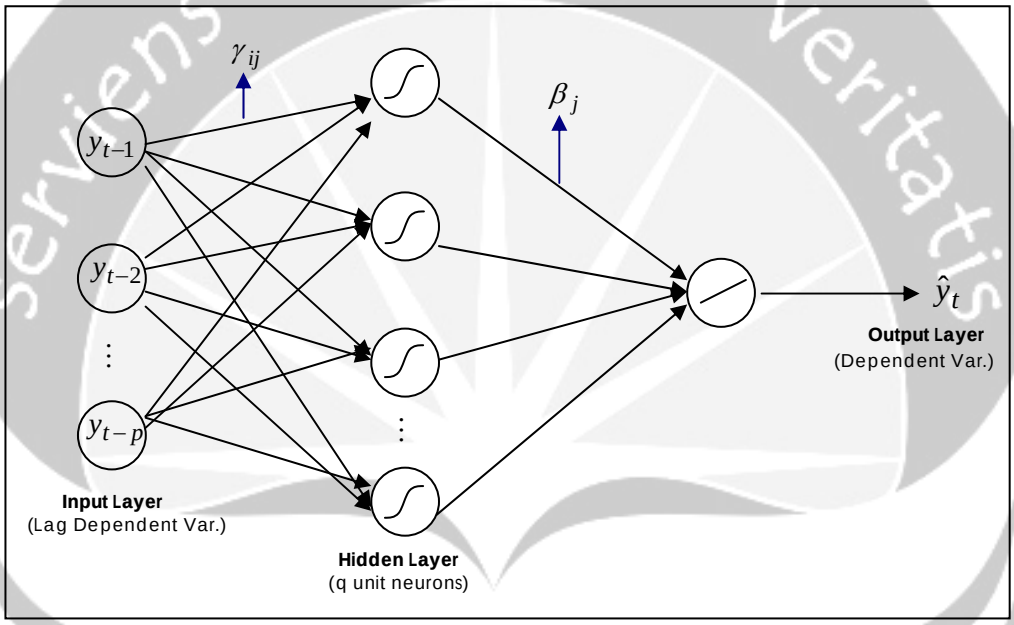


Figure 1. Architecture of neural network model with single hidden layer

Functionally, the FFNN expressed in equation (4) is equivalent to a nonlinear AR model. This simple structure of the network model has been shown to be capable of approximating arbitrary function (see e.g. [13, 15, 17, 18, 35]). However, few practical guidelines exist for building a FFNN for a time series, particularly the specification of FFNN architecture in terms of the number of input and hidden nodes is not an easy task.

3. Research Methodology

The purpose of this research is to provide empirical evidence on the comparative study between FFNN and traditional econometrics time series model for forecasting inflation in Indonesia. The major research questions we investigate are:

- Does FFNN have a better result on the accuracy for forecasting inflation in Indonesia than traditional econometrics time series model ?
- How to build the best FFNN model for forecasting inflation in Indonesia ?

3.1. Data

The Indonesian inflation data that used in this empirical study contain 76 month observations, started in January 1999 and ended in April 2005. The first 72 data observations are used for model selection and parameter estimation (training data in term of NN model) and the last 4 points are reserved as the test for forecasting evaluation and comparison (testing data). Figure 2 (see appendices) plots representative time series of this data. It is clear that the series has an stationary condition with little seasonal variations.

3.2. Research Design

In this research, four type models for forecasting inflation in Indonesia are applied and compared by implementing statistical package MINITAB and SAS for econometrics time series models and using S-Plus and MATLAB for FFNN models. Those models are ARIMA, Combination between Intervention and Variation Calendar Models, FFNN with input as ARIMA and FFNN with input as Combination Intevention and Variation Calendar Models.

To determine the best model, an experiment is conducted with the basic cross validation method. The available training data is used to estimate the parameters (weights) for any specific model. The testing set is the used to select the best model among all models considered. In this study, the number of hidden nodes for FFNN model varies from 1 to 6 with an increment of 1.

The FFNN model used in this empirical study is the standard FFNN with single-hidden-layer shown in Figure 1. The initial value is set to random with replications in each model to increase the chance of getting the global minimum. We did not use the standard data preprocessing in NN by transform data to [-1,1] and N(0,1) scale, because data inflation varies around 0. The performance of in-sample fit and out-sample forecast is judged by the commonly used error measures. They are the mean squared error (MSE) and the root mean square error (RMSE).

4. Empirical Results

In this section the empirical results for ARIMA, Combination Intervention and Variation Calendar (for simplicity we write ARIMAX) and FFNN models are presented and discussed.

4.1. Results of ARIMA Model

The identification step shows that the autocorrelation function (ACF) cuts off after lag 1 and significant at lag 11 and 12, while the partial autocorrelation function (PACF) also cuts off after lag 1 and significant at lag 10, 11 and 13. This suggests

that seasonal ARIMA model should be used for the data. We estimate eight ARIMA models with seasonal length 11 and 12.

The results of forecast comparison by using MSE and RMSE criteria show that ARIMA(1,0,0)(1,0,0)^{11} is the best model for out-sample forecast (testing data), while ARIMA(0,0,1)(0,0,1)^{12} is the best model for in-sample forecast (training data), as shown in table 1 (in appendices). From table 1, we can also observe that out-sample forecast of ARIMA models yield greater errors than in-sample forecast.

4.2. Results of ARIMAX Model

Table 2 (in appendices) shows the results of three ARIMAX models that satisfy adequate model by testing parameter model and diagnostic check of residual model. From table 2, we can conclude that intervention variable and Islamic calendar significantly influence the increasing of forecast accuracy, particularly in out-sample forecast.

These three models contain the effect of increasing BBM price and Islamic calendar to inflation data plus ARIMA model for the errors; those are ARIMA(0,0,[1,12]), ARIMA(0,0,1)(0,0,1)^{12} and ARIMA(1,0,0)(0,0,1)^{12} for model 1, 2 and 3 respectively. For example, model 1 can be written as

$$y_t = 0.4506 + 0.885556 I_t + 0.856335 C_t + (1 + 0.51111 B + 0.29758 B^{12}) a_t \qquad (6)$$

where I_t is intervention variable (increasing of BBM price), C_t is Islamic calendar variable and B is backshift operator. This model shows that increasing of BBM and Islamic calendar have positive effect to inflation in Indonesia.

4.3. Results of FFNN Model

In this paper, building process for FFNN model particulary determination of inputs are based on the inputs of ARIMA dan ARIMAX models. Table 3 (in appendices) summarizes the results of FFNN forecasting with input lags based on ARIMA and ARIMAX models.

The results show that the more complex of FFNN architecture (it means the more number of unit nodes in hidden layer) always yields better result in training data, but the opposite result happened in testing data. Moreover, FFNN models with input lags based on ARIMAX model give better forecast than based on ARIMA model. It can be clearly seen from the reduction of MSE and RMSE particularly in testing data.

4.4. Results of Comparison Study

We concentrate on the dynamic forecasts (testing data) to choice the best model for forecasting inflation in Indonesia. The comparison study uses MSE testing data of the best model in each approach and also ratio of forecast errors of each model to the forecast error of the FFNN model with lags input based on ARIMAX model. The results are presented in table 4.

Table 4. Summary of Dynamic Forecasting Performance Comparison

Best Model	MSE (testing data)	Ratio MSE (to FFNN based on ARIMAX)
▪ ARIMA	0.6826480	3.02
▪ ARIMAX	0.2407240	1.07
▪ FFNN with input based on ARIMA	0.4711709	2.08
▪ FFNN with input based on ARIMAX	0.2261001	1.00

In table 4, numbers greater than one on the ratio column indicate poorer forecast performance than comparable FFNN with inputs based on ARIMAX model and vice versa for numbers less than one. Based on the result at this table, we can conclude that FFNN with inputs based on ARIMAX model, that is input lags 1, 12, I_t, C_t and 4 unit neurons in hidden layer, gives the best dynamic forecast (testing data) for inflation data.

5. Conclusions

Based on the results at the previous section, we can conclude that the FFNN with inputs such as the Combination Intervention and Calendar Variation Model gives the best result for forecasting inflation in Indonesia. Our result also shows that the best FFNN model in training data tends to yield overfitting on testing. This condition gives a chance to do further research by implementing some NN model selection procedures as explained in [1].

References

[1] Anders, U. and O. Korn (1999), Model selection in neural network, *Neural Networks*, **12**, 309–323.

[2] Anglingkusumo, R. (2005), *Money – Inflation Nexus in Indonesia: Evidence form a P-Star Analysis*, Timbergen Inst. Discussion Paper, TI 2005-054/4.

[3] Arief, S. (1995), *An Econometric Study on the Relative Effectiveness of Monetary and Fiscal Policies in Indonesia*, UGM, Indonesia.

[4] Barnett, W.A., A. Medio and A. Serletis (2003), *Nonlinear and Complex Dynamics in Economics*, Working Paper.

[5] Bell, W.R. and S. Hillmer (1983), Modeling time series with calendar variation, *Journal of American Statistical Association*, **78**, 526-534.

[6] Bhattacharya, M.N and A.P. Layton (1979), Effectiveness of Seat Belt Legislation on Queensland Road Toll: An Australian Case Study in Intervention Analysis, *Journal of American Statistical Association*, **74**, pp.367.

[7] Bokil, M. and A. Schimmelpfennig (2005), *Three Attempts at Inflation Forecasting in Pakistan*, IMF Working Paper, WP/05/105.

[8] Box, G.E.P., G.M. Jenkins and G.C. Reinsel (1994), *Time Series Analysis, Forecasting and Control,* 3rd edition, Englewood Cliffs: Prentice Hall.

[9] Box, G.E.P and C.G. Tiao (1975), Intervention Analysis with Applications to Economic and Environmental Problems, *Journal of American Statistics Association*, **70**, pp. 70-79.

[10] Brock, W.A. and C.H. Hommes (1997), A Rational Route to Randomness, *Econometrica,* **65**, 1059-1095.

[11] Chen, X., J. Racine and N.R. Swanson (2001), *Semiparametric ARX Neural Network Models with an Application to Forecasting Inflation*, Working Paper, Economics Department, Rutgers University.

[12] Cryer, J.D. (1986), *Time Series Analysis*, Boston: Publishing Company.

[13] Cybenko, G. (1989), Approximation by superpositions of a sigmoidal function. Mathematics of Control, *Signals and Systems*, **2**, 304–314.

[14] Enders, W., T. Sandler and J. Cauley (1990), Assessing the Impact of Terrorist Thwarting Policies: An Intervention Time Series Approach, *Defense Economics,* **2**, 1-18.

[15] Funahashi, K. (1989), On the approximate realization of continuous mappings by neural networks, *Neural Networks*, **2**, 183–192.

[16] Hamid, S.A. and Z. Iqbal (2004), Using neural networks for forecasting volatility of S&P 500 Index futures prices, *Journal of Business Research,* **57**, pg. 1116–1125.

[17] Hornik, K., M. Stinchombe and H. White (1989), Multilayer feedforward networks are universal approximators, *Neural Networks*, **2**, 359–366.

[18] Hornik, K., M. Stinchombe and H. White (1990), Universal approximation of an unknown mapping and its derivatives using multilayer feedforeard networks, *Neural Networks*, **3**, pp. 551–560.

[19] Kaashoek, J. F., and H.K. Van Dijk (2001), *Neural Networks as Econometric Tool*, Report EI 2001–05, Econometric Institute Erasmus University Rotterdam.

[20] Kaashoek, J.F. and Van Dijk, H.K. (2002), Neural Network Pruning Applied to Real Exchange Rate Analysis, *Journal of Forecasting*, **21**, pp. 559–577.

[21] Kabundi, A.N., D.J. Marais and L. Greyling (2003), *Forecasting South African Inflation: Neural Networks Versus Econometric Models*, Research Paper, Department of Economics, Rand Afrikaans University.

[22] Leonard, M. (2001), *Promotional Analysis and Forecasting for Demand Planning: A Practical Time Series Approach*, Cary, NC, USA : SAS Inst. Inc.

[23] McNelis, P. and P. McAdam (2004), *Forecasting Inflation with Thick Models and Neural Networks*, Working Paper Series, No. 352, European Central Bank.

[24] Montgomery, D.C. and G. Weatherby (1980), Modeling and Forecasting Time Series Using Transfer Function and Intervention Methods, *AIIE Transactions*, pg. 289-307.

[25] Moshiri, S. and N.E. Cameron (1997), *Neural Network vs. Econometric Models in Forecasting Inflation*, Working Paper, Department of Economics, University of Manitoba.

[26] Refenes, A.P. and H. White (1998), Neural Networks and Financial Economics, *International Journal of Forecasting*, **17**.

[27] Stock, J.H. and M.W. Watson (1998), *A Comparison of Linear and Non-linear Univariate Models for Forecasting Macroeconomic Time Series*, NBER Work Paper 6607.

[28] Stock, J.H. and M.W. Watson (1999), Forecasting Inflation, *Journal of Monetary Economics*, **44**, pg. 293-335.

[29] Suhartono and B.S. Sampurno (2002), The Comparison Study between Transfer Function and Intervention-Calendar Variation Model for Forecasting The Number of Plane and Train Passengers, *Jurnal Matematika atau Pembelajarannya*, Special Edition, Universitas Negeri Malang, Indonesia.

[30] Suhartono and E. Hariroh (2003), Analysis of The Effect of WTC New York Bomb to The Fluctuation of World Stock Market by Using Intervention Model, *Proceeding National Seminar of Mathematics and Statistics*, Institut Teknologi Sepuluh Nopember, Indonesia.

[31] Suhartono and I.N.A.W.W. Putra (2005), The Effect of Bali Bomb to The Number of Hotel Occupancy in Bali: An Application Study of Intervention Model for Tourism, *Proceeding National Conference Mathematics XII*, UDAYANA, Bali, Indonesia.

[32] Swanson, N. R. and H. White (1997), A model-selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks, *Review of Economic and Statistics*, **79**, 540–550.

[33] Versace, M., R. Bhatt, O. Hinds, and M. Shiffer (2004), Predicting the exchange fund DIA with a combination of genetic algorithms and neural networks, *Expert Systems with Applications,* **27**, pg. 417–425.

[34] Wei, W.W.S. (1990), *Time Series Analysis: Univariate and Multivariate Methods,* Addison-Wesley Publishing Co., USA.

[35] White, H. (1990), Connectionist nonparametric regression: Multilayer feed forward networks can learn arbitrary mapping, *Neural Networks*, **3**, 535–550.

SUHARTONO: Ph.D student at Department of Mathematics, Universitas Gadjah Mada, Bulaksumur, Yogyakarta 55281, Indonesia.
Department of Statistics, Institut Teknologi Sepuluh Nopember, Kampus ITS Keputih Surabaya 60111, Indonesia.
E-mail: suhartono@statistika.its.ac.id

SUBANAR: Department of Mathematics, Universitas Gadjah Mada, Bulaksumur, Yogyakarta 55281, Indonesia.
E-mail: subanar@yahoo.com

Appendices:

A. Time series plot of the Indonesian inflation data, January 1999 – April 2005

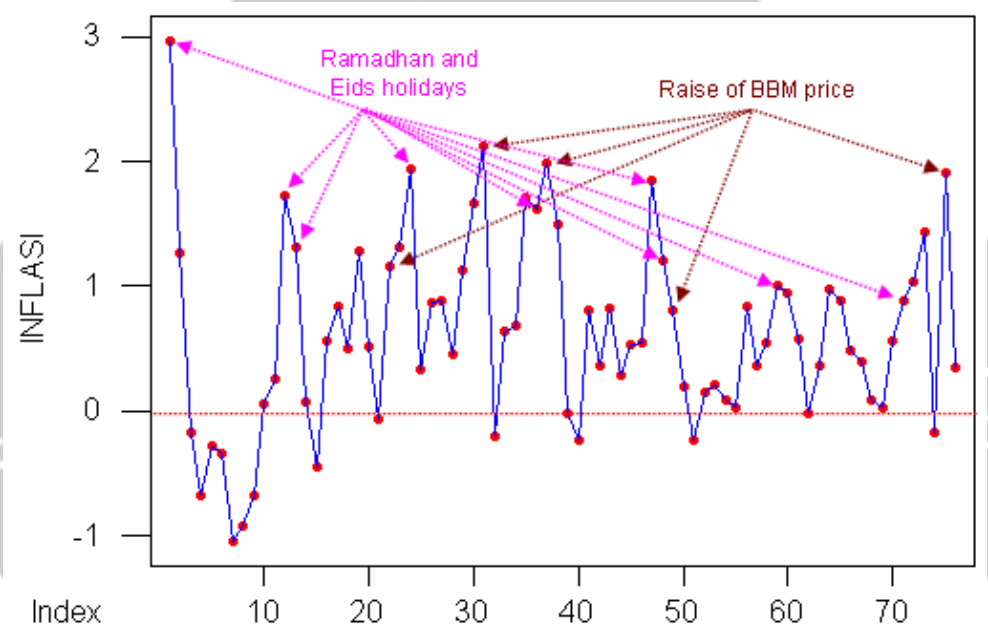


Figure 2. Time series plot of the Indonesian inflation, January 1999 – April 2005

B. The results of ARIMA models, both in training and testing data

Table 1. The results of ARIMA models, both in training and testing data

Model	MSE		RMSE	
	Training data	Testing data	Training data	Testing data
▪ ARIMA(1,0,0)(1,0,0)[11]	0.3576	0.682648	0.597997	0.826225
▪ ARIMA(0,0,1)(0,0,1)[12]	0.2624	0.827925	0.512250	0.909904

C. The results of ARIMAX models, both in training and testing data

Table 2. The results of ARIMAX models, both in training and testing data

Model	MSE		RMSE	
	Training data	Testing data	Training data	Testing data
▪ Model 1	0.28626167	0.289602	0.535034	0.538147
▪ Model 2	0.29634263	0.240724	0.544374	0.490636
▪ Model 3	0.29359180	0.319303	0.541841	0.565069

D. The results of FFNN models, both in training and testing data

Table 3. The results of FFNN models, both in training and testing data

Model	Input lags	Number of neurons	Training data		Testing data	
			MSE	RMSE	MSE	RMSE
▪ FFNN with input lags based on ARIMA	1,12	1	0.3105369	0.557258	0.7257537	0.85191
		2	0.3031623	0.550602	0.7064180	0.84049
		3	0.2854341	0.534260	0.7579172	0.87058
		4	0.2057219	0.453566	1.0984050	1.04805
	1,11,12	1	0.2069178	0.454882	0.8657498	0.93046
		2	0.1891711	0.434938	0.8337273	0.91309
		3	0.1736372	0.416698	0.4711709	0.68642
		4	0.1418450	0.376623	0.8205497	0.90584
		5	0.1235492	0.351496	1.3148560	1.14667
▪ FFNN with input lags based on ARIMAX	1,12 I_t, C_t	1	0.2229641	0.472191	0.3670807	0.605872
		2	0.2040528	0.451722	0.3122488	0.558792
		3	0.1499683	0.387257	0.2601240	0.510024
		4	0.1366765	0.369698	0.2261001	0.475500
		5	0.1210808	0.347967	0.2973856	0.545331
	1,11,12 I_t, C_t	1	0.2950905	0.543222	0.3461342	0.588332
		2	0.1531187	0.391304	0.3603422	0.600285
		3	0.1471724	0.383631	0.4064297	0.637518
		4	0.2202060	0.469261	0.3210728	0.566633
		5	0.1224852	0.349979	0.5476139	0.740009

NEURAL NETWORKS FOR CLASSIFICATION PROBLEMS

Sri Rezeki[a], Subanar[b] and S. Guritno[b]

[a]Universitas Islam Riau, Pekanbaru, Indonesia
[b]Universitas Gadjah Mada, Yogyakarta, Indonesia

Abstract. This paper presents a definitive description of neural network methodology for classification problem and provides an evaluation of its advantages and disadvantages relative to statistical procedure, particularly toward logit models. Neural network models are closely related to generalized linear models. Additionally, neural network models do not require the restrictive assumptions about the relationship between the independent variables and dependent variable(s). Consequently, these models have already been very successfully applied in many diverse disciplines, including biology, psychology, statistics, mathematics, bussiness, insurance, and computer science. By using two examples of classification problems, we demonstrate that neural network provide better predictive power than logit models.

Key words: neural networks, logit model, classification

1. Introduction

The successful application of neural networks (NN) to practical problems has been shown in numerous papers. In the medical field, this application range from diagnosis of myocardial infarction over classification of EEG signals and PET scans to prediction of mechanisms of action in cancer drug development [10]. The use of NN is now quite common in economics and business and these models are used to describe and forecast many important variables. For example, they have been used to predict and explain patronage behavior [5] and to model brand choice [4], [12], [13].

Implementation of NN for classification problems are such that an analysis based on a logistic regression model [7] with a standard approach for a statistical analysis. A perceptron with a logistic activation function represent the most simple neural network, they consist only in an input layer and an output layer. These relationships between the inputs and the outputs are identical to the logistic regression. By adding one or more hidden layers, it is enables the the network to map interactions and more generally nonlinear relationships.

Although NN is feasible alternative to logit models, its forecasting ability compared to logit models are not clear. In general there is continuing debate about the comparative performance of the NN and traditional approach in several different contex. In marketing research, there is some evidence that an NN yields more useful insight than a logistic regression model. For example, West *et al* [13] compare an NN with discriminant analysis and logistic regression. They conclude

that an NN can outperform the two statistical techniques when the underlying choice rule is known and can give better out-of-sample forecasts when the choice rule is not known. The same research which was done by Dasgupta *et al* [6] conclude that the superiority of the NN is not statistically significant. Comparison between NN and logit models can also be seen at some articles (see e.g. [1], [3], [8], [10], [14]).

Based on the progresses at classification problems, we feel that a thorough study of logit models and NN still deserved attention. This research is done and focused on the comparison between the logit models and NN particularly about classification correct rate.

2. Logit Models

Logit modelling has been extensively described in the literature [9], and we shall only describe briefly the characterisitcs of these models. However, in order to compare logit models to NN, it is important to mention their fundamental equations and properties.

The binomial logit model has response variable Y which contain two categories (binary or dichotomous). This response is usually distinguished to 'success' and 'fail' denoted Y=1 (success) and Y=0 (fail) [2]. Let the explanatory variables X' = (X_1, X_2, …, X_k) that pair to response variable Y. Probability Y=1 denoted by $\pi(x)$. Logistic regression function between $\pi(x)$ and x is:

$$P(Y=1)=\pi(x)=\frac{\exp[g(x)]}{1+\exp[g(x)]} \qquad (1)$$

where $g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$

The multinomial logit model is an extension of the binomial logit model, where variable Y has polychotomous response with outcome possibility 1, 2, … , J (J is integer, usually a small number). The multinomial logit model computes the probability $P_i(Y=j)$ of choosing category j on choice occasion *i* as a function of all other categories [4]. As for the binomial logit, it assumes that a linear combination of the explanatory variables is related to the choice probability $P_i(Y=j)$ as described by

$$P_i(Y=j)=\frac{\exp[g_{ij}(x)]}{\sum_{j=1}^{J}\exp[g_{ij}(x)]} \qquad (2)$$

where

$P_i(Y=j)$ is the probability of choosing category j on choice occasion i

$$g_{ij}(x) = \sum_{k \in T} b_k x_{ijk} \quad \text{with}$$

b_k : kth parameter estimation
x_{ijk} : represents the explanatory variable
J : number of categories considered
T : set of variables

Three interesting properties of the multinomial logit model arise from equation (2). First, g(x) function is undetermined to the extent of an additive constant. Second, if there are only two alternatives, equation (2) will be reduced to equation (1). Finally, $P_i(Y=j)$ is S-shaped in $g_{ij}(x)$ when $g_{ik}(x)$ are held constant. Therefore, as for the binomial logit, large or small values for $g_{ij}(x)$ make $P_i(Y=j)$ flat and insensitive to changes in $g_{ij}(x)$.

The parameters of model are estimated by Maximum Likelihood Estimation (MLE). The likelihood of observing actual choice, given input vector X and model parameter vector β, can be expressed by

$$L(Y|X,\beta) = \prod_{1}^{N} P_i^{Y_i} (1-P_i)^{1-Y_i} = \prod_{i=1}^{N} \prod_{j=1}^{J} \left(\frac{\exp(g_{ij}(x_{ij}, b))}{\sum_{h=1}^{J} \exp(g_{ih}(x_{ih}, b))} \right)^{Y_{ij}} \quad (3)$$

where
Y = dependent variable (polychotomous data)
X = $(x_{11}, \ldots, x_{1n}, \ldots, x_{kn})$ represent the explanatory variables
β = $(b_1, \ldots, b_k)$ represent the model parameters
P_i is the predicted probability for choice occasion i
g_{ij} is value of function for the jth category on the ith choice occasion
N is the number of choice occasions and J is the number of categories

In practice, however, it is preferable to maximise the logarithm of the likelihood as expressed in:

$$LogL = \sum_{i=1}^{N} \sum_{j=1}^{J} Y_{ij} \ln(Pi(Y = j|x, \beta)) \quad (4)$$

with $P_i(Y = j|x, \beta) = \dfrac{\exp g_{ij}(x_{ij}, b)}{\sum_{h=1}^{J} \exp g_{ih}(x_{ih,b})}$

3. FFNN as an extension of Logit Models

The type of NN that are considered in this paper is the multilayer feedforward neural networks (FFNN) trained with the standar backpropagation algorithm. Details of the algorithm can be found in any textbook on NN and will not be addressed in this paper. However, it is interesting to illustrate a parallel between NN with sigmoidal outputs and logistic regression models (respectively the binomial and multinomial logit model)

Analogy between FFNN and binomial logit is FFNN can be considered as nonlinear regression models, the complexity of which can be changed. At their lowest complexity level, they consist only in an input layer and an output. These relationships between the input and the output are identical to binomial logit. Parameters of FFNN are usually estimated by minimizing the mean square error produced by the model. This is equivalent to maximum likelihood estimation when the dependent variable is a continuous function of the inputs with additive Gaussian noise, as is the case for regression problems. For classification problems, however, the dependent variable is binary and a Gaussian distribution for the errors is inappropriate and therefore the error function is a priori not mean square error. The entropy function is a more appropriate error function. Equation (5) defines the entropy function E:

$$E = -\sum_{i=1}^{M} (Y_i \ln(S_i) + (1-Y_i)\ln(1-S_i)) \tag{5}$$

Where S_i is the network output and Y_i the desired output.

Apart from its sign, this function is identical to the log-likelihood function used to estimate the coefficients of the binomial logit model (equation (4)). The FFNN with sigmoidal output is therefore exactly identical in form and estimation procedure to the binomial logit model. It models the choice probability based on a linear utility (g(x)) function and therefore does not take into account potential interaction effects between explanatory variables. It is possible to change the network complexity in order to take into account non-linearities, in particular interaction effects. This is done by adding one or more hidden layer(s) to the network. Hence, an FFNN can be viewed as a generalization of the binomial logit model.

Just as the sigmoidal output network generalizes the binomial logit model, the softmax output network with shared weights generalizes the multinomial logit model. The used error function (relative entropy), given by equation (6), is identical (apart sign) to the log-likelihood function used to derive the multinomial logit (equation4).

$$E = -\sum_{i=1}^{M}\sum_{j=1}^{J} Y_{ij} \ln(S_{ij}) \tag{6}$$

The multinomial logit model corresponds to a partially connected FFNN with shared weight and softmax output. By adding hidden neurones to the network, it is

possible to model choices resulting from nonlinear $g_{ij}(x)$ function. Therefore, NN models can be viewed as generalized multinomial logit model.

The architecture of FFNN for classification problems is illustrated in figure 1.

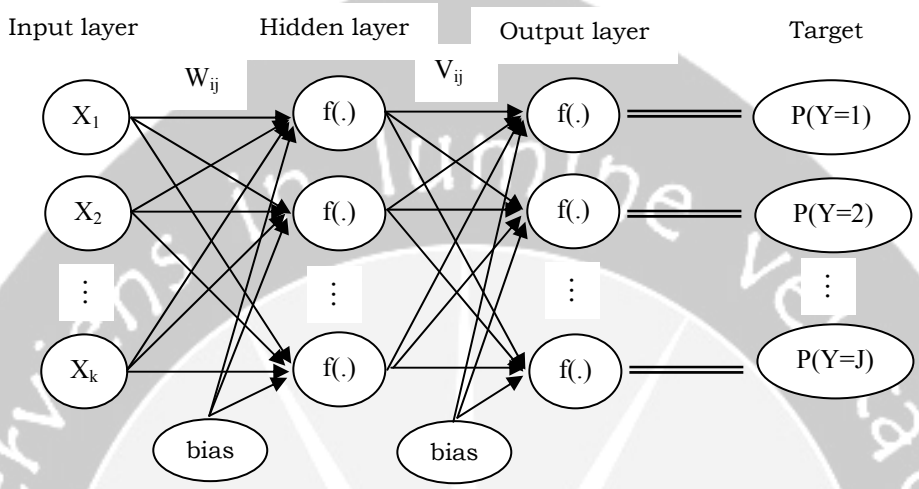


Figure 1. FFNN with single hidden layer

Advantages and disadvantages of NN compared with the logit models are:

1. NN afford to model nonlinear preferences with few (if any) a priori assumption, while logit models can suffer from a specification bias if the assumptions are not fulfilled [4], [8].

2. NN are very flexible and able to model highly complex relationships. However they are very difficult to interpret and have been perceived as 'black boxes' which can neither explain how they reach an outcome nor provide an explicit representation of the relationship that they estimate. On the other hand, logit models provide easily interpretable coefficients and significance statistics [4], [8].

3. NN to be relatively easier in terms of analytical and computational effort, whereas advantage of logit models are its ability to provide close form solutions for the choice probabilities [1].

4. NN approach is parsimonious, produces better classification, and is stronger at interpolation [1], [8].

The predictive power of two models can be measured by percentage correct of classification. Suppose the following result were obtained:

Tabel 1. Classification of polychotomous response data

Data	Predicted			
Observed	1	2	...	J
1	f_{11}	f_{12}	...	f_{1J}
2	f_{21}	f_{22}	...	f_{2J}
⋮	⋮	⋮	⋱	⋮
J	f_{J1}	f_{J2}	...	f_{JJ}

$$\text{Percentage Correct} = \frac{\sum_{i=1}^{J} f_{ii}}{\sum_{i=1}^{J}\sum_{j=1}^{J} f_{ij}} \qquad [7]$$

4. Implementation of The Logit Models and NN

This paper used synthetic dataset to illustrate the Logit models and NN. For binomial response case, let us consider the response of consumer toward a new product, namely instant food (whether they buy or not) and three decision variables i.e. consumer novelty seeking (X_1), normative influence (X_2) and informational influence (X_3). The number of object are 193 observation, 60% for training set and 40% for testing set. The second case for multinomial response also rest on a shyntetic dataset, those are the choice of house types (for example A, B, and C type), while the explanatory variables are monthly income (X_1) and married age (X_2). The sample size is 150 observation where data partition for training and testing set is the respective 50%

The determination of NN architecture is done by trial and error until obtained the optimal number of hidden neuron, which is the smallest misclassification rate and should consider parsimony model. In order to get the stable result, we have to carry out resampling technique. Then, data are divided into training and testing set in arbitrary ratio (usually the number of training set are greater than the one of testing set. The mean of classification correct rate after ten times replications can be summarized as follow:

A. Binomial Response Case

Table 2. The Mean of Classification correct rate (%)

Model	Training Set	Testing Set
The Binomial Logit Model	68.83	57.37
Neural Network (3-2-2)	70.43	66.47

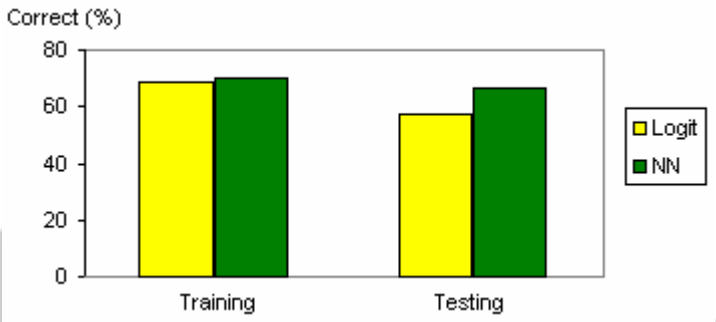


Figure 2. Comparison of classification correct average

B. Multinomial Response Case

Table 3. The Mean of Classification correct rate (%)

Model	Training Set	Testing Set
The Multinomial logit Model	97.6	94.94
Neural Network (2-3-3)	99.74	95.46

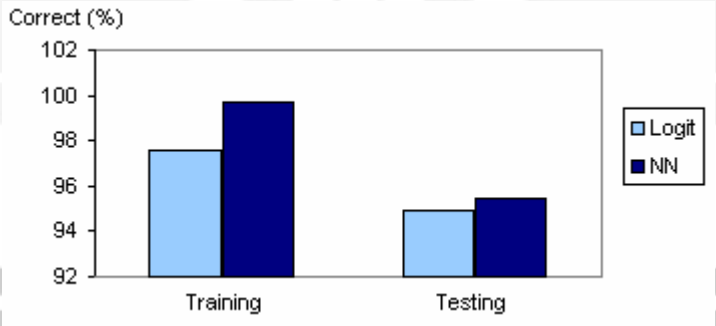


Figure3. Comparison of classification correct average

Table 2. and 3, also Figure 2. and 3. show that ability of NN for classifying correctly are better than the Logit model. NN are more powerful than the logit model both in training and testing set. This is consistent with the previous finding in the literature that NN approach is parsimonious, produces better classification, handle complex underlying relationships better, and is stronger at interpolation. On the other hand, the logistic regression technique has a superior solution methodology (close form versus heuristic) and better interpretability [8].

5. Conclusion

Since we have discussed the logistic perceptron as the most simple neural network is equivalent to a logistic regression model and the weights in neural network can be interpreted as regression coeffisients in the corresponding logistic model. In this paper we have compared the logistic regression model with FFNN. There are some obvious extensions of the logistic perceptron allowing for more than one output unit and/or having unit in a hidden layers.

The advantages and disadvantages of NN approach compared to logit models, one of which is a restriction of the other. Choosing between the two models implicitly requires the user to decide whether the complexity of the data justifies the use of the more complex one. Results in this paper indicate that for classification problems, NN performance relatively better than logit models in two case, dichotomous and polychotomous response. Interaction effects have not been investigated in this paper. This case should be done for further research.

As the respective strengths and weaknesses of the models are complementary, it may be possible to combine them in a single framework in order to take into account potential non-linearities as well as to improve the model interpretability. In certain field, interpretability is at least as important as predictive power.

References

[1] Agrawal, D. & C. Schorling (1996), Market Share Forecasting: An Empirical Comparison of Artificial Neural Networks and Multinomial logit Model, *Journal of Retailing*, **72**(4), pp. 383-407.

[2] Agresti, A. (1990), *Categorical Data Analysis*. John Wiley and Sons, Inc,: New York

[3] Arana, E., P. Delicado & L. Marti-Bonmati (1999), Validation Procedure In Radiological Diagnostic Models, Neural Network and Logistic Regression. To appear in *Investigative Radiology*, October. Department of Radiology, Hospital Casa de Salud, Valencia.

[4] Bentz, Y. & D. Merunka (2000), Neural Networks and the Multinomial logit for Brand Choice Modelling: a Hybrid Approach, *Journal of Forecasting*, **19**, 177 – 200.

[5] Chiang, W.K., D. Zhang, & L. Zhou (2004), Predicting and Explaining Patronage Behavior toward Web and Traditional Stores Using Neural Networks: a Comparative Analysis with Logistic Regression, *Decision Support System*, **xx**.

[6] Dasgupta, C.G., G.S. Dispensa, & S. Ghose (1994), Comparing the predictive performance of a neural network model with some traditional market response models, *International Journal of Forecasting*, **10**, 235-254.

[7] Hosmer, D.W. & S. Lemeshow (1989), *Applied Logistic Regression*, John Wiley & Sons Ltd., New York.

[8] Kumar, A., Rao, V.R., and Soni, H., (1995), An empirical comparison of neural networks and logistic regression models', *Marketing Letters*, **6** (4), 251-263.

[9] McCullagh, P. & J. Nelder (1989), *Generalized Linear Models*, Second edition, Chapman and Hall, New York.

[10] Schumacher, M., R. Roβner, & W. Vach (1996), Neural Network and Logistic Regression: Part I. *Computational Statistic & Data Analysis*, **21**: 661-682.

[11] Venables, W.N. & B.D. Ripley (1998), *Modern Applied Statistic with S-Plus*, second edition, Springer-Verlag, New York.

[12] Vroomen, B., P.H. Franses, & E. Nierop (2001), Modelling Consideration Sets and Brand Choice Using Artificial Neural Networks. *Erasmus Research Institute of Management (ERIM)*, Erasmus Universiteit Rotterdam, www.erim.eur.nl

[13] West, P.M., P.L. Brockett, & L.L. Golden (1997), A comparative analysis of neural networks and statistical methods for predicting consumer choice. *Marketing Science*, **16**, 370–391.

[14] Zhou, L., G. Ersheng, & J. Pihua (1997), Comparison between the Logistic regression and Back Propagation Neural Network, Department of Health Statistic, Shanghai Medical University, Shanghai Cina.

SRI REZEKI: Ph.D student at Department of Mathematics, Universitas Gadjah Mada
Jl. C. Simanjuntak Sekip Utara – Yogyakarta 55281, Indonesia.
Phone/Fax: +62 +274 522 443
Department of Mathematics Education, FKIP Universitas Islam Riau,
Jl. Kaharudin Nasution Marpoyan Pekanbaru, Indonesia.
E-mail: sri_rezeki_uir@yahoo.com

SUBANAR: Department of Mathematics, Universitas Gadjah Mada,
Jl. C. Simanjuntak Sekip Utara – Yogyakarta 55281, Indonesia.
Phone/Fax: +62 +274 522 443
E-mail: subanar@yahoo.com

S. GURITNO: Department of Mathematics, Universitas Gadjah Mada,
Jl. C. Simanjuntak Sekip Utara – Yogyakarta 55281, Indonesia.
Phone/Fax: +62 +274 522 443
E-mail: suryoguritno@ugm.ac.id

Strong Suppression of Radiation States in a Slab Waveguide Sandwiched between Omnidirectional Mirrors

H.J.W.M. Hoekstra[1], D. Yudistira[1] and R. Stoffer[2]

[1)] IOMS group, MESA+ Institute, University of Twente, The Netherlands
[2)] Phoenix B.V., The Netherlands

Abstract: Structures in channel or slab waveguides, applied deliberately or due to imperfections, may lead to strong modal losses, corresponding to the excitation of radiation modes. As an example, losses are generally very large in slab photonic crystal (PhC) impurity waveguides (WGs) due to the combined effect of field enhancement and fabrication errors. In the presentation it is shown that for a silicon slab in air such radiation losses can be strongly reduced, by approximately one order of magnitude, by structural optimization of such a slab sandwiched between two omni-directional mirrors. The effect can be used for the production of low loss PhC impurity WGs, high Q-cavities and low-loss transitions between different WG sections.

Keywords: photonic crystals, photonic bandgap, high Q, cavities, low loss.

IMPLICIT SCHEME FOR NUMERICAL INTEGRATION OF THE NONLINEAR PARTIAL DIFFERENTIAL EQUATION

M. Nurhuda

Physics Department, Brawijaya University, Malang 65144, Indonesia

Abstract. An implicit scheme for integrating the nonlinear partial differential will be presented and discussed. The method is based on the splitting the propagator in term belonging to the homogenous part and non-homogenous part (nonlinearity). The homogenous part is solved using Crank-Nicholson scheme, while non-homogenous part is solved using power expansion of the propagator, assuming that within a finite interval Δz, the nonlinearity is considerably constant. The method is employed for the 1D propagation of intense electromagnetic field in a nonlinear medium. Comparison of the results with that obtained using other method, i.e. Runge Kuta, will also be presented and discussed. It turns out the method is effective, efficient and unconditionally stable.
Key-words: ICAM05, numerical method

1 Introduction

This paper elaborates a scheme for numerical solution of the nonlinear partial differential equation (NLPDE) emerging in physics. Due to broad scope of physical problems that can be associated with NLPDE, we limit our self on the problem of pulse propagation in the nonlinear medium, as we are extensively involved in the last four years [1]. We expect that the method can be extended to solve other nonlinear physical problems, since the basic algorithm is generic for almost all NLPDE.
We start by invoking Maxwell equation describing the electric wave propagation in the nonlinear medium (see e.g. [2]):

$$\nabla^2 E(x,y,z,t) - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}E(x,y,z,t) = \frac{1}{c^2}\frac{\partial^2}{\partial t^2}P_{NL}(x,y,z,t) \qquad (1)$$

where c is the speed of light and $P_{NL}(x,y,z,t)$ is the source of the nonlinearity. We assume that the electric field is pulsed, possesses cylindrical symmetry and is propagating along z direction:

$$E(x,y,z,t) = \mathcal{E}(r_\perp,z,t)e^{i(kz-\omega_0 t)} = \int_{-\infty}^{\infty} \mathcal{E}_0(r_\perp,z)e^{i(kz-\omega t)}d\omega \qquad (2)$$

with $r_\perp = \sqrt{x^2+y^2}$ and $k = \omega/c$ is the wave number. Accordingly, the nonlinear term can be written in the same manner as Eq. (2). In general, the nonlinearity consists of the instantaneous- and delayed terms originated from rotational

atomic/molecular medium and time integrated term, i.e. plasma when the medium is ionized.

$$P_{NL}(r_\perp, z, t) = \left(\Delta\chi(|E(r_\perp, z, t)|^2) - \frac{\omega_p^2}{\omega_0^2}\right)E(r_\perp, z, t) \tag{3}$$

where $\omega_p = \sqrt{e^2\rho(t)/m_e\epsilon_0}$ is the plasma frequency and $\rho(t)$ accounts for the plasma density given by:

$$\frac{\partial\rho(t)}{\partial t} = N_0(1-\rho(t))\Gamma(|E^2(r_\perp, z, t)|^2) \tag{4}$$

with N_0 being the atomic/molecular density of the nonlinear medium, and Γ is the ionization rate. Inserting Eq. (4) in to Eq. (3), and Eq. (3) in to Eq. (1) one can immediately see that the NLPDE in Eq. (1) is of the type of highly coupled nonlinear partial differential equation.
To solve Eq. (1), it is usually assumed that the field is propagating forward along z direction. This assumption neglects the re-scattering of the field (paraxial approximation). Moreover, by applying the moving frame $t = t + z/c$, one can get the envelope equation:

$$\nabla_\perp^2\mathcal{E}(r_\perp, z, t) + \frac{2}{c}\left(i\omega_0 - \frac{\partial}{\partial t}\right)\frac{\partial}{\partial z}\mathcal{E}(r_\perp, z, t) + k_0''\frac{\partial^2}{\partial t^2}\mathcal{E}(r_\perp, z, t) = e^{i\omega_0 t}\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\mathcal{P}_{NL}(r_\perp, z, t)e^{-i\omega_0 t} \tag{5}$$

with $k_0'' = d^2k/d\omega^2|_{\omega_0}$. Note that the dispersion is usually neglected when the electric field is strong, since its effect is too small compared to that contributed by the nonlinearity and ionization.

2 Numerical scheme

The numerical scheme is based on the splitting operator, where for *one step* forward propagation the solution is iteratively obtained from the previous solution by approximating the kernel propagator in the following term:

$$\mathcal{E}(r_\perp, z + \Delta z, t) = G_H(r_\perp, \Delta z, t)G_{NL}(r_\perp, \Delta z, t)\mathcal{E}(r_\perp, z, t) + O(\Delta z^2) \tag{6}$$

where G_H is the homogenous propagator and G_{NL} accounts for the propagator of the respective nonlinear term. Splitting the solution as Eq. (6) is equivalent with simultaneously solving the two coupled differential equations.

$$\nabla_\perp^2\mathcal{E}(r_\perp, z, t) + k_0''\frac{\partial^2}{\partial t^2}\mathcal{E}(r_\perp, z, t) = \frac{2}{c}\left(-i\omega_0 + \frac{\partial}{\partial t}\right)\frac{\partial}{\partial z}\mathcal{E}(r_\perp, z, t) \tag{7}$$

$$\frac{2}{c}\left(i\omega_0 - \frac{\partial}{\partial t}\right)\frac{\partial}{\partial z}\mathcal{E}(r_\perp, z, t) = e^{i\omega_0 t}\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\mathcal{P}_{NL}(r_\perp, z, t)e^{-i\omega_0 t} \tag{8}$$

Note that factorization as described in Eqs. (7) and (8) is not a necessity; one may alternatively put the dispersion term in Eq. (8). The homogenous part Eq. (7) is easily solved in spectral domain with the help of Fourier transform:

$$\nabla_\perp^2 \mathcal{E}(r_\perp, z, \omega-\omega_0) - k_0'' \Delta\omega^2 \mathcal{E}(r_\perp, z, \omega-\omega_0) \quad = \quad 2ik\frac{\partial}{\partial z}\mathcal{E}(r_\perp, z, \omega-\omega_0) \quad (9)$$

The forward propagation then can be obtained using well known Crank Nicholson scheme [3]:

$$\mathcal{E}(r_\perp, z+\Delta z, \omega-\omega_0) = \frac{1-\frac{i\Delta z}{4k}\left(\nabla_\perp^2 - k_0''\Delta\omega^2\right)}{1+\frac{i\Delta z}{4k}\left(\nabla_\perp^2 - k_0''\Delta\omega^2\right)}\mathcal{E}(r_\perp, z, \omega-\omega_0) \quad (10)$$

For the nonlinear part, $\mathcal{P}_{NL}(r_\perp, z, t)$ has to be solved in the temporal domain. However, the presence of the time derivative operator in the left hand side Eq. (8) could lead into complicated operations. To overcome this problem, we introduce the following approximation for Eq. (8):

$$\begin{aligned}\frac{\partial}{\partial z}\mathcal{E}(r_\perp, z, t) &= \left[\frac{2i\omega_0}{c}\left(1+\frac{i\partial}{\omega_0\partial t}\right)\right]^{-1} e^{i\omega_0 t}\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\mathcal{P}_{NL}(r_\perp, z, t)e^{-i\omega_0 t} \\ &\approx -\frac{ic}{2\omega_0}\left(1-i\frac{\partial}{\omega_0\partial t}\right)e^{i\omega_0 t}\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\mathcal{P}_{NL}(r_\perp, z, t)e^{-i\omega_0 t} \quad (11)\end{aligned}$$

Eq. (11) basically can be solved using standard solver of ordinary differential equation, e.g. implicit second order Runge Kuta method. However, due to existence of nonlinear term, the differential equation frequently becomes stiff and hence requires that Δz must be set to be very small. To overcome this stiffness problem, we utilize the following trick. Define a new function $\tilde{\mathcal{X}}_{NL}$ as:

$$\tilde{\mathcal{X}}_{NL}(r_\perp, z, t) = \frac{-\frac{ic}{2\omega_0}\left(1-i\frac{\partial}{\omega_0\partial t}\right)e^{i\omega_0 t}\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\mathcal{P}_{NL}(r_\perp, z, t)e^{-i\omega_0 t}}{\mathcal{E}(r_\perp, z, t)} \quad (12)$$

then Eq. (12) reads

$$\frac{\partial}{\partial z}\mathcal{E}(r_\perp, z, t) \quad = \quad \tilde{\mathcal{X}}_{NL}(r_\perp, z, t)\mathcal{E}(r_\perp, z, t) \quad (13)$$

Assuming that within small interval Δz $\tilde{\mathcal{P}}_{NL}(r_\perp, z, t)$ is constant, one can immediately see that the solution due to effect of nonlinearity can be written as:

$$\mathcal{E}(r_\perp, z+\Delta z, t) = e^{\tilde{\mathcal{X}}_{NL}(r_\perp, z, t)\Delta z}\mathcal{E}(r_\perp, z, t) \quad (14)$$

The scheme Eq. (14) is not only simple in nature, but also cheap in operation, since of each propagation step, we only need one time to evaluate the nonlinearity. This is important, since for large system, the evaluation of the nonlinearity is sometime very costly.

3 Implementation

In order to demonstrate how efficient is the nonlinear solver described in Eq. (8), we performed the following simulation. For the purpose of this presentation, we omitted the homogenous transverse Laplacian in Eq. (7) such that the differential equations reduces to 1D. The pulse is modeled as a Gaussian function and the corresponding intensity has width 60 femtosecond at full wave at half maximum (FWHM). The field amplitude is set to be 0.0534 a.u. while the central wavelength is 789 nm. Data for nonlinearity and ionization rate are depicted in Fig. 1. The

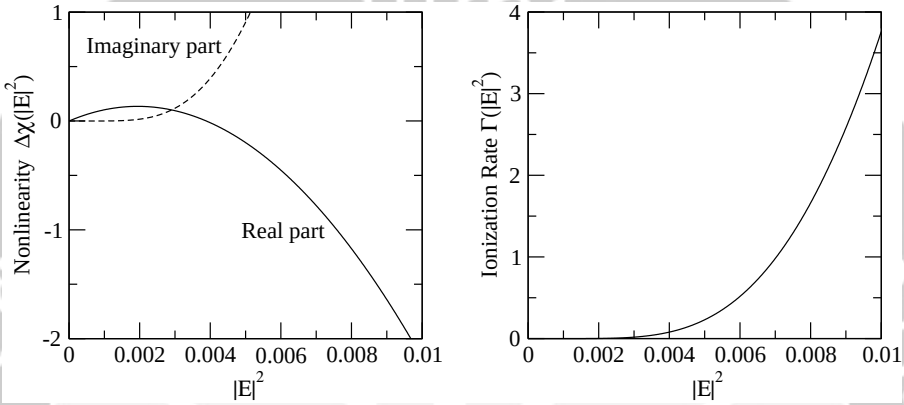


Figure 1: The complex nonlinearity as a function of intensity $|E|^2$ (left) and in the right side is the plot of ionization rate. Both nonlinearity and ionization rate data are of argon atom.

pulse is propagated from distance z=0 to $z = 10$ cm by using various propagation step Δz (in unit of mm). The results are displayed in the spectral domain in Fig. 2; top panels for the results obtained using power expansion method and bottom panels tho those obtained using RK method. It is shown that in general, the present method is unconditionally stable, even with quite extensive propagation step. On the other hand, performing the integration with Runge Kuta method requires very small Δz due to its stability and convergence, as seen in Fig 2 (bottom panel) that for $\Delta z = 0.3$ mm that the result start to deviate from the correct one, while for larger Δz, the method is fail to yield converged result.

4 Conclusions

We have described an efficient method for solving the NLPDE. The non-homogenous term is solved using power expansion method. It is seen from the displayed results that the method is unconditionally stable, and can be used to integrate the NLPDE with quite large propagation step.
We believe that the present method can be extended to another kind of NLPDE, such as water wave, heat and mass transfer problems, and other physical problems.

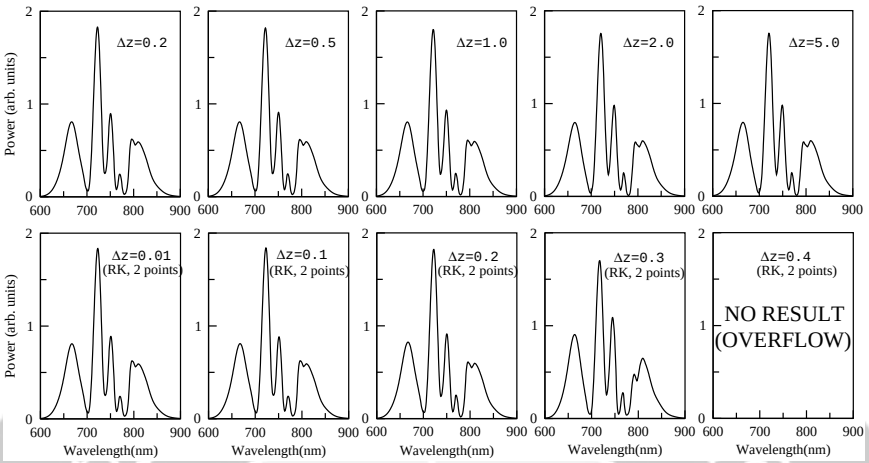


Figure 2: The spectral profiles obtained by integrating the nonlinear term using power expansion method (top) and Runge Kuta method with two points for various propagation step. Unit of Δz is mm.

References

[1] M. Nurhuda, A. Suda, K. Midorikawa, Phys. Rev. A. **60**, 2002; M. Nurhuda, A. Suda, and K. Midorikawa, Phys. Rev. A **66**, p. 041802 (1-4) (2002); M. Nurhuda, A. Suda, and K. Midorikawa, RIKEN Rev. **48**,p. 40-43 (2002); M. Nurhuda, A. Suda, M. Hatayama, K. Nagasaka, and K. Midorikawa, J. Opt. Soc. Am. B **20**, p. 2002-2011 (2003); M. Nurhuda, A. Suda, and K. Midorikawa, JNOPM, World Scientific, **13** p. 301-313 (2004); M. Nurhuda, and E. van Groesen, Phys. Rev. E, June 2005. M. Nurhuda, A. Suda, K. Midorikawa, and H. Budiono, JOSA B, August 2005.

[2] J. D. Jackson, *Classical Electrodynamics*, John Wiley & Sons (1998).

[3] W.H. Press, S.A. Teukolsky, W.T. Vetterling and B.P. Flannery, *Numerical Recipes in Fortran*, Cambridge, (1992).

M. Nurhuda: Physics Department, Brawijaya University, Malang 65144, Indonesia
E-mail: mnurhuda@brawijaya.ac.id

FINITE ELEMENT ANALYSIS OF PHOTONIC CRYSTAL FIBERS

H.P. Uranus, H.J.W.M. Hoekstra, E. van Groesen

University of Twente, The Netherlands

Abstract. A finite-element-based vectorial optical mode solver, furnished with Bayliss-Gunzburger-Turkel-like transparent boundary conditions, is used to rigorously analyze photonic crystal fibers (PCFs). Both the real and imaginary part of the modal indices can be computed in a relatively small computational domain. The leakage loss, the dispersion properties, the vectorial character, as well as the degeneracy of modes of the fibers can be studied through the finite element results. Results for PCFs with either circular or non-circular microstructured holes, solid- or air-core will be presented, including the air-core air-silica Bragg fiber. Using the mode solver, the single-modeness of a commercial endlessly single-mode PCF was also investigated.
Key-words: finite element analysis, photonic crystal fibers, transparent boundary conditions, leaky modes.

1 Introduction

Since the introduction of the photonic crystal fiber (PCF) [7], various waveguiding structures that utilize the arrangement of microstructured holes [16] or thin layers [4] have been realized. The large variety of possible hole shapes and arrangements demand the use of numerical methods that can handle arbitrary cross-sectional shapes to analyze this kind of structures. Besides, the existence of interfaces with high index-contrast between the solid host material and air holes calls for the use of the vectorial wave equation to accurately model the structure. Finite element method (FEM) is suitable for such analysis as it can handle complicated structure geometries and solve vectorial equations transparently. By incorporating proper boundary conditions, it also can model the leaky behavior of the realistic PCFs.

In this paper, we apply a vectorial optical mode solver based on Galerkin FEM [17], which is furnished with a 1st-order Bayliss-Gunzburger-Turkel-like (BGT-like) transparent boundary conditions (TBC) to rigorously model various kinds of PCFs [18]. Thanks to the boundary conditions, the structure can be analyzed in a relatively small computational domain for its complex-valued modal indices and field profiles. The structures being considered include those with either solid material or air as the core; circular or non-circular microstructured holes arranged around the core. Through the FEM results, we studied the leakage loss, dispersion properties, vectorial character, as well as the degeneracy of modes and single-modeness of particular kinds of PCFs.

2 Formulation of the method

The detail discussions on the formulation of the mode solver has been given elsewhere [17], but for convenience will be briefly reviewed here.

2.1 Finite element formulation

Using the H-field-based vectorial wave-equation, $\nabla \times \bar{\bar{\varepsilon}}_r^{-1} \nabla \times \vec{H} = k_0^2 \vec{H}$, for longitudinally-invariant structures composed of non-magnetic anisotropic materials with diagonal permittivity tensors and exp(jωt) time dependence of the field; it is possible to get a vectorial wave-equation expressed only in terms of the transverse components of the magnetic field as follows:

$$\begin{bmatrix} \partial_y \left[\frac{1}{n_{zz}^2} \left(\partial_x H_y - \partial_y H_x \right) \right] \\ -\partial_x \left[\frac{1}{n_{zz}^2} \left(\partial_x H_y - \partial_y H_x \right) \right] \end{bmatrix} - \begin{bmatrix} \frac{1}{n_{yy}^2} \partial_x \left(\partial_x H_x + \partial_y H_y \right) \\ \frac{1}{n_{xx}^2} \partial_y \left(\partial_x H_x + \partial_y H_y \right) \end{bmatrix} + k_0^2 n_{\text{eff}}^2 \begin{bmatrix} \frac{1}{n_{yy}^2} H_x \\ \frac{1}{n_{xx}^2} H_y \end{bmatrix} = k_0^2 \begin{bmatrix} H_x \\ H_y \end{bmatrix}.$$

Here, the *x* and *y* denote the transverse Cartesian coordinates associated with the structure cross-section, k_0 the vacuum wavenumber, n_{eff} the complex modal index, H_x and H_y the *x* and *y* components of the magnetic field $\vec{H}$, while n_{xx}^2, n_{yy}^2, and n_{zz}^2 the non-zero entries located at the diagonal of the relative permittivity tensor $\bar{\bar{\varepsilon}}_r$ associated with the *x*, *y*, and *z* components of the electric field, respectively. Using the Galerkin procedure and discretizing the computational domain into triangular elements lead to the following discretized weak formulation:

$$\begin{aligned} &\sum_{\substack{BoundaryElement \\ e}} \left\{ -\int_{\Gamma_e} \frac{1}{n_{zz}^2} w_y \left(\partial_x H_y - \partial_y H_x \right) dy - \int_{\Gamma_e} \frac{1}{n_{zz}^2} w_x \left(\partial_x H_y - \partial_y H_x \right) dx \right. \\ &\qquad \left. -\int_{\Gamma_e} \frac{1}{n_{yy}^2} w_x \left(\partial_x H_x + \partial_y H_y \right) dy + \int_{\Gamma_e} \frac{1}{n_{xx}^2} w_y \left(\partial_x H_x + \partial_y H_y \right) dx \right\} \\ &+ \sum_{\substack{InterfaceElement \\ e}} \left\{ -\int_{\Gamma_{\text{int},e}} \frac{1}{n_{yy}^2} w_x \left(\partial_x H_x + \partial_y H_y \right) dy + \int_{\Gamma_{\text{int},e}} \frac{1}{n_{xx}^2} w_y \left(\partial_x H_x + \partial_y H_y \right) dx \right\} \\ &\quad + \sum_{\substack{TriangularElement \\ e}} \iint_{\Omega_e} \left\{ \frac{1}{n_{zz}^2} \left(\partial_x w_y - \partial_y w_x \right) \left(\partial_x H_y - \partial_y H_x \right) + \left[\partial_x \left(\frac{1}{n_{yy}^2} w_x \right) + \partial_y \left(\frac{1}{n_{xx}^2} w_y \right) \right] \left(\partial_x H_x + \partial_y H_y \right) \right. \\ &\qquad \left. + k_0^2 n_{\text{eff}}^2 \left(\frac{1}{n_{yy}^2} w_x H_x + \frac{1}{n_{xx}^2} w_y H_y \right) - k_0^2 \left(w_x H_x + w_y H_y \right) \right\} dx\, dy = 0 \end{aligned} \tag{1}$$

with w_x and w_y denoting the weight functions, Ω_e the area in each triangular element, $\Gamma_{\text{int},e}$ the line element at the interface between different materials, and Γ_e the line element at the computational boundaries.

Approximating the fields using quadratic nodal-based basis functions will lead to a sparse generalized matrix eigenvalue equation, which can be solved using an eigenvalue solver to obtain the eigenvalues related to the modal indices (n_{eff}) and eigenvectors associated with the transverse components of the magnetic field $\left[H_x, H_y \right]^T$ of the corresponding modes.

2.2 Boundary conditions

The derivatives of the fields occurring in the boundary term in Eq. (1) will be handled through the 1st-order BGT-like [1] TBC to mimic the properties of the fields in the exterior domain properly. We use a vector radiation function

$$\vec{H}(r,\theta)\big|_{\Gamma} = \left[\begin{matrix} H_{\mathrm{x}} \\ H_{\mathrm{y}} \end{matrix}\right]\Bigg|_{\Gamma} = \sum_{p=0}^{\infty} \frac{1}{r^{p+1/2}} \left[\begin{matrix} H_{\mathrm{x,p}}(\theta)\exp\left(-jk_{\mathrm{r,x}}r\right) \\ H_{\mathrm{y,p}}(\theta)\exp\left(-jk_{\mathrm{r,y}}r\right) \end{matrix}\right] \tag{2}$$

along the computational boundary Γ, which leads to a 1st-order operator on the boundary fields as follows:

$$B_1\left(\left[\begin{matrix} H_{\mathrm{x}} \\ H_{\mathrm{y}} \end{matrix}\right]\right)\Bigg|_{\Gamma} = \left\{\left(\partial_{\mathrm{r}} + \frac{1}{2r}\right)\left[\begin{matrix} H_{\mathrm{x}} \\ H_{\mathrm{y}} \end{matrix}\right] + j\left[\begin{matrix} k_{\mathrm{r,x}}H_{\mathrm{x}} \\ k_{\mathrm{r,y}}H_{\mathrm{y}} \end{matrix}\right]\right\}\Bigg|_{\Gamma} = O\left(r^{-5/2}\right). \tag{3}$$

In Eqs. (2) and (3), r and θ are the polar coordinates of the cross-section whereby the center of the core of the waveguide has been taken as the origin, and $k_{\mathrm{r,x}}$ and $k_{\mathrm{r,y}}$ are the complex transverse wavenumbers associated with the x and y components of the field. Solving the wave-equation at the elementwise homogeneous anisotropic exterior domain leads to

$$k_{\mathrm{r,x}}\big|_{\Gamma} = k_0\sqrt{n_{\mathrm{xx}}^2 - n_{\mathrm{eff}}^2}\Big|_{\Gamma} \quad \text{and} \quad k_{\mathrm{r,y}}\big|_{\Gamma} = k_0\sqrt{n_{\mathrm{yy}}^2 - n_{\mathrm{eff}}^2}\Big|_{\Gamma}$$

with $\mathrm{Re}(k_{\mathrm{r}})>0$ associated with the outward leaking case (the leaky-mode being considered in this paper) and $\mathrm{Im}(k_{\mathrm{r}})<0$ associated with evanescently decaying case (the guided-mode case). By neglecting the angular dependence of the field at each line element,

$$\partial_{\mathrm{n}}H_{\mathrm{x}}\big|_{\Gamma} = -\hat{r}\bullet\hat{n}\left(jk_{\mathrm{r,x}} + \frac{1}{2r}\right)H_{\mathrm{x}}\Big|_{\Gamma} + O\left(r^{-5/2}\right) \tag{4}$$

$$\partial_{\mathrm{n}}H_{\mathrm{y}}\big|_{\Gamma} = -\hat{r}\bullet\hat{n}\left(jk_{\mathrm{r,y}} + \frac{1}{2r}\right)H_{\mathrm{y}}\Big|_{\Gamma} + O\left(r^{-5/2}\right) \tag{5}$$

Dirichlet to Neumann (DtN) map can be obtained and used for approximating the derivative operators within the boundary terms of Eq. (1), hence allows a proper truncation of the FEM mesh. In Eqs. (4) and (5), the caret (^) notation denotes the unit vector, while n denotes the normal direction. Note that, these boundary conditions induce non-linearity to the eigenvalue problem due to the appearance of n_{eff} (the eigenvalue itself) within the DtN. In this work, we have employed linearization by simple iteration technique to enable the use of linear eigenvalue solver and search only for eigenvalues (n_{eff}) of interest, i.e those related to low-loss, localized leaky modes within expected range [20].

3 Study of PCF properties through the FEM results

The computed complex-valued mode indices and field profiles associated with the eigenvalues and eigenvectors of the eigenvalue equation discussed in the previous

section can be used to study the properties of the PCFs. The dispersion parameter D of the PCFs can be obtained from the wavelength dependence of the real part of the mode indices as follows

$$D = -\frac{\lambda}{c}\frac{\partial^2}{\partial \lambda^2}\mathrm{Re}\left(n_{\mathrm{eff}}\right),$$

while the attenuation due to leakage loss can be deduced from the imaginary part as follows

$$\alpha = Loss/L = -20k_0\,\mathrm{Im}\left(n_{\mathrm{eff}}\right)\log(e).$$

As will be shown in the next section, the vectorial character of the modes will show up from the discrepancies of results for modes which are often regarded as the same mode in scalar analysis for weakly guiding structures [5].

Although a rigorous study on the degeneracy and classification of modes requires knowledge on group theory [11], we can also use the FEM results to recognize the degeneracy or non-degeneracy of modes of PCFs. Instead of using mesh refinement [9] or symmetry-preserving mesh generation [13], here we will simply use a visual inspection on the vector plots of the transverse modal field to recognize the degeneracy of a pair of modes. A symmetry operated modal field can be expressed as a linear combination of the orthogonal set of degenerate modes as follows

$$\mathbb{S}\vec{\Psi}_{\mathrm{k}} = \sum_{i=1}^{p} a_{\mathrm{i}}\vec{\Psi}_{\mathrm{i}} \tag{6}$$

where $\mathbb{S}$ is such a symmetry operation, $\left\{\vec{\Psi}_1..\vec{\Psi}_{\mathrm{p}}\right\}$ is the orthogonal degenerate modal fields set with $\vec{\Psi}_{\mathrm{k}} \in \left\{\vec{\Psi}_1..\vec{\Psi}_{\mathrm{p}}\right\}$. A non-degenerate mode requires p=1 meaning that its symmetry operated modal field will always be linearly dependent with the initial modal field. A failure to fulfill this requirement is already enough to indicate that a mode is degenerate. A pair of two-fold degenerate modes (which is the case for degenerate modes in optical waveguides [11] with rotational structural symmetry C_{m} or C_{mv}) requires p=2, meaning that we can reconstruct the modal profile of a mode from 2 symmetry-operated modal fields taken from its degenerate pair, provided that these symmetry-operated modal fields are not linearly dependent.

4 Examples

Here, we will demonstrate the application of the FEM leaky mode solver to study the properties of various kinds of PCFs, including those of solid and hollow core, circular and non-circular microstructured holes.

4.1 Solid-core PCF with circular microstructured holes

First, we studied the most widely used kind of PCFs, i.e. those with cladding made up of circular holes arranged in a triangular lattice with core residing in the region at the center formed by missing of hole(s) [18]. We found that adding more rings of holes will be influential only to the leakage loss, while giving practically similar dispersion properties. Hence, for the sake of efficiency, we took structure with only

6 circular holes in the cladding for most of the computations and took more rings of holes to show their effect on the leakage loss, when necessary. Fig. 1 shows the fiber cross sections together with the FEM mesh and size of computational window. Note that, by the help of the boundary conditions, the computation can be carried out in a relatively small computational domain. Also note that by taking advantages of the structure mirror symmetry, we only need a quarter of the structure as the computational domain [20]. The diameter of the holes is d=5μm with pitch length of Λ=6.75μm. Pure SiO_2 is considered as the host solid material with its refractive index n_{bg} taken from the Sellmeier's equation [10], while the refractive index of the air holes n_{hole} is 1. Fig. 2 shows the real part of n_{eff} and the dispersion parameter as function of wavelength for the structure of 6 circular holes. Note that, by using the Sellmeier's equation, the material dispersion effect has been rigorously taken into account in the plots here and in all other wavelength dependent plots in this paper. In this paper we have used similar hybrid mode notation as in ordinary fiber with additional superscript a and b to denote the results obtained using perfect magnetic conductor (PMC) and perfect electric conductor (PEC), respectively, as the symmetry boundary conditions at the horizontal symmetry plane.

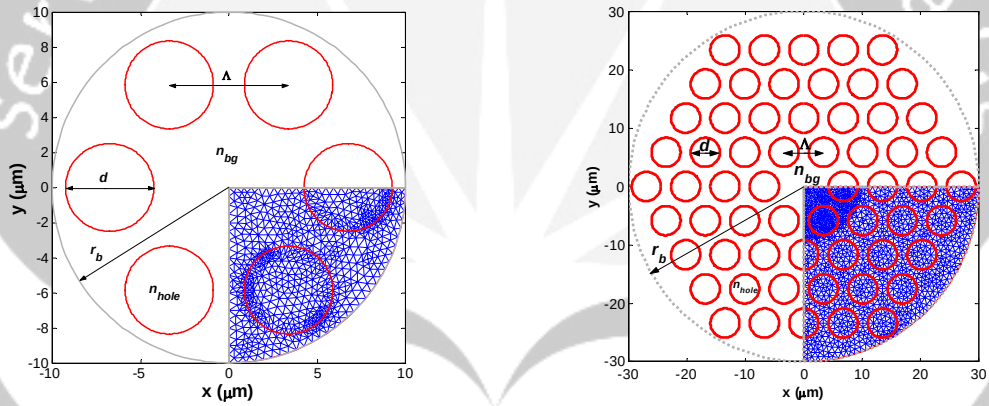


Fig. 1. The PCF with (a) 6 circular holes and (b) 60 circular holes forming their cladding, the FEM mesh and their computational window.

The curves in Fig. 2 show three groups of modes, which correspond to the first-three *LP*-like modes [5]. By using the FEM mode solver, it is possible to distinguish modes within the same group. The differences between the curves of modes associated with the same group, i.e. the vectorial properties of the modes, are more pronounced for longer wavelength, where the dimension of the structure becomes more comparable to the wavelength. The divergence between curves for HE_{31}^{a}- and HE_{31}^{b}-like modes indicates their non-degeneracy, a property which can be intuitively understood by visual inspection on the vector plot of their transverse modal field profile as shown in Fig. 3, whereas the symmetry-operated (e.g. rotated with $2\pi m/6$ rotation angle with m integer) modal field is always linearly dependent with the initial modal field. Note that these modes are degenerate in ordinary fiber,

a property which can also be understood using the same intuitive procedure. Fig. 2(b) also shows that HE_{11}-like mode has zero dispersion wavelength at shorter wavelength than the ordinary fiber. Note that this zero-dispersion wavelength can be engineered by playing with the hole sizes [8], making this kind of fibers attractive for applications like dispersion compensation [3], supercontinuum light generation [14], ultra-flat and ultra-low dispersion [15], etc.

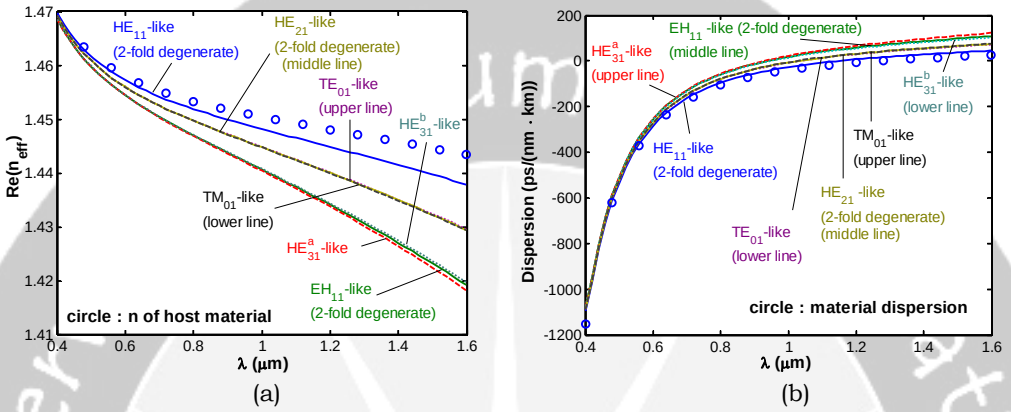


Fig. 2. (a). The real part of n_{eff} and (b) the dispersion parameter of the structure with 6 circular holes forming its cladding.

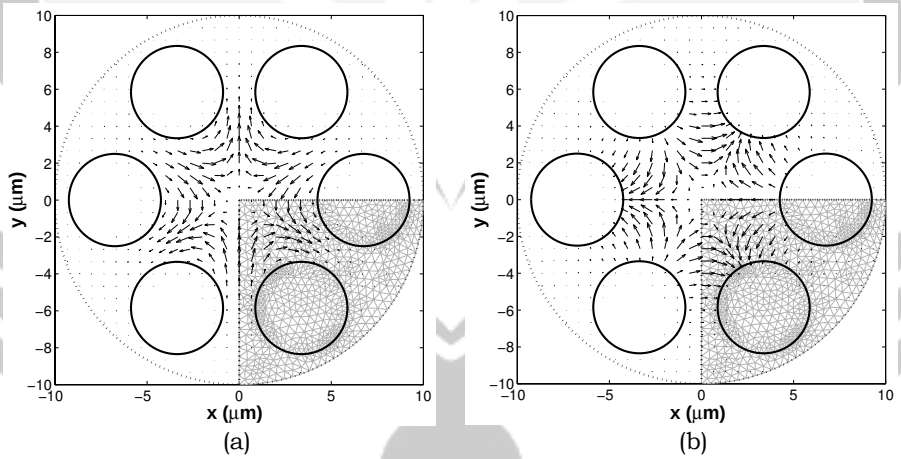


Fig. 3. Vector plot of (a) HE_{31}^{a} - and (b) HE_{31}^{b} -like modes.

Fig. 4(a) shows the leakage loss as function of wavelength. The vectorial behavior of the modes is even more pronounced through the obvious divergence of curves within the same *LP*-like modes. As wavelength increases, the modes become less quasi-confined; consequently, the leakage loss increases. Fig. 4(b) shows the effect

of adding more rings of holes around the central core. In this case; 2-ring, 3-ring, and 4-ring structures correspond to 18, 36, and 60 holes (see Fig. 1b) in the cladding arranged in triangular lattice setting. The figure shows that adding rings of holes will reduce the leakage loss exponentially.

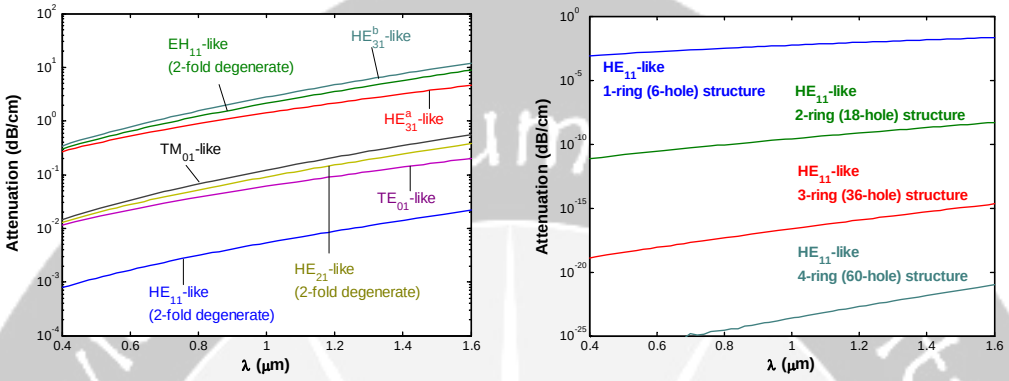


Fig. 4. The leakage loss of the structure with circular holes in the cladding. (a). Leakage loss of the first-ten modes in the 1-ring (6-hole) structure. (b). The effect of adding more rings of holes in the cladding.

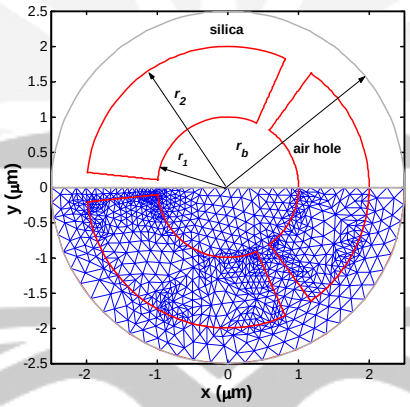


Fig. 5. The cross-section of the PCF with 3 annular-sector-shaped holes in its cladding, the mesh definition, and the computational window.

4.2 Solid-core PCF with annular-sector-shaped holes

Next, we consider a PCF with cladding consists of three annular-sector-shaped holes as shown in Fig. 5 together with its mesh definition and the size of the computational window. The annular-sector-shaped holes has an inner radius of

r_1=1μm and outer radius r_2=2μm. Again here, we took the pure SiO_2 as the host material. Fig. 6 shows the real n_{eff} and related dispersion parameter of the structure as function of wavelength. Fig. 7 shows the imaginary part of the n_{eff} and its related leakage loss. Since this structure has much smaller core size and larger (local) air filling fraction than the previous example, the effect of the air core is stronger, leading to a shorter zero-dispersion wavelength (see arguments given in [8]). Also, the vectorial character of the modes is more pronounced as indicated by more divergent curves of TE_{01}-, TM_{01}-, and HE_{21}-like modes (which are all associated to LP_{11}-like mode in scalar analysis), both in their real and imaginary part of the n_{eff}. Fig. 6(a) shows that at around 1.483μm, the real part of HE_{21}- and TM_{01}-like modes cross over, leading to a rather dissimilar dispersion properties in Fig. 6(b).

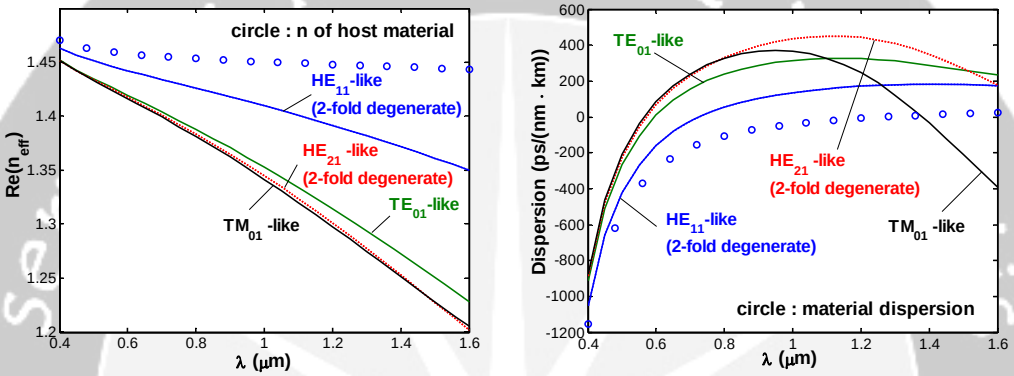


Fig. 6. (a). The real part of n_{eff} and (b) the dispersion parameter of the modes of the PCF with 3 annular-sector-shaped holes.

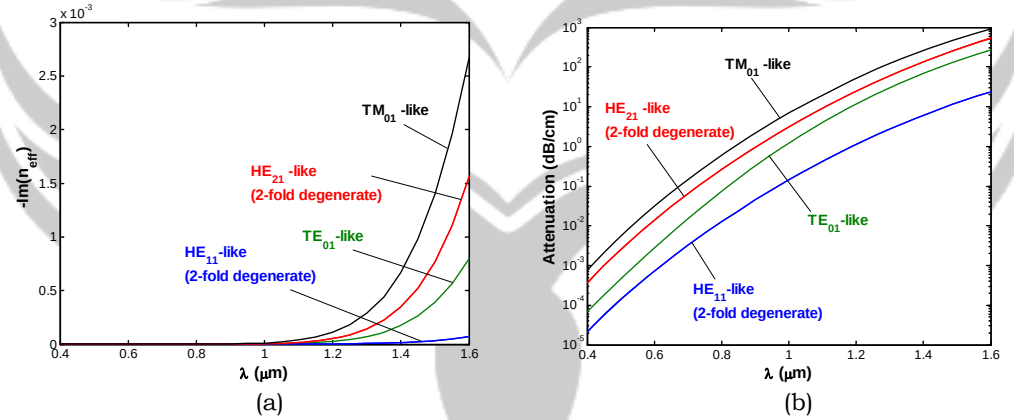


Fig. 7. (a). The imaginary part of the effective indices and (b). the leakage loss of the modes of the PCF with 3 annular-sector-shaped holes.

4.3 Hollow-core PCF with rounded-annular-sector-shaped holes

For hollow-core PCF, we take the one with rounded-annular-sector-shaped holes in the cladding, i.e. the air-silica Bragg fiber as proposed by Vienne *et al.* [21]. The structure is shown in Fig. 8, where r_{core}=10μm, $t_{annular}$=2.3μm, r_{corner}=$t_{annular}$/4, t_{ring}=0.2μm, and t_{bridge}=45nm. The structure has 24, 34, and 44 holes at the first, second, and third ring of holes. We used computational domain with r_b=19μm and wavelength 1.06μm. The refractive index of the host silica material is taken from its Sellmeier's equation [10].

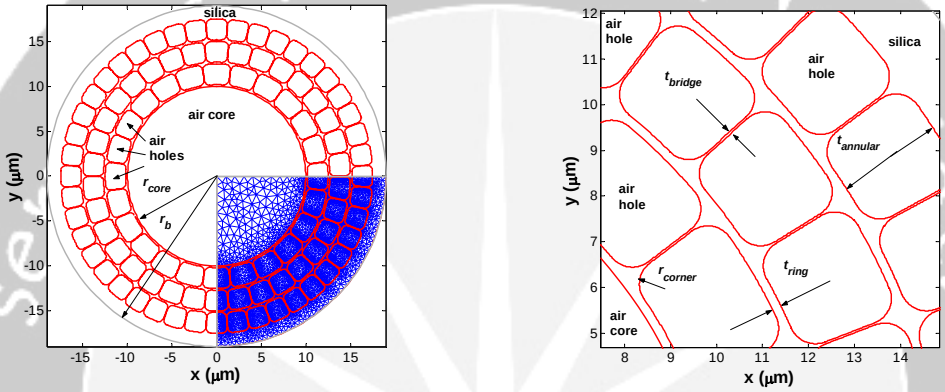


Fig. 8. The model of the air-silica Bragg fiber with 3 rings of annular-sector-shaped holes in the cladding. Also shown are the mesh definition and the computational window size.

Fig. 9 shows the vector plot of the transverse component of the magnetic field of modes of the PCF. In this type of waveguide, the leakage loss of TE_{01}-like mode (computed to be 0.015 dB/cm) was found to be lower than that of the HE_{11}-like mode (computed to be 0.44 dB/cm). This is a typical property of a Bragg fiber (as the micro-structured cladding can be regarded as alternating air and silica Bragg 'layers'). This lower loss for *TE*-like mode comes from the fact that the Fresnel reflection coefficient for *TE* is higher than *TM* polarization. As modes other than *TE*-like modes have some component with *TM*-like polarization, they will exhibit higher leakage loss due to the lower reflection coefficient of the cladding. Fig. 10 shows the longitudinal component of the time averaged Poynting vector of HE_{11}- and TE_{01}-like modes. Small spots at the solid material in the cladding for HE_{11}-like mode indicate the onset of the anti-crossing of this mode with a cladding resonance mode. For hollow-core structure, since the n_{eff} of interest is below 1, the transverse wavenumber $k_t = k_0\sqrt{n_{solid}^2 - n_{eff}^2}$ is rather large, which enables resonances to take place even in thin solid material of the cladding. This explains why cladding resonance modes are easily observable near to our modes of interest in such hollow-core PCFs.

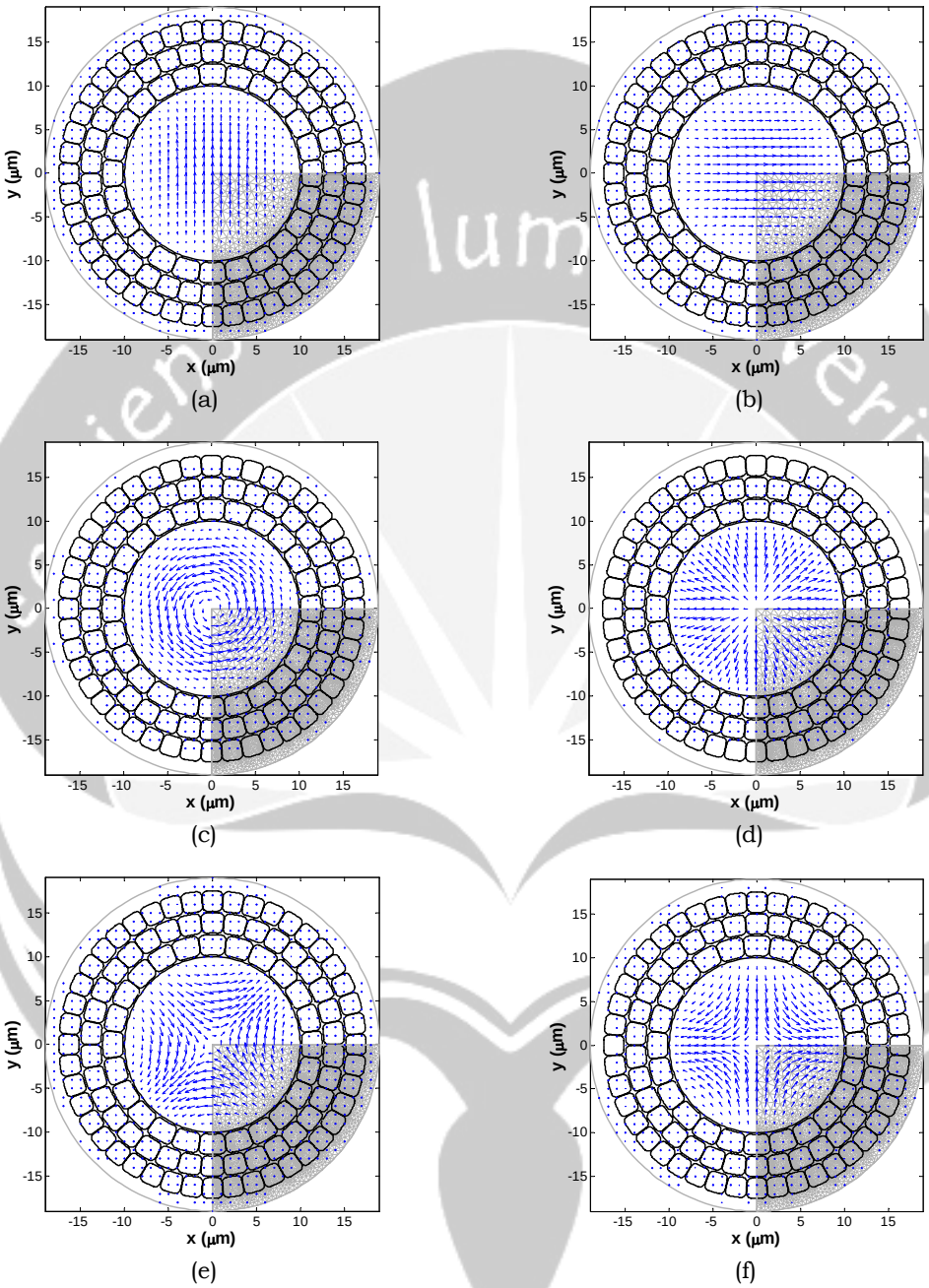


Fig. 9. The transverse component of the magnetic field of (a). HE_{11}^{a} -, (b). HE_{11}^{b} -, (c). TM_{01} -, (d). TE_{01} -, (e). HE_{21}^{a} -, and (f). HE_{21}^{b} -like modes of the air-silica Bragg fiber.

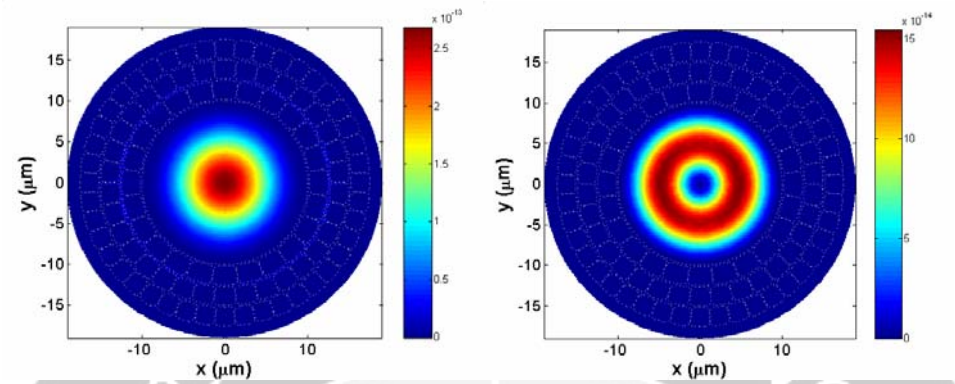


Fig. 10. The longitudinal component of the time averaged Poynting vector of (a). HE_{11}^{a} - and (b). TE_{01} -like modes of the air-silica Bragg fiber.

4.4 Modal degeneracy in PCF

Here, we will demonstrate the use of visual inspection of the vector plot of the transverse modal field of the modes of PCF to recognize their degeneracy or non-degeneracy. As an example, we took the HE_{21}-like modes of the PCF with 3 annular-sector-shaped holes in the cladding as discussed in Section 4.2. Fig. 11 shows the modal profile of the two HE_{21}-like modes. As shown by Eq. (6) for p=1, a non-degenerate mode requires linear dependency of a symmetry-operated modal field with its initial modal field. Visual inspection on Fig. 11(a) or (b) shows that rotating the modal field by $2\pi/3$ (which is a symmetry operation due to the C_{3v} structure symmetry) will not result in a linear dependent modal field with the initial modal field. This fact is already enough as a proof of the degeneracy of such modes.

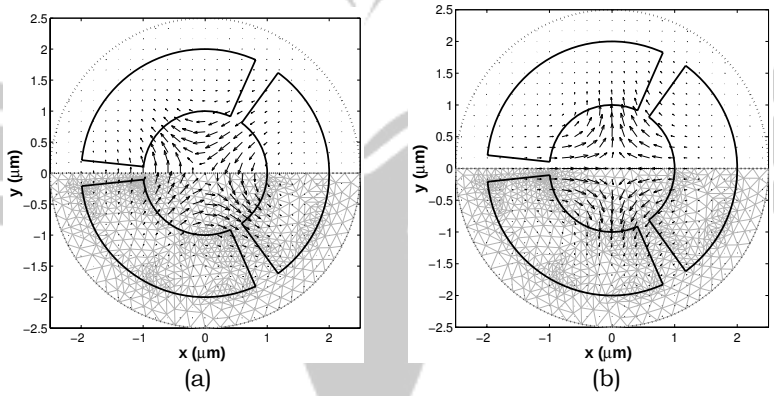


Fig. 11. Vector plot of the transverse component of magnetic field of (a). HE_{21}^{a} - and (b). HE_{21}^{b} -like modes of the PCF with 3 annular-sector-shaped holes.

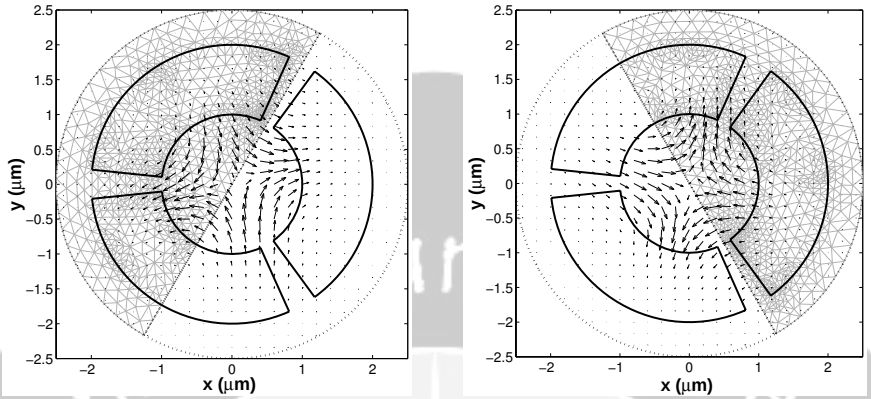


Fig. 12. Fields obtained by rotating the transverse modal field of HE_{21}^{a} by (a). $2\pi/3$ and (b). $4\pi/3$ radiant.

To further show that the HE_{21}^{a}-like mode (Fig. 11(a)) is the degenerate pair of the HE_{21}^{b}-like mode (Fig. 11(b)), we will show that it is possible to construct the modal field of Fig. 11(b) from two symmetry-operated modal fields taken from Fig. 11(a), which is a consequence of Eq. (6) for p=2 in the case that these two symmetry-operated modal fields are not linear dependent each other. Fig. 12(a) and (b) show the fields obtained by rotating Fig. 11(a) by $2\pi/3$ and $4\pi/3$ radiant, respectively. Note that these rotation operations are symmetry operations as they leave the structure unchanged. Fig. 13 shows the field obtained by multiplying the field of Fig. 12(a) by -1 and adding the result to the field of Fig. 12(b). Visual inspection on Fig. 13 and Fig. 11(b) shows that they are linear dependent. This fact proves that HE_{21}^{a}- and HE_{21}^{b}-like modes are degenerate pair.

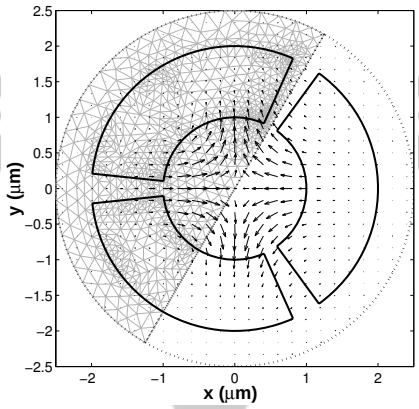


Fig. 13. Fields obtained by multiplying field of Fig. 12(a) by -1 and adding the result to field of Fig. 12(b).

4.5 Single-modeness of an endlessly single-mode PCF

Finally, we will consider the single-modeness of a PCF which is specially designed to have single-mode properties over a wide wavelength range. This type of PCF is known as the endlessly-single-mode (ESM) PCF [2]. An intuitive explanation of this property can be given by the modified-total-internal-reflection model [16] of the PCF. At shorter wavelength, the effective refractive index of the cladding becomes closer to the refractive index of the silica host material. This dispersive property will decrease the effective index contrast between the core and cladding, which in turn compensates the effect of the decrease of the wavelength and keep the structure to be single-moded over a wide wavelength range.

In this section, we will use the FEM leaky mode solver to study such ESM-PCF rigorously using the leakage loss as a measure of its single-modeness [19]. Here, we picked up a commercial ESM-PCF, which is ESM-12-01 fiber made by BlazePhotonics [6]. The structure cross-section of the fiber is shown by Fig. 14, with d=3.68μm and Λ=8μm. We assume pure SiO_2 as the host material with its refractive index taken from Sellmeier's equation [10].

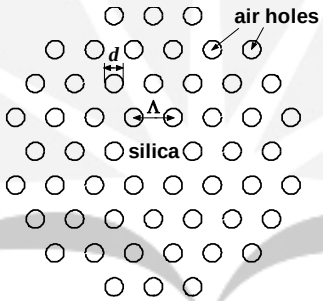


Fig. 14. The cross-section of the ESM-12-01 PCF.

Fig. 15 shows the real and imaginary part of the n_{eff} of the first-few modes of the PCF with attenuation constant of smaller than 10 dB/cm. The figure shows that the curve for HE_{11}-like mode is well separated from other modes, indicating the single-mode behavior, which is stronger for longer wavelength. The figure also shows the existence of cladding and core-cladding resonance modes, similar to the one usually found in hollow-core PCF. As the ESM-PCF is designed using small d/Λ to obtain low effective index contrast, it has large solid medium between holes, hence small transverse wavenumber is already enough for resonance within this cladding region.

Fig. 16 shows the dispersion parameter and the attenuation constant of the modes. The curves in Fig 16(a) almost coincide, indicating the dominance of the material dispersion of the bulk silica as the air holes effect is weak due to the small d/Λ ratio. Fig. 16(b) shows that due to the low effective index contrast, the vectorial

property of the modes is not so pronounced as indicated by the very similar loss profile of the TE_{01}-, TM_{01}-, and HE_{21}-like modes, which are often regarded as LP_{11}-like scalar mode. The figure also shows that the HE_{11}-like mode exhibits the lowest loss, hence can be regarded as the most dominant mode.

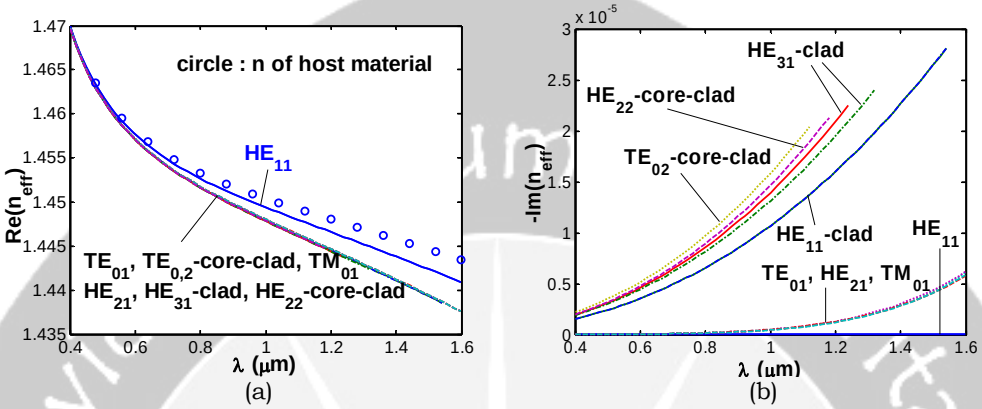


Fig. 15. The (a). real and (b). imaginary part of n_{eff} of the first-few modes of the ESM-12-01 PCF. In (b) the curve for HE_{11}-like mode almost coincides with the horizontal axis.

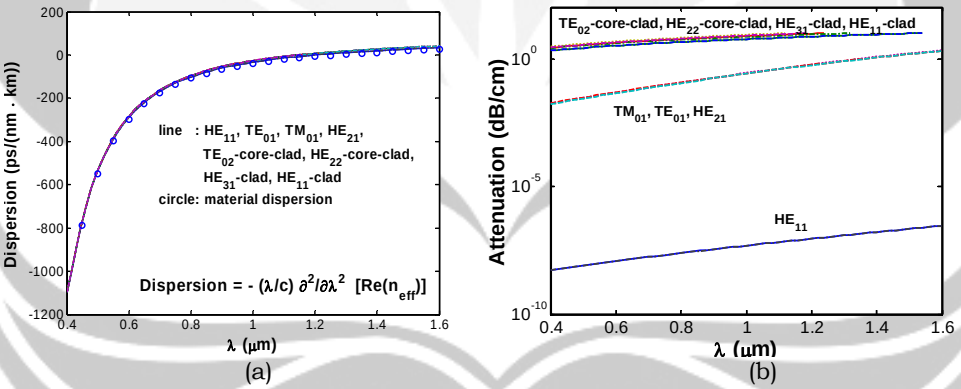


Fig. 16. The (a). dispersion parameter and (b). attenuation constant of the first-few modes of the ESM-12-01 PCF. The curves in (a) almost coincide, indicating the weak waveguide dispersion.

The single-modeness of the ESM-PCF can be rigorously measured through the ability of the PCF to discriminate the dominant mode from the nearest higher-order mode in the sense of leakage loss. For this purpose, we define a quantity called multi-mode rejection ratio (*MMRR*) as follows

$$MMRR = 10\log\left(P_0/P_1\right) = \left(\alpha_1 - \alpha_0\right)L$$

with P_0 and P_1 denoting the power of the dominant fundamental and the nearest higher-order mode (which are assumed to be equally excited), while α_0, α_1, and L denoting the attenuation constant of the dominant mode, the nearest higher-order mode, and the length of the fiber, respectively. By putting 20 dB as the minimum *MMRR* (meaning that the power of the fundamental mode is at least 100 times larger than the nearest higher order mode) as the single-modeness criterion, we get the minimum length of the fiber for single mode operation are 11.97m, 3.86m, and 0.78m for wavelength of 0.4μm, 0.6328μm, and 1μm, respectively, as shown in Fig. 17. Allowing the power of the fundamental mode to be only at least 10 times the nearest higher order mode, the minimum length is just half of those of the previous criterion, a length which is still considerably long for applications like gas/liquid sensing when operated at short wavelength region. Hence, although this fiber geometry does not fulfill the ESM criterion of Mortensen *et al.* [12], it still can be regarded as an ESM-PCF for long fiber-length applications. While, for short fiber-length applications, especially for short wavelength region, the endlessly single-modeness should be considered with some precaution. Although the attenuation of the fundamental mode is 6 orders lower (in dB scale) than the nearest higher order modes, the low attenuation of these higher order modes can make them to be quite significant for these particular applications. This fact suggests the requirement of ESM-PCF which is specially designed for short fiber-length applications. As these applications can tolerate higher attenuation, the use of smaller d/Λ and less rings of air holes can be incorporated. Otherwise, some manner to strip off higher order modes might be required.

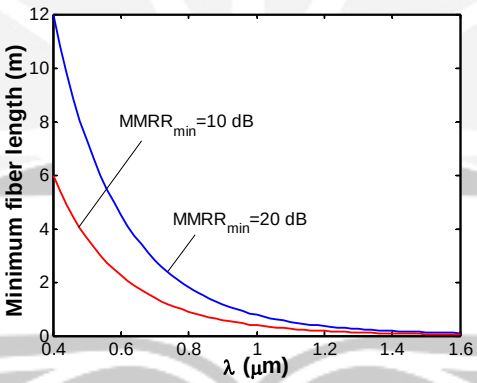


Fig. 17. Minimum fiber length for single-mode operation by the loss discrimination criterion for minimum *MMRR* of 10 and 20 dB

5 Conclusions

We demonstrate the use of FEM leaky mode solver to rigorously analyze various kinds of PCFs, ranging from those of solid core to hollow core, circular to non-circular microstructured holes in the cladding. The dispersion properties, leakage loss, vectorial character, mode degeneracy, and single-modeness of the structures can be well studied through the FEM results.

Acknowledgment

This work is supported by STW Technology Foundation through project TWI.4813 and TOE.6596.

References

[1] Bayliss, A., M. Gunzburger, and E. Turkel (1982), Boundary conditions for the numerical solution of elliptic equations in exterior regions, *SIAM J. Appl. Math.*, **42**, 430-451.

[2] Birks, T.A, J.C. Knight, and P.S.J. Russell (1997), Endlessly single-mode photonic crystal fiber, *Opt. Lett.*, **22**, 961-963.

[3] Birks, T.A. *et al.* (1999), Dispersion compensation using single-material fibers, *Photon. Technol. Lett.*, **11**, 674-676.

[4] Fink, Y. *et al.* (1999), Guiding optical light in air using an all-dielectric structure, *J. Lightwave Technol.*, **17**, 2039-2041.

[5] Gloge, D. (1971), Weakly guiding fibers, *Applied Opt.*, **10**, 2252-2258.

[6] http://www.blazephotonics.com/pdf/ESM%20-%2012%20-%2001.pdf

[7] Knight, J.C., T.A. Birks, P.S.J. Russell, and D.M. Atkin (1996), All-silica single-mode optical fiber with photonic crystal cladding, *Optics Lett.*, **21**, 1547-1549.

[8] Knight, J.C. *et al.* (2000), Anomalous dispersion in photonic crystal fiber, *Photon. Technol. Lett.*, **12**, 807-809.

[9] Koshiba, M., and K. Saitoh (2001), Numerical verification of degeneracy in hexagonal photonic crystal fibers, *Photonics Technol. Lett.*, **13**, 1313-1315.

[10] Malitson, I.H. (1965), Interspecimen comparison of the refractive index of fused silica, *J. Opt. Soc. Am.*, **55**, 1205-1209.

[11] McIsaac, P.R. (1975), Symmetry-induced modal characteristics of uniform waveguides – I: summary of results, *Trans. Microwave Theory and Tech.*, **MTT-23**, 421-429.

[12] Mortensen, N.A. *et al.* (2003), Modal cutoff and the V parameter in photonic crystal fibers, *Opt. Lett.*, **28**, 1879-1881.

[13] Peyrilloux, A. *et al.* (2003), Theoretical and experimental study of the birefringence of a photonic crystal fiber, *J. Lightwave Technol.*, **21**, 536-539.

[14] Ranka, J.K., R.S. Windeler, and A.J. Stentz (2000), Visible continuum generation in air silica microstructure optical fibers with anomalous dispersion at 800nm, *Opt. Lett.*, **25**, 25-27.

[15] Reeves, W.H. *et al.* (2002), Demonstration of ultra-flattened dispersion in photonic crystal fibers, *Opt. Express*, **10**, 609-613.

[16] Russell, P. (2003), Photonic Crystal Fibers, *Science*, **299**, 358-362.

[17] Uranus, H.P., H.J.W.M. Hoekstra, and E. van Groesen (2004), Galerkin finite element scheme with Bayliss-Gunzburger-Turkel-like boundary conditions for vectorial optical mode solver, *J. Nonlinear Opt. Phys. and Materials*, **13**, 175-194.

[18] Uranus, H.P. and H.J.W.M. Hoekstra (2004), Modelling of microstructured waveguides using a finite-element-based vectorial mode solver with transparent boundary conditions, *Opt. Express*, **12**, 2795-2809.

[19] Uranus, H.P., H.J.W.M. Hoekstra, and E. van Groesen (2004), Modes of an endlessly single-mode photonic crystal fiber: a finite element investigation, *Proc. 9th Annual Symposium IEEE/LEOS Benelux*, Ghent, Belgium, 311-314.

[20] Uranus, H.P. (2005), *Guiding light by and beyond the total internal reflection mechanism*, Ph.D. thesis, University of Twente, Enschede.

[21] Vienne, G. *et al.* (2004), First demonstration of air-silica Bragg fiber, Post deadline paper, *Optical Fiber Conference*, Los Angeles.

H.P. URANUS: IOMS group, MESA+ Research Institute, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
E-mail: h.p.uranus@ewi.utwente.nl

H.J.W.M. HOEKSTRA: IOMS group, MESA+ Research Institute, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
E-mail: h.j.w.m.hoekstra@ewi.utwente.nl

E. VAN GROESEN: AAMP group, MESA+ Research Institute, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
E-mail: groesen@math.utwente.nl

APPLICATION OF THE MULTIGRID METHOD IN IMAGE PROCESSING: OVERVIEW AND IMPROVED MULTIGRID PHASE UNWRAPPING METHOD

Andriyan Bayu Suksmono
Institut Teknologi Bandung, Indonesia

Abstract. This paper is an overview on the multigrid method for image processing. Multigrid is a powerful numerical method for solving elliptic differential equations with $O(N)$ computational complexity. In image processing, it can be employed to solve phase unwrapping (PU) problems, which is essentially seeking a solution of $\nabla^2\phi=\rho$, where ϕ indicates *unknown* unwrapped phase and ρ is modulo 2π Laplacian estimate of "phase source" obtained from a given wrapped phase. PU is an *ill-posed* inverse problem, where a noisy wrapped phase (interferogram) is to be unwrapped to obtain an "absolute" phase values. Engineering applications of PU is abundant, ranging from optics, biomedical imaging (MRI), to coherent radar imaging (InSAR for DEM construction). One of the uniqueness in applying the multigrid method to PU problems that distinguish it from other numerical problems is the existence of phase noise. In this paper, application of the multigrid method to PU is overviewed. A new PU strategy by embedding a gradient re-estimator in the multigrid cycle is also introduced.

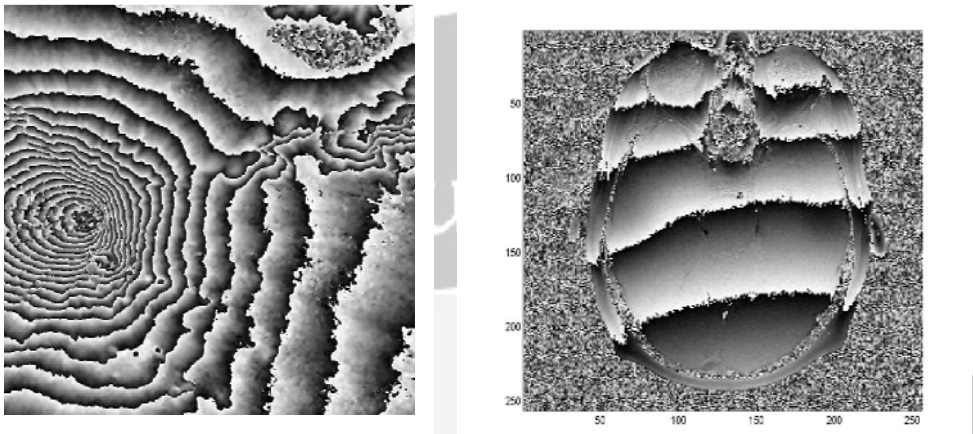
Key-words: elliptic differential equation, boundary value problem, phase unwrapping, image processing, multigrid method, multi resolution analysis, InSAR, MRI, complex-valued image

1 Introduction

Mathematically, a (digital) image is a two dimensional array of numbers that represents a meaningful information of real-world objects captured by an imaging device. The information content may be perceived by direct observation of the image or be extracted after some processing stages in an automated process. Image processing assumes a particular underlying model of the image and a suitable manipulating algorithm. The image processing algorithm includes; among others, enhancement, restoration, transformation and feature extraction. Recently, various kinds of innovative image processing algorithms based on differential equation models are explored. The most popular ones are active contour model and energy-minimization-based algorithms.

Solving a partial differential equation (PDE) is one of central themes in numerical computing. Based on their characteristics or curves of the information propagation, it can be distinguished into three classes: hyperbolic, parabolic and elliptic [2]. The linear one dimensional wave propagation's PDE belongs to the hyperbolic; the diffusion equation is categorized as a parabolic one, while the Poisson equation exemplified the elliptic PDE. Both the hyperbolic and parabolic PDEs correspond to

the initial value problems while the elliptic PDE is related to the boundary value problem.



(a) **(b)**

Fig.1 Phase image of (a) InSAR and (b) MRI

Most of the real-life image is an array of real numbers (or integers in a digital image). But, some imaging devices—such as InSAR (Inteferometric Synthetic Aperture Radar) and MRI (Magnetic Resonance Imaging) provide a complex-valued image. In this case, there is a pair of image representing the real part and the imaginary part, or equivalently, the magnitude and the phase image. Usually, the amplitude image looks just like an ordinary image captured by a camera. But the phase image is different, both its meaning and processing approach. In the InSAR, such as in Fig.1 (a), the phase image corresponds to terrain elevation, while in the MRI, such as in Fig.1 (b), it represents phase shifting corresponds to water/fat in tissues or temperature in thermal imaging or fluid flow (blood) in angiography. Since the phase values are obtained through *arctan* function, they are wrapped to a domain of $[-\pi, \pi)$. The physical meaning of the information is known after a particular phase processing called phase unwrapping (PU) has been conducted.

The PU is a process to obtain an "absolute" phase from a wrapped one. There are two main PU methods: the local (path-following) method and the global (least-squares) method. The local method solves the PU by connecting SP's based on a particular rule and then unwraps the phase along a line that avoid the connecting lines. The global method formulates the PU problem in a PDE expression, and then solves it by numerical computation. They both have their drawbacks and advantages. Generally speaking, the local method provides an accurate solution, but some area with dense SP is not solvable. Additionally, it is difficult to place the paths, even after reduction of SP by phase filtering. On the other hand, the global method gives a solution for the entire image, but with a less precision. This paper discusses the global method, particularly the multigrid PU (MGPU), as a PDE-based image processing method.

The rest of the paper is organized as follows. Section 2 explains the formulation of global PU method and some related algorithms. The multigrid method and its application to PU are reviewed in Section 3. Section 4 explains an improvement of the multigrid method, and Section 5 concludes the paper.

2 Global Phase Unwrapping: PDE Formulation and Numerical Computations

The global PU method is based on discretization of (finite-differencing) the elliptic differential (Poisson) equation

$$\nabla^2\phi = \frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} = \rho(x,y) \tag{1}$$

on a phase field ϕ with source ρ derived from a known wrapped phase ϕ^W. Let $\phi^W(n_1,n_2)$ be an observed (noisy) wrapped phase image, $\phi^U(n_1,n_2)$ be the desired unwrapped one and n_i indicates (discrete) spatial coordinates. The discrete gradients of the wrapped phase are defined as

Row-directional gradient: $\phi^W(n_1+1,n_2) - \phi^W(n_1,n_2) \equiv d^W_{n_1}(n_1,n_2)$ (2.a)

Column-directional gradient: $\phi^W(n_1,n_2+1) - \phi^W(n_1,n_2) \equiv d^W_{n_2}(n_1,n_2)$ (2.b)

and the gradients of the estimated unwrapped phase are

Row-directional gradient: $\phi^U(n_1+1,n_2) - \phi^U(n_1,n_2) \equiv d^U_{n_1}(n_1,n_2)$ (2.c)

Column-directional gradient: $\phi^U(n_1,n_2+1) - \phi^U(n_1,n_2) \equiv d^U_{n_2}(n_1,n_2)$ (2.d)

where operations applied to a wrapped phase image are performed in modulo 2π. The global PU method is based on the least square estimate of the phase gradient, which is achieved by minimizing the following error value E with respect to $\phi^U(n_1,n_2)$:

$$E = \sum_{n_1,n_2}\left[\left\{d^U_{n_1}(n_1,n_2) - d^W_{n_1}(n_1,n_2)\right\}^2 + \left\{d^U_{n_2}(n_1,n_2) - d^W_{n_2}(n_1,n_2)\right\}^2\right] \tag{3}$$

By the calculus of variation, this step leads to the discrete Poisson equation

$$\begin{aligned}\left\{\phi^U(n_1+1,n_2) - 2\phi^U(n_1,n_2) + \phi^U(n_1-1,n_2)\right\} + \\ \left\{\phi^U(n_1,n_2+1) - 2\phi^U(n_1,n_2) + \phi^U(n_1,n_2-1)\right\} = \rho^W(n_1,n_2)\end{aligned} \tag{4.a}$$

where

$$\rho^W(n_1,n_2) = d^W_{n_1}(n_1+1,n_2) - d^W_{n_1}(n_1,n_2) + d^W_{n_2}(n_1,n_2+1) - d^W_{n_2}(n_1,n_2) \tag{4.b}$$

As shown by (4.b), the Laplacian $\rho^W(n_1, n_2)$ is computed from the wrapped phase $\phi^W(n_1, n_2)$. Now, the problem is how to calculate $\phi^U(n_1, n_2)$ in equation (4) by assuming a periodic boundary condition on the grid. Popular solutions include, among others, matrix inversion, Fourier transform, relaxation, and the multigrid method.

Matrix Inversion Method. Equation (4) can be expressed in a matrix form as $\mathbf{A}.\boldsymbol{\phi} = \boldsymbol{\rho}$, where $\mathbf{A}$ is typically a tridiagonal (with fringes) matrix, $\boldsymbol{\rho}$ is vector of ρ^W and $\boldsymbol{\phi}$ is desired solution. From basic matrix algebra, $\boldsymbol{\phi}$ is obtained directly by multiplying the inverse of $\mathbf{A}$, i.e. $\mathbf{A}^{-1}$, with $\boldsymbol{\rho}$. In most cases, the matrix inversion method is non-practical because of the difficulty in inversing $\mathbf{A}$ due to its large size.

Fourier Transform Method. In the Fourier-based PU, a periodic boundary condition of the Poisson equation is assumed and it is obtained by performing a mirror reflection. Then, the phase image is extended to a periodic field as

$$\phi^U(n_1,n_2) \rightarrow \tilde{\phi}^U(n_1,n_2), \quad \text{and} \quad \phi^W(n_1,n_2) \rightarrow \tilde{\phi}^W(n_1,n_2) \tag{5}$$

where

$$\tilde{\phi}(n_1,n_2) = \begin{cases} \phi(n_1,n_2), & 0 \le n_1 \le N_1, \quad 0 \le n_2 \le N_2 \\ \phi(2N_1 - n_1, n_2), & N_1 + 1 \le n_1 \le 2N_2, \quad 0 \le n_2 \le N_2 \\ \phi(n_1, 2N - n_2), & 0 \le n_1 \le N_1, \quad N_1 + 1 \le n_2 \le 2N_2 \\ \phi(2N_1 - n_1, 2N_2 - n_2), & N_1 + 1 \le n_1 \le 2N_2, \quad N_1 + 1 \le n_2 \le 2N_2 \end{cases} \tag{6}$$

Using these conditions, the problem can be solved by the FFT by computing [2]:

$$\Phi(t_1,t_2) = \frac{P(t_1,t_2)}{\left(2\cos\frac{\pi\ t_1}{N_1}\right) + \left(2\cos\frac{\pi\ t_2}{N_2}\right) - 4} \tag{7}$$

where $\Phi(t_1,t_2)$ and $P(t_1,t_2)$ are the FFT's of $\tilde{\phi}^U(n_1,n_2)$ and $\tilde{\rho}(n_1,n_2)$ respectively, while $\tilde{\rho}(n_1,n_2)$ denotes the Laplacian of $\tilde{\phi}^W(n_1,n_2)$. The complexity of this algorithm is O(Nlog(N)).

Relaxation Method. The relaxation method involves splitting the sparse matrix arises from (4), and then iterating until a solution is found. In a Gauss-Siedel relaxation, (4) is re-arranged into:

$$\phi^U(n_1,n_2) = \frac{1}{4}\left\{\phi^U(n_1+1,n_2) + \phi^U(n_1-1,n_2) + \phi^U(n_1,n_2+1) + \phi^U(n_1,n_2-1)\right\} - \frac{1}{4}\rho^W(n_1,n_2) \tag{8}$$

It makes use of updated values of ϕ^U on the right-hand side of (8) as soon as they become available, i.e., the updating is done "in place" instead of being "copied" from an earlier timestep to a later one. The weakness of this method is its slow convergence, which is then improved by multigrid method.

PU adds one more important parameter determining the quality of a solution on a grid, named as weighting factor. The weighting factor corresponds to a degree of confidence or quality factor of a pixel value in the (phase) image, since phase noise affects image quality severely. Then, instead of minimizing (3), one should perform on its weighted version

$$E = \sum_{n_1,n_2} \begin{bmatrix} w_{n_1}(n_1,n_2)\left\{d^U_{n_1}(n_1,n_2) - d^W_{n_1}(n_1,n_2)\right\}^2 + \\ w_{n2}(n_1,n_2)\left\{d^U_{n_2}(n_1,n_2) - d^W_{n_2}(n_1,n_2)\right\}^2 \end{bmatrix} \tag{9}$$

where $w_{n_1}(n_1,n_2)$ and $w_{n_2}(n_1,n_2)$ are defined by

$$\begin{aligned} w_{n_1}(n_1,n_2) &= \min\left(w^2_{n_1}(n_1,n_2), w^2_{n_1}(n_1-1,n_2)\right) \quad \text{and} \\ w_{n_2}(n_1,n_2) &= \min\left(w^2_{n_2}(n_1,n_2), w^2_{n_2}(n_1,n_2-1)\right) \end{aligned} \tag{10}$$

Minimization of (9) gives weighted version of (4)

$$\begin{aligned} &w_{n_1}(n_1+1,n_2)\left(\phi^U(n_1+1,n_2) - \phi^U(n_1,n_2)\right) - w_{n_1}(n_1,n_2)\left(\phi^U(n_1,n_2) - \phi^U(n_1-1,n_2)\right) + \\ &w_{n_2}(n_1,n_2+1)\left(\phi^U(n_1,n_2+1) - \phi^U(n_1,n_2)\right) - w_{n_2}(n_1,n_2)\left(\phi^U(n_1,n_2) - \phi^U(n_1,n_2-1)\right) \\ &= \hat{\rho}^W(n_1,n_2) \end{aligned} \tag{11.a}$$

where

$$\begin{aligned} \hat{\rho}^W(n_1,n_2) = {} & w_{n_1}(n_1+1,n_2)\,d^W_{n_1}(n_1+1,n_2) - w_{n_1}(n_1,n_2)\,d^W_{n_1}(n_1,n_2) + \\ & w_{n_2}(n_1,n_2+1)\,d^W_{n_2}(n_1,n_2+1) - w_{n_2}(n_1,n_2)\,d^W_{n_2}(n_1,n_2) \end{aligned} \tag{11.b}$$

Unlike the unweighted case, this equation cannot be solved directly (eg. by FFT), instead, an iterative method should be applied. The Gauss-Siedel relaxation can be used to do this task by iterating weighted version of (8)

$$\phi^U(n_1,n_2) = \left\{ \begin{aligned} & w_{n_1}(n_1+1,n_2)\,\phi^U(n_1+1,n_2) + w_{n_1}(n_1,n_2)\,\phi^U(n_1-1,n_2) \\ & + w_{n_2}(n_1,n_2+1)\,\phi^U(n_1,n_2+1) + w_{n_2}(n_1,n)\,\phi^U(n_1,n_2-1) - \hat{\rho}^W(n_1,n_2) \end{aligned} \right\} \Big/ v(n_1,n_2) \tag{12.a}$$

where

$$v(n_1,n_2) = w_{n_1}(n_1+1,n_2) + w_{n_1}(n_1,n_2) + w_{n_2}(n_1,n_2+1) + w_{n_2}(n_1,n_2) \tag{12.b}$$

The Gauss-Siedel relaxation obtains only high-frequency component of a solution in earlier stages, while the low-frequency are left in the residual image. It happens because the relaxation on a point in the computational grid is influenced by its nearest neighbors only. In the multigrid method, various sizes of computational domains corresponding to the difference equation of the problem are defined. Effectively, the large grid extract high frequency component, while the small grid gathered the low frequency component of the solution. Then, a transfer mechanism among grids improved final solution.

3 The Multigrid Phase Unwrapping Method

The multigrid algorithm, firstly introduced in 1970s by Brandt [1], is a fast iterative numerical method for solving an elliptic differential equation with complexity O(*N*), where *N* is the gridsize [2]. The first application of the multigrid method for PU problem is described in [3] by Pritt. It applies Gauss-Siedel relaxation on smaller (coarser) grids and transferring the intermediate results to the larger (finer) one as illustrated in Fig.2. The intermediate results are then used as initial solutions of Gauss-Siedel relaxation on the finer grids.

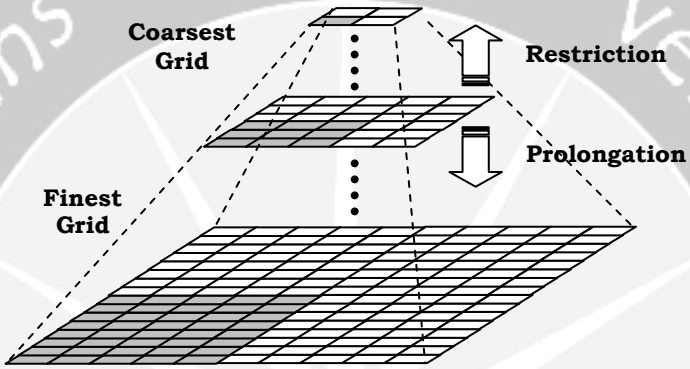


Fig.2 Restriction and prolongation in a multigrid method. One may look the prolongation as an upsampling process, while restriction is downsampling.

Usually, restriction operation uses the injection operator $c(i,j) = f(2i,2j)$ where $f(i,j)$ is the original field/function and $c(i,j)$ is the results or, a better one, is the full weighting operator defined as

$$c(i,j) = \frac{1}{16}\left(f(2i-1,2j-1) + f(2i+1,2j-1) + f(2i-1,2j+1) + f(2i+1,2j+1)\right) + \frac{1}{8}\left(f(2i,2j-1) + f(2i,2j+1) + f(2i+1,2j)\right) + \frac{1}{4}f(2i,2j) \quad (13)$$

The prolongation operator, that is the complement of the full weighting, is the bilinear interpolation,

$$\begin{aligned} f(2i,2j) &= c(i,j), \qquad f(2i+1,2j) = \frac{1}{2}\left(c(i,j) + c(i+1,j)\right) \\ f(2i,2j+1) &= \frac{1}{2}\left(c(i,j) + c(i,j+1)\right) \\ f(2i+1,2j+1) &= \frac{1}{2}\left(c(i,j) + c(i+1,j) + c(i,j+1) + c(i+1,j+1)\right) \end{aligned} \quad (14)$$

Two main multigrid methods are widely used in numerical computing, namely the (ordinary) multigrid method and the full multigrid method (FMG). In the first method, the defined grid acts as a temporary computation template. In the second one, the PDE is discretized into different-size sets of finite-difference-equation on various resolutions. Involvement of the weighting factor further splits those two multigrid methods into four, i.e., un-weighted multigrid, weighted multigrid, un-weighted FMG and weighted FMG.

Multigrid procedure is explained as follows. Consider a residual equation that is defined as

$$\mathbf{Ae} = \boldsymbol{\rho}^W - \mathbf{A}\hat{\boldsymbol{\varphi}}^U \equiv \mathbf{r} \tag{15}$$

where **e** is unknown error, and $\hat{\boldsymbol{\varphi}}^U$ is the estimated solution. Then, relaxing (12) with an initial guess $\hat{\boldsymbol{\varphi}}^U$ is equivalent to relaxing residual equation **Ae**=**r** in with an initial guess **e** = **0**. After a Gauss-Siedel relaxation, a solution is obtained accompanied by a residual error where most information resides initially. Then, this residual error is restricted and relaxed, giving a solution in the coarser grid accompanied by the next coarser residual error. The coarse solution is then prolonged to finer grid and superposed by previous solution, while the coarse residual goes to the next coarse grid and processed iteratively. This procedure yields a class of multigrid algorithm called the V-cycle. Another class that is related to the FMG is the so called FMG-cycle that uses V-cylce as its elemental computational procedure. Details of these algorithms can be found in [3].

4 Improved Multigrid Method

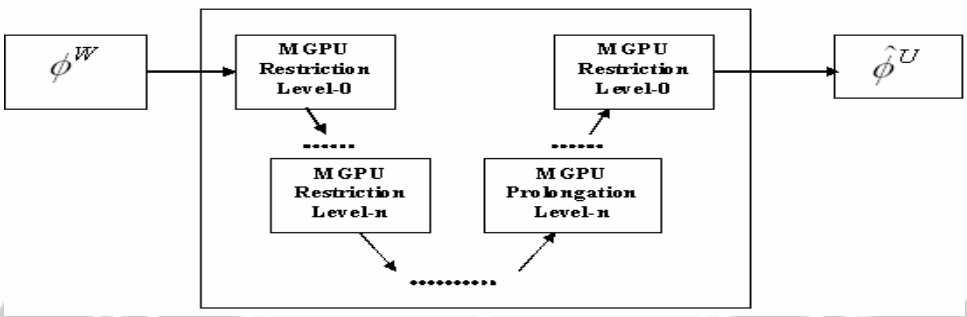
The original multigrid method incorporated the noise level in the weighting factor. Actually, the noise affects the result much earlier in PU stages, i.e. during Laplacian computation. An improvement of this algorithm is possible if the Laplacian can be estimated more precisely. A closer observation on the iteration shows that the residual fringes become wider in the later stages. Then, re-estimating the Laplacian/gradient value at later stages will be more reliable. In this Section, this idea is illustrated with unweighted V-cycle algorithm. Block diagram comparing the original and improved V-cycle is shown in Fig.3.

The residual wrapped phase image error is defined as:

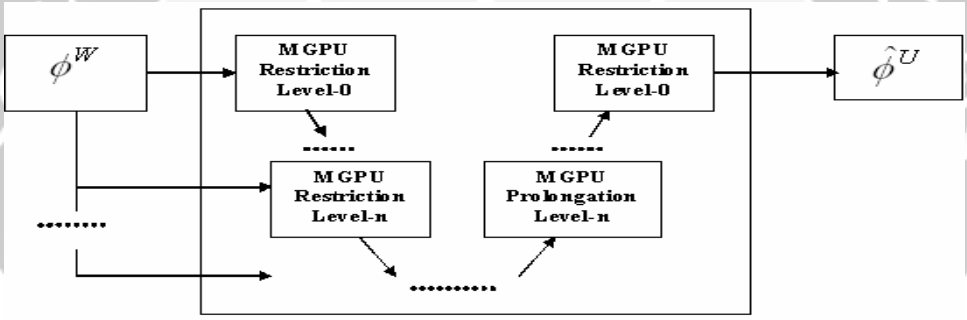
$$e_k^W = \phi^W - W\left(\hat{\phi}_k^U\right) \tag{16}$$

where the arithmetic is performed in modulo 2π, W is rewrapping operator and k is index of the error reprocessing cycle (k =0 is the first reprocessing stage, k=1 the second …etc.). The modification of the multigrid PU is performed through re-estimation of unwrapped phase image from the residual wrapped image (16). Since estimated solution $\hat{\phi}_0^U$ is obtained by solving a (matrix-formulated) discrete Poisson equation, there is an underlying assumption that (16) is non-zero because of imperfect estimation of ρ^W in. In fact, it is the case due to singular points

generation causing rotational field emergences in phase image that cannot be extracted with the ∇^2 operator. At least there are two sources of this defect; namely phase noise and insufficient sampling caused by, for example, layover/shadowing.



(a)



(b)

Fig. 3 Block diagram of (a) original V-cycle MGPU and (b) improved method.

Conventional multigrid V-cycle algorithm calculates the Laplacian from wrapped phase ϕ^W only once, as shown in Fig.3 (a). Gradient values in subsequent stages are based on this computation, both during restriction and prolongation as well. On the other hand, the improved method always involves original wrapped phase image ϕ^W on subsequent stages (Fig.3(b)).

Figure 4 shows detail block diagram of the improved method. In the figure, G-S Relax denotes Gauss-Siedel Relaxation, Pro (*i*) indicates prolongation to the *i*-th level, Wrap is a (re-) wrapping operator and Res(*i*) is the restriction to the *i*-th level. The figure only shows three-level processing, however, the generalization for *n*-th level is straightforward. The main difference of this method with the conventional one is that the original wrapped phase is involved in the gradient estimation process at all grid levels. The subtraction of the original wrapped phase with the intermediate solution effectively enlarges fringe periodicities. Then, lowpass filtering inherent in the restriction operator (and by interferogram filter if necessary) will become more effective in reducing SPs and consequently improves gradient

estimation result. Therefore, a better solution will be obtained as shown in the experiments.

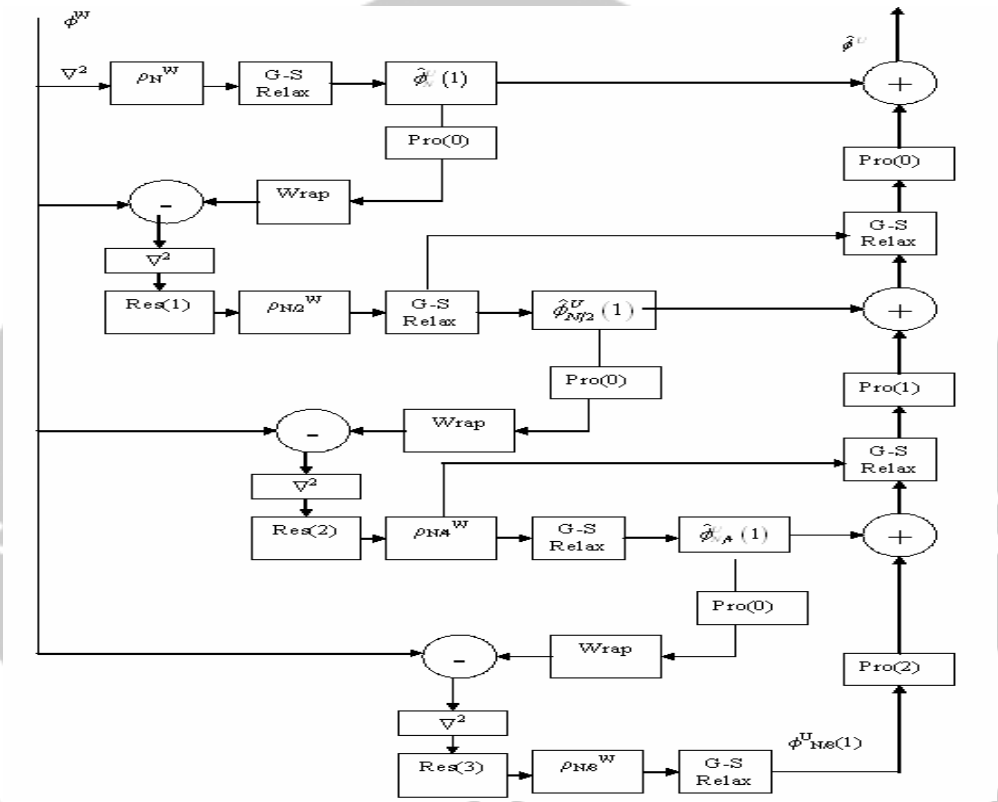


Fig. 4 Detailed diagram of the improved method: The main feature in the improved method is that the original wrapped image ϕ^W is always involved in the gradient estimation process. Subtraction of the intermediate result from ϕ^W effectively enlarges fringe periodicities and consequently improve SP elimination by low-pass filtering inherent in the restriction operator (or possibly inserted interferogram filter).

PU of MRI phase image is used in an experiment. Fig.5 shows snapshots of temporary solutions during restriction process for: (a) conventional and (b) improved method. Although at a glance both of methods gives almost similar results, a closer evaluations reveals the difference in details. The proposed method shows more details are added in the solutions than the conventional ones. As a result, final solution of the proposed method will be a better one, shown in Fig.6: (a) original phase image, (b) rewrapped PU result by conventional method and (c) rewrapped PU result of the improved method. It demonstrates the improvement obtained by embedding gradient re-estimation.

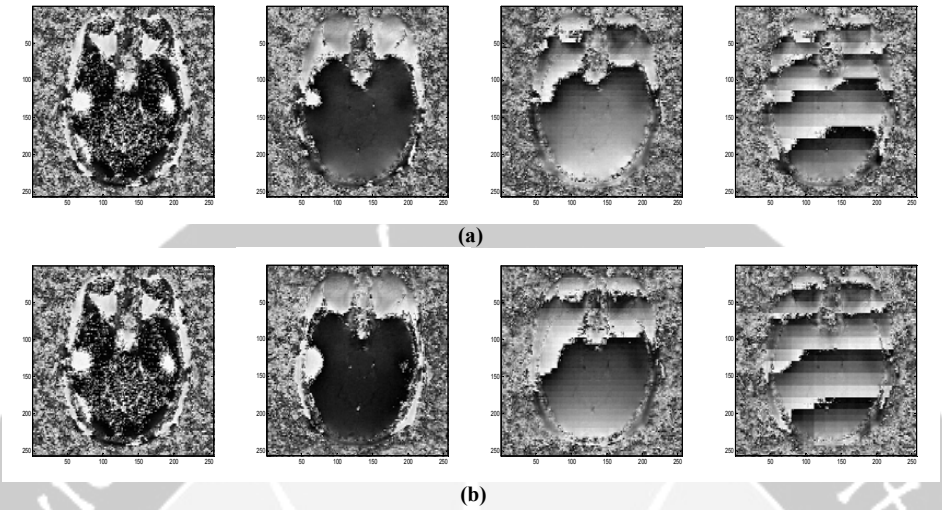


Fig.5 Snapshots of rewrapped PU results in subsequent stages: (a) Conventional method and (b) Improved method. Some differences, with more details in (b), emerge during the PU process. All grids are up-sampled to original image's size for clarity.

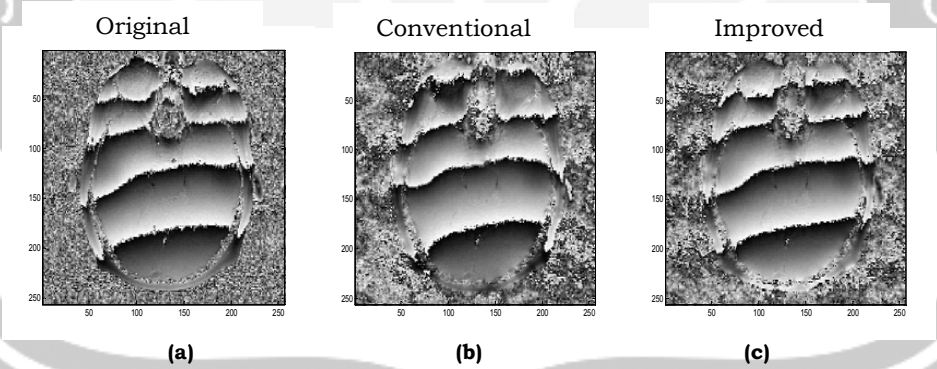


Fig.6 Comparison of PU results using (b) conventional and (b) improved method. Compared to the original in (a), the improved method is better than the conventional one. It is observed on region around the eyes (middle-top).

5 Epilog

A review on PDE-based image processing, particularly the PU method, focusing on the usage of multigrid method has been presented. Particularity of the problem may lead to a new computational strategy. As shown in this paper, noise—that has no counterparts in other PDE related problem, leads to the improved multigrid method. Some experimental examples have been shown to clarify the ideas.

Acknowledgments

The author would like to acknowledge Dr. Masanobu Shimada of JAXA (previously EORC-NASDA), Japan for supplying the InSAR image and Prof. J. Pauly of Information System Laboratory, Stanford University, USA, for a permission to use the MRI image.

References

[1] A. Brandt, *Mathematics of Computation*, Vol. 31, 1970, pp. 333-390.
[2] W.H. Press, S.A. Teukolsky, W.T. Vetterling and B.P. Flannery, *Numerical Recipes in C, The Art of Scientific Programming*, Cambridge Univ. Press, 2nd Ed., 1997.
[3] MD. Pritt, "Phase unwrapping by means of multigrid techniques for interferometric SAR," *IEEE Trans. On Geoscience and Remote Sensing*, Vol. 34, No. 3, May 1996, pp. 728-738.
[4] A.B. Suksmono, "Improving the multigrid phase-unwrapping algorithm by re-processing the residual error and its application to MRI phase image processing," *Proc. of APT Workshop 2005*, Multimedia University, Malaysia.
[5] A.B. Suksmono and A. Hirose, "Recursive transform-based phase unwrapping," *Proc. of ICIP 2003*.
[6] D.E.O. Dewi, A.B. Suksmono and TLR Mengko "Progressive Multigrid V-Cycle Phase Unwrapping for MRI Phase Images," Proc. of Healthcom 2005, Busan-Korea.
[7] A.B. Suksmono, "An Improved Multigrid Phase Unwrapping Method with Embedded Gradient Re-Estimation for MRI Phase Images, " Proc. of Healthcom 2005, Busan-Korea.
[8] A.B. Suksmono, "Improving the multigrid phase-unwrapping algorithm by reprocessing the residual error and its application to MRI phase image processing," *Proc. of the Third AP Workshop*, Multimedia University, Malaysia, 2005.

ANDRIYAN B. SUKSMONO: Imaging & Image Processing Research Group and
The Radio Communication & Microwave Laboratory, Department of Electrical Engineering, Institut Teknologi Bandung, Jl. Ganesha No. 10, Bandung, Indonesia, Phone:+62-22-2501661, Fax:+62-22-2534133
Email: suksmono@ltrgm.ee.itb.ac.id, suksmono@yahoo.com

APPLICATIONS OF HSLO(3)-FDTD ON DIRECT-DOMAIN AND TEMPORARY-DOMAIN APPROACHES FOR MAXWELL EQUATIONS

M. K. Hasan[a], M. Othman[b], Z. Abbas[b], J. Sulaiman[c], R. Johari[b]

[a] Universiti Kebangsaan Malaysia, Bangi, Malaysia.
[b] Universiti Putra Malaysia, Serdang, Malaysia
[c] Universiti Malaysia Sabah, Kota Kinabalu, Malaysia

Abstract. In this paper, a numerical simulation by a new high-speed low order (3)-finite difference time domain (HSLO(3)-FDTD) method on direct-domain and temporary-domain approaches will be conducted to simulate one dimensional free space wave propagation represented by a Gaussian pulse. The simulation will be conducted on 2 meter of solution domain with perfectly electric conducting (PEC) boundary condition. The efficiency of the new schemes are analyzed and compared with the standard finite difference time domain (FDTD) method in terms of processing time, amplitude, phase velocity and global error. Results obtained using HSLO(3)-FDTD with both approaches show very good matching to result simulated by FDTD and solve faster than the standard method.
Key-words: HSLO(3)-FDTD, direct-domain approach, temporary-domain approach, FDTD, wave propagation, Maxwell equations.

1 Introduction

In the era of computer technology, numerical simulation plays an important role in the development of science and technology. The method facilitate to enhance research and industrial development in many fields. The demands of advanced wireless communication devices to fulfil high-technology lifestyle has increase the demand of tools that facilitate research and development in the field of electromagnetic.

The finite difference time domain (FDTD) method is one of the most popular tool in simulating electromagnetic problem, such as antennas,wireless and wired communication, high speed electronic circuit, biomedical, semiconductors and etc[21]. The wave propagated in free space from a transmitter to a receiver is the ultimate event of wireless communication. All of those problems are solved via Maxwell equations ([8], [21]). Beside FDTD, Transmission Line Method (TLM)[6] is an alternative method that can be implemented in the time-domain.

2 Some Researches on Improving the Speed of FDTD Method

In computational electromagnetic (CEM), FDTD refers to second order temporal and spatial finite difference approximation to the Faraday's and Ampere's laws.

The method was first proposed by [28] to solve Maxwell equations in isotropic media. Yee used an electric field (E) which was offset both spatially and temporally from a magnetic field grid to obtain update equations that yield the present fields throughout the computational domain in term of past fields. The method was further developed by [23] to solve electromagnetic scattering from a dielectric cylinder. This method is the most commonly used to solve problem in time-domain because of its simplicity and directly adapted to homogeneous problem.

Since then, FDTD has become one of the most powerful Maxwell equations solver of electrodynamics. It has been implemented on various applications ([2], [5], [7], [10], [11], [12], [13], [18], [20], [22], [24], [25] & [26]). The algorithm simplicity, robustness, and potential for high complexity afforded by FDTD have prompted an extraordinary level of interest in this technique. However, there are drawbacks in the method. One of the drawback is it needs a long processing time to simulate problems [22].

To improve the speed of the method, some researchers apply higher-order scheme in FDTD on coarser grid. Lan et. al. [9] have developed a second order accurate in time and fourth order in space. This new scheme is then compared to FDTD by modelling plane-wave pulse propagating through free-space. Result show that the higher-order method reduces the numerical dispersion and has improved stability. Georgekapoulus and his research members [4] apply the FDTD(2,4) to a wave propagating problem. In the paper, they used the same gridding concept in [3]. The only difference is that they implement FDTD(2,4) at the coarser grid. Propokidis and Tsiboukis [17] have implement FDTD(2,4) to simulate a lossy dielectric problem. The implementation of higher-order truncation to Maxwell equations increase the complexity of the method, however by solving the problem in coarser grid will increase the speed of the processing time.

The advancement of multiprocessor technology machine also influence the development of high speed FDTD algorithms. Perlik et. al. [16] develop parallel FDTD algorithm to predict scattering of electromagnetic fields on a Connection Machine (CM). Rodohan and Saunders [19] implement a parallel FDTD algorithm on network of Transputers to simulate electromagnetic waves of a rectangular antenna. Meanwhile, Jensen et. al. [7] proposed a new design of parallelism, in both spatial and time. They solve circular scatterer but the design is merely analyzed theoretically. [14] solve electromagnetic scattering problem using domain decomposition FDTD with heterogeneous network of workstation which composes of 4 SUN SPARCstation and four IBM RS/6000. Zhenghui et. al. [29] describe the strategy used for parallel implementation of the FDTD algorithm on cluster of workstation with two heterogeneous PC (Pentium-II 266/64Mb and Pentium-II 200/64Mb) and a workstation by PVM parallel software to solve one dimensional free space problem. Yang and his research members [27] proposed new design of decomposition for implementing parallel FDTD on domain decomposition using MPI library to analyze coupling model of pulse into slot. The researcher decompose the whole domain into several sub-domains according to features of the

problem. Moreover, each sub-domain may have its own lattices independently to suit the special shapes.

In this research, we reduce FDTD processing time via different approach. The new method which is called the High Speed Low Order FDTD (HSLO-FDTD) method will be implement through direct-domain and temporary-domain approach to solve a one dimensional wave propagating in free space problem on the same mesh size used by the standard FDTD by a single processor machine.

3 Free space Maxwell Equations

Let's consider the Maxwell equations for free space below.

$$\frac{\partial E}{\partial t} = -\frac{1}{\varepsilon_0}\nabla \times H \tag{1}$$

$$\frac{\partial H}{\partial t} = -\frac{1}{\mu_0}\nabla \times E \tag{2}$$

where E,H,ε_0 and μ_0 are the electric fields, magnetic fields, electric permittivity and magnetic permeability, respectively. For the one-dimensional case using E_x and H_y, the Eqs. (1) and (2) become

$$\frac{\partial E_x}{\partial t} = -\frac{1}{\varepsilon_0}\frac{\partial H_y}{\partial z} \tag{3}$$

$$\frac{\partial H_y}{\partial t} = -\frac{1}{\mu_0}\frac{\partial E_x}{\partial z} \tag{4}$$

These are the equations of a plane wave with the electric field oriented in the x-direction, magnetic field in the y- direction, and travelling in the z-direction. For further details, see [21].

4 High Speed Low Order Finite-Difference Time-Domain Method

The HSLO-FDTD method was developed by borrowing the concept implemented in Modified Explicit Group(MEG) introduced recently by Othman and Abdullah [15], which is the extension of the half-sweep iterative method propose by Abdullah [1] through the Explicit Decoupled Group (EDG) iterative method. Both MEG and EDG are classified as iterative method and was used to solve eliptic type of problem.

In this research, we modify the concept used in MEG method to develop HSLO-FDTD for solving the free space Maxwell equations in time-domain. The iterative concept in MEG is ignored because there is no matrix in HSLO-FDTD method to be solved. By taking central difference approximations as below,

$$\frac{\delta F(i)}{\delta x} = \frac{F^n(i+\frac{m}{2}) - F^n(i-\frac{m}{2})}{m\Delta x} + O(\Delta x^2) \tag{5}$$

for spatial derivatives and

$$\frac{\delta F(i)}{\delta t} = \frac{F^{n+\frac{1}{2}}(i) - F^{n-\frac{1}{2}}(i)}{\Delta t} + O(\Delta t^2) \tag{6}$$

for temporal derivatives in (3) and (4) yields

$$\frac{E_x^{n+\frac{1}{2}}(k) - E_x^{n-\frac{1}{2}}(k)}{\Delta t} = -\frac{H_y^n(k+\frac{m}{2}) - H_y^n(k-\frac{m}{2})}{m\varepsilon_0\Delta x} \tag{7}$$

$$\frac{H_y^{n+1}(k+\frac{m}{2}) - H_y^n(k+\frac{m}{2})}{\Delta t} = -\frac{E_x^{n+\frac{1}{2}}(k+m) - E_x^{n+\frac{1}{2}}(k)}{m\mu_0\Delta x} \tag{8}$$

By rearranging Eqs. (7) and (8) above the same way as standard FDTD scheme, yields

$$\widetilde{E}_x^{n+\frac{1}{2}}(k) = \widetilde{E}_x^{n-\frac{1}{2}}(k) - \frac{D_t^*}{m}\left(H_y^n(k+\frac{m}{2}) - H_y^n(k-\frac{m}{2})\right) \tag{9}$$

$$H_y^{n+1}(k+\frac{m}{2}) = H_y^n(k+\frac{m}{2}) - \frac{D_t^*}{m}\left(\widetilde{E}_x^{n+\frac{1}{2}}(k+m) - \widetilde{E}_x^{n+\frac{1}{2}}(k)\right) \tag{10}$$

where

$$D_t^* = \frac{\Delta t}{\sqrt{\varepsilon_0\mu_0}\Delta x}$$

which m is odd number. See that (9) and (10) is a generalized form of FDTD method where when $m = 1$, it is the standard FDTD, when $m = 3$, it is the HSLO(3)-FDTD, when $m = 5$, it is the HSLO(5)-FDTD, and HSLO(7)-FDTD for $m = 7$. By using Eqs. (9) and (10), we only calculate $\frac{1}{m}$ of node points in the entire solution domain from 0 to T time steps. In this paper, we will only consider HSLO(3)-FDTD.

In this paper, Eqs. (9) and (10) with $m = 3$, will be used to solve problem in solution domain given by Fig.1(b) with the black square and circle is the magnetic and electric fields respectively that have to be solved in the main HSLO(3)-FDTD algorithm loop. The uncalculated node, will be solved later only at T after all black node have been calculated. The standard FDTD will be executed on solution domain given by Fig. 1(a).

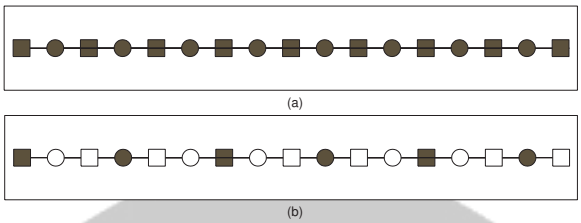


Figure 1: (a) Solution domain for standard FDTD method and (b) Solution domain for HSLO(3)-FDTD method

5 Direct-Domain and Temporary-Domain Algorithm

The HSLO(3)-FDTD can be implemented in direct-domain(DD) and temporary-domain(TD) algorithms. The difference between both algorithms are as shown in Algo. 1 & 2. The algorithm for HSLO(3)-FDTD will be illustrated in Algo. 1 & 2 with $D_t^* = \frac{\Delta t}{\sqrt{\varepsilon_0 \mu_0}\Delta x} = 0.5$. From both algorithms (Algo. 1 & 2), we can see that both algorithms have the same complexity which is $\theta(\frac{N_p N_t}{3})$, where N_p is the number of grid points and N_t is the number of time step. We further analyze the complexity of both algorithms by its number of arithmetic operation for addition & substraction (ADD/SUB) and multiplication & division (MUL/DIV). The arithmetic calculation for FDTD, HSLO(3)-FDTD(DD) and HSLO(3)-FDTD(TD)are summarize in Table 1.

Transform Actual Solution Domain,N_p to
Temporary Solution Domain,$\frac{N_p}{3}$
Loop for T from 0 to N_t
Loop for i from 0 to $\frac{N_p}{3}$
$E_x = E_x - \frac{D_t^*}{3} * (H_{y_{i+1}} - H_{y_i})$
End Loop
Gaussian Pulse,$E_x = e^{-0.5*(\frac{t_0 - T}{\sigma})^2}$
Setting PEC Boundary
Loop for i from 0 to $\frac{N_p}{3}$
$H_{y_i} = H_{y_i} - \frac{D_t^*}{3} * (E_{x_i} - E_{x_{i-1}})$
End Loop
End Loop
Transform Temporary Solution Domain,$\frac{N_p}{3}$ to
Actual Solution Domain,N_p
Calculate remaining grid point in solution domain

Algo. 1. HSLO(3)-FDTD Temporary Domain Algorithm.

Loop for T from 0 to N_t
  Loop for i from 1,$(i+3)$ to N_p
    $E_x = E_x - \frac{D_t^*}{3} * (H_{y_{i+2}} - H_{y_{i-1}})$
  End Loop
  Gaussian Pulse,$E_x = e^{-0.5*(\frac{t_0 - T}{\sigma})^2}$
  Setting PEC Boundary
  Loop for i from 0,$(i+3)$ to N_p
    $H_{y_i} = H_{y_i} - \frac{D_t^*}{3} * (E_{x_{i+4}} - E_{x_{i+1}})$
  End Loop
End Loop
Calculate remaining grid point in solution domain

Algo. 2. HSLO(3)-FDTD Direct Domain Algorithm.

Table 1: Comparison of FDTD, HSLO(3)-FDTD(DD) and HSLO(3)-FDTD(TD) arithmetic complexity

Method	**Arithmetic Operation**	
	ADD/SUB	MUL/DIV
FDTD	$4N_pN_t$	$2N_pN_t$
HSLO(3)-FDTD(DD)	$\frac{4N_pN_t}{3} + \frac{2N_p}{3}$	$\frac{2N_pN_t}{3} + \frac{2N_p}{3}$
HSLO(3)-FDTD(TD)	$\frac{4N_pN_t}{3} + \frac{2N_p}{3}$	$\frac{2N_pN_t}{3} + \frac{2N_p}{3} + 1$

From Table 1, we can calculate relative gain by both HSLO(3)-FDTD algorithms to standard FDTD. The formulation of gain for both method are as below. Relative gain (ADD/SUB) for HSLO(3)-FDTD(DD),

$$G_r(ADD/SUB) = \frac{2}{3} - \frac{1}{6N_t}$$

and relative gain (ADD/SUB) for HSLO(3)-FDTD(TD),

$$G_r(ADD/SUB) = \frac{2}{3} - \frac{1}{6N_t}$$

and taking the limit $N_t \to \infty$, we obtain the percentage gain of 67% for both HSLO(3)-FDTD algorithms and relative gain (MUL/DIV) for HSLO(3)-FDTD(DD),

$$G_r(MUL/DIV) = \frac{2}{3} - \frac{1}{3N_t}$$

and relative gain (MUL/DIV) for HSLO(3)-FDTD(TD),

$$G_r(MUL/DIV) = \frac{2}{3} - \frac{1}{3N_t} - \frac{1}{2N_pN_t}$$

and again taking the limit $N_t \to \infty$, we obtain the percentage gain of 67% for both HSLO(3)-FDTD algorithms. As the complexity is the major contributor to processing time, we predict that the maximum relative reduction in processing time for both HSLO(3)-FDTD algorithms are 67%.

6 Numerical Experiment and Results

The effectiveness of HSLO(3)-FDTD for both direct-domain and temporary-domain algorithms are analyzed by generating a one dimensional free space wave propagation problem with Gaussian pulse as the point source at the middle of the solution domain of 2 meter, truncated with perfectly electric conducting (PEC) boundary condition. To ensure the accuracy of the simulated result, the solution domain is discretize into 600 grid points with cell size of 0.0033 meter and time slice size of 5.5×10^{-12} ns. The experiment was run on Intel Pentium 3 of Mobile CPU 1 GHz 727 MHz 256 MB of RAM with LINUX operating system.

The result of simulations are given in Figures 2 and 3. From those figures, both HSLO(3)-FDTD algorithms results are very similar with standard FDTD. These figures (Fig. 2 & 3) shows the behavior of wave propagation in free space from the point source until reflected by the PEC boundary. However, there exist a small reduction in accuracy of approximation by HSLO(3)-FDTD method. The reduction in accuracy are shown in Fig. 4 as global error in power density unit. From the figure, we found that HSLO(3)-FDTD(DD) has better accuracy than HSLO(3)-FDTD(TD) but HSLO(3)-FDTD(TD) method approximate amplitude similar to standard FDTD (refer Fig. 5)than HSLO(3)-FDTD(DD).

The comparison of execution time is given in Fig. 6. and we can see that both new schemes simulate the problem faster than the standard FDTD scheme with 49.99% to 66.21% reduction in processing time for HSLO(3)-FDTD(DD) and 44.50% to 60.15% reduction in processing time for HSLO(3)-FDTD(TD). As mention before, it is expected that the maximum reduction in processing time gain by both HSLO(3)-FDTD are 67%.

7 Conclusion and Future Research

The HSLO(3)-FDTD method give us the opportunity to solve $\frac{1}{3}$ grid point of the solution domain in the main loop of HSLO(3)-FDTD and the remaining point only at the required time step. This approach will increase the speed of FDTD algorithm. The Performance of this scheme was tested for problem in one dimensional free space propagation with PEC boundary condition. The major advantages of this scheme is that it requires less processing time and less complexity algorithm than the existing FDTD scheme but there exist a small reduction in its accuracy. It is clearly shown that both HSLO(3)-FDTD approaches are better alternative than FDTD in one-dimension for free space wave propagating simulation.

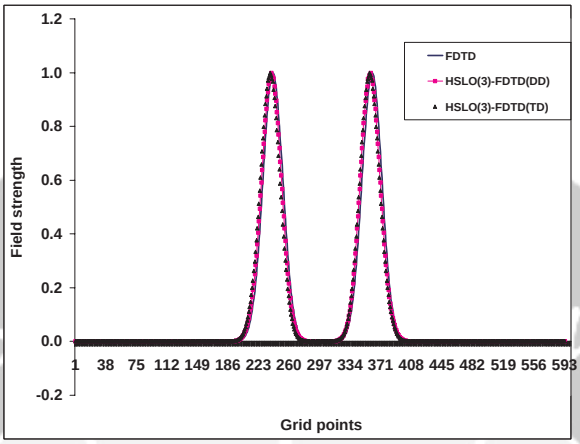


Figure 2: Wave propagation from the center of solution domain at 1ns

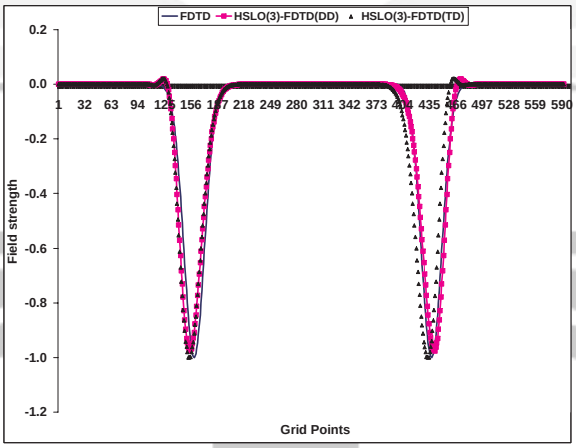


Figure 3: Wave propagation from the center of solution domain at 3.47 ns

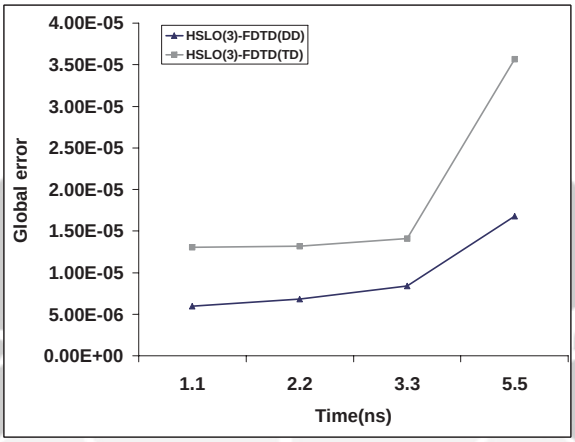


Figure 4: Global error between HSLO(3)-FDTD and standard FDTD methods

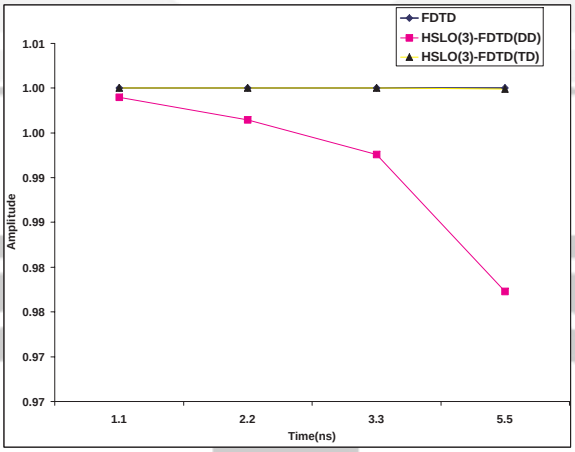


Figure 5: Amplitude in volts per meter for standard FDTD and HSLO(3)-FDTD

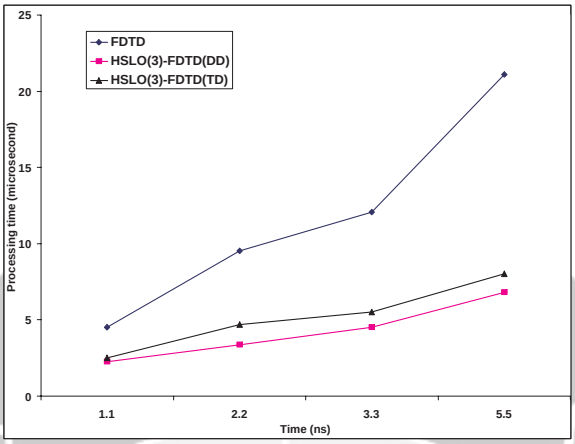


Figure 6: Comparison of HSLO(3)-FDTD and standard FDTD processing time

In this paper, we demonstrate the effectiveness of the new method on free space wave propagation with perfectly electric conducting boundary via direct-domain and temporary-domain approaches. Both approaches show tremendous result in simulating the problem. In the near future, we will apply this method for two dimensional problem.

References

[1] Abdullah, A.R. (1991), The four point explicit decoupled group (EDG) method: a fast poisson solver, *International Journal Computer Mathematics.*, **38**, 61–70.

[2] Anantha, V. & A. Taflove (1998), Calculation of diffraction coefficients of three-dimensional infinite conducting wedges using FDTD, *IEEE Trans. Antennas and Prop.*, **46**(11), 1755–1756.

[3] Chevalier, M.W., R.J. Luebbers & V.P. Cable (1997), FDTD local grid with material traverse, *IEEE Trans. Antennas and Prop.*, **45**(3), 411–421.

[4] Georgakopoulos, S.V., C.R. Birtcher, C.A. Balanis & R.A. Renaut (2002), Higher-order finite difference schemes for electromagnetic radiation, scattering, and penetration, part 1: theory, *IEEE Antennas Prop. Mag.*, **44**(1), 134–142.

[5] Hockanson, D.M., J.L. Drewniak, T.H. Hubing & T.P. Doren (1996), FDTD modelling of common-mode radiation from cables, *IEEE Trans. Electromagnetic and Compatibility*, **38**(3), 376–387.

[6] Hoefer, W.J.R. (1985), The transmission-line matrix method and applications, *IEEE Trans. Micro. Theory and Tech.*, **33**(10), 882–893.

[7] Jensen, M.A., A. Fijany & Y. Rahmat-Samii (1994), Time-parallel computational strategy for FDTD solution of maxwell's equations, *AP-S. Digest*, **1**, 380–383.

[8] Kunz, K.S. & R.J. Luebbers (1993), *The Finite Difference Time Domain Method for Electromagnetics*, CRC Press, Florida.

[9] Lan,K., Y. Liu & W. Lin (1999), Higher order (2,4) scheme for reducing dispersion in FDTD algorithm, *IEEE Tran. on Electromag. Comp.*, **41**(2), 160–165.

[10] Luebbers, R.J., J. Beggs & K. Chamberlain (1993), Finite difference time-domain calculation of transients in antennas with nonlinear loads, *IEEE Trans. Antennas and Prop.*, **41**(5), 566–573.

[11] Luebbers, R.J. & K. Kunz (1992), Finite difference time domain calculations of antenna mutual coupling, *IEEE Trans. Electromagnetic Compatibility*, **34**(3), 357–359.

[12] Luebbers, R.J. & H.S. Langdon (1996), A simple feed model that reduce time steps needed for fDTD antenna and microstrip calculations, *IEEE Trans. Antennas and Prop.*, **44**(7), 1000–1005.

[13] Maloney, J.G., G.S. Smith & W.R.J. Scott (1990), Accurate computation of the radiation from simple antennas using the finite-difference time-domain method, *IEEE Trans. Antenna and Prop.*, **38**(7), 1059–1068.

[14] Nguyen, S.T., B.J. Zook & X. Zhang (1994), Distributed computation of electromagnetic scattering problems using finite-difference time-domain decomposition, *IEEE Proc.-High Perform. Distrib. Comp.*, 85–93.

[15] Othman, M. & A.R. Abdullah (2000), An efficient four points modified explicit group poisson solver, *Intern. Jour. Comp. Math.*, **76**, 203–217.

[16] Perlik, A.T., T. Opsahl & A. Taflove (1989), Predicting scattering of electromagnetic fields using FDTD on a connection machine, *IEEE Trans. Magnetics*, **25**(4), 2910–2912.

[17] Propokidis, K.P. & T.D. Tsiboukis (2003), Higher-order FDTD(2,4) scheme for accurate simulations in lossy dielectrics, *Electronic Letters*, **39**(11), 835–836.

[18] Radisic, V., Y. Qian & T. Itoh (1998), Novel architechtures for high-efficiency amplifiers for wireless applications, *IEEE Trans. Antenna and Prop.*, **46**(11), 1901–1909.

[19] Rodohan, D.P. & S.R. Saunders (1994), Parallel implementations of the finite difference time domain (FDTD) method, 2^{nd} *Int. Conf. Comp. Electromagnetics*, 367–370.

[20] Schneider, J.B. & S. Hudson (1993), The finite-difference time-domain method applied to anisotropic material, *IEEE Trans. Antennas and Prop.*, textbf41(7), 994–999.

[21] Taflove, A. (1995), *Computational Electrodynamics: The Finite-Difference Time-Domain Method*, 1^{st} edition, Artech House, Boston.

[22] Taflove, A. (1980), Application of the finite-difference time-domain method to sinusoidal steady state electromagnetic-penetration problems, *IEEE Trans. Electromagnetic Compatibility*, **22**(3), 191–202.

[23] Taflove, A. & M. Brodwin (1975), Numerical solution of steady state electromagnetic scattering problems using the time-dependent maxwell's equations, *IEEE Tran. Micro. Theo. Tech.*, **23**(8), 623–730.

[24] Taflove, A., K. Umashankar & T.G. Jurgens (1985), Validation of FDTD modelling of the radar cross section of three-dimensional structures spanning up to nine wavelengths, *IEEE Trans. Antenna Prop.*, **33**(6), 662–666.

[25] Tirkas, P.A. & C.A. Balanis (1992), Finite-difference time-domain method for antenna radiation, *IEEE Trans. Antennas and Prop.*, **40**(3), 334–340.

[26] Umashankar, U. & A. Taflove (1982), A novel method to analyze electromagnetic scattering og complex objects, *IEEE Trans. Electromagnetic Compatibility*,**24**(4), 398–405.

[27] Yang, D., C. Liao, L. Jen & J. Xiong (2003), A parallel FDTD algorithm based on domain decomposition method using the MPI library, *Proc. Intern. Conf. on Par. and Distrib. Comp., App. and Tech., 2003*, 730–733.

[28] Yee, K.S. (1966), Numerical solution of initial boundary value problems involving maxwell's equations in isotropic media,*IEEE Tran. Antennas Prop.*, **14**(3), 302–307.

[29] Zhenghui, X., G. Benqing, & Z. Zejie (2002), A strategy for parallel implementation of the FDTD algorithm, *2002 3rd International Symposium on Electromagnetic Compatibility*, 259–263.

M.K. Hasan: Department of Industrial Computing, Faculty of Information Science and Technology, 43600 Universiti Kebangsaan Malaysia Bangi, Selangor, Malaysia & PHD student at Department of Communication Technology and Network, 43400 Universiti Putra Malaysia, Serdang, Selangor, Malaysia.
E-mail: khatim@ftsm.ukm.my;khatim71@hotmail.com

M. Othman : Department of Communication Technology and Network, 43400 Universiti Putra Malaysia, Serdang, Selangor, Malaysia.
E-mail: mothman@fsktm.upm.edu.my

Z. Abbas : Department of Physic, 43400 Universiti Putra Malaysia, Serdang, Selangor,

Malaysia.
E-mail: za@fsas.upm.edu.my

J. Sulaiman : School of Science and Technology, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia.
E-mail: jumat@ums.edu.my

R. Johari : Department of Communication Technology and Network, 43400 Universiti Putra Malaysia, Serdang, Selangor, Malaysia.
E-mail: rozita@fsktm.upm.edu.my

Finite Difference Modelling of Two-Dimensional Elastic Wave Propagation in Media Containing a Large Number of Skew Small Crack

Wono Setya Budhi, Marwan Wirianto

Department of Mathematics, Institut Teknologi Bandung, Indonesia

Abstract: If a wave is propagating in the media containing a large number of small crack then we will have a scatter wave. We will use a finite-difference technique to analyze the scatter wave.

The cracks are characterized by an explicit boundary condition. The sizes of the cracks are small enough such that the cracks are not represented in the finite-difference mesh. In order to that we will derive the formula of scatter wave because of skew cracks. We compare the method with an accurate integral representation of the solution and conclude that finite-difference technique is accurate and computationally fast.

Keywords : finite-difference methods, wave propagation

SIMULATING SEISMIC WAVE PROPAGATION IN TWO-DIMENSIONAL MEDIA USING DISCONTINUOUS SPECTRAL ELEMENT METHODS

Pranowo[a], F. Soesianto[b] & Bambang Suhendro[b]

[a] Atma Jaya Yogyakarta University, Indonesia
[b] Gadjah Mada University, Indonesia

We introduce a discontinuous spectral element method for simulating seismic wave in 2-dimensional elastic media. The methods combine the flexibility of a discontinuous finite element method with the accuracy of a spectral method. The elastodynamic equations are discretized using high-degree of Lagrange interpolants and integration over an element is accomplished based upon the Gauss-Lobatto-Legendre integration rule. This combination of discretization and integration results in a diagonal mass matrix and the use of discontinuous finite element method makes the calculation can be done locally in each element. Thus, the algorithm is simplified drastically. We validated the results of one-dimensional problem by comparing them with finite-difference time-domain method and exact solution. The comparisons show excellent agreement.

Keyword: seismic wave propagation, discontinuous spectral element, elastic media

1 Introduction

Simulation of seismic wave propagation played an important role in geophysics for imaging the structure of the earth interior and understanding the geodynamic phenomena [1]. The elastodynamic equation has been used intensively to model the seismic wave propagation in the earths. Because of analytical solutions of the equations are rare, the equations are solved numerically. The chalenge is to develop high performance numerical methods that are capable of solving the elastodynamic equations accurately and that can deal with complicated computational domain [2].

Continuous efforts have been devoted for developing numerical methods. During the last two decades, finite-difference time-domain (FDTD) methods have used extensively in modeling a large variety seismic wave propagation problems [3] [4] [5]. The FDTD methods are relatively easy to implement in computer code and do not require too much memory and CPU time. FDTD methods directly simulate the physical systems by making discrete approximation for the time and spatial derivatives via Taylor expansion to turn the partial differential equations into a system of algebraic equations. Yee introduced the first FDTD methods in 1966. This method compute electromagnetic fields that are staggered in space and time and can be interpreted as standard leapfrog method and well known as Yee's FDTD method. The Yee's FDTD methods suffer from poor numerical dispersion, which makes it difficult to run simulation for long time without introducing excessive errors and they have only second order accuracy in time and space. Some new schemes have also

started with Yee's scheme but were extended for greater accuracy rather than for geometry. High-order staggered finite-difference schemes, including compact schemes, are developed to improve the FDTD's accuracy. Pranowo et al. [6] developed multiresolution time-domain (MRTD) methods to simulate elastic wave fields. In the MRTD methods, the field components are expanded by using scaling and wavelet function then tested with using scaling and wavelet function through Galerkin's procedure. They show that computational effort can be reduced via wavelet thresholding. It is found that the implementation of MRTD on the boundaries is not easy task.

Finite volume methods (FVM), intensively used to solve fluid dynamics problems, have been adopted for elastodynamic equations [7] [1]. Le Veque [8] calculated the flux of the wave fields based Riemann solver succesfuly. Contrary to the FDTD methods, the FV methods allow one to deal with complicated geometries. The FV methods have second order accuracy and it is difficult to increase the order accuracy.

Finite element methods (FEM), based on variational formulation, can handle complicated geometries and heterogeneous material properties easily. The FE method exhibit poor dispersion properties for simulating wave propagation. Recently, least square Galerkin (LSG) [9] [10] and Discontinuous Galerkin (DG) methods [11] [12] have been developed to overcome the dispersion problems.

Spectral methods, that have high-order accuracy, have been adopted for elastodynamic equations. Spectral methods can not handle complex geometries easily. Komatitsch [13] used tensorial formulation approach for modelling curved interface. This approach can overcome the drawback, but with an increase of the computational cost.

Spectral element methods (SEM) are high-order Finite element methods which solve the variational formulations of the equations using spectral functions as basis functions. Komatistsch [2] used Legendre functions to solve the elastodynamic equations and Priolo [14] used Chebyshev functions. The Spectral element methods generate large global matrix from the elemental matrix, the methods require too much computer memory and CPU time.

In this paper we introduce a discontinuous spectral element (DSEM) method for simulating seismic wave in rectangular domain with free surface boundary conditions and internal material discontinuity. The DSEM methods combine the flexibility of a discontinuous finite element (Discontinuous Galerkin) methods with the accuracy of a spectral methods. The DSEM methods allow more general mesh (structured or unstructured mesh) configuration and inter-element continuity is not required. The basis function is discontinuous across mesh boundaries. Through a proper choices of flux computation points, the method only requires communication between mesh that have common faces. No global matrix invertion is required and the problem can be solved locally in each mesh. They are also suitable for both *h*- and *p*-type adaptivity [15] .

2 Elastodynamic Equations

Our approach of treating seismic waves numerically is based on the theory elastodynamics. We use the velocity-stress formulation as the governing equations:

$$\frac{\partial \hat{q}}{\partial t} + A\frac{\partial \hat{q}}{\partial x} + B\frac{\partial \hat{q}}{\partial y} = f \tag{1}$$

where

$$\hat{q} = \begin{pmatrix} \tau_{xx} \\ \tau_{xy} \\ \tau_{yy} \\ v_x \\ v_y \end{pmatrix} \qquad f = \begin{pmatrix} 0 \\ 0 \\ 0 \\ f_x \\ f_y \end{pmatrix}$$

$$A = \begin{pmatrix} 0 & 0 & 0 & (\lambda + 2\mu) & 0 \\ 0 & 0 & 0 & 0 & \mu \\ 0 & 0 & 0 & \lambda & 0 \\ \frac{1}{\rho} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\rho} & 0 & 0 & 0 \end{pmatrix} \qquad B = \begin{pmatrix} 0 & 0 & 0 & 0 & \lambda \\ 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & (\lambda + 2\mu) \\ 0 & \frac{1}{\rho} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\rho} & 0 & 0 \end{pmatrix}$$

In which v_x and v_y are the components of the velocity vector, τ_{xx}, τ_{yy} and τ_{xy} are the elements of the stress tensor and (f_x, f_y) is body force vector. The medium is described by the density $\rho(x, y)$ and the Lame coefficients $\lambda(x, y)$ and $\mu(x, y)$.

3 Discontinuous Spectral Element Methods

In this section we adopted Stanescu's notations [15] to describe the DSEM discretization. The domain is divided into non-overlapping rectangular elements within which N^{th} order Legendre polynomial (L_N) expansion is used. We mapped the global coordinates (x, y) onto local coordinates (ξ, η) in each element of the mesh. Under the mapping, equation (1) becomes

$$\frac{\partial q}{\partial t} + A\frac{\partial q}{\partial \xi} + B\frac{\partial q}{\partial \eta} = f$$

where $q = J\hat{q}$ are the transformed components of the velocity vector and stress tensor and J is the Jacobian of the transformation.

$$J = \mathbf{det}\left(\frac{\partial(x,y)}{\partial(\xi,\mu)}\right) = \Delta x \Delta y \text{ for rectangular mesh.}$$

The two-dimensional basis is constructed by taking a product of the one-dimensional basis which can be thought of as one-dimensional tensors. The one-dimensional global coordinate is transformed into elemental nodes as:

$$x_i = x_m + \frac{x_{m+1} - x_m}{2}(1+\xi_i), i \in \{0,...,N\} \tag{2}$$

where $\Omega_m = [x_m, x_{m+1}] = \Delta x$ represents the current element, ξ_i are roots of $(1-\xi^2)L'_N(\xi) = 0$ and L'_N denotes the derivative of L_N.

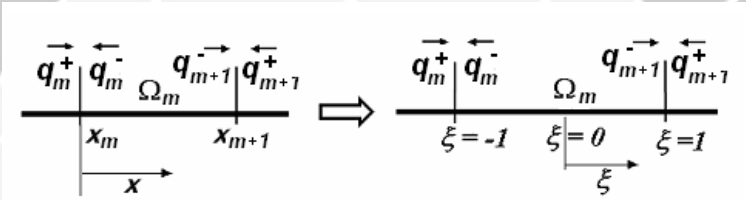


Figure 1. Local coordinate and flux.

The elemental Lagrangian interpolants $h_i(\xi)$ are chosen as a basis, it can be constructed as follows:

$$h_i(\xi) = -\frac{(1-\xi^2)L'_N(\xi)}{N(N+1)L_N(\xi_i)(\xi - \xi_i)} \tag{3}$$

The vector q is expanded using tensor product of equation (3) as follows:

$$q(t,\xi,\mu) = \sum_{i=0}^{N}\sum_{j=0}^{N} q_{ij}(t)h_i(\xi)h_j(\eta) \tag{4}$$

where $q_{ij}(t)$ denote pointwise value of q at time t. After we sample Galerkin procedure using the same trial function within each element, we obtain the following equation:

$$\left(\frac{\partial q}{\partial t}, \phi_{ij}\right) + \left(A\frac{\partial q}{\partial \xi}, \phi_{ij}\right) + \left(B\frac{\partial q}{\partial \eta}, \phi_{ij}\right) = (f, \phi_{ij}) \tag{5}$$

We can simplified the equation (5) as:

$$\left(\frac{\partial q}{\partial t},\phi_{ij}\right)+\left(\nabla_{\xi}\mathbf{F},\phi_{ij}\right)=\left(f,\phi_{ij}\right) \tag{6}$$

where $\mathbf{F}=Aq\vec{\mathbf{i}}+Bq\vec{\mathbf{j}}$ is the flux vector. Here $(.,.)$ represent the usual L^2 inner product, and $\phi_{ij}=h_i(\xi)h_j(\eta)$ are the trial function. Using the divergence theorem, equation is recast as:

$$\left(\frac{\partial q}{\partial t},\phi_{ij}\right)+\int_{\partial\Omega}\phi_{ij}\,\mathbf{F}\bullet\mathbf{n}\,dS=\left(\mathbf{F},\nabla_{\xi}\phi_{ij}\right)+\left(f,\phi_{ij}\right) \tag{3}$$

where $\mathbf{n}$ is normal to the of element interface.

The Gauss-Lobatto-Legendre (GLL) quadrature is applied to integrate the integrals. The GLL quadrature is defined as follows:

$$\int_{-1}^{1}\Phi(\xi)d\xi=\sum_{i=0}^{N}\Phi(\xi_i)\omega_i \tag{4}$$

The points $\xi_0=-1,\ \ \xi_N=1\quad L'_N(\xi_i)=0\quad \forall i\in\{1,2,...,N-1\}$ are called GLL Points, Φ is arbitrary polynomial. As long as Φ is a polynomial of degree less than $(2N-1)$ this quadrature rule is exact.

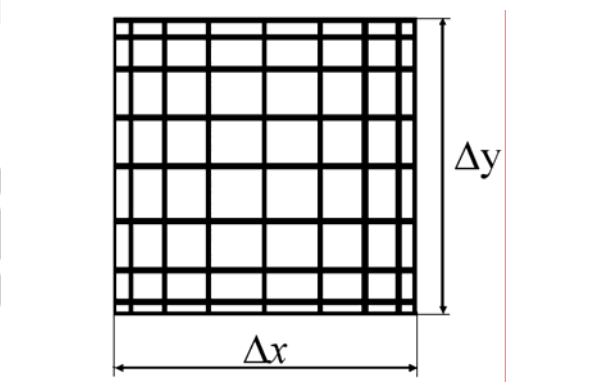


Figure 2. Element and GLL points.

After expanding the boundary integral and and performing some algebraic manipulation, we obtain the semi discrete form of the equations at the GLL points.

$$\frac{\partial q_{ij}}{\partial t} = \frac{1}{\Delta x}\left[-\mathbf{D}^{\xi}\,\mathbf{F} + \left(\frac{1}{\omega_N}\mathbf{F}^*\left(1,\eta_j\right)h_i(1)\right) - \frac{1}{\omega_0}\mathbf{F}^*\left(-1,\eta_j\right)h_i(-1)\right] + \frac{1}{\Delta y}\left[-\mathbf{D}^{\eta}\,\mathbf{F} + \left(\frac{1}{\omega_N}\mathbf{F}^*\left(\xi,1\right)h_j(1)\right) - \frac{1}{\omega_0}\mathbf{F}^*\left(\xi,-1\right)h_j(-1)\right] + f_{ij} \tag{5}$$

Notation F^* denotes numerical flux at the interface between elements and it can be approximated by using average flux:

$$\mathbf{F}^* = \frac{1}{2}\left(\mathbf{F}^+ + \mathbf{F}^-\right) \tag{7}$$

The differential matriks $\mathbf{D}$ can be written as:

$$\mathbf{D} = \frac{\partial h_j(\xi_i)}{\partial \xi}\begin{cases} \dfrac{L_m(\xi_i)}{L_m(\xi_j)(\xi_i - \xi_j)} & \text{if } i \neq j \\ 0 & \text{if } i = j, i \neq 0, m \\ \dfrac{-m(m+1)}{4} & \text{if } i = j = 0 \\ \dfrac{m(m+1)}{4} & \text{if } i = j = m \end{cases} \tag{8}$$

For simplicity, we use explicit staggered leapfrog method which has second order accuracy for temporal discretization.

4 Numerical results and discussion

4.1. One-dimensional problems

The methodology described above has been validated by comparison with both exact solution and FDTD methods for one-dimensional problem [16]. The following initial conditions are taken to perform numerical simulations:

$$v_y(x,0) = 0$$
$$\tau_{xy}\left(x,\frac{\Delta t}{2}\right) = \exp\left(-\ln 2x^2/9\right) \quad ; \text{-100} \le \text{x} \le \text{100}$$

$$\tau_{xy}(x,t) = \left(\mathbf{exp}\left(-\mathbf{ln}\,2\left((x-t)^2\right)/9\right) + \mathbf{exp}\left(-\mathbf{ln}\,2\left((x+t)^2\right)/9\right)\right)/2 \tag{9}$$

In both DSEM and FDTD, we performance the calculations by taking

$$\Delta t = 0.05 \text{ and } CFL = 0.075, 0.01, 0.125, 0.15$$

Courant Friedrich Lewy Number *CFL* is calculated as

$$CFl = v\frac{\Delta t}{\Delta x} \quad \text{for FDTD}$$

and

$$CFl = v\frac{\Delta t}{\Delta x_{\min}} \text{ for DSEM}$$

For DSEM, we take fixed order of polynomial $N = 8$. Figure 3 shows the exact solution for the stress at time=20.025.

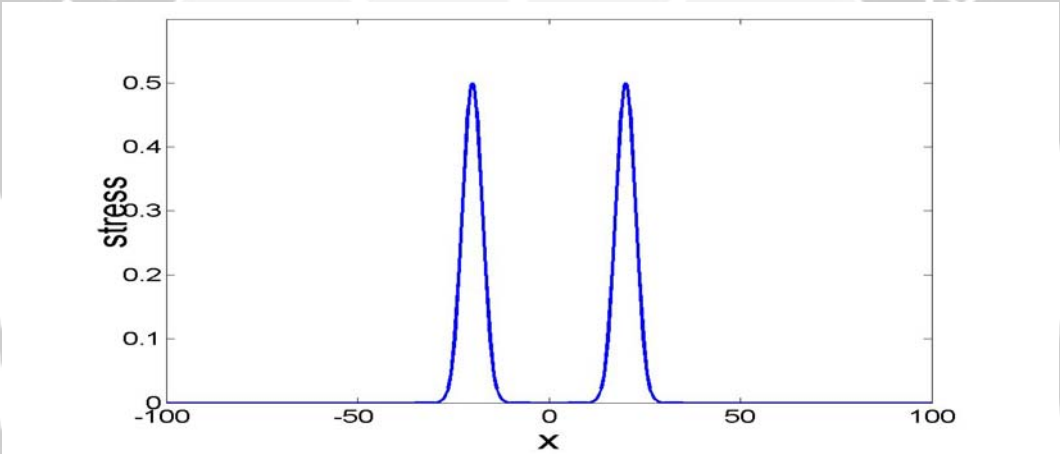


Figure 3. Exact solution for the stress at *t*=20.025

Since this problem has well-defined (infinitely smooth), we begin by computing the true error $\|\tau_{exact} - \tau_{numerical}\|_{\infty}$ for both DSEM and FDTD. In figure 4, the discrete maksimum error $\|\tau_{exact} - \tau_{numerical}\|_{\infty}$ is plotted versus the degrees of freedom (*dofs*). From the figure, we can see that DSEM need number of *dofs* approximately less than a half of FDTD's *dofs* to achieve the same accuracy.

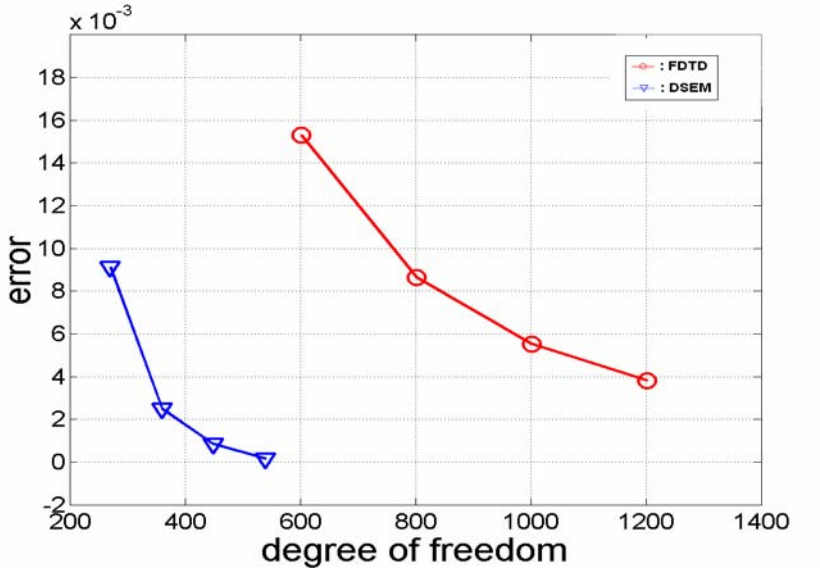


Figure 4. Comparison number of dofs between DSEM and FDTD

The accuracy of the solution in both DSM and FDTD is illustrated in figure 5 for various low CFL numbers. We can see that maksimum error $\left\|\tau_{exact} - \tau_{numerical}\right\|_{\infty}$ for DSEM decreased faster than FDTD as the CFL number increases.

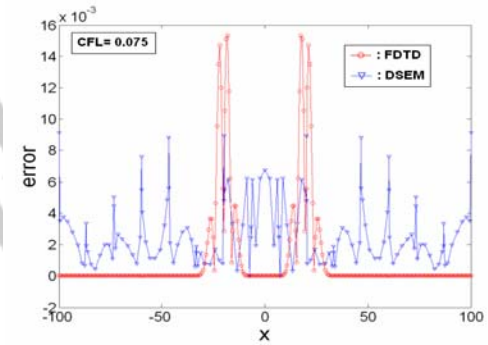


Figure 5a. Comparison error between DSEM and FDTD for *CFL*=0.075

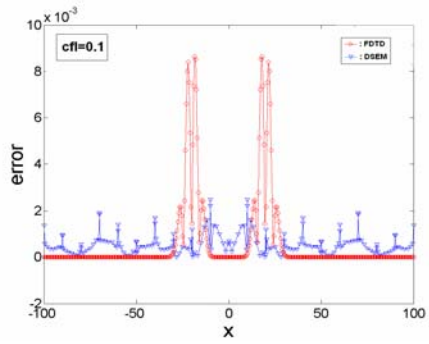


Figure 5a. Comparison error between DSEM and FDTD for *CFL*=0.1

High-order basis of DSEM can suppress the error of DSEM, but the use of average flux makes the error spread out in entire domain. If the basis is constant, it will make DSEM equal to central difference schemes of Finite Difference methods which are not stable. Staggered grid can reduce the error, it is shown in figure 5. The error of FDTD is localized only in region where large gradient of the fields, i.e. sharp wave front, occurs.

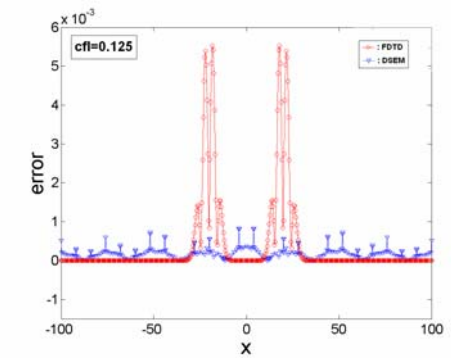


Figure 5c. Comparison error between DSEM and FDTD for *CFL*=0.125

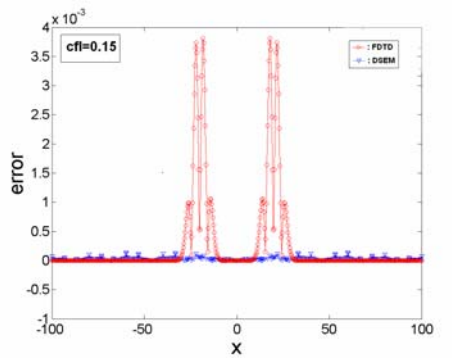


Figure 5d. Comparison error between DSEM and FDTD for *CFL*=0.15

4.1. Two-dimensional problems

The model we consider is two layered for heterogeneus media. The medium has a horizontal internal boundary that divides it into two layers. The upper layer is characterized by a *P*-wave velocity of 2000 m.s^{-1}, an *S*-wave velocity of 1300 m.s^{-1}, and a mass density of 1000 kg.m^{-2}. The lower layer elastic parameters are a *P*-wave velocity of 2800 m.s^{-1}, an *S*-wave velocity of 1473 m.s^{-1}, and a mass density of 1500 kg.m^{-2} [2]. A strong contrast both in velocity and in Poisson's ration is hence modeled, with ν =0.13 for the lower layer and ν =0.38 for the upper layer. The source is explosive and located inside the upper layer.

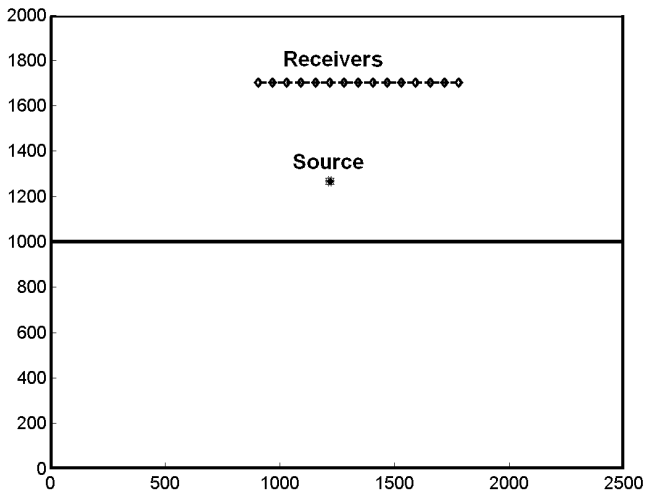


Figure 6. Two layered heteregeneous elastic media

The numerical model has a width of 2500 m amd a height of 2000 m. The source position is $(x, y) = (1218.75, 1366.67)\,\mathrm{m}$. The line of receivers goes from $x = 906.25\,\mathrm{m}$ to $x = 1781.25\,\mathrm{m}$ at $y = 1700\,\mathrm{m}$. The mesh is composed of 30×40 elements, with polynomial of order $N = 12$. The explosive source is a Ricker wavelet in time with central frequency of 14.5. The time step $\Delta t = 2.5\,\mathrm{m\,sec}$. Figure 6 shows the description of the model. The four sides of the model are assigned to be free surfaces.

Figure 7 shows the snapshots of $P - SV$ wave propagation in two-layered media at $t = 0.1875, t = 0.2625, t = 0.3375, t = 0.4125, t = 0.4875, t = 0.5625\,\mathrm{sec}$. The entire wavefields are composed of direct phases *(P,S)*, reflected waves from internal boundary *(PPr, PSr, SPr, SSr)* or the free surface *(PP,PS)*. Mode conversions of wave reflected at the internal boundary as well as at the top free surface are clearly visible.

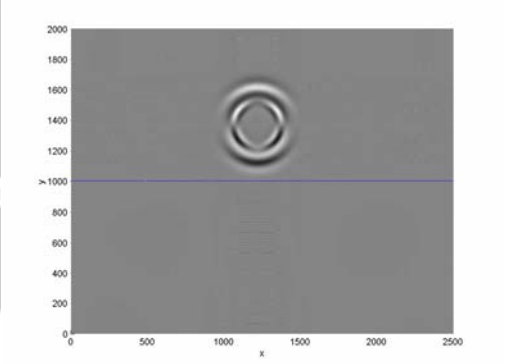


Figure 7a. τ_{yy} field at t = 0.1875 sec

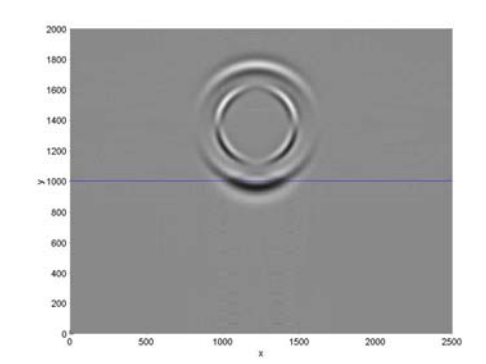


Figure 7b. τ_{yy} field at t = 0.2675 sec

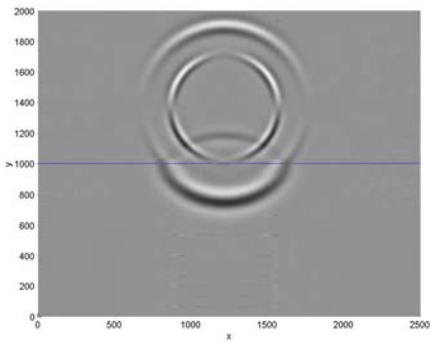


Figure 7c. τ_{yy} field at t = 0.3375 sec

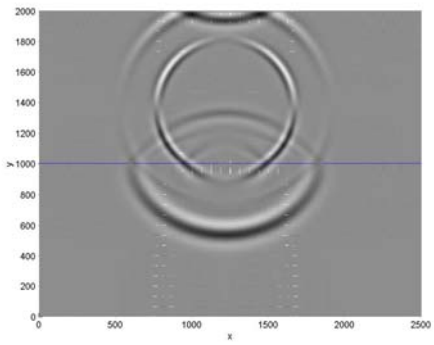


Figure 7d. τ_{yy} field at t = 0.4125 sec

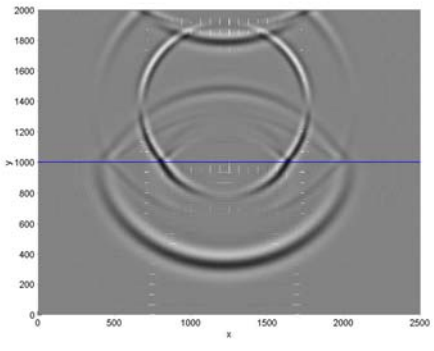


Figure 7e. τ_{yy} field at t = 0.4875 sec

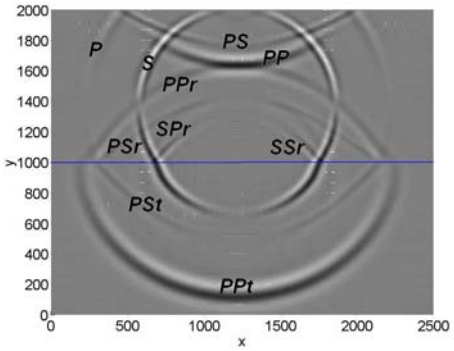


Figure 7f. τ_{yy} field at t = 0.5625 sec

Figure 8 shows the numerical time response of $P-SV$ waves in heterogeneous medium recorded at 15 receivers placed horizontally inside the medium.

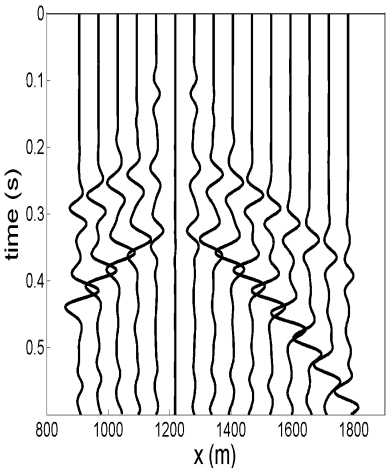


Figure 8. Seismogram of v_x

5 Conclusion

We have presented discontinuous spectral element methods for simulation of seismic wave propagation. Comparison with the FDTD methods for one-dimensional problem shows that DSEM need dofs less than FDTD methods for the same accuracy. We demonstrated that heteregeneous media, that contain material discontinuity, can be handled easily by using DSEM. Mode conversion of reflected waves can be captured well.

For future research, we plan to extend the DG method for solving problems with irregular domain and apply *hp adaptive* technique to increase the accuracy and to reduce computational costs. The numerical flux will be calculated based on Riemann solver.

Acknowledgment

We are very grateful to Dr. Kaser and Prof. Stanescu for sending us their papers. This research is partially supported by Atma Jaya Yogyakarta University.

References

[1] Kaser, M.A. (1999), *Simulation of seismic wave propagation on irregular grids*, Diplomarbeit, Ludwig-Maximilians-Universitat, Muenchen.

[2] Komatitsch, D. & J. P.Villote (1998), The spectral element Method: An efficient tool to simulate response of 2D and 3D geological structure , *Bulletin of the Seismological Society of America*, Vol. 2, 368-392.

[3] Virieux, J. (1986), P-SV wave propagation in heteregeneous media: velocity-stress finite difference method, *Geophysics* 51, 889-901.

[4] Meyer, T.F.N., (2001), *Numerical simulation of 3-D seismic wave propagation through subduction Zones*, Diplomarbeit, Ludwig-Maximilians-Universitat, Muenchen.

[5] Ewald, M.A, (2001), *Numerical simulation of site effects with application to the Cologne basin*, Diplomarbeit, Ludwig-Maximilians -Universitat, Muenchen.

[6] Pranowo, F. Soesianto & B. Y. Andiyanto (2003), The Multiresolution time-domain method based on Haar wavelets for numerical simulation of elastic wave propagation, *Proceedings of International Seminar on Aerospace Technology,* Yogyakarta, Indonesia.

[7] Dormy, E. & A. Tarantola (1996), Numerical simulation of elastic wave propagation using a finite volume method, J. Geophysics. Res., 100, 2123-2133.

[8] Le Veque, R.J. (2004), *Finite-Volume methods for hyperbolic problems*, Cambridge University Press, Cambridge.

[9] Thompson, L.L. (1994), *Design and analysis of space-time and Galerkin least-squares finite element methods for fluid structure interaction in exterior domain*, Ph.D thesis, Stanford University.

[10] Hulbert, G. (1989), Space-time finite element methods for second order hyperbolic equations, Ph.D thesis, Stanford University.

[11] Li, X.D. (1996), *Adaptive Finite Element Procedures in Structural Dynamics*, Ph.D Thesis, Chalmers University of Tecnology.

[12] Ekevid, T. (2003), Computational solid wave prpagation: Numerical technique and industrial applications, Ph.D. Thesis, Chalmers University of Tecnology.

[13] Komatitsch, F. Coutel & P. Mora (1996), Tensorial formulation of the wave equation for modelling curved interface, *J. Int. Geophys*., 127, 156-168.

[14] Priolo, E. (2001), Earthquake ground motion simaulation through the 2-d spectral element method, *J. Computational Acoustics*, Vol. 9 no. 4, 127, 1561-1581.

[15] Stanescu, D., M.Y. Hussaini & F. Farassat (2002), Aircraft noise scattering – A Discontinuous spectral element approach, *Proceedings of the 40th AIAA Aerospace Sciences Meeting*, Reno, NV, USA.

[16] Pranowo, F. Soesianto & B. Suhendro (2004), High-order discontinuous galerkin for numerical simulation of elastic wave propagation, *Proceedings of Quality in Research*, Jakarta, Indonesia.

PRANOWO: Ph D student at Department of Electrical Engineering, Gadjah Mada University,
Jl. Grafika 2 Yogyakarta 55281, Indonesia.
Department of Informatics, Atma Jaya Yogyakarta University.
Jl. Babarsari 43 Yogyakarta 55281, Indonesia.
E-mail: pran@mail.uajy.ac.id}

F. SOESIANTO: Department of Electrical Engineering, Gadjah Mada University,
Jl. Grafika 2 Yogyakarta 55281, Indonesia.

BAMBANG SUHENDRO: Department of Civil Engineering, Gadjah Mada University,
Jl. Grafika 2 Yogyakarta 55281, Indonesia.

NUMERICAL SOLUTIONS TO STATIC ELASTICITY PROBLEMS OF INHOMOGENEOUS ISOTROPIC MATERIALS

Mohammad Ivan Azis

Hasanuddin University, Makassar, Indonesia

Abstract. A boundary element method is derived for the solution of static elasticity problems of inhomogeneous isotropic elastic materials. Some particular problems are considered to illustrate the application of the method.
Key-words: Boundary Element Method, static elasticity, inhomogeneous, isotropic materials

1 Introduction

Following the early work of Rizzo in [5] a large number of authors have used the boundary element method to effectively obtain numerical solutions to a variety of elastic problems for homogeneous isotropic elastic materials (see for example Brebbia and Dominguez [1]).

In contrast the application of the method to problems for inhomogeneous isotropic elastic materials is very limited due to the difficulty in obtaining appropriate Green's functions for the kernels of the relevant boundary integral equations. Recently Manolis and Shaw in [4] obtained a suitable Green's function for the vector wave equation in a mildly heterogeneous isotropic continuum. Their Green's function was obtained for a particular variation in the material parameters and in particular is restricted to the case when the Lamé parameters λ and μ are equal. This leads to a Poisson's ratio of 0.25 which restricts the application of the method but as Manolis and Shaw in [4] pointed out, this particular value of Poisson's ratio is a common value for rock materials (see Turcotte and Schubert [3]).

This paper builds on the work of Manolis and Shaw [4] to develop a perturbation procedure for the solution of plane static problems for isotropic inhomogeneous media with Lamé parameters given by

$$\begin{aligned} \lambda(\mathbf{x}) &= \lambda^{(0)} g(\mathbf{x}) + \epsilon \lambda^{(1)}(\mathbf{x}) \\ \mu(\mathbf{x}) &= \mu^{(0)} g(\mathbf{x}) + \epsilon \mu^{(1)}(\mathbf{x}) \end{aligned}$$

where $\mathbf{x} = (x_1, x_2, x_3)$ is a vector in R^3, $g(\mathbf{x})$ is a function which must satisfy particular constraint, $\lambda^{(0)} = \mu^{(0)}$ are constants and ϵ is a small parameter. Within this constraint these forms permit a wide choice of variations for the elastic parameters $\lambda(\mathbf{x})$ and $\mu(\mathbf{x})$. Boundary integral equations are obtained for the solution of problems for materials with Lamé parameters of this form and these integral equations are used to solve some particular boundary value problems.

Parameters Lamé λ and μ can be expressed in elastic modulus E and Poisson ratio ν as $\lambda = \frac{\nu E}{(1+\nu)(1-2\nu)}, \mu = \frac{E}{2(1+\nu)}$. Conversely, elastic modulus E and Poisson ratio ν can be expressed in parameters Lamé λ and μ as $E = \frac{\mu(3\lambda+2\mu)}{\mu+\lambda}, \nu = \frac{\lambda}{2(\mu+\lambda)}$. Sometimes μ is written as the rigidity modulus or shear modulus G.

2 Basic equations

Referred to a Cartesian frame $Ox_1x_2x_3$ the equilibrium equations in an elastic material in the absence of body force may be written in the form

$$\sigma_{ij,j} = \mathbf{0} \tag{1}$$

where σ_{ij} for $i, j = 1, 2, 3$ denotes the stress tensor, the indexed commas indicate partial differentiation with respect to the spatial coordinates x_j and the repeated suffix summation convention (summing from 1 to 3) is employed. The stress-displacement relations are

$$\sigma_{ij} = \lambda\delta_{ij}u_{k,k} + \mu\left(u_{i,j} + u_{j,i}\right) \tag{2}$$

where u_k for $k = 1, 2, 3$ denotes the displacement and δ_{ij} the Kronecker delta. Also in (2) $\lambda(\mathbf{x})$ and $\mu(\mathbf{x})$ with $\mathbf{x} = (x_1, x_2, x_3)$ denote the Lamé parameters which are taken to be twice differentiable functions of the spatial variables x_1, x_2 and x_3. Substitution of (2) into (1) yields

$$\left[\lambda\delta_{ij}u_{k,k} + \mu\left(u_{i,j} + u_{j,i}\right)\right]_{,j} = \mathbf{0} \tag{3}$$

3 Statement of the boundary value problem

An inhomogeneous isotropic elastic material occupies the region Ω in R^3 with boundary $\partial\Omega$ which consists of a finite number of piecewise smooth closed surfaces. On $\partial\Omega_1$ the displacement u_i is specified and on $\partial\Omega_2$ the stress vector $P_i = \sigma_{ij}n_j$ is specified where $\partial\Omega = \partial\Omega_1 \cup \partial\Omega_2$ and $\mathbf{n} = (n_1, n_2, n_3)$ denotes the outward pointing normal to $\partial\Omega$. It is required to find the displacement and stress throughout the material. Thus a solution to (3) is sought which is valid in Ω and satisfies the specified boundary conditions on $\partial\Omega$.

4 Reduction to a constant coefficient equation

In this section the procedure developed in Manolis and Shaw [4] is used to obtain a boundary element integral method for particular classes of coefficients $\lambda(\mathbf{x})$ and $\mu(\mathbf{x})$. This derivation is achieved by introducing a transformation of the dependent variable $u_i(\mathbf{x})$ to transform (3) to a constant coefficients equation. The coefficients $\lambda(\mathbf{x})$ and $\mu(\mathbf{x})$ are required to take the form

$$\lambda(\mathbf{x}) = \lambda^{(0)}g(\mathbf{x}) \qquad \mu(\mathbf{x}) = \mu^{(0)}g(\mathbf{x}) \tag{4}$$

where $\lambda^{(0)}$ and $\mu^{(0)}$ are constants. Use of (4) in (3) yields

$$\left\{g\left[\lambda^{(0)}\delta_{ij}u_{k,k}+\mu^{(0)}\left(u_{i,j}+u_{j,i}\right)\right]\right\}_{,j}=\mathbf{0} \tag{5}$$

Let

$$\psi_i(\mathbf{x})=g^{1/2}(\mathbf{x})\,u_i(\mathbf{x}) \tag{6}$$

so that (5) may be written in the form

$$\left\{g\left[\lambda^{(0)}\delta_{ij}\left(g^{-1/2}\psi_k\right)_{,k}+\mu^{(0)}\left(\left(g^{-1/2}\psi_i\right)_{,j}+\left(g^{-1/2}\psi_j\right)_{,i}\right)\right]\right\}_{,j}=\mathbf{0}$$

Thus

$$\lambda^{(0)}\left[g\left(g^{-1/2}\psi_k\right)_{,k}\right]_{,i}+\mu^{(0)}\left[g\left(g^{-1/2}\psi_i\right)_{,j}\right]_{,j}+\mu^{(0)}\left[g\left(g^{-1/2}\psi_j\right)_{,i}\right]_{,j}=\mathbf{0} \tag{7}$$

Now

$$\begin{aligned}&\left[g\left(g^{-1/2}\psi_k\right)_{,k}\right]_{,i}\\ =\;&\frac{1}{4}g^{-3/2}g_{,i}g_{,k}\psi_k-\frac{1}{2}g^{-1/2}g_{,ki}\psi_k-\frac{1}{2}g^{-1/2}g_{,k}\psi_{k,i}+\frac{1}{2}g^{-1/2}g_{,i}\psi_{k,k}+g^{1/2}\psi_{k,ki}\\ =\;&-g^{1/2}_{,ki}\psi_k+g^{1/2}\psi_{k,ki}-\frac{1}{2}g^{-1/2}g_{,k}\psi_{k,i}+\frac{1}{2}g^{-1/2}g_{,i}\psi_{k,k}\end{aligned} \tag{8}$$

Similarly

$$\left[g\left(g^{-1/2}\psi_i\right)_{,j}\right]_{,j}=-g^{1/2}_{,jj}\psi_i+g^{1/2}\psi_{i,jj} \tag{9}$$

$$\left[g\left(g^{-1/2}\psi_j\right)_{,i}\right]_{,j}=-g^{1/2}_{,ij}\psi_j+g^{1/2}\psi_{j,ij}-\frac{1}{2}g^{-1/2}g_{,i}\psi_{j,j}+\frac{1}{2}g^{-1/2}g_{,j}\psi_{j,i} \tag{10}$$

Substitution of (8), (9) and (10) into (7) yields

$$\begin{aligned}&g^{1/2}\left[\lambda^{(0)}\delta_{ij}\psi_{k,k}+\mu^{(0)}\left(\psi_{i,j}+\psi_{j,i}\right)\right]_{,j}-\left[\lambda^{(0)}\psi_k g^{1/2}_{,ki}+\mu^{(0)}\psi_i g^{1/2}_{,jj}+\mu^{(0)}\psi_j g^{1/2}_{,ij}\right]\\ &-\left(\lambda^{(0)}-\mu^{(0)}\right)\left[\frac{1}{2}g^{-1/2}\right]\left[g_{,k}\psi_{k,i}-g_{,i}\psi_{k,k}\right]=\mathbf{0}\end{aligned} \tag{11}$$

If $g(\mathbf{x})$ assumes the form

$$g(\mathbf{x})=\left(\gamma_0+\gamma_1x_1+\gamma_2x_2+\gamma_3x_3\right)^2 \tag{12}$$

where γ_t, $t=0,1,2,3$ are constants and also

$$\lambda^{(0)}=\mu^{(0)} \tag{13}$$

so that $\lambda(\mathbf{x}) = \mu^{(0)} \left(\gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_3\right)^2 = \mu(\mathbf{x})$ then (11) reduces to

$$\left[\lambda^{(0)} \delta_{ij} \psi_{k,k} + \mu^{(0)} \left(\psi_{i,j} + \psi_{j,i}\right)\right]_{,j} = \mathbf{0} \quad (\text{with } \lambda^{(0)} = \mu^{(0)}) \tag{14}$$

Thus if ψ_i is any solution of the equations of equilibrium in displacement form for a homogeneous isotropic elastic material with Lamé constants $\lambda^{(0)}$ and $\mu^{(0)}$ then a corresponding solution of the equations of equilibrium for an inhomogeneous isotropic elastic material with Lamé parameters $\lambda(\mathbf{x})$ and $\mu(\mathbf{x})$ given by the multiparameter form (4) may be written, from (6), in the form

$$\begin{aligned} u_i(\mathbf{x}) &= g^{-1/2}(\mathbf{x})\, \psi_i(\mathbf{x}) \\ &= \left(\gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_3\right)^{-1} \psi_i(\mathbf{x}) \end{aligned}$$

The corresponding stresses obtained from (2) are given by

$$\sigma_{ij} = -\psi_k\, \sigma^{[g]}_{ijk} + g^{1/2}\, \sigma^{[\psi]}_{ij}$$

where

$$\begin{aligned} \sigma^{[g]}_{ijk} &= \lambda^{(0)} \delta_{ij}\, g^{1/2}_{,k} + \mu^{(0)} \left(\delta_{ki}\, g^{1/2}_{,j} + \delta_{kj}\, g^{1/2}_{,i}\right) \\ \sigma^{[\psi]}_{ij} &= \lambda^{(0)} \delta_{ij}\, \psi_{k,k} + \mu^{(0)} \left(\psi_{i,j} + \psi_{j,i}\right) \end{aligned}$$

and the stress vector

$$P_i = -\psi_k P^{[g]}_{ik} + g^{1/2} P^{[\psi]}_i \tag{15}$$

where

$$\begin{aligned} P^{[g]}_{ik} &= \sigma^{[g]}_{ijk} n_j \\ P^{[\psi]}_i &= \sigma^{[\psi]}_{ij} n_j \end{aligned} \tag{16}$$

A boundary integral equation for the solution of (14) is given in Brebbia and Dominguez [1] in the form

$$\eta(\mathbf{x}_0)\psi_j(\mathbf{x}_0) = \int_{\partial\Omega} \left[\Gamma_{ij}(\mathbf{x}, \mathbf{x}_0)\, \psi_i(\mathbf{x}) - \Phi_{ij}(\mathbf{x}, \mathbf{x}_0)\, P^{[\psi]}_i(\mathbf{x})\right] ds(\mathbf{x}) \tag{17}$$

where $\mathbf{x}_0$ is the source point, $\eta = 0$ if $\mathbf{x}_0 \notin \Omega \cup \partial\Omega$, $\eta = 1$ if $\mathbf{x}_0 \in \Omega$ and $\eta = \frac{1}{2}$ if $\mathbf{x}_0 \in \partial\Omega$ and $\partial\Omega$ has a continuously turning tangent at $\mathbf{x}_0$. The Φ_{ij} in (17) is any solution of the equation

$$\left[\lambda^{(0)} \delta_{ij} \Phi_{km,k} + \mu^{(0)} \left(\Phi_{im,j} + \Phi_{jm,i}\right)\right]_{,j} = -\delta_{im} \delta(\mathbf{x} - \mathbf{x}_0)$$

where δ_{im} is known as the Kronecker delta and the Γ_{ij} is given by

$$\Gamma_{im} = \left[\lambda^{(0)} \delta_{ij} \Phi_{km,k} + \mu^{(0)} \left(\Phi_{im,j} + \Phi_{jm,i}\right)\right] n_j$$

For the three dimensional case

$$\Phi_{ij} = \frac{1}{16\pi\mu^{(0)}(1-\nu)d}\left[(3-4\nu)\delta_{ij}+d_{,i}d_{,j}\right] \tag{18}$$

$$\Gamma_{ij} = -\frac{1}{8\pi(1-\nu)d^2} \left[\frac{\partial d}{\partial n}\left\{(1-2\nu)\delta_{ij}+3d_{,i}d_{,j}\right\}+(1-2\nu)\left(n_i d_{,j}-n_j d_{,i}\right)\right] \tag{19}$$

and for two dimensional case

$$\Phi_{ij} = \frac{1}{8\pi\mu^{(0)}(1-\nu)}\left[(3-4\nu)\log\frac{1}{d}\delta_{ij}+d_{,i}d_{,j}\right] \tag{20}$$

$$\Gamma_{ij} = -\frac{1}{4\pi(1-\nu)d} \left[\frac{\partial d}{\partial n}\left\{(1-2\nu)\delta_{ij}+2d_{,i}d_{,j}\right\}+(1-2\nu)\left(n_i d_{,j}-n_j d_{,i}\right)\right] \tag{21}$$

where $d=\|\mathbf{x}-\mathbf{x}_0\|$, $\nu=\lambda^{(0)}/(2(\mu^{(0)}+\lambda^{(0)}))$ and $\partial d/\partial n = d_{,k}n_k$.

Use of (6) and (15) in (17) yields

$$\eta(\mathbf{x}_0)\, g^{1/2}(\mathbf{x}_0)\, u_j(\mathbf{x}_0) = \int_{\partial\Omega}\Big\{u_i(\mathbf{x})\left[g^{1/2}(\mathbf{x})\,\Gamma_{ij}(\mathbf{x},\mathbf{x}_0)-P^{[g]}_{ki}(\mathbf{x})\,\Phi_{kj}(\mathbf{x},\mathbf{x}_0)\right]- P_i(\mathbf{x})\left[g^{-1/2}(\mathbf{x})\,\Phi_{ij}(\mathbf{x},\mathbf{x}_0)\right]\Big\}\,ds(\mathbf{x}) \tag{22}$$

This equation provides a boundary integral equation for determining u_i and σ_{ij} at all points of Ω.

5 A perturbation method

The boundary element procedure described in the previous section provides an effective numerical method for determining $u_i(\mathbf{x})$ when $g(\mathbf{x})$ takes the form (12) and the parameters $\lambda^{(0)}$ and $\mu^{(0)}$ satisfy the relation (13). In this section a procedure is obtained for the case when the coefficients $\lambda(\mathbf{x})$ and $\mu(\mathbf{x})$ are perturbed about $\lambda^{(0)}g(\mathbf{x})$ and $\mu^{(0)}g(\mathbf{x})$ respectively while retaining equations (12) and (13).
The coefficients $\lambda(\mathbf{x})$ and $\mu(\mathbf{x})$ are required to take the form

$$\lambda(\mathbf{x}) = \lambda^{(0)}g(\mathbf{x})+\epsilon\lambda^{(1)}(\mathbf{x}) \tag{23}$$

$$\mu(\mathbf{x}) = \mu^{(0)}g(\mathbf{x})+\epsilon\mu^{(1)}(\mathbf{x}) \tag{24}$$

with

$$\lambda^{(0)}=\mu^{(0)} \quad \text{and} \quad g^{1/2}_{,ij}=0$$

and where $\lambda^{(1)}$ and $\mu^{(1)}$ are twice differentiable functions. Therefore from (3)

$$\left\{g\left[\lambda^{(0)}\delta_{ij}u_{k,k}+\mu^{(0)}(u_{i,j}+u_{j,i})\right]\right\}_{,j} = -\epsilon\left[\lambda^{(1)}\delta_{ij}u_{k,k}+\mu^{(1)}(u_{i,j}+u_{j,i})\right]_{,j} \tag{25}$$

Use of the transformation (6) and following the analysis used to derive (11) from (5) gives

$$\left[\lambda^{(0)}\delta_{ij}\psi_{k,k} + \mu^{(0)}(\psi_{i,j} + \psi_{j,i})\right]_{,j} = -\epsilon g^{-1/2}\left[\lambda^{(1)}\delta_{ij}u_{k,k} + \mu^{(1)}(u_{i,j} + u_{j,i})\right]_{,j} \tag{26}$$

A solution to equation (26) is sought in the form

$$\psi_i(\mathbf{x}) = \sum_{r=0}^{\infty} \epsilon^r \psi_i^{(r)}(\mathbf{x}) \tag{27}$$

From (6) and (27) the displacement u_k may also be written in a series form as follows

$$u_k(\mathbf{x}) = \sum_{r=0}^{\infty} \epsilon^r u_k^{(r)}(\mathbf{x}) \tag{28}$$

where $u_k^{(r)}$ corresponds to $\psi_k^{(r)}$ according to the relationship

$$\psi_k^{(r)} = g^{1/2}\, u_k^{(r)} \tag{29}$$

Substitution of (27) into (26) and equating the coefficients of powers of ϵ yields

$$\left[\lambda^{(0)}\delta_{ij}\psi_{k,k}^{(r)} + \mu^{(0)}\left(\psi_{i,j}^{(r)} + \psi_{j,i}^{(r)}\right)\right]_{,j} = h^{(r)} \qquad \text{for } r = 0, 1, \ldots, \tag{30}$$

where

$$h^{(0)}(\mathbf{x}) = \mathbf{0}, \tag{31}$$

$$h^{(r)}(\mathbf{x}) = -g^{-1/2}\left[\lambda^{(1)}\delta_{ij}u_{k,k}^{(r-1)} + \mu^{(1)}(u_{i,j}^{(r-1)} + u_{j,i}^{(r-1)})\right]_{,j} \qquad \text{for } r = 1, 2, \ldots \tag{32}$$

The integral equation for (30) is

$$\begin{aligned}\eta(\mathbf{x}_0)\,\psi_j^{(r)}(\mathbf{x}_0) \;=\; & \int_{\partial\Omega}\left[\Gamma_{ij}(\mathbf{x},\mathbf{x}_0)\;\psi_i^{(r)}(\mathbf{x}) - \Phi_{ij}(\mathbf{x},\mathbf{x}_0)\;P_i^{[\psi^{(r)}]}(\mathbf{x})\right] ds(\mathbf{x}) \\ & + \int_{\Omega} h_i^{(r)}(\mathbf{x})\;\Phi_{ij}(\mathbf{x},\mathbf{x}_0)\;dS(\mathbf{x}) \qquad \text{for } r = 0, 1, \ldots\end{aligned} \tag{33}$$

where

$$P_i^{[\psi^{(r)}]} = \left[\lambda^{(0)}\delta_{ij}\psi_{k,k}^{(r)} + \mu^{(0)}\left(\psi_{i,j}^{(r)} + \psi_{j,i}^{(r)}\right)\right] n_j$$

Also

$$P_i^{[\psi^{(r)}]} = g^{1/2}P_i^{(r)} + u_k^{(r)}P_{ik}^{[g]} \qquad \text{for } r = 0, 1, \ldots \tag{34}$$

where

$$P_i^{(r)}(\mathbf{x}) = \left[\lambda^{(0)}\delta_{ij}u_{k,k}^{(r)} + \mu^{(0)}\left(u_{i,j}^{(r)} + u_{j,i}^{(r)}\right)\right] n_j \tag{35}$$

and $P_{ik}^{[g]}$ is given by (16). Thus the integral equation (33) may be written in the form

$$\eta(\mathbf{x}_0)\, g^{1/2}(\mathbf{x}_0)\, u_j^{(r)}(\mathbf{x}_0) = \int_{\partial\Omega} \Big\{ u_i^{(r)}(\mathbf{x}) \Big[g^{1/2}(\mathbf{x})\, \Gamma_{ij}(\mathbf{x}, \mathbf{x}_0) - P_{ki}^{[g]}(\mathbf{x})\, \Phi_{kj}(\mathbf{x}, \mathbf{x}_0) \Big] - P_i^{(r)}(\mathbf{x}) \Big[g^{1/2}(\mathbf{x})\, \Phi_{ij}(\mathbf{x}, \mathbf{x}_0) \Big] \Big\}\, ds(\mathbf{x}) + \int_{\Omega} h_i^{(r)}(\mathbf{x})\, \Phi_{ij}(\mathbf{x}, \mathbf{x}_0)\, dS(\mathbf{x}) \tag{36}$$

Now, the corresponding value of P_i may be written as

$$P_i = g P_i^{(0)} + \sum_{r=1}^{\infty} \epsilon^r \left(g P_i^{(r)} + G_i^{(r)} \right) \tag{37}$$

where

$$G_i^{(r)}(\mathbf{x}) = \left[\lambda^{(1)} \delta_{ij} u_{k,k}^{(r-1)} + \mu^{(1)} (u_{i,j}^{(r-1)} + u_{j,i}^{(r-1)}) \right] n_j$$

To satisfy the boundary conditions in Section 3 it is required that

$$\begin{aligned} u_i^{(0)} &= u_i && \text{on } \partial\Omega_1 \\ P_i^{(0)} &= g^{-1} P_i && \text{on } \partial\Omega_2 \end{aligned}$$

where u_i takes on its specified value on $\partial\Omega_1$ and P_i takes on its specified value on $\partial\Omega_2$. It then follows from (28) and (37) that for $r = 1, 2, \ldots$

$$\begin{aligned} u_i^{(r)} &= \mathbf{0} && \text{on } \partial\Omega_1 \\ P_i^{(r)} &= -g^{-1} G_i^{(r)} && \text{on } \partial\Omega_2 \end{aligned}$$

The integral equation (36) may now be used to find the numerical values of the unknowns on the boundary $\partial\Omega$ and the numerical values of $u_i^{(r)}$ and derivatives in the domain Ω for $r = 0, 1, \ldots$. Equations (28) and (37) then provide the values of u_i and P_i throughout the domain Ω.

6 Numerical results

In this section some particular boundary value problems in plane strain and plane stress are solved numerically by employing the integral equations obtained in Sections 4 and 5. In implementing this method to obtain numerical solutions standard boundary element procedure is employed (see for example Clements [2]). For the chosen variations in the elastic parameters of the forms (23) and (24) the right hand side of (32) is small so that it is only necessary to retain two terms in the expression (28).

To switch from plane strain problems to plane stress problems the elastic modulus E and the Poisson raion ν must be transformed as follows

$$E \Longleftrightarrow E \left(1 - \frac{\nu^2}{(1+\nu)^2} \right) \qquad \nu \Longleftrightarrow \frac{\nu}{(1+\nu)}$$

Problem 1 : Extension of a constrained slab
Consider the boundary value problem given in Figure 1 for a material with elastic coefficients

$$\lambda(\mathbf{x}) = 1.2\lambda^{(0)}(1+0.1x_1')^2 \quad (38)$$
$$\mu(\mathbf{x}) = \lambda^{(0)}(1+0.1x_1')^2 \quad (39)$$

where $\lambda^{(0)}$ is a reference elastic modulus and $x_1' = x_1/l$. The elastic coefficients

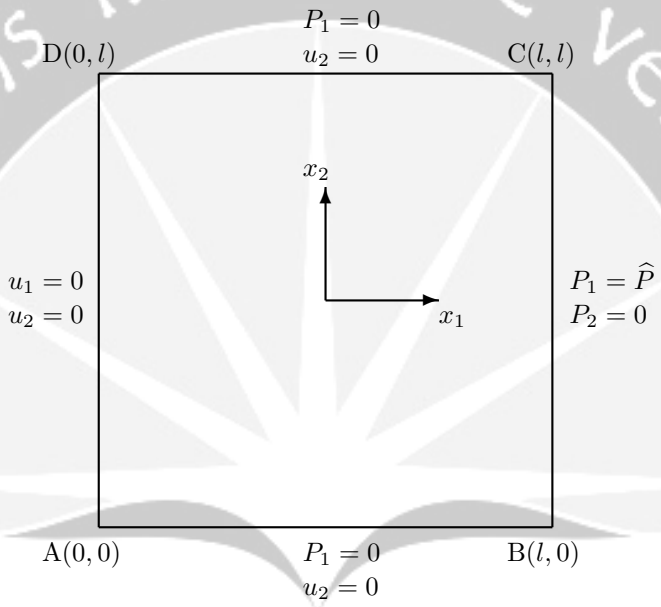


Figure 1: The geometry for Problems 1

(38) and (39) take the forms (23) and (24) with $g(\mathbf{x}) = (1+0.1x_1')^2$, $\mu^{(0)} = \lambda^{(0)}$, $\lambda^{(1)} = \lambda^{(0)}(1+0.1x_1')^2$, $\mu^{(1)} = 0$ and $\epsilon = 0.2$. The boundary conditions (see Figure 1) are

$P_1/\widehat{P} = 0$	$u_2/\widehat{u} = 0$	on AB
$P_1/\widehat{P} = 1$	$P_2/\widehat{P} = 0$	on BC
$P_1/\widehat{P} = 0$	$u_2/\widehat{u} = 0$	on CD
$u_1/\widehat{u} = 0$	$u_2/\widehat{u} = 0$	on AD

where $\widehat{u}$ is a reference displacement and $\widehat{P} = \lambda^{(0)}\widehat{u}/l$.

This problem admits the analytical solution $u_1/\widehat{u} = x_1'/[3.2(1+0.1x_1')]$, $u_2 = 0$ with the stress given by $\sigma_{11}/\widehat{P} = 1$, $\sigma_{12}/\widehat{P} = 0$ and $\sigma_{22}/\widehat{P} = 0.375$.

Table 1 – Table 4 show the analytical and BEM results for some points in the domain Ω and for the cases when the boundary $\partial\Omega$ is divided into 40, 80 and 160 segments. The results converge to the known solution as the number of segments increases. The displacement displays fourth figure and the stress third figure accuracy when 160 boundary segments are used.

Table 1: Displacements for Problem 1

Position	BEM 40 segments		BEM 80 segments		BEM 160 segments	
(x_1', x_2')	$u_1/\widehat{u}$	$u_2/\widehat{u}$	$u_1/\widehat{u}$	$u_2/\widehat{u}$	$u_1/\widehat{u}$	$u_2/\widehat{u}$
(0.1,0.5)	0.0289	0.0000	0.0299	0.0000	0.0304	0.0000
(0.3,0.5)	0.0881	-0.0002	0.0896	-0.0001	0.0903	0.0000
(0.5,0.5)	0.1451	-0.0002	0.1470	0.0000	0.1479	0.0000
(0.7,0.5)	0.2002	-0.0001	0.2023	0.0000	0.2033	0.0000
(0.9,0.5)	0.2530	-0.0001	0.2557	0.0000	0.2568	0.0000

Table 2: Displacements for Problem 1

Position	Analytical	
(x_1', x_2')	$u_1/\widehat{u}$	$u_2/\widehat{u}$
(0.1,0.5)	0.0309	0.0000
(0.3,0.5)	0.0910	0.0000
(0.5,0.5)	0.1488	0.0000
(0.7,0.5)	0.2044	0.0000
(0.9,0.5)	0.2580	0.0000

Table 3: Stresses for Problem 1

Position	BEM 40 segments			BEM 80 segments		
(x_1', x_2')	$\sigma_{11}/\widehat{P}$	$\sigma_{12}/\widehat{P}$	$\sigma_{22}/\widehat{P}$	$\sigma_{11}/\widehat{P}$	$\sigma_{12}/\widehat{P}$	$\sigma_{22}/\widehat{P}$
(0.1,0.5)	0.9900	0.0002	0.3767	0.9956	0.0000	0.3756
(0.3,0.5)	0.9907	0.0002	0.3822	0.9953	0.0001	0.3782
(0.5,0.5)	0.9926	0.0005	0.3822	0.9960	0.0001	0.3784
(0.7,0.5)	0.9952	0.0003	0.3809	0.9972	0.0001	0.3778
(0.9,0.5)	0.9467	0.0000	0.4195	0.9978	0.0001	0.3787

Problem 2 : Another extension of a constrained slab
Now consider the boundary value problem given in Figure 2 with the coefficients $\lambda(\mathbf{x})$ and $\mu(\mathbf{x})$ again given by (38) and (39). The boundary conditions (see Figure

Table 4: Stresses for Problem 1

Position	BEM 160 segments			Analytical		
(x_1', x_2')	$\sigma_{11}/\widehat{P}$	$\sigma_{12}/\widehat{P}$	$\sigma_{22}/\widehat{P}$	$\sigma_{11}/\widehat{P}$	$\sigma_{12}/\widehat{P}$	$\sigma_{22}/\widehat{P}$
(0.1,0.5)	0.9978	0.0000	0.3751	1.0000	0.0000	0.3750
(0.3,0.5)	0.9975	0.0000	0.3764	1.0000	0.0000	0.3750
(0.5,0.5)	0.9978	0.0000	0.3766	1.0000	0.0000	0.3750
(0.7,0.5)	0.9984	0.0000	0.3763	1.0000	0.0000	0.3750
(0.9,0.5)	0.9992	-0.0002	0.3764	1.0000	0.0000	0.3750

2) are

$$\begin{array}{lll} u_1/\widehat{u} = x_1' & u_2/\widehat{u} = 0 & \text{on AB} \\ u_1/\widehat{u} = 1 & u_2/\widehat{u} = 0 & \text{on BC} \\ u_1/\widehat{u} = x_1' & u_2/\widehat{u} = 0 & \text{on CD} \\ u_1/\widehat{u} = 0 & u_2/\widehat{u} = 0 & \text{on AD} \end{array}$$

There is no explicit analytical solution to this particular problem.

Table 5 – Table 7 show the BEM results for some points in the domain Ω and for the cases when the boundary $\partial\Omega$ is divided into 40, 80 and 160 segments. As for Problem 1 the results converge as the number of boundary segments increases.

Table 5: Displacements for Problem 2

Position	BEM 40 segments		BEM 80 segments		BEM 160 segments	
(x_1', x_2')	$u_1/\widehat{u}$	$u_2/\widehat{u}$	$u_1/\widehat{u}$	$u_2/\widehat{u}$	$u_1/\widehat{u}$	$u_2/\widehat{u}$
(0.1,0.5)	0.1036	0.0000	0.1058	0.0000	0.1069	0.0000
(0.3,0.5)	0.3159	-0.0008	0.3175	-0.0002	0.3182	-0.0001
(0.5,0.5)	0.5210	-0.0008	0.5218	-0.0002	0.5220	0.0000
(0.7,0.5)	0.7188	-0.0006	0.7189	-0.0001	0.7185	0.0000
(0.9,0.5)	0.9091	-0.0004	0.9093	0.0000	0.9084	-0.0001

Table 6: Stresses for Problem 2

Position	BEM 40 segments			BEM 80 segments		
(x_1', x_2')	$\sigma_{11}/\widehat{P}$	$\sigma_{12}/\widehat{P}$	$\sigma_{22}/\widehat{P}$	$\sigma_{11}/\widehat{P}$	$\sigma_{12}/\widehat{P}$	$\sigma_{22}/\widehat{P}$
(0.1,0.5)	3.4965	0.0000	1.3159	3.4827	0.0000	1.2940
(0.3,0.5)	3.4481	0.0009	1.3084	3.4259	0.0003	1.2793
(0.5,0.5)	3.3983	0.0024	1.3066	3.3763	0.0005	1.2807
(0.7,0.5)	3.3489	0.0014	1.3045	3.3285	0.0002	1.2827
(0.9,0.5)	3.1544	0.0005	1.3932	3.2825	0.0006	1.2601

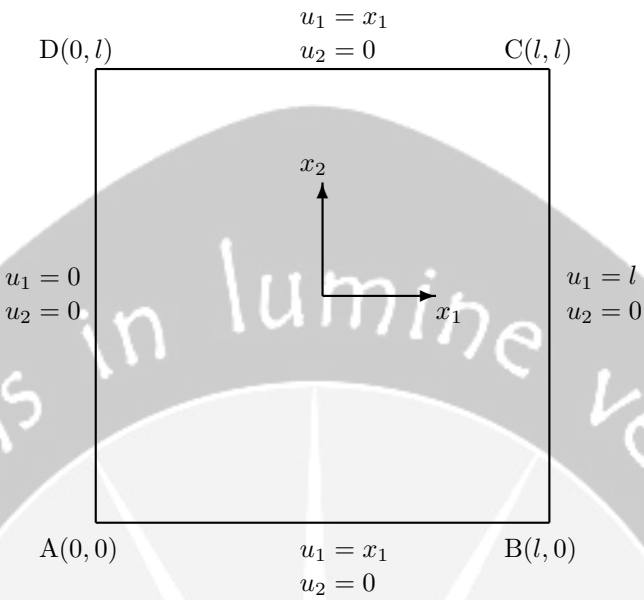


Figure 2: The geometry for Problems 2

Problem 3 : Compression of a slab
Consider the boundary value problem for an isotropic inhomogeneous material of square shape as shown by Figure 3. The square is equally loaded along the top side, and is clamped along the bottom side. The left-hand and right-hand sides are free. The elastic coefficients of the material are given by

$$\lambda'(\mathbf{x}) = 1.5(1 + \alpha x_1' + \beta x_2')^2 \quad (40)$$
$$\mu'(\mathbf{x}) = (1 + \alpha x_1' + \beta x_2')^2 \quad (41)$$

and the boundary conditions are

$u_1' = 0$	$u_2' = 0$	pada AB
$P_1' = 0$	$P_2' = 0$	pada BC
$P_1' = 0$	$P_2' = -1$	pada CD
$P_1' = 0$	$P_2' = 0$	pada AD

The elastic coefficients (40) and (41) dictate the form (23) and (24) with $g(\mathbf{x}) = (1 + \alpha x_1' + \beta x_2')^2$, $\lambda^{(0)} = \mu^{(0)} = \overline{\lambda}$, $\lambda^{(1)} = \overline{\lambda}(1 + \alpha x_1' + \beta x_2')^2$, $\mu^{(1)} = 0$ and $\epsilon = 0.5$. If $\overline{\lambda} = 2.49 \times 10^3$ ksi then the material's elastic coefficients under consideration are the coefficients for a magnesium alloy.

Four cases concerning the material's elastic coefficients λ and μ will be considered. The first case is the case for a homogeneous material (ie. when $\alpha = \beta = 0$). The

Table 7: Stresses for Problem 2

Position	BEM 160 segments		
(x_1', x_2')	$\sigma_{11}/\widehat{P}$	$\sigma_{12}/\widehat{P}$	$\sigma_{22}/\widehat{P}$
(0.1,0.5)	3.4699	0.0000	1.2852
(0.3,0.5)	3.4143	0.0001	1.2670
(0.5,0.5)	3.3643	0.0002	1.2688
(0.7,0.5)	3.3166	0.0002	1.2715
(0.9,0.5)	3.2698	0.0000	1.2498

other cases are inhomogeneous material cases which are when $\alpha = 0, \beta = 0.1$; $\alpha = 0.1, \beta = 0$; $\alpha = 0.1, \beta = 0.1$.

Figure 4 shows results of the deformation of the square boundary and figure 6 shows results of the deformation of the region $-0.25 \leq x_1' \leq 0.25$, $-0.25 \leq x_2' \leq 0.25$ inside the square. These two figures indicate the effect of the inhomogeneity function g on the displacements. The new coordinate system $OX_1'X_2'$ in figures 4 and 6 is the system for deformed object, where the coordinate variables are defined by $X_i' = x_i' + u_i'$ for $i = 1, 2$.

7 Summary

Boundary element methods for static elasticity problems of a class of inhomogeneous isotropic materials has been derived. The methods are generally easy to implement to obtain numerical values for particular problems. They can be applied to a wide class of important practical problems for inhomogeneous isotropic materials. The numerical results obtained using the methods indicate that they can provide accurate numerical solutions.

References

[1] C.A. Brebbia & J. Dominguez (1989), *Boundary Elements An Introductory Course*, Computational Mechanics Publications, Boston.

[2] D.L. Clements (1981), *Boundary Value Problems Governed by Second Order Elliptic Systems*, Pitman.

[3] D.L. Turcotte & P. Schubert (1982), *Geodynamic Applications of Continuum Physics to Geological Problems*, Wiley, New York.

[4] Manolis, R. P. and Shaw, R. P. (1996), Green's function for the vector wave equation in a mildly heterogeneous continuum, *Wave Motion*, **24**, 59–83.

[5] Rizzo, F. J. (1967), An Integral Equation Approach to Boundary Value Problems of Classical Elastostatics, *Quarterly of Applied Mathematics*, **25**, 83–95.

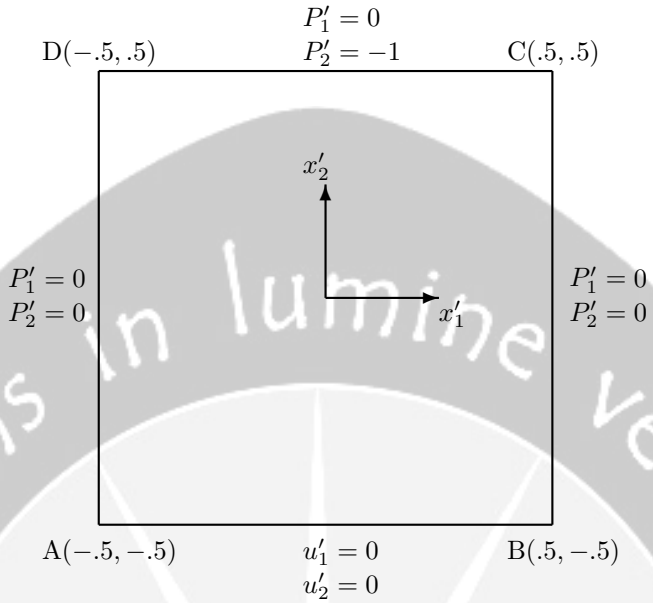


Figure 3: The geometry for Problem 3

M.I. Azis: Department of Mathematics, Hasanuddin University, Jl. P. Kemerdekaan km. 10 Tamalanrea Makassar 90245, Indonesia.
Phone/Fax: +62 +411 585 643
E-mail: ivan@unhas.ac.id Website: www.unhas.ac.id\~ivan

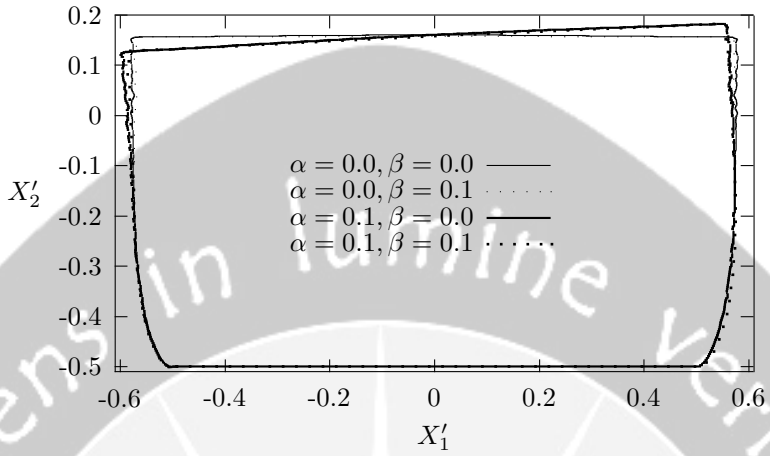


Figure 4: Deformation along the boundary of the square

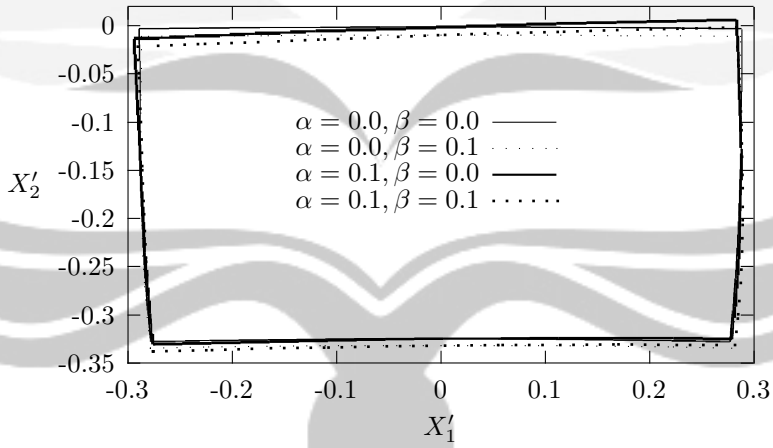


Figure 5: Deformation along the boundary of the region $-.25 \leq x_1' \leq .25$, $-.25 \leq x_2' \leq .25$ inside the square

THE WAVEFORM-RELAXATION METHOD FOR SOLVING FORWARD-BACKWARD STOCHASTIC DIFFERENTIAL EQUATIONS

Bevina D. Handari[a]

[a] The University of Indonesia, Indonesia

Abstract. One of the most appealing features of Forward-Backward Stochastic Differential Equations (FBSDEs) is that they can be applied to finance problems and give deep insights into them. However, the availability of numerical methods for solving these problems is still very limited. In this paper we propose the Waveform-Relaxation (WR) method for solving FBSDEs problems which is surely convergent and demonstrate their performance by windowing on a number of simulations. In this method, the adapted solution as the result of the application of the Four Step Scheme on the problem is needed in order to implement the WR scheme.

Key-words: forward-backward stochastic differential equations, four step scheme, waveform-relaxation method, windowing technique.

1 Introduction

This paper will focus on Forward-Backward stochastic differential equations (FBSEDs) problems and their efficient numerical methods. The areas of applications of FBSDEs include applied and theoretical areas such as stochastic control, mathematical finance, differential geometry, etc. [7]. One of the most appealing features of FBSDEs is that they can be applied to finance problems and give deep insights into them [7]. However, the availability of numerical methods to solve FBSDEs problems is still very limited; whilst, generally the explicit solution for this problem is not available. Thus, the aim of this paper is to introduce the appropriate numerical method for FBSDEs problems.

An example of an application of FBSDEs is the Stock-Sale Advertising Response model by Grosset [4]. FBSDEs consist of a system of forward SDE and backward SDE. Backward SDEs (BSDEs) are terminal value problems of SDEs involving the Itô stochastic integral. A thorough discussion on BSDEs can be found in [8]. An established method for solving a (coupled) FBSDE, known as the Four Step Scheme [7].

To be specific, this paper is organized as follows: At second section, the theory of FBSDEs will be reviewed, including material on adapted and unique solutions of FBSDEs and the possibility that FBSDEs are not solvable. Afterwards, an established method for solving FBSDEs, the Four Step Scheme, will be discussed in section 3. Section 4 discusses the Waveform- Relaxtion(WR) scheme for FBSDEs. Section 5 discusses the convergence of the WR method for FBSDEs. Numerical results will be given in section 6 and finally the conclusion in the last section.

2 Forward-Backward SDEs (FBSDEs)

A system of FBSDEs is a generalisation of the two-point boundary value problem for an ordinary differential system. FBSDEs are defined on a complete filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, P)$ on which an m-dimensional standard Wiener process is defined and $\{\mathcal{F}_t\}_{t\geq 0}$ is the natural filtration of the Wiener process. A FBSDE has the form [8]

$$\left\{\begin{array}{lcl} dx(t) &=& b(t,x(t),y(t),z(t))dt + \sigma(t,x(t),y(t),z(t))dW(t), \\ dy(t) &=& h(t,x(t),y(t),z(t))dt + z(t)dW(t), \\ x(0) &=& x_0, \quad y(T) = g(x(T)), \end{array}\right. \tag{1}$$

where x, y, z are unknown processes with non random functions b, σ, h, g given. In FBSDEs case, the process $x(t)$ satisfies a forward SDE and the process $y(t)$ satisfies a BSDE. The process $z(t)$ is needed to find the $\{\mathcal{F}_t\}_{t\geq 0}$-adapted process (x, y).

Definifion 2.1. *A triple of stochastic processes $(x, y, z) \in \hat{\mathcal{M}}[0, T]$ is called an (adapted) solution of (1) if it satisfies*

$$\left\{\begin{array}{lcl} x(t) &=& x + \int_0^t b(s,x(s),y(s),z(s))ds + \int_0^t \sigma(s,x(s),y(s),z(s))dW(s), \\ y(t) &=& g(x(T)) - \int_t^T h(s,x(s),y(s),z(s))ds - \int_t^T z(s)dW(s), \forall t \in [0,T], \mathbf{P}-a.s., \end{array}\right.$$

where $\hat{\mathcal{M}}[0,T] \triangleq L^2_{\mathcal{F}_t}(\Omega; C([0,T]; R^n)) \times L^2_{\mathcal{F}_t}(\Omega; C([0,T]; R^k)) \times L^2_{\mathcal{F}_t}(0,T; R^{k\times m})$.

A unique solution of the FBSDEs (1) is any two adapted solutions (x, y, z) and $(\tilde{x}, \tilde{y}, \tilde{z})$ which satisfy

$$P\{(x(t), y(t)) = (\tilde{x}(t), \tilde{y}(t)), \ \forall t \in [0,T] \ \text{ and } \ z(t) = \tilde{z}(t), \ a.e.\ t \in [0,T]\} = 1.$$

Since FBSDEs are a generalisation of the two-point boundary value problem for an ordinary differential system, then a FBSDE is not necessarily solvable. This condition is supported by the following proposition.

Proposition 2.1. *Let the following two-point boundary value problem for a system of linear ordinary differential equations admit no solutions:*

$$\left\{\begin{array}{l} \begin{pmatrix} \dot{x}(t) \\ \dot{y}(t) \end{pmatrix} = \mathcal{A} \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}, \\ x(0) = x_0, \quad y(T) = Sx(T), \end{array}\right.$$

where $\mathcal{A}$ and S are certain matrices. Then, for any bounded $\sigma : [0,T] \times R^n \times R^k \times R^{k\times m} \to R^{n\times m}$, the FBSDE

$$\left\{\begin{array}{l} d\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \mathcal{A} \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} dt + \begin{pmatrix} \sigma(t,x(t),y(t),z(t)) \\ z(t) \end{pmatrix} dW(t), \\ x(0) = x_0, \quad y(T) = Sx(T), \end{array}\right.$$

does not have an adapted solution.
This proposition and its proof can be found in [8].

The FBSDE problem chosen as a test problem is given by [7]

$$\begin{cases} dx(t) &= \frac{x(t)}{(z(t)-y(t))^2+1}dt + x(t)dW(t), \\ dy(t) &= \frac{z(t)}{(z(t)-y(t))^2+1}dt + z(t)dW(t), \\ x(0) &= x_0, \\ y(T) &= x(T), \end{cases} \tag{2}$$

with analytic solution given by

$$x(t) = y(t) = z(t) = \exp\{W(t) + t/2\}x_0, \forall t \in [0,T].$$

The general form of FBSDEs (2) is given by

$$\begin{cases} dx(t) &= b(t,x(t),y(t),z(t))dt + \sigma(t,x(t))dW(t), \\ dy(t) &= h(t,x(t),y(t),z(t))dt \ + z(t)dW(t), \\ x(0) &= x_0, \\ y(T) &= g(x(T)). \end{cases} \tag{3}$$

The following theorem assures when the FBSDE (3) has a unique adapted solution.

Theorem 2.2.
Suppose that the conditions (FB1),(FB2) and (FB3) hold. Then (3) admits a unique adapted solution (x,y,z).

The three conditions (FB1), (FB2) and (FB3) are given by
FB1: Consider that $m = n$ in (3) and the functions b, σ, h and g are smooth functions taking values in $R^n, R^{n\times n}, R^k$ and R^k, respectively, and their first-order derivatives in x, y, z are all bounded uniformly by some constant $L > 0$.
FB2: There exists positive constants ν, μ such that

$$\nu I \le \sigma(t,x)\sigma(t,x)^T \le \mu I, \ \ \forall (t,x) \in [0,T] \times R^n,$$

$$|\ b(t,x,y,z)\ |, |\ h(t,x,0,0)\ | \le \mu, \forall (t,x,y,z) \in [0,T] \times R^n \times R^k \times R^{k\times n}.$$

FB3: The function g is bounded in $C^{2+\alpha}(R^n, R^k)$ for some $\alpha \in (0,1)$.
Theorem 2.2 and proof can be found in [8].

3 The Four Step Scheme

A method to solve the FBSDE (1) over any time duration $[0,T]$ is the Four Step Scheme. This method was derived by X.Y. Zhou [8] to find an adapted solution (x,y,z) of (1). This method is given as follows:
Assume that y and x are related by

$$y(t) = \theta(t,x(t)), \ \ \forall t \in [0,T], \text{a.s. } \mathbf{P},$$

where the function θ will be determined. The value of x, y, z in (1) can be obtained by the following steps:
Step 1: Find a function z that satisfies

$$z(t,x,y,p) = p\sigma(t,x,y,z(t,x,y,p)), \tag{4}$$

$\forall (t,x,y,p) \in [0,T] \times R^n \times R^m \times R^{m\times n}$.
Step 2: Use z to solve this quasilinear partial differential equation

$$\left\{\begin{array}{rcl} \theta_t^k & + & \frac{1}{2}tr[\theta_{xx}^k(\sigma\sigma^T)(t,x,\theta,z(t,x,\theta,\theta_x))] \\ & + & < b(t,x,\theta,z(t,x,\theta,\theta_x)),\theta_x^k > \\ & - & h^k(t,x,\theta,z(t,x,\theta,\theta_x)) = 0, \\ & & (t,x) \in [0,T) \times R^n, \ \ 1 \le k \le m \\ \theta(T,x) & = & g(x), \ \ x \in R^n. \end{array}\right. \tag{5}$$

Here, $< A,B > \stackrel{\triangle}{=} tr\{AB^T\}, \ \forall \ A,B \in R^{k\times m}$.
Step 3: Use θ and z to solve the forward SDE

$$x(t) = x + \int_0^t \tilde{b}(s,x(s))ds + \int_0^t \tilde{\sigma}(s,x(s))dW(s), \tag{6}$$

where

$$\left\{\begin{array}{rcl} \tilde{b}(t,x) & = & b(t,x,\theta(t,x),z(t,x,\theta(t,x),\theta_x(t,x))), \\ \tilde{\sigma}(t,x) & = & \sigma(t,x,\theta(t,x),z(t,x,\theta(t,x),\theta_x(t,x))). \end{array}\right.$$

Step 4: Set

$$\left\{\begin{array}{rcl} y(t) & = & \theta(t,x(t)), \\ z(t) & = & z(t,x(t),\theta(t,x(t)),\theta_x(t,x(t))), \end{array}\right. \tag{7}$$

where (x,y,z) is an adapted solution.
The theorem that assures if the above scheme is realizable then the adapted solution will be a unique solution can be found in [8].

Now, we want to apply the Four Step Scheme to the following FBSDE problem which is a generalized form of (2):

$$\left\{\begin{array}{rcl} dx(t) & = & b(t)x(t)dt + x(t)dW(t), \\ dy(t) & = & b(t)z(t)d(t) + z(t)dW(t), \\ x(0) & = & x_0, \\ y(T) & = & g(x(T)), \end{array}\right. \tag{8}$$

where $g(x(T)) = ax(T) + b$, a and b are constants. The result of the application will be used later when solving the test problem numerically.

Step 1: Comparing (8) with (1), then the function z that satisfies this step is

$$z(t,x,y,p) = px(t), \ \ \forall (t,x,y,p) \in [0,T] \times R \times R \times R, \tag{9}$$

since $\sigma(t,x,y,z(t,x,y,p)) = x(t)$ in (8).
Step 2: From (8), we have

$$\sigma(t,x,\theta,z(t,x,\theta,\theta_x)) = x(t),$$

$$b(t,x,\theta,z(t,x,\theta,\theta_x)) = b(t)x(t),$$
$$h(t,x,\theta,z(t,x,\theta,\theta_x)) = b(t)z(t).$$

We use these values and z in (9) to form the quasilinear partial differential equation

$$\begin{cases} \theta_t + \frac{1}{2}x^2\theta_{xx} = 0, \quad (t,x) \in [0,T) \times R, \\ \theta|_{t=T} = ax + b, \quad x \in R. \end{cases} \tag{10}$$

To solve this system, we let $\xi \stackrel{\triangle}{=} ln\ x$ and $\varphi(t,\xi) \stackrel{\triangle}{=} \theta(t,e^{\xi})$. By using the chain rule, equation (10) can be written as

$$\begin{cases} \varphi_t + \frac{1}{2}(\varphi_{\xi\xi} - \varphi_{\xi}) = 0, \quad (t,x) \in [0,T) \times R, \\ \varphi|_{t=T} = ae^{\xi} + b, \quad \xi \in R. \end{cases} \tag{11}$$

Now, let $s \stackrel{\triangle}{=} \gamma t$ and $\psi(s,\xi) \stackrel{\triangle}{=} e^{-\frac{\alpha s}{\gamma} - \beta\xi}\varphi(\frac{s}{\gamma},\xi)$, then by the chain rule equation (11) can be written as

$$\begin{cases} e^{\frac{\alpha s}{\gamma}+\beta\xi}\{\ \gamma\psi_s & + & \psi\alpha + \frac{1}{2}(\psi_{\xi\xi} + 2\psi_{\xi}\beta + \beta^2\psi) \\ & - & \frac{1}{2}(\psi_{\xi} + \psi\beta)\ \} = 0, \quad (s,\xi) \in [0,\gamma T) \times R, \\ \psi|_{s=\gamma T} & = & e^{-\alpha T - \beta\xi}(ae^{\xi} + b), \quad \xi \in R. \end{cases} \tag{12}$$

To transform (12) into the terminal value problem, equation (12) must satisfy the following conditions:

$$\begin{cases} \gamma\psi_s + \frac{1}{2}\psi_{\xi\xi} = 0, \\ \psi_{\xi}(\beta - \frac{1}{2}) = 0, \\ \psi(\alpha + \frac{\beta^2}{2} - \frac{\beta}{2}) = 0. \end{cases} \tag{13}$$

Conditions (13) will be satisfied if the following conditions hold:

$$\alpha = 1/8, \quad \beta = 1/2 \ \text{ and } \ \gamma = 1/2. \tag{14}$$

Assuming (14) holds, the terminal value problem of (12) can be written as

$$\begin{cases} \psi_s + \psi_{\xi\xi} = 0, \quad (s,\xi) \in [0,\gamma T) \times R, \\ \psi|_{s=\gamma T} = e^{-\alpha T - \beta\xi}(ae^{\xi} + b), \quad \xi \in R. \end{cases} \tag{15}$$

Now, we have to transform the terminal value problem (15) into the initial value problem by using the transformation $\bar{t} = \gamma T - s$ and let $\psi^{(1)}(\bar{t},\xi) = \psi(s,\xi)$ such that equation (15) becomes the initial value problem

$$\begin{cases} \psi^{(1)}_{\bar{t}} = \psi^{(1)}_{\xi\xi}, \quad (\bar{t},\xi) \in [0,\gamma T) \times R, \\ \psi|_{\bar{t}=0} = e^{-\alpha T - \beta\xi}(ae^{\xi} + b), \quad \xi \in R. \end{cases} \tag{16}$$

The solution of (16) is given by

$$\psi^{(1)}(\bar{t},\xi) = \int_{-\infty}^{\infty} K(\bar{t},\xi - \tau)e^{-\alpha T - \beta\tau}(ae^{\tau} + b)d\tau, \tag{17}$$

where K is defined by

$$K(t,y) = \frac{1}{\sqrt{4\pi t}} \exp\ \left\{-\frac{y^2}{4t}\right\},\ t > 0, \tag{18}$$

The solution (17) is obtained based on the following theorem [3].

Theorem 2.3. *For all piecewise-continuous f that satisfy*

$$\|f(y)\| \leq C_1 exp\{C_2\|y\|^{1+\alpha}\},\ \ 0 \leq \alpha < 1,$$

where C_1 and C_2 are positive constants with the function $K(t,y)$ defined by (18)

$$u(t,y) = \int_{\infty}^{\infty} K(t, y-\tau) f(\tau) d\tau,\ \ t > 0$$

is a solution of the initial-value problem

$$\left\{\begin{array}{rcl} u_t & = & u_{yy},\ \ -\infty < y < \infty,\ \ 0 < t, \\ u(0,y) & = & f(y),\ \ -\infty < y < \infty. \end{array}\right.$$

Now, we want to solve (17) by substituting the transformation $\bar{t} = \gamma T - s$ back into (17) and then (17) can be written as

$$\psi(s,\xi) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{4\pi(\gamma T - s)}} \exp\left\{-\frac{(\xi-\tau)^2}{4(\gamma T - s)}\right\} e^{-\alpha T - \beta\tau}(ae^{\tau} + b) d\tau,$$

which gives

$$\begin{array}{rcl} \psi(s,\xi) & = & \int_{-\infty}^{\infty} \frac{ae^{-\alpha T}}{\sqrt{4\pi(\gamma T - s)}} \exp\left\{-\frac{(\xi-\tau)^2}{4(\gamma T-s)} - \tau(\beta-1)\right\} d\tau \\ & + & \int_{-\infty}^{\infty} \frac{be^{-\alpha T}}{\sqrt{4\pi(\gamma T - s)}} \exp\left\{-\frac{(\xi-\tau)^2}{4(\gamma T-s)} - \tau\beta\right\} d\tau. \end{array} \tag{19}$$

First, we want to evaluate the first integral of the r.h.s of (19), that is

$$\psi_1(s,\xi) = \int_{-\infty}^{\infty} \frac{ae^{-\alpha T}}{\sqrt{4\pi(\gamma T - s)}} \exp\left\{-\frac{(\xi-\tau)^2}{4(\gamma T - s)} - \tau(\beta-1)\right\} d\tau. \tag{20}$$

After some algebraic arrangements, we can verify that the solution of (20) is

$$\psi_1(s,\xi) = a\ \exp\ (-\alpha T - (\beta-1)\xi + (\beta-1)^2(\gamma T - s)). \tag{21}$$

Since $\psi_1(s,\xi) \stackrel{\triangle}{=} e^{-\frac{\alpha s}{\gamma} - \beta\xi} \varphi_1(\frac{s}{\gamma}, \xi)$ and $s \stackrel{\triangle}{=} \gamma t$, equation (21) can be written as

$$\varphi_1(t,\xi) = ae^{\xi} e^{(t-T)(\alpha - \gamma(\beta-1)^2)}. \tag{22}$$

By substituting variables $\alpha = 1/8,\ \beta = 1/2$, and $\gamma = 1/2$ back into equation (22), this gives

$$\varphi_1(t,\xi) = ae^{\xi}. \tag{23}$$

Finally, since $\xi \stackrel{\triangle}{=} ln\ x$ and $\varphi_1(t,\xi) \stackrel{\triangle}{=} \theta_1(t,e^{\xi})$, equation (23) is given by

$$\theta_1(t,x) = ax. \tag{24}$$

Now, we want to evaluate the second integral of the r.h.s of (19), that is

$$\psi_2(s,\xi) = \int_{-\infty}^{\infty} \frac{be^{-\alpha T}}{\sqrt{4\pi(\gamma T - s)}} \exp\left\{ -\frac{(\xi-\tau)^2}{4(\gamma T - s)} - \tau\beta \right\} d\tau. \tag{25}$$

By comparing equation (25) with equation (20), the exponential term in (20) involves the term $\tau(\beta-1)$ and the term $\tau(\beta)$ in (25). Based on this difference and refering to equation (21), equation (25) can be written as

$$\psi_2(s,\xi) = b\ \exp(-\alpha T - \beta\xi + \beta^2(\gamma T - s)). \tag{26}$$

Since $\psi_2(s,\xi) \stackrel{\triangle}{=} e^{-\frac{\alpha s}{\gamma} - \beta\xi}\varphi_2(\frac{s}{\gamma},\xi)$ and $s \stackrel{\triangle}{=} \gamma t$, equation (26) can be written as

$$\varphi_2(t,\xi) = b\ \exp(t-T)(\alpha - \gamma\beta^2). \tag{27}$$

By substituting variables $\alpha = 1/8,\ \beta = 1/2$, and $\gamma = 1/2$ back to (27), gives

$$\varphi_2(t,\xi) = b, \tag{28}$$

and also since $\xi \stackrel{\triangle}{=} ln\ x$ and $\varphi_2(t,\xi) \stackrel{\triangle}{=} \theta_2(t,e^{\xi})$, equation (28) can be written as

$$\theta_2(t,x) = b. \tag{29}$$

Thus, the complete solution of (10) is given by

$$\theta(t,x) = \theta_1(t,x) + \theta_2(t,x), \tag{30}$$

which is equal to

$$\theta(t,x) = ax + b. \tag{31}$$

Step 3: Here, we use θ and z to solve the forward SDE

$$x(t) = x + \int_0^t \tilde{b}(s,x(s))ds + \int_0^t \tilde{\sigma}(s,s(s))dW(s), \tag{32}$$

where

$$\begin{cases} \tilde{b}(t,x) &= b(t,x,\theta(t,x),z(t,x,\theta(t,x),\theta_x(t,x))), \\ &= b(t)x(t), \\ \tilde{\sigma}(t,x) &= \sigma(t,x,\theta(t,x),z(t,x,\theta(t,x),\theta_x(t,x))), \\ &= x(t). \end{cases}$$

Thus, we look for the x solution, by solving the forward SDE

$$\begin{array}{rcl} dx(t) &=& b(t)x(t)dt + x(t)dW(t), \\ x(0) &=& x_0, \end{array} \tag{33}$$

which has the exact solution [5]

$$x(t) = x_0 \exp\left(\int_0^T (b(s) - \frac{1}{2})ds + \int_0^t dW(s)\right). \tag{34}$$

Step 4: Set

$$\begin{cases} y(t) &= \theta(t, x(t)), \\ &= ax(t) + b \\ z(t) &= z(t, x(t), \theta(t, x(t)), \theta_x(t, x(t))), \\ &= x(t)\theta_x(t, x(t)), \\ &= ax(t). \end{cases} \tag{35}$$

where (x, y, z) is an adapted solution. Note that from (35) we have $y(t) = g(x(t))$ and $z(t) = g'(x(t))x(t)$. The important result from the Four Step Scheme application is that we know the function z. This information is very useful when we want to implement a Waveform Relaxation method for solving FBSDEs.

4 The Waveform Relaxation (WR) scheme for FBSDEs

The proposed Waveform Relaxation (WR) method for solving FBSDE (8) has the form

$$\begin{cases} dx^{(k+1)}(t) &= b^{(k)}(t)x^{(k+1)}(t)dt + x^{(k+1)}(t)dW(t), \\ dy^{(k+1)}(t) &= b^{(k)}(t)z^{(k+1)}(t)dt + z^{(k+1)}(t)dW(t), \\ x^{(k+1)}(0) &= x_0, \\ y^{(k+1)}(T) &= g(x^{(k+1)}(T)). \end{cases} \tag{36}$$

where $b^{(k)}(t) = \frac{1}{(z^{(k)}(t) - y^{(k)}(t))^2 + 1}$. However, when we consider the iteration scheme

$$dy^{(k+1)}(t) = b^{(k)}(t)z^{(k+1)}(t)dt + z^{(k+1)}(t)dW(t),$$

in (36), the variable z is still implicit at each iteration. To solve this problem, we use the adapted solution (x, y, z) of the generalized form (8) in Section 3. From (35), the adapted solution for the generalized form (8) is given by

$$\begin{cases} x(t) &= x_0 \exp\left(\int_0^t (b(s) - \frac{1}{2})ds + \int_0^t dW(s)\right), \\ y(t) &= ax(t) + b, \\ z(t) &= ax(t). \end{cases}$$

From this adapted solution, we know that $z(t) = g'(x(t))x(t)$. By using this information, we can implement the WR scheme (36) which can rewritten as,

$$\begin{cases} dx^{(k+1)}(t) &= b^{(k)}(t)x^{(k+1)}(t)dt + x^{(k+1)}(t)dW(t). \\ dy^{(k+1)}(t) &= b^{(k)}(t)z^{(k+1)}(t)dt + z^{(k+1)}(t)dW(t). \\ x^{(k+1)}(0) &= x_0, \\ y^{(k+1)}(T) &= g(x^{(k+1)}(T)), \end{cases} \tag{37}$$

where

$$b^{(k)}(t) = \frac{1}{(z^{(k)}(t) - y^{(k)}(t))^2 + 1}$$

and

$$z^{(k+1)}(t) = g'(x^{(k+1)}(t))x^{(k+1)}(t).$$

Thus, the WR scheme for problem (2) where $g(x(t)) = x(t)$ is equation (37) where

$$z^{(k+1)}(t) = x^{(k+1)}(t)$$

and initial waveforms are

$$x^{(0)}(t) = x_0, \;\; z^{(0)}(t) = x_0 \quad \text{and} \quad y^{(0)}(t) = x_0, \;\; \forall t \in [0, T]$$

with the terminal condition $y^{(k+1)}(T) = x^{(k+1)}(T)$.
A thorough discussion on the WR on SDEs can be found in ([2]).

5 The Convergence of the WR method for FBSDEs

Here we want to investigate the convergence properties of the WR scheme (36). In the WR scheme (36), the value of $x^{(k+1)}$ has to be obtained before we calculate $y^{(k+1)}$. Thus, we first consider the WR scheme of the form

$$\left\{ \begin{array}{lcl} dx^{(k+1)}(t) & = & b^{(k)}(t)x^{(k+1)}(t)dt + x^{(k+1)}(t)dW(t), \\ x^{(k+1)}(0) & = & x_0. \end{array} \right. \tag{38}$$

This is the WR scheme of the forward SDE in equation (8), namely

$$\left\{ \begin{array}{lcl} dx(t) & = & b(t)x(t)dt + x(t)dW(t), \\ x(0) & = & x_0. \end{array} \right. \tag{39}$$

This multiplicative linear SDE has an exact solution [5]

$$x(t) = x_0 \, \exp \left(\int_0^t (b(s) - \frac{1}{2})ds + \int_0^t dW(s) \right). \tag{40}$$

Since (39) has equation (40) as an exact solution, the solution of the WR scheme (38) can be written as

$$x^{(k+1)}(t) = x_0 \, \exp \, (\int_0^t (b^{(k)}(s) - \frac{1}{2})ds + \int_0^t dW(s) \,), \tag{41}$$

where $b^{(k)}(s) = \frac{1}{(z^{(k)}(s) - y^{(k)}(s))^2 + 1}$.

It is trivial to show that the iteration scheme (41) converges to the exact solution (40).
Now, we want to look at the convergence property of the WR scheme

$$\begin{cases} dy^{(k+1)}(t) &= b^{(k)}(t)z^{(k+1)}(t)dt + z^{(k+1)}(t)dW(t), \\ y^{(k+1)}(T) &= g(x^{(k+1)}(T)). \end{cases} \tag{42}$$

By discretizing the interval $[0,T]$ with the equidistant stepsize $h = T/N$, for some integer N and $t_n = t_0 + nh,\ (n = 0, 1, 2, \cdots, N)$, the solution of the backward scheme (42) in each iteration is given by

$$y_t^{(k+1)} = y_{t+1}^{(k+1)} - h\ b^{(k)}(t)z^{(k+1)}(t) - \Delta W z^{(k+1)}(t)$$

or in a forward scheme as

$$y_{t+1}^{(k+1)} = y_t^{(k+1)} + h\ b^{(k)}(t)z^{(k+1)}(t) + \Delta W z^{(k+1)}(n). \tag{43}$$

The equation (43) is known as the explicit Euler method which approximates the following SDE [6]

$$y_{t+1} = y_t + \int_{t_n}^{t_{n+1}} f(t,y)dt + \int_{t_n}^{t_{n+1}} g(t,y)dW(t),$$

where $\Delta W(t)g(t,y)$ is calculated at the left point value of function $g(t,y)$. It is clear that numerical solutions of the explicit Euler method converges to the Itô SDE.
Since the WR scheme (38) and (42) are convergent then the WR scheme (36) is convergent.

6 Numerical Results

In this section some numerical simulations of the WR method for FBSDE problems will be presented. There are three purposes in this numerical simulation. Firstly, we want to show numerical simulations of the solutions $x,\ y$ and z compare to the exact solution when there is no windowing technique.

Secondly, we want to know more about the convergence properties of the method, so we calculate the absolute error at some particular points, at $x(T), y(0), z(T)$, for particular stepsize and tolerance. The reason we choose these points is as follows: we solve variables x and z in a forward direction, so the error is calculated at the end of interval T. However, the variable y works in a backward direction so the error is calculated at the beginning of interval (the point 0).

Thirdly, we want to know the effect of the windowing technique and the stepsize to the convergence.

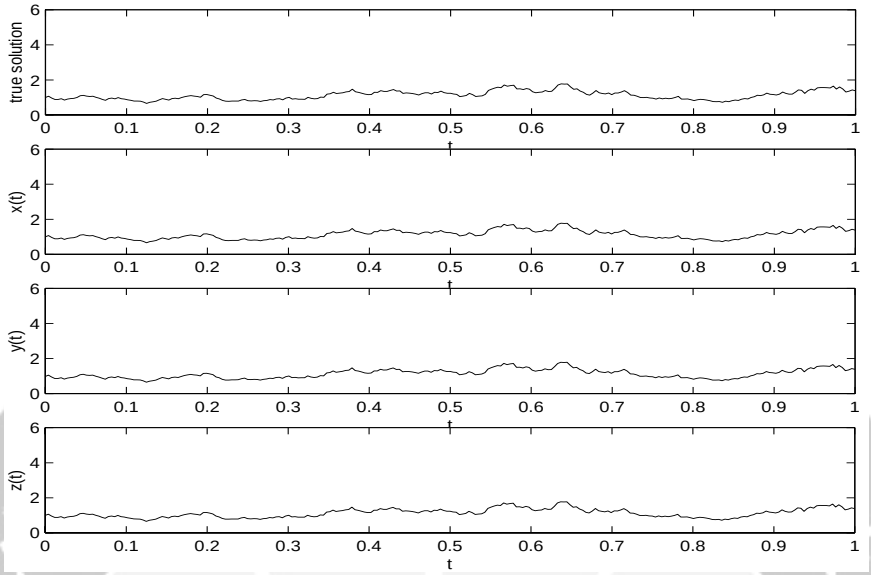


Figure 1: (a) The true solution path of x, y and z ,where $y(T) = x(T)$; the approximate solution paths of (b) $x(t)$; (c) $y(t)$; (d) $z(t)$, where $h = 2^{(-8)}, tol = 10^{(-3)}$ in 1 window.

The numerical simulation of the FBSDE test problem (see Figure 1) shows that even without windowing technique, $h = 2^{(-8)}$ and $tol = 10^{(-3)}$, numerical solutions of x, y and z variables are close to the numerical solution of the true solution.

Figure 2 shows the absolute errors at points $x(T), y(0), z(T)$ in 500 simulations without the windowing technique. We see that some of the absolute errors at $x(T)$ and $z(T)$ are large (around 2). However, the mean error is quite good.

For the third aim, we present some tables show the effect of the windowing technique and the steplength to the convergence of solutions (see Tables 1, 2). Since the results for variable z are equal to the results of variable x then we do not present the table for z.

We see from the tables that the reduced stepsize really gives a better convergence. Event though the results without the windowing technique is quite good when the stepsize is small enough, the results still can be improved when we increase the number of windows. The performance of this WR scheme on other problem can be found in [1].

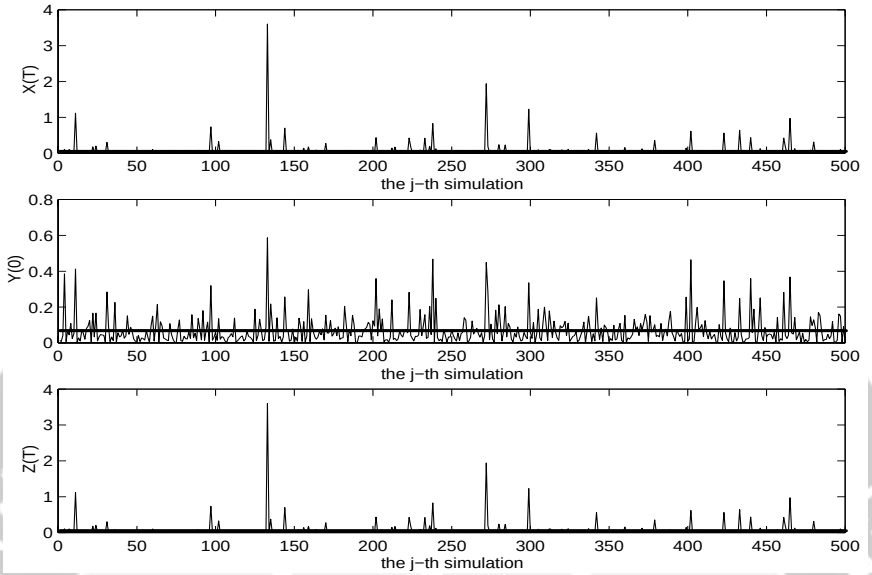


Figure 2: The absolute error of the problem where $y(T) = x(T)$ at the following points: (a) $x(T)$; (b) $y(0)$; (c) $z(T)$, where $h = 2^{(-8)}, tol = 10^{(-3)}$, in 1 window and 500 simulations. The bold lines denote the mean of absolute errors of $x(T), y(0)$ and $z(T)$, where $T = 1$. The means are $0.0513, 0.0687, 0.0513$, respectively.

the stepsize	1 window	2 windows	4 windows
$2^{(-4)}$	0.3786	0.2294	0.1248
$2^{(-6)}$	0.1343	0.0730	0.0482
$2^{(-8)}$	0.0513	0.0249	0.0122

Table 1: The mean error of the problem with $y(T) = x(T)$ at $x(T)$ in $1, 2, 4$ windows, where $tol = 10^{(-3)}$ and $T = 1$ in 500 simulations.

the stepsize	1 window	2 windows	4 windows
$2^{(-4)}$	0.2508	0.1419	0.0838
$2^{(-6)}$	0.1277	0.0722	0.0412
$2^{(-8)}$	0.0555	0.0310	0.0198

Table 2: The mean error of the problem with $y(T) = x(T)$ at $y(0)$ in $1, 2, 4$ windows, where $tol = 10^{(-3)}$ and $T = 1$ in 500 simulations.

7 Conclusion

The adapted solution as the result of the application of the Four Step Scheme to the generalised form is needed in order to implement the WR scheme. The performance of the WR method for solving FBSDE test problem is quite satisfactory and the performance is getting better when the stepsize is getting smaller. This implies that the numerical solution will converge to the exact solution when the stepsize is small enough. The performance is also getting better when the windowing technique is applied.

References

[1] Handari, B.D.(2002), *Numerical methods for SDEs and their dynamics*, Phd thesis, The Department of Mathematics at The University of Queensland, Brisbane, Australia.

[2] Burrage, K., and Burrage, P. (2002), Numerical methods for stochastic differential equations with applications, *preprint*.

[3] Cannon, J.R.,(1984), *The one-dimensional heat equation*, Addison-Wesley Publishing Company, Advanced Book Program, Reading, Massachusset.

[4] Grosset, L. (2000),Stock sale-advertising response model, *Dipartimento di Matematica, Universita di Pandova*.

[5] Kloeden, P.E., Platen, E., & Schurz, H.(1993), *Numerical solutions of stochastic differential equations through computer experiments*, Springer-Verlag, New York.

[6] Tianhai, T.(2001), *Implicit numerical methods fo stiff stochastic differential equations and numerical simulations of stochastic models*, Phd thesis, The Department of Mathematics at The University of Queensland, Brisbane, Australia.

[7] Yong, J., & Ma, J.(1999), *Forward-Backward stochastic differential equations and their applications*, Springer-Verlag, Berlin.

[8] Yong, J., & Zhou, X.(1999), *Stochastic controls. Hamiltonian systems and HJB equations*, Springer-Verlag, New York.

Bevina D.Handari: Department of Mathematics, Faculty of Mathematics and Natural Sciences, The University of Indonesia.
Phone/Fax: +62 +21 7863439
E-mail: bevina1@cso.ui.ac.id

THE DEVELOPMENT OF FILTERED–U GLOBAL LEAST MEAN SQUARE ALGORITHM FOR ACTIVE NOISE CONTROL APPLICATION.

Agustinus
Universitas Kristen Maranatha, Bandung, Indonesia

Abstract. A transversal finite impulse response (FIR) filter is a simple structure for adaptive filtering and has extensively developed for active noise control (ANC) application. Recently researchers had attempted to utilize the recursive infinite impulse response (IIR) filter structure because it performs better for the same number of filter's coefficients. Additionally, adaptive IIR filter is also able to characterizing a pole–zero transfer function. Perhaps the most popular ANC algorithm that uses the IIR filter is the Filtered–U Least Mean Square (FULMS). However, some major drawbacks that inherent to the adaptive IIR filter are slow convergence, possible convergence to a bias or unacceptable sub-optimal solution, and the need for stability monitoring. Therefore, using the global optimization method proposed by Edmondson, this paper introduces the development of Filtered–U Global Least Mean Square (FUGLMS) algorithm. The discussion also includes a comparative evaluation with the use of FULMS and FUGLMS in the experiment. It is shown that FUGLMS algorithm gives not only a minimum mean square error and a minimum peak error, but also a smooth parameter evolution.

Key-words: active control, adaptive filter, least mean square, acoustic noise

1 Introduction

ANC involves an electro–acoustic or electro–mechanical system [1] that reduces the intensity of unwanted noise based on the principle of superposition, specifically the destructive acoustic interference. This phenomenon happens when an anti–noise of equal amplitude and opposite phase is generated and combined with the unwanted noise, thus resulting in the cancellation of both noises. Since the characteristics of noise source and the environment are time varying, besides that the frequency, amplitude, phase, and sound velocity of the undesired noise are non–stationary, an ANC system must therefore be adaptive in order to cope with these variations. Many examples of ANC system use adaptive filter to produce an estimate of the noise, and can be realized as transversal FIR, recursive IIR, lattice, and transform–domain filters. The coefficients of adaptive filter are adjusted using adaptation rule that minimize the difference between the noise and the anti–noise. Here, the least mean square (LMS) algorithm that originally proposed by Widrow [2] had grown to be the most popular adaptation rule for ANC system [3] because of its simplicity.

Adaptive filters based upon the FIR structure have matured to a point of practical implementations. A major drawback is that certain applications will require a large number of coefficients to achieve good performance, and it consequently increasing computational costs. This becomes evident when an adaptive FIR filter is applied to identify a process that actually has to be represented as a pole–zero model.

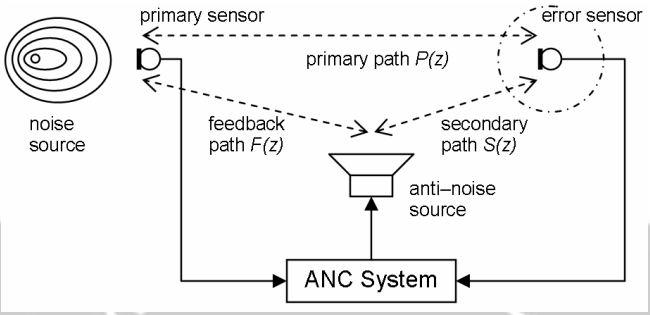


Figure 1. *Common arrangement of single channel ANC.*

On the other hand, the adaptive IIR filters have the advantage of approximating a pole–zero process more accurately with an equivalent–order of IIR filter. Thus it reduces the computational requirement in terms of the number of coefficients to be adjusted. Although adaptive IIR filters require fewer coefficients to be estimated, the system may become unstable because of the possibility that some poles of the filter will move outside the unit circle during adaptation. In practice, it is important to observe system stability at each of iteration.

Another problem area, in addition to slow convergence rate, is that the objective function for an adaptive IIR filter can be non–convex, which implies the existence of multiple local minima [4]. Logically, conventional gradient–based algorithms, such as LMS, can easily be trapped at an unacceptable local optimum. Therefore global optimization methods should be exploited in adaptive IIR filtering to overcome this problem and to preserve stability throughout adaptation.

Edmonson had suggested the use of stochastic approximation with convolution smoothing (SAS) for adaptive IIR filtering [5]. The technique of SAS is based on a random perturbation to find the absolute optimum of an arbitrary cost function. In particular, this SAS approach has been successfully used as a global optimization algorithm in some applications and empirically proven to be efficient in converging to the global optimum in terms of computation and accuracy. Based on his work, this paper presents the introduction of global least mean square (GLMS) algorithm in the IIR based ANC system.

2 Filtered–U Least Mean Square

An adaptive IIR filter structure was proposed for use in ANC by Eriksson. This new approach considers the feedback as a part of overall plant and the adaptive IIR filter deals with it directly as part of the problem. The IIR control system allows the system to dynamically track changes in the secondary and feedback paths during the cancellation operation. Respectively, these paths are the anti–noise propagation ways that lie from the anti–noise source to the location of error sensor and primary sensor. Figure 1 shows the locations of sources, sensors, and paths. Here, the third path exists between primary sensor and error sensor. This is the propagation path of unwanted noise and it usually called the primary path.

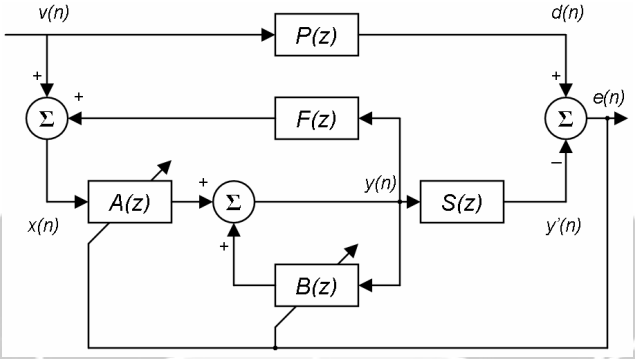


Figure 2. ANC system using an adaptive IIR filter.

Although there are a number of adaptive IIR algorithms, the recursive algorithm called FULMS was selected for reason that it adapts an IIR filter in a manner similar to the LMS algorithm for FIR filter. This algorithm had gained popularity because it is easy and simple. Followed here is the derivation of FULMS algorithm that is taken from [6]. The residual error *e(n)* in figure 2 is

$$e(n) = d(n) - s(n) * y(n)$$

so the output of the IIR filter *y(n)* is computed as

$$y(n) = a^T(n)x(n) + b^T(n)y(n-1) \tag{1}$$

where *a(n)* is the weight vector of *A(z)* at time *n* defined as

$$a(n) \equiv \begin{bmatrix} a_0(n) & a_1(n) & \cdots & a_{L-1}(n) \end{bmatrix}^T$$

and the signal vector *x(n)* is defined as

$$x(n) \equiv \begin{bmatrix} x(n) & x(n-1) & \cdots & x(n-L+1) \end{bmatrix}^T$$

The weight vector of *B(z)* at time *n* defined as

$$b(n) \equiv \begin{bmatrix} b_1(n) & b_2(n) & \cdots & b_M(n) \end{bmatrix}^T$$

and

$$y(n-1) \equiv \begin{bmatrix} y(n-1) & y(n-2) & \cdots & y(n-M) \end{bmatrix}^T$$

is the output signal vector delayed by one sample. *L* and *M* is the number of zeros in *A(z)* and poles in *B(z)*, respectively. Let define a new overall weight vector

$$q(n) \equiv \begin{bmatrix} a(n) \\ b(n) \end{bmatrix}$$

and a generalized reference vector

$$f(n) \equiv \begin{bmatrix} x(n) \\ y(n-1) \end{bmatrix}$$

so equation (1) can then be simplified to

$$y(n) = \mathbf{q}^T(n)\mathbf{f}(n)$$

The objective of the adaptive system is to determine an optimum values of $\mathbf{q}(n)$ to minimize a performance criterion that is based on *e(n)*. LMS algorithm updates the coefficient vector in the negative gradient direction of quadratic error surface with step size μ,

$$\mathbf{q}(n+1) = \mathbf{q}(n) - \frac{\mu}{2}\nabla\hat{x}(n) \tag{2}$$

where

$$\nabla\hat{x}(n) = \nabla e^2(n) = 2[\nabla e(n)]e(n)$$

is an instantaneous gradient estimate of the MSE gradient at time *n*. Since *e(n)* not a function of *a(n)* and *b(n)*, the error gradient is calculated as

$$\begin{aligned}\nabla e(n) &= \left[\frac{\partial e(n)}{\partial a_0(n)} \quad \cdots \quad \frac{\partial e(n)}{\partial a_{L-1}(n)} \quad \frac{\partial e(n)}{\partial b_1(n)} \quad \cdots \quad \frac{\partial e(n)}{\partial b_M(n)}\right]^T \\ &= -s(n)*\left[\frac{\partial y(n)}{\partial a_0(n)} \quad \cdots \quad \frac{\partial y(n)}{\partial a_{L-1}(n)} \quad \frac{\partial y(n)}{\partial b_1(n)} \quad \cdots \quad \frac{\partial y(n)}{\partial b_M(n)}\right]^T\end{aligned} \tag{3}$$

Assuming that the step size μ is small for slow convergence so that

$$\frac{\partial y(n-j)}{\partial a_l(n)} \approx \frac{\partial y(n-j)}{\partial a_l(n-j)}$$

then

$$\begin{aligned} y_{a_l}(n) &\equiv \frac{\partial y(n)}{\partial a_l(n)} \\ &= x(n-l) + \sum_{j=1}^{M} b_j(n)\frac{\partial y(n-j)}{\partial a_l(n)} \\ &\approx x(n-l) + \sum_{j=1}^{M} b_j(n) y_{a_l}(n-j) \qquad l = 0, 1, \cdots, L-1 \end{aligned} \tag{4}$$

Similary

$$y_{b_m}(n) \approx y(n-m) + \sum_{j=1}^{M} b_j(n) y_{b_m}(n-j) \qquad m = 1, 2, \cdots, M \tag{5}$$

Substituting equations (4) and (5) into equation (3) to obtain

$$\nabla e(n) \approx -s(n)*\left[y_{a_0}(n) \quad \cdots \quad y_{a_{L-1}}(n) \quad y_{b_1}(n) \quad \cdots \quad y_{b_M}(n)\right]^T \tag{6}$$

With assumption that the recursion based on the old output gradients is negligible,

$$y_{a_l}(n-j) = y_{b_m}(n-j) = 0 \qquad j = 1, 2, \cdots, M$$

equation (6) can be simplified as

$$\begin{aligned}\nabla e(n) &= -\hat{s}(n) * [x(n) \quad \cdots \quad x(n-L+1) \quad y(n-1) \quad \cdots \quad y(n-M)]^T \\ &= -\hat{s}(n) * f(n)\end{aligned}$$

Accordingly equation (2) reduces to

$$q(n+1) = q(n) + \mu f'(n) e(n) \tag{7}$$

where

$$f'(n) = \hat{s}(n) * f(n)$$

Equation (7) is the adaptation rule of FULMS algorithm. Here $\hat{s}(n)$ denote the best available estimate of the actual secondary path impulse response *s(n)* that can be acquired using the off–line identification methods.

3 Global Least Mean Square

In order to reach the global optimum, the objective of convolution smoothing can be viewed as filtering out the noise and performing minimization on the smoothed convex function. Since in general the optimum of the smoothed convex function does not coincide with the global function minimum, a sequence of optimization steps are required with the amount of smoothing eventually reduced to zero in the neighborhood of the global optimum.

According to [7], the smoothing process is performed by averaging $f(q)$ over some region of the parameter space R^n using the proper weighting function $\hat{h}(h, b)$,

$$\begin{aligned}\hat{f}(q, b) &= \int_{R^n} \hat{h}(h, b) f(q - h) dh \\ &= \int_{-\infty}^{\infty} \hat{h}(q - h, b) f(h) dh\end{aligned}$$

Hence

$$\hat{f}(q, b) = E_h[f(q - b)]$$

where $\hat{f}(q, b)$ is the smoothed approximation of $f(q)$, $h \in R^n$ is a vector of random perturbation, $E_h[f(q-b)]$ is the expectation with respect to the random variable h, and q denote weight vector that has a similar form with $q(n)$. It can be inferred that b actually represent a parameter that controls the standard deviation and defines the degree of smoothing. Therefore, an unbiased estimator $\hat{f}(q, b)$ is

$$\hat{f}(q, b) = \frac{1}{N} \sum_{i=1}^{N} f(q - h^i)$$

The smoothing function $\hat{h}(h,b)$ is a kernel function whose properties follow:

1. $\hat{h}(h,b) = \frac{1}{b^n} h(h,b)$,
2. $\lim_{b \to 0} \hat{h}(h,b) = d(h)$,
3. $\hat{h}(h,b)$ is a probabilistic distribution function,

and under this condition,

$$\lim_{b \to 0} \hat{f}(h,b) = \int_{R^n} d(h) f(q-h) dh$$
$$= f(q)$$

Therefore, the objective is to solve the optimization problem of minimizing the smoothed function, such as

$$\min_{q \in R^n} \hat{f}(q,b) \text{ as } b \to 0$$

In practice, to avoid adaptation process trapped in local minima, the initial value of b should be sufficiently large, so the smoothing function would able to eliminate all possible local minima of $\hat{f}(q,b)$, and later it gradually reduced to a very small value so the true function minimum could be achieved.

Under the assumption that the gradient of functional $f(q-b)$ is known, the unbiased single and double sided gradient estimate of the smoothed functional $\hat{f}(q,b)$ can be represented as

$$\nabla_q \hat{f}(q,b) = \frac{\sum_{i=1}^{N} \nabla_q f(q - bh_i)}{N}$$

and

$$\nabla_q \hat{f}(q,b) = \frac{\sum_{i=1}^{N} \left[\nabla_q f(q + bh_i) + \nabla_q f(q - bh_i)\right]}{2N}$$

The key to implementing a practical algorithm for adaptive IIR filter is to develop an on–line gradient estimate $\nabla_q \hat{e}(q,b)$, which is

$$\nabla_q \hat{e}(q,b) = \frac{1}{N} \sum_{i=1}^{N} \nabla_q e(q - bh_i)$$

By setting *N=1*, the one-sample gradient estimate can be obtained as

$$\nabla_q \hat{e}(q,b) = \nabla_q e(q - bh) \quad (8)$$

Equation (8) is iterated for each input sample, and theoretically shows that the on–line version of the SAS is given by the gradient value at the randomly selected neighborhood of the present operating point. The variance of the neighborhood is controlled by b, which decreases along with the adaptation.

In the implementation, equation (8) needs one filter for computing the input–output relationship and another one for computing the gradient estimate at the perturbing point $(q - bh)$. This is impractical for a large order system, and therefore some simplification have to be made. Using Taylor series around the operating point, an alternative representation of the gradient estimate function can be expressed as

$$\nabla_q \; e(q - bh) = e'(q) + bh \; e''(q) + \cdots$$

By assuming a diagonal Hessian form the first two terms, the first approximation of gradient estimate is

$$\nabla_q \; e(q - bh) = e'(q) - bh$$

This extreme approximation assumes that the second derivative of the gradient vector is independent of q so that its variance is constant throughout the adaptation process. The second term bh can be interpreted as a perturbing noise, which is the important term to avoid convergence to the local minimum.

The development of GLMS algorithm involve evaluating the MSE objective function that can be described as

$$\begin{aligned} x(n,q) &= \frac{1}{2} E\left[e^2(q)\right] \\ &= \frac{1}{2} E\left[(d(n) - y(n))^2\right] \\ &= \frac{1}{2} E\left[\left(d(n) - q^T(n) f(n)\right)^2\right] \end{aligned}$$

Use the instantaneous value as the expectation of

$$E[e^2(n)] \approx e^2(n)$$

such that

$$x(n,q) = \frac{1}{2}[d(n) - q^T(n) f(n)]^2$$

The gradient estimate vector with respect to the parameters q is

$$\nabla x(n,q) = \nabla_q \frac{1}{2}\left[e^2(n,q)\right] = e(n,q) \; \nabla_q \; e(n,q) = -e(n,q) \begin{bmatrix} \dfrac{\partial e(n,q)}{\partial a_i} \\ \dfrac{\partial e(n,q)}{\partial b_i} \end{bmatrix}$$

Thus the GLMS adaptation rule can be obtained using equation (2) as follow

$$q(n+1) = q(n) - \mu \, \Xi(n) - b(n) h \qquad (9)$$

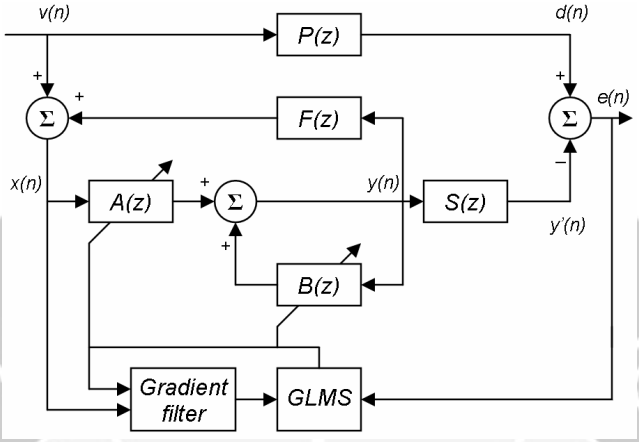


Figure 3. Block diagram of FUGLMS ANC system.

where $\Xi(n)$ represent a gradient estimate vector $[e(n)A, e(n)\Gamma]$, with

$$A = \{a_i(n)\} = \left\{\frac{\partial y(n)}{\partial a_i}\right\}$$

and

$$\Gamma = \{g_i(n)\} = \left\{\frac{\partial y(n)}{\partial b_i}\right\}$$

The Filtered–U Global Least Mean Square (FUGLMS) is developed using the GLMS adaptation rule in equation (9). Figure 3 illustrate a configuration of single channel FUGLMS ANC system.

4 Experiment setup and analysis

Based on assumptions that the length of adaptive IIR filter used as the controller is sufficient to characterizing the primary propagation path, and the estimation of the actual secondary path impulse response *s(n)* is suitable, the experiment conducted using a recorded version of pure tone and random noises samples. As stated in [8], for adaptive filter with length *L*, the minimum number of noise sample *N* that meet coefficient's miss–adjustment *M* can be calculated using

$$M = L/N$$

Written using Matlab v.6.5, the structure of experimental model consist primary, secondary, and feedback propagation paths that are derived from a real acoustic plant, and also the FULMS and FGLMS controllers as shown in figure 2 and 3. One typical example of experiment result are reported in figure 4, while the average MSE and TER value that show controllers performance are tabulated in table 1.

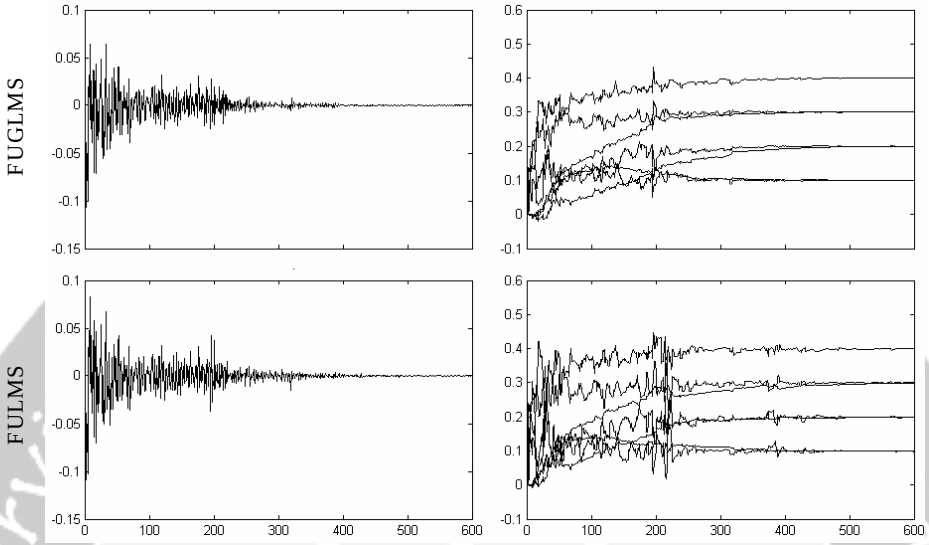


Figure 4. Evolution of random noise residue and filter's coefficients.

This example shows that FUGLMS and FULMS algorithms, with $L = 5$ and $M = 4$, are able to handle random noise. Both algorithm's performance are relatively equal – notice that envelope shapes of noises residue look almost the same with transient time around 500 samples – except that the coefficients evolution resulted from the FUGLMS seems to be more calm. This behavior gives higher confidence that the proposed algorithm keeps the adaptation process far away from any local minima.

Table 1. Average MSE and TER

Noise	FULMS		FUGLMS	
	MSE	**TER (dB)**	**MSE**	**TER (dB)**
Pure Tone	2.17×10^{-6}	23.89	1.85×10^{-6}	24.60
Random	1.47×10^{-5}	16,04	1.21×10^{-5}	15,29

Correspondingly, table 1 reveals comparative performance of FULMS and FUGLMS in a quantitative sense. The reported value is the average MSE and TER that came from experiment using a number of noises samples. Here, MSE and TER, or total energy reduction, are computed using

$$MSE = \frac{1}{N}\sum_n e(n)^2$$

and

$$TER = 10 \times \left[{}^{10}\log\left(\mathrm{var}\left(e_0(n)\right)\right) - {}^{10}\log\left(\mathrm{var}\left(e_1(n)\right)\right)\right]$$

where $e_0(n)$ and $e_1(n)$ is noise residue samples before and after the activation of controller, respectively. These values confirm that FUGLMS controller gives better performance than FULMS controller in reducing unwanted noise intensity.

Briefly stated, the developed FUGLMS algorithm seems to meet all expectations by providing smooth parameter evolution and maximum noise reduction. However, in practice, the FUGLMS controller required more memory space to accommodate gradient observer. Without losing controller performance, further treatment should be studied to manage memory requirement that rises significantly with the increase of filter order.

5 Conclusion

The development of Filtered–U Global Least Mean Square Algorithm for Active Noise Control System had been presented. As reflected from the experiment result, the proposed algorithm performs better than its counterpart. While FULMS offers its simplicity, and therefore could save memory space and ease processor tasks, the FUGLMS assures convergence and gives benefit of greater noise reduction.

Acknowledgment

The author would like to thank Dr. Bambang Rijanto who gave an introduction to the active noise control system, and also Mr. Juliady who had participated actively during the preliminary study of Global Least Mean Square.

References

[1] Sen M. Kuo, Dennis R. Morgan (1999), Active Noise Control: A Tutorial Review, *Proceedings of the IEEE*, Vol. 87, No. 6, 943 – 973.

[2] Simon Haykin (2002), *Adaptive Filter Theory*, Prentice Hall, Pearson Education International, New Jersey.

[3] Xuan Kong, Pu Liu, Sen M. Kuo (1998), Multiple Channel Hybrid Active Noise Control System, *IEEE Trans. on Control Sys. Tech.*, Vol.6, No.6, 719–729.

[4] P. A. Regalia (1995), Adaptive IIR Filtering in Signal Processing and Control. Marcel Dekker Incorporation, New York.

[5] William Edmonson, Jose Principe, Kannan Srinivasan, Chuan Wang (1998), A Global Least Mean Square Algorithm for Adaptive IIR Filtering, *IEEE Trans. on Circuits and Sys.–II: Analog & Digital Signal Processing*, Vol.45, No.3, 379–384.

[6] Sen M. Kuo, Dennis R. Morgan (1996), Active Noise Control: Algorithms and DSP Implementations, John Wiley & Sons Incorporation, New York.

[7] Ching-An Lai (2002), Global Optimization Algorithms for Adaptive Infinite Impulse Response Filters, Dissertation Report, University Of Florida.

[8] Bernard Widrow, M Kamenetsky (2003), On the Statistical Efficiency of the LMS Family of Adaptive Algorithm, Stanford University.

AGUSTINUS: Department of Electrical Engineering, Universitas Kristen Maranatha, Jl. Surya Sumantri No. 65, Bandung 40164, Indonesia. Phone: +62 22 2006543, Fax: +62 22 201 7622, e-mail: august@maranatha.edu.

High and Low Rank MIMO Channel Capacity on MIMO-Wireless Communication Systems

Rina Pudji Astuti[a], Tati L.R Mengko[b], Sugihartono[c], Andriyan B. Suksmono[d]

[a,b,c,d] ITB, Bandung, Indonesia

Abstract.

The impact of a rich scattering environment and channel knowledge in channel performance will determine the MIMO-wireless systems capacity. It is also depend on MIMO inter antennas elements correlation. To achieve a high MIMO system capacity, the orthogonality (un-correlation) of signals from MIMO antenna elements must be guaranteed. The condition will cause a change on MIMO channel matrix, $\mathbf{H}$, become high-rank or low-rank channels. The high-rank or low-rank channels will decide MIMO-wireless systems capacity.

In general, both of the MIMO systems and the MIMO-STBC systems with the no channel knowledge at the receiver have more capacity than the systems with the channel knowledge at the receiver. However with the small S/N , the systems with the channel knowledge at the receiver have better capacity performance than the other.

Both of the MIMO channel and the MIMO-STBC channel with the same number of receive antennas and the large S/N, double of the number of transmit antennas has a large increase capacity performance. In high-rank and low rank correlation**,** the MIMO channel has a better capacity performance than the MIMO with STBC.

Key-words: MIMO, STBC, channel capacity, high and low rank

1 Introduction

The steep progress of wireless communication services has impact on the need of a huge bandwidth to support new wireless application or services. On the other hand, frequency bandwidth is the limited resources. So, a system which has high frequency bandwidth usage efficiency is needed to increase an average capacity. MIMO (multiple inputs and multiple outputs) is one of the solutions to handle the condition.

The basic idea of MIMO systems is *space time signal processing.* The technique utilizes combination of signal processing in time dimension, which is a natural dimension of digital communication data, and space dimension which is a usage of array antenna in space distribution. The large spectral efficiencies associated with MIMO channels are based on multipath propagation with a rich scattering environment provides independent transmission path from each transmit antenna. On the other hand MIMO takes advantage of random fading [Foshini'96,'98] for increasing transfer rates.

Channel capacity characteristic of MIMO system is depend on readiness of channel model and estimation in both of transmission and receiver side. The MIMO channel capacity is also depend on MIMO inter antennas elements correlation. To achieve a high MIMO system capacity, the orthogonality (un-correlation) of signals from MIMO antenna elements must be guaranteed. The condition will cause a change on MIMO channel matrix, $\mathbf{H}$, become high-rank or low-rank channels. The high-rank or low-rank channels will determine MIMO-wireless systems capacity.

Therefore, our objective is to determine the impact of a rich scattering environment and channel knowledge in channel performance which is indicated in wireless systems capacity. The focus of this paper is to compare the channel performance of MIMO system with the MIMO-STBC system.

This paper is organized as follows. In section 2, MIMO system model and channel model are described. In this section, MIMO-wireless systems with space time block code (STBC) is applied in a rich scattering environment, such as Rayleigh and Rician fading models. Next, section 3 describes information theoretic capacity of MIMO systems. The capacity of MIMO systems includes MIMO with and without STBC system. Next, section 4 describes the results. In this section is discussed the effect of parameter propagation and channel knowledge changes in MIMO systems capacity. Finally, section 5 summarizes altogether the capacity results.

2 MIMO system model

Multiple input and multiple output (MIMO) systems, employing several transmit and receive antenna at both ends, are able to increase in capacity compared to traditional single antenna systems. However, this increasing in capacity is dependent upon the fact that the channels from a transmitter to a receiver follow independent path. If a severe correlations present at the transmitter and/or the receiver side, for example, the capacity of the MIMO systems can be shown to degrade[1].

In order to make the availability of independent channel, we consider to applied space time block code (STBC) such as an inner coding. STBC, however, are not designed to provide significantly coding gain. Hence, powerful outer codes, combining of convolution code and block interleaving, can be concatenated with STBC to have a required coding gain. STBC are applied according to H-BLAST model (as horizontal encoder) which are fixed to each antennas arm.

The channel model of the MIMO systems can be illustrated as follow:

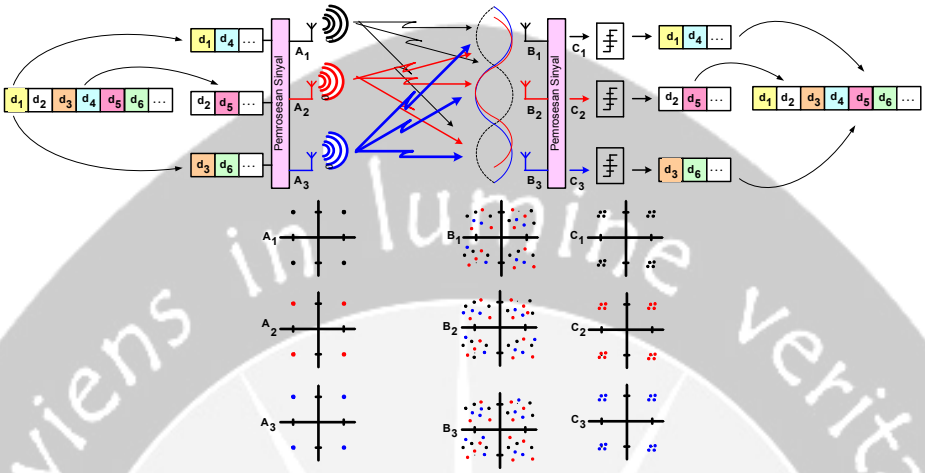


Figure 1. Spatial multiplexing scheme of the MIMO systems with spectral efficiency

A high data bit streams is decomposed into 3 independent of 1/3 rate bit sequences. Then, the data streams are transmitted simultaneously using multiple antennas with spectral efficiency, thus consuming 1/3 of the nominal spectrum. At the receiver, the mixing channel matrix is identified by training symbols. Later, the individual bit streams are separated and estimated. This occurs in the same way as three unknowns are resolved from a linear system of three equations.

The separation of MIMO Channels is possible only if the equations are independent which can be interpreted by each antenna a sufficiently different channel. Wherein case the bit streams can be detected and merged together to obtain the original high data rate signal.

The received signal of the MIMO system at n^{th} antenna receiver from m^{th} antenna transmitter with input signal, $s_m(t)$, is as follow:

$$y_n(t) = \sum_{m=1}^{M_T} h_{m,n}(\tau,t) * s_m(t) \qquad (1)$$

Where:

$n=1, 2, \ldots, N_R$

$s_m(t)$: is input signal of m^{th} antenna transmitter, T_X

$h_{m,n}(\tau,t)$: is channel impulse response of m^{th} antenna transmitter to n^{th} antenna receiver which be offered by propagation channel

characteristic, pulse shaping at the transmitter and matched filter at the receiver.

MIMO-wireless channel model in flat fading condition, where M_T antennas in transmitter and N_R antennas in receiver is a matrix $\mathbf{H}$ with $N_R \text{x} M_T$ elements, has signal equation in receiver as follow:

$$\mathbf{y} = \mathbf{Hs} + \mathbf{n} \tag{2}$$

Where: $\mathbf{y}$ is receive signal vector, N_R x 1

$s = [s_1 s_2 \ldots s_{M_T}]$ is transmit signal vector, M_T x 1

$\mathbf{H}$ is MIMO channel matrix, $N_R \text{x} M_T$

$\mathbf{n}$ is *additive white Gaussian noise vector*, N_R *x 1*

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1M_T} \\ h_{21} & h_{22} & \cdots & h_{2M_T} \\ \vdots & \ddots & \ddots & \vdots \\ h_{N_R} & \cdots & \cdots & h_{N_R M_T} \end{bmatrix} \tag{3}$$

Early MIMO-wireless system model with space time block code (STBC) using two transmit antennas and one receive antenna is suggested by Alamouti[6]. This scheme supports maximum-likelihood (ML) detection based only on linear processing at the receiver. Furthermore, Vahid Tarokh [6] develops the system for general configuration systems. In this scheme, a number of code symbols equal to the number of T_X antennas are generated and transmitted simultaneously, one symbol from each antenna. These symbols are generated by the space time encoder, the diversity gain and/or the coding gain is maximized.

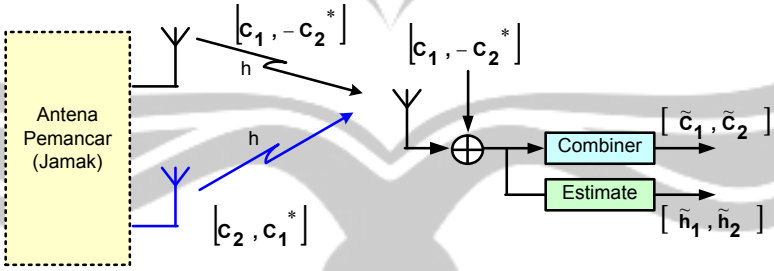


Figure 2. MIMO-STBC system of Alamouti scheme

The Alamouti STBC models, the input symbols to the space time block encoder are divided into groups of two symbols each. At a given symbol period, the two symbols in each group (c_1, c_2) are transmitted simultaneously from the two antennas. The signal transmitted from antenna 1 is c_1 and the signal transmitted from antenna 2

is c_2. In next symbol period, the signal $-c_2^*$ is transmitted from antenna 1 and the signal c_1^* is transmitted from antenna 2. Let h_1 and h_2 be the channel from the first and second T_X antennas to the R_X antenna, respectively [1]. It is assumed a receiver with a single R_X antenna, and denotes the received signal over two consecutive symbol periods as x_1 and x_2. The received signals can be expressed as follow:

$$x_1 = h_1 c_1 + h_2 c_2 + n_1 \quad (4)$$

$$x_2 = -h_1 c_2^* + h_2 c_1^* + n_2 \quad (5)$$

The received signals can be rewritten in a matrix form as:

$$\mathbf{x} = \mathbf{c}.\mathbf{H} + \mathbf{n} \quad (6)$$

Where:

$$\mathbf{x} = [x_1 \quad x_2]^T \quad (7)$$

$$\mathbf{c} = \begin{bmatrix} c_1 & c_2 \\ -c_2^* & c_1^* \end{bmatrix} \quad (8)$$

The individual rows correspond to time diversity and the individual columns correspond to space (antenna) diversity.

$$\mathbf{H} = [h_1 \quad h_2]^T \quad (9)$$

$$\mathbf{n} = [n_1 \quad n_2]^T \quad (10)$$

Equation (6)-(8) can be rewritten as follow:

$$\mathbf{x} = \mathbf{H}.\mathbf{c} + \mathbf{n} \quad (11)$$

$$\mathbf{x} = \left[x_1 \quad x_2^*\right]^T \quad (12)$$

$$\mathbf{H} = \begin{bmatrix} h_1 & h_2 \\ h_2^* & -h_1^* \end{bmatrix} \quad (13)$$

$$\mathbf{c} = [c_1 \quad c_2]^T \quad (14)$$

$$\mathbf{n} = \left[n_1 \quad n_2^*\right]^T \quad (15)$$

A virtual MIMO matrix, $\mathbf{H}$, with space (columns) and time (rows) dimensions, is not to be confused with the purely spatial MIMO channel matrix in previous sections.

3 Information theoretic capacity of MIMO systems

3.1 Introduction

The spectral efficiencies related with MIMO channels are based on a rich scattering environment provides independent transmission paths each transmit antenna to each receive antenna. The capacity linearly with $\min(M_T, N_R)$ relative to a system with just one transmit antenna and one receive antenna. This capacity increase

requires a scattering environment such that the matrix of channel gains has full rank and independent entries and that perfect estimates of these gain are available at the transmitter and the receiver [2].

MIMO channel capacity depends greatly on the statistical properties and antenna correlations of the channel. Antenna correlation varies drastically as a function of the scattering environment, the distance between transmitter and receiver, the antenna configurations, and the Doppler spreads [2]. In real-world cases, correlated channel path gains occur.

When there is a rich scattering environment and when T_X and R_X arrays are relatively **near one another**, **high rank MIMO channels** occur. It is occur when there is **little correlation** among the path gains. MIMO channels have a **diversity gain** defined by the rank of $\mathbf{HH}^*$ which is achieve the maximum diversity if $rank(HH^*) = \min(M_T, N_R)$. The upper limit for the capacity of MIMO channels and the maximum diversity gain is represented by the orthogonal channel case.

The **low rank MIMO channel** is equivalent to a single antenna channel with the same total power. It is occur under **scatter-free** or **long-distance links**. It is occur when there is **strong correlation** between the channel path gains. The correlation characteristics decide the rank of $\mathbf{HH}^*$, which in turn determines the diversity advantage. A full correlated $\mathbf{H}$ matrix is a scaled version of the one's matrix with dimensions (M_T, N_R) and provides no diversity gain over the single antenna case.

On MIMO channel capacity in Shannon theoretic sense, the Shannon (ergodic) capacity of a single user time invariant channel is defined as the maximum mutual information between the channel input and output. This maximum mutual information is shown by Shannon's capacity theorem to be the maximum data rate that can be transmitted over the channel with random small error probability. While the channel is time-varying channel capacity has multiple definitions, depending on what is channel knowledge about the channel distribution or its state at the transmitter and/or receiver and whether the capacity is measured based on minimum rate or maintaining a constant fixed or averaging the rate over all channel distributions/states.

3.2 MIMO fading channel capacity

3.2.1 Information theoretic definition

The entropy of a random variable is a measure of uncertainty of the random variable. It is a measure of the amount of information required on the average to describe the random variable [4].

The mutual information can be described as the reduction in the uncertainty of one random variable due to the knowledge of the other [2]. It will depend on the properties of the wireless channel used to express information from the transmitter

to the receiver. With $h_{de}(.)$ denoting differential entropy (entropy of a continuous random variable), the mutual information can be expressed as

$$I(\mathbf{S};\mathbf{Y}) = h_{de}(\mathbf{Y}) - h_{de}(\mathbf{Y}\mid\mathbf{S}) \tag{16}$$
$$= h_{de}(\mathbf{Y}) - h_{de}(\mathbf{HS}+\mathbf{N}\mid\mathbf{S})$$
$$= h_{de}(\mathbf{Y}) - h_{de}(\mathbf{N}\mid\mathbf{S})$$
$$= h_{de}(\mathbf{Y}) - h_{de}(\mathbf{N}) \tag{17}$$

It will be assumed that $\mathbf{N} \sim N(0,\mathbf{K}^n)$, where $\mathbf{K^n} = E\{\mathbf{NN^H}\}$ is the noise covariance matrix.

Because the normal distribution maximizes the entropy over all distributions with the same covariance (i.e. the power constraint), the mutual information is maximized when $\mathbf{Y}$ represents a multivariate Gaussian random variable, i.e. $\mathbf{Y} = N(0,\mathbf{K}^y)$. With the assumption that $\mathbf{S}$ and $\mathbf{N}$ are uncorrelated, the received covariance matrix of the desired signal can be expressed as follow:

$$\mathbf{K^y} = E\{\mathbf{YY^Y}\} = E\{(\mathbf{HS}+\mathbf{N})(\mathbf{HS}+\mathbf{N})^H\} \tag{18}$$
$$= \mathbf{HK}^S\mathbf{H}^H + \mathbf{K}^n \tag{19}$$

where $\mathbf{K}^S = E\{\mathbf{SS}^H\}$ (20)

3.2.2 Capacity of MIMO channel

When perfect channel knowledge at the receiver by assuming maximum ratio combining at the receiver and the transmitter has no knowledge of the channel, it is optimal to consistently distribute the available power P_T between the transmit antennas [4], i.e. $\mathbf{K}^S = \frac{P_T}{M_T}\mathbf{I}_{n_S}$. We assume that the noise is uncorrelated among branches, the noise covariance matrix $\mathbf{K}^n = \sigma_n^2\mathbf{I}_{n_R}$. Therefore, the MIMO fading channel capacity can be written as:

$$C = h_{de}(\mathbf{Y}) - h_{de}(\mathbf{N}) \tag{21}$$
$$= \log_2[\det(\pi e(\mathbf{HK}^S\mathbf{H}^H + \mathbf{K}^n))] - \log_2[\det(\pi e\mathbf{K}^n)]$$
$$= \log_2[\det(\mathbf{HK}^S\mathbf{H}^H\mathbf{K}^n)] - \log_2[\det\mathbf{K}^n]$$
$$= \log_2[\det(\mathbf{HK}^S\mathbf{H}^H + \mathbf{K}^n)(\mathbf{K}^n)^{-1})]$$
$$= \log_2[\det(\mathbf{HK}^S\mathbf{H}^H(\mathbf{K}^n)^{-1} + \mathbf{I}_{n_R})]$$
$$= \log_2\left[\det\left(\mathbf{I}_{N_R} + (\mathbf{K}^n)^{-1}\mathbf{HK}^S\mathbf{H}^H\right)\right]$$
$$= \log_2\left[\det\left(\mathbf{I}_{N_R} + \frac{P_T}{\sigma_n^2 M_T}\mathbf{HH}^H\right)\right] \tag{22}$$

If we assume that $\mathbf{H}$ is a random process, we can identify the channel at the receiver by using a training sequence with assuming the training sequence does not cost any capacity. By condition it is no channel knowledge at the receiver, the MIMO channel capacity takes expectation over channel instantiations, as follow:

$$C = E\left\{\log_2 \det\left(I_{N_R} + \frac{P_T}{\sigma_n^2 M_T}\mathbf{H}\mathbf{H}^H\right)\right\} \tag{23}$$

3.2.3 MIMO with STBC fading channel capacity

For a 2 x 2 MIMO channel, accompany equation (6), the received signal can be expressed as

$$\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} = \begin{bmatrix} c_1 & c_2 \\ -c_2^* & c_1^* \end{bmatrix}\begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} + \begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix} \tag{24}$$

Which can be reorganized and written as[3][4] :

$$\underbrace{\begin{bmatrix} x_{11} \\ x_{21}^* \\ x_{12} \\ x_{22}^* \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} h_{11} & h_{21} \\ h_{21}^* & -h_{11}^* \\ h_{12} & h_{22} \\ h_{22}^* & -h_{12}^* \end{bmatrix}}_{\mathcal{H}} \underbrace{\begin{bmatrix} c_1 \\ c_2 \end{bmatrix}}_{c} + \underbrace{\begin{bmatrix} n_{11} \\ n_{21}^* \\ n_{12} \\ n_{22}^* \end{bmatrix}}_{n} \tag{25}$$

x_{11} and x_{12} represent the received symbols at the first transmit antenna and the second transmit antenna at time index t and also x_{21} and x_{22} signify the received symbols the first transmit antenna and the second transmit antenna at time index $t + T_S$.

With assuming that the MIMO-STBC channel is perfect channel knowledge at the receiver, example with matched filtering at the receiver, the received signal after matched filtering may be expressed as:

$$\mathbf{y} = \mathcal{H}^H \mathbf{x} \tag{26}$$

$$= \mathcal{H}^H \mathcal{H}\mathbf{c} + \mathcal{H}^H \mathbf{n}$$

$$= \|\mathbf{H}\|_F^2 \mathbf{c} + \mathcal{H}^H \mathbf{n} \tag{27}$$

Where:

$$\mathcal{H}^H \mathcal{H} = \begin{bmatrix} h_{11}^* & h_{21} & h_{12}^* & h_{22} \\ h_{21}^* & -h_{11} & h_{22}^* & -h_{12} \end{bmatrix}\begin{bmatrix} h_{11} & h_{21} \\ h_{21}^* & -h_{11}^* \\ h_{12} & h_{22} \\ h_{22}^* & -h_{12}^* \end{bmatrix} \tag{28}$$

$$= \begin{bmatrix} |h_{11}|^2 + |h_{12}|^2 + |h_{21}|^2 + |h_{22}|^2 & 0 \\ 0 & |h_{11}|^2 + |h_{12}|^2 + |h_{21}|^2 + |h_{22}|^2 \end{bmatrix}$$

$$= \|\mathbf{H}\|_F^2 . \mathbf{I}_2 \quad (29)$$

$\|\mathbf{H}\|_F^2$ is the squared Frobenius norm of the matrix $\mathbf{H}$.

The received signal after matched filtering can be written individually as:

$$y_1 = \|\mathbf{H}\|_F^2 c_1 + \mathcal{H}^H \mathbf{n} \quad (30)$$

$$y_2 = \|\mathbf{H}\|_F^2 c_2 + \mathcal{H}^H \mathbf{n} \quad (31)$$

In general, it can be written as:

$$y_l = \|\mathbf{H}\|_F^2 c_l + \mathcal{H}^H \mathbf{n} \quad (32)$$

Then, the capacity of a MIMO fading channel using STBC can be written as:

$$C = \frac{S}{T} . \log_2 \left(1 + \frac{P_T}{\sigma_n^2 M_T} \|\mathbf{H}\|_F^2 \right) \quad (33)$$

Where the symbols, S, of STBC system are transmitted in the time slots, T. The $\frac{S}{T}$ denotes the rate of the STBC.

4 Results

Added analysis of the MIMO channel capacity is possible by diagonalizing the product matrix $\mathbf{H}\mathbf{H}^H$ also by eigenvalue decomposition or singular value decomposition. Eigenvalue decomposition of matrix product is $\mathbf{H}\mathbf{H}^H = \mathbf{E}\Lambda\mathbf{E}^H$, where $\mathbf{E}$ is the eigenvector matrix with orthonormal columns and Λ is a diagonal matrix with the eigenvalues on the main diagonal. While singular value decomposition of the channel matrix is $\mathbf{H} = \mathbf{U}\sum\mathbf{V}^H$, where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices of left and right singular vectors respectively. $\sum$ is a diagonal matrix with singular values on the main diagonal.

With the singular value decomposition approach, the capacity of the MIMO system, equation (22) may be expressed as follow:

$$C = \log_2 \left[\det \left(\mathbf{I}_{N_R} + \frac{P_T}{\sigma_n^2 M_T} \mathbf{U} \sum \sum^H \mathbf{U}^H \right) \right] \quad (34)$$

$$= \log_2 \left[\det \left(\mathbf{I}_{N_R} + \frac{P_T}{\sigma_n^2 M_T} \mathbf{U}^H \mathbf{U} \sum^2 \right) \right]$$

$$= \log_2\left[\det\left(\mathbf{I}_{N_R} + \frac{P_T}{\sigma_n^2 M_T}\sum{}^2\right)\right]$$

$$= \log_2\left[\left(1+\frac{P_T}{\sigma_n^2 M_T}\sigma_1^2\right)\left(1+\frac{P_T}{\sigma_n^2 M_T}\sigma_2^2\right)\cdots\left(1+\frac{P_T}{\sigma_n^2 M_T}\sigma_k^2\right)\right] \quad (35)$$

$$= \sum_{l=1}^{k}\log_2\left(1+\frac{P_T}{\sigma_n^2 M_T}\sigma_l^2\right) \quad (36)$$

Otherwise, with an eigenvalue decomposition of the matrix product $\mathbf{H}\mathbf{H}^H$, the MIMO system capacity may be written as follow:

$$C = \sum_{l=1}^{k}\log_2\left(1+\frac{P_T}{\sigma_n^2 M_T}\lambda_l\right) \quad (37)$$

Where :

$$k = rank(\mathbf{H}) \le \min(M_T, N_R) \quad (38)$$

$\sum$ is a real matrix

$\det(\mathbf{I}_{AB} + \mathbf{AB}) = \det(\mathbf{I}_{BA} + \mathbf{BA})$

λ_l are the eigenvalue of matrix Λ.

Although STBC provide full diversity over the coherent, flat fading channel, at a low computational cost. It can be shown that they incur a loss in capacity because the convert the matrix channel into a scalar AWGN channel whose capacity is smaller than the true channel capacity. The MIMO system capacity, equation (34), may be written as follow:

$$C = \log_2\left(1+P\sum_{l=1}^{k}\sigma_l^2 + P^2\sum_{l_1\neq l_2}^{l_1<l_2}\sigma_{l_1}^2\sigma_{l_2}^2 + \cdots + P^k\prod_{l=1}^{k}\sigma_l^2\right) \quad (39)$$

$$= \log_2\left(1+P\|\mathbf{H}\|_F^2 + P^2\sum_{l_1\neq l_2}^{l_1<l_2}\sigma_{l_1}^2\sigma_{l_2}^2 + \cdots + P^k\prod_{l=1}^{k}\sigma_l^2\right) \quad (40)$$

$$\ge \log_2(1+P\|\mathbf{H}\|_F^2)$$

$$\ge \frac{S}{T}\log_2\left(1+P\|\mathbf{H}\|_F^2\right) \quad (41)$$

(the MIMO-STBC system capacity)

Where: $P = \dfrac{P_T}{\sigma_n^2 M_T}$ (42)

The MIMO system capacity above is explicitly equal or large than the MIMO-STBC system capacity. The capacity difference is a function of channel singular values. This can used to decide under which conditions STBC is optimal in terms of capacity.

In practice, the realization of high MIMO capacity is responsive not only to the fading correlation between individual antennas but also to the rank performance of the channel. High rank behavior has been heavily linked to the existence of a rich scattering environment and a little correlation between the channel path gains, for example in Rayleigh fading model. The MIMO channels have a diversity gain defined by the rank of $\mathbf{HH}^*$. The maximum achievable diversity gain is $rank(\mathbf{HH}^*) = \min(M_T, N_R)$. The maximum diversity gain and the upper limit for the capacity of MIMO channels is represented by the orthogonal channel gain case. With assuming M_T columns of $\mathbf{H}$ are orthogonal and the entries of $\mathbf{H}$ are normalized to unit power, so that the eigenvalues of $\mathbf{HH}^H$ are $\lambda_l = N_R$ for $l = 1,2,\cdots,M_T$. The capacity of the high-rank MIMO channel can be rewritten from equation (36) and (37) as:

$$C_{high\_rank} = \sum_{l=1}^{k=\min(M_T,N_R)} \log_2\left(1+\frac{P_T}{\sigma_n^2 M_T}\lambda_l\right) \tag{43}$$

$$\approx \underbrace{\min(M_T, N_R)}_{\substack{array\_capacity \\ advantage}} \underbrace{\log_2\left(1+\frac{P_T}{\sigma_n^2 M_T}N_R\right)}_{\substack{receiver\_antenna \\ SNR\_advantage}} \tag{44}$$

On the other hand, Low rank performance has been closely linked to the being of a poor scattering environment and a strong correlation between the channel path gains, for example in Rician fading model with large K factor. The correlation characteristics determine the rank of $\mathbf{HH}^H$, which in revolve determines the diversity advantage. A full correlated $\mathbf{H}$ matrix provides no diversity gain over the single antenna case and gives the all one's matrix with dimension $M_T x N_R$. With assuming high correlation, all gains h_{ij} are approximately equal, and $\mathbf{H}$, a multiple of the all-one matrix, has a single non zero singular value

$$\lambda_l = N_R \sum_{l=1}^{M_T} E(h_{ll}^H h_{ll}) \approx M_T N_R \tag{45}$$

The capacity of the low-rank MIMO channel can be expressed as :

$$C_{low\_rank} = \sum_{l=1}^{k=\min(M_T,N_R)} \log_2\left(1+\frac{P_T}{\sigma_n^2 M_T}\lambda_l\right) \tag{46}$$

$$\approx \log_2\left(1+\frac{P_T}{\sigma_n^2}N_R\right) \tag{47}$$

The low-rank MIMO channel behaves like a point to point channel or a single antenna channel with N_R times the received signal power due to the antenna array. It is achieved by simple maximum-ratio combining at the receiver.

If Low-SNR MIMO channel, $\left(P_T / \sigma_n^2\right)$, is low with a Taylor series approximation of $\log(1+x) \approx x$ for small values of x, both of the capacity of the high-rank MIMO channel for small SNR, C_{high_SNR}, and the capacity of the low-rank MIMO channel for small SNR, C_{low_SNR}, may expressed as:

$$C_{high\_SNR} \approx \min(N_R, M_T) \frac{P_T}{\sigma_n^2 M_T} N_R \tag{48}$$

$$C_{low\_SNR} \approx \frac{P_T}{\sigma_n^2} N_R \tag{49}$$

Altogether, from equation (22), (23), (33), (39),(41), (44), (47), (48), and (49) the MIMO channel and the MIMO-STBC channel capacity can be illustrated as two figures follow.

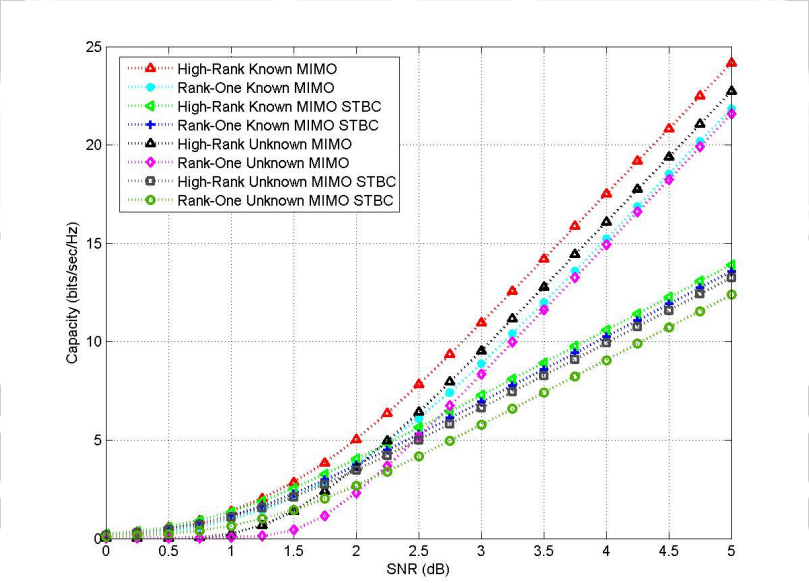


Figure 3. The capacity performance comparison between both of the MIMO systems and the MIMO-STBC systems with channel known and unknown, 2x2 antennas for high-rank and low-rank correlation

Figure 3 shows MIMO system and MIMO-STBC system capacity for **high-rank and low rank correlation** with the assumption of an average power constraint P_{Tx}, no

channel knowledge at transmitter, **channel knowledge and no channel knowledge at the receiver** with 2x2, antennas of $T_X xR_X$ system. With the large S/N, in relation to Figure 4 that both of the MIMO systems and the MIMO-STBC systems with the no channel knowledge at the receiver have more capacity than the systems with the channel knowledge at the receiver. However with the small S/N , the systems with the channel knowledge at the receiver have better capacity performance than the other.

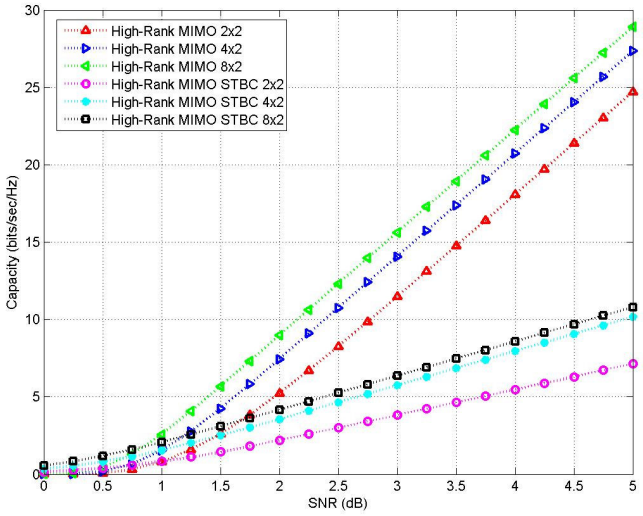


Figure 4. The capacity performance comparison between both of the MIMO systems and the MIMO-STBC systems with channel unknown, 2x2, 4x2, 8x2 antennas for high-rank correlation

Figure 4 shows MIMO system and MIMO-STBC system capacity for **high rank correlation** with the assumption of an average power constraint P_{Tx}, no channel knowledge at transmitter and **no channel knowledge at the receiver** with 2x2, 4x2, 8x2 antennas of $T_X xR_X$ system. (note : the ratio symbols of time slots of 4x2 and 8x2 antennas of $T_X xR_X$ system is ¾). With the same number of receive antennas and the large S/N, as shown in Figure 4 that for both of the MIMO channel and the MIMO-STBC channel, double of the number of transmit antennas has a large increase capacity performance. Generally, the MIMO channel capacity has better capacity performance than the MIMO-STBC channel. But in the small S/N, the MIMO channel capacity has bad capacity performance than the MIMO-STBC channel.

5 Conclusions

In general, both of the MIMO systems and the MIMO-STBC systems with the no channel knowledge at the receiver have more capacity than the systems with the

channel knowledge at the receiver. However with the small S/N , the systems with the channel knowledge at the receiver have better capacity performance than the other.

So, for both of the MIMO channel and the MIMO-STBC channel, with the same number of receive antennas and the large S/N, double of the number of transmit antennas has a large increase capacity performance.

Therefore, with the same number of receive antennas and the large S/N, the MIMO channel has a better capacity performance than the MIMO with STBC in **high-rank and low rank correlation.** While in the small S/N, the MIMO channel capacity has bad capacity performance than the MIMO-STBC channel.

References

[1] Gesbert, D, M. Shafi, D.S. Shiu, P. Smith and A. Naguib (2003), From Theory to Practice : An Overview of MIMO Space Time Coded Wireless Systems, *IEEE Journal on Selected Areas in Comm*., pt. I, **21**, 281-302.

[2] Goldsmith, A., A.A Jafar, N. Jindal, and S. Vishwanath (2003), Capacity limits of MIMO channels, *IEEE Journal on Selected Areas in Comm.,* pt. II, **21**, 684-700.

[3] Hassibi, B and B.M. Hochwald (2002), High-rate codes that are linear in space and time, *IEE trans. On information theory,* no.7, **48**, 1805-1823

[4] Holter, B (2002), Capacity of multiple-input multiple output (MIMO) systems in wireless communications, *NTNU*

[5] Schlegel, C and Z.B. Schlegel (2002), MIMO channels and space time coding, *WOC*

[6] Tarokh, V, N. Seshadri and A.R. Calderbank (1998), Space Time Codes for High data rate Wireless Comm. Performance Criterion and Code Construction, *IEEE Trans. Inform. Theory,* **44**, 744-765

RINA PUDJI ASTUTI: Ph D student at Department of Electrical Engineering, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia. E-mail: rina.p.a@plasa.com

TATI L.R MENGKO: Department of Electrical Engineering, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia. E-mail: tmengko@itb.ac.id

SUGIHARTONO: Department of Electrical Engineering, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia. E-mail: sugi_sugihartono@yahoo.com

ANDRIYAN B. SUKSMONO: Department of Electrical Engineering, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia. E-mail: suksmono@yahoo.com

COMPARATIVE STUDY OF ITERATIVE SEARCH METHOD FOR ADAPTIVE FILTERING PROBLEMS

N.A. Ahmad

Universiti Sains Malaysia, Malaysia

Abstract. Adaptive filtering problem refers to a class of application in signal processing that deals with adaptation of a system so as to adjust itself with the phenomenon that is taking place in its surrounding. Examples of such problem include adaptive system modeling, noise cancellation, equalization and prediction. The adaptive filtering problem may be mathematically formulated as an adaptive least squares problem in which a set of parameter values are updated so that the time varying sum of squared error cost function is minimized. Numerous algorithms are available for solving adaptive filtering problems and they may be classified into two categories: iterative methods and direct methods. In this paper, we perform a comparative study of four iterative search methods, namely, the method of steepest descent, the Newton's method, the Conjugate Gradient method and the Direction Set method. The methods are implemented and applied in system modeling and they are assessed in terms of rate of convergence, computational complexity, misadjustments and their sensitivity to spectral condition number or the eigenvalue spread. Our main objective is to provide a comprehensive understanding of the adaptive implementation of the methods, their performance according to assessment criteria mentioned above and also provide possible modifications to improve the performance.

Key-words: Adaptive least squares problem, adaptive filtering, adaptive algorithms

1 Introduction

Adaptive filtering problems has received considerable attention from the engineering community during the past several decades due to its application in many diverse fields such as system identification, equalization, prediction and noise cancellation. From mathematical perspective, adaptive filtering problem may be viewed as an adaptive least squares problem. The standard least squares problem can be reduced to solving a linear system of equations whereas in the adaptive least squares problem, a time varying linear system is resulted at each state.

As a consequence, many of the algorithms available for solving adaptive filtering problem, are derived from iterative methods for solving linear system of equations. For example, the most widely implemented adaptive algorithms in practice, namely the Least Mean Square (LMS) algorithm is derived from the method of steepest descent. Other search methods which have found their place in solving adaptive filtering problem are the direction set method and the conjugate gradient method [3, 4, 5].

In this paper, we will review several algorithms which are derived from four different iterative search methods, namely the method of steepest descent, the Newton's method, the direction set method and the conjugate gradient method. Our objectives are 1) to provide a clear understanding of their implementation in solving adaptive filtering problem; 2) to evaluate their performance in terms of rate of convergence, misadjustment, computational complexity and sensitivity towards eigenvalue spread, and, 3) to recommend improvements to the algorithms.

2 Adaptive Filtering Problems

Adaptive filtering problem is a filter design technique which allows for adjustable coefficients that can be optimized to minimize some measure of error. Mathematical formulation for adaptive filtering problem usually takes the form of an adaptive least squares problem, where the value of the filter coefficients are adjusted so that they are optimized in the least squares sense. In contrast with the standard least squares problem, the sum of squared error function in the adaptive least squares problem is a time varying function which adapts itself with the time varying input data.

2.1 Adaptive Filters

A schematic diagram of an adaptive filter is given in Fig. 1 below where for every input signal $u(n)$, the filter produces output $y(n)$. This output is compared with a desired signal $s(n)$to produce an error signal $e(n)=y(n)-s(n)$ which is the difference between the output and the desired signal. The objective in the design an adaptive filter is to adjust its parameters so that the error $e(n)$ is minimized.

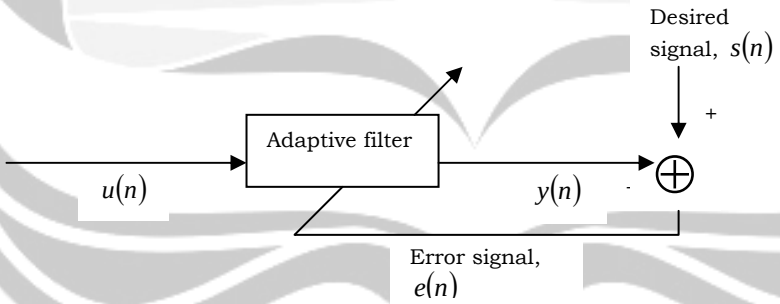


Fig.1 Block diagram for an adaptive filter

2.2 Mathematical Fomulations for Adaptive Filtering Problems

For a linear linear transversal filter, the output to filter at time n is given by,

$$y_n = \sum_{j=0}^{N-1} x_j u_{n-j}$$

where N is the filter order and $x(j)$ is the j th coefficient of the filter. The n th state of the sum of squared error function is the sum of squared errors from time 0 up to n, which is given by the equation below,

$$J_n(\mathbf{x}) = \sum_{i=0}^{n} \lambda_i^{(n)} \left(\mathbf{a}_i^T \mathbf{x} - s_i \right)^2 \tag{1}$$

where $\mathbf{a}_i = [u_i \quad u_{i-1} \quad \dots \quad u_{i-N+1}]$, $x = [x_0 \quad x_1 \quad \cdots \quad x_{N-1}]^T$ and $\lambda_i^{(n)}$ can be in two forms, either $\lambda_i^{(n)} = \frac{1}{n}$ or $\lambda_i^{(n)} = \lambda^{n-i}$ where $0 < \lambda < 1$. The first choice of $\lambda_i^{(n)}$ gives rise to an average sum of squared error whereas the second choice gives an exponentially weighted sum of squared error and the weighting factor λ is referred to as the forgetting factor which is intended to ensure that the past data are "forgotten" in order to track the statistical variations of the data in nonstationary environment.

The adaptive least squares problems is the problem of minimizing the cost function (1) with respect to filter coefficients x_j, $j = 0, \dots, N-1$. In matrix form, the adaptive least squares problem may be represented as follows

$$\min_{\mathbf{x}} J_n(\mathbf{x}) = \min_{\mathbf{x}} \left(\mathbf{b}_n{}^T \mathbf{b}_n - 2\mathbf{x}^T \mathbf{A}_n \mathbf{b}_n + \mathbf{x}^T \mathbf{A}_n \mathbf{A}_n^T \mathbf{x} \right) \tag{2}$$

where

$$\mathbf{A}_n = \left[\sqrt{\lambda_0^{(n)}} \mathbf{a}_0, \sqrt{\lambda_1^{(n)}} \mathbf{a}_1, \dots \sqrt{\lambda_n^{(n)}} \mathbf{a}_n \right]^T \text{ and } \mathbf{b}_n = \left[\sqrt{\lambda_0^{(n)}} s_0, \sqrt{\lambda_1^{(n)}} s_1, \dots, \sqrt{\lambda_n^{(n)}} s_n \right].$$

The quantity $\mathbf{R} = \mathbf{A}_n \mathbf{A}_n^T$ is identified as the autocorrelation matrix of the input signal and the vector $\mathbf{p} = \mathbf{A}_n \mathbf{b}_n$ is the vector of cross-correlation between the input and the desired signal.

3 Iterative Search Methods and Application in Adaptive Filtering Algorithms

We note that the minimization problem given in (2) has an exact solution at every state n, which is the solution to a system of $N \times N$ linear equation

$$\mathbf{R}\mathbf{x} = \mathbf{p}$$

so that at every state n, the optimum solution is given by

$$\mathbf{x}_{opt}(n) = \mathbf{R}(n)^{-1} \mathbf{p}(n).$$

Hence the adaptive least squares problem can be viewed as an adaptive search for the solution to a (time varying) system of $N \times N$ linear equations.

In this section we will review three classes of algorithms which are based on four iterative search methods, the steepest descent method, the Newton's method, the direction set method and the Conjugate Gradient method.

3.1 LMS and LMS-Newton Algorithm

In this section we will be discussing the most widely used adaptive algorithm and that is the Widrow-Hoff Least Mean Square (LMS) algorithm derived in 1959 [8]. It is based on the method of steepest descent where the coefficient vector $\mathbf{x}$ is updated along the direction of steepest descent, i.e. the negative gradient. This gives the following recursion formula for the coefficient vector,

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \mu\nabla \tag{4}$$

where ∇ denotes the gradient vector. For the least squares problem, the gradient vector is given by $\nabla = 2(\mathbf{R}\mathbf{x} - \mathbf{p})$. The LMS algorithm is obtained by replacing the gradient vector with its instantaneous value and that is $-2e_n\mathbf{x}_n$ where $e_n = \mathbf{y_n} - s_n$ is the instantaneous error.

In the Newton's method, the gradient vector in (4) is scaled with a factor $\lambda_{ave}\mathbf{R}^{-1}$ where λ_{ave} is the average of the eigenvalues of $\mathbf{R}$, giving

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \mu\lambda_{ave}\mathbf{R}^{-1}\nabla$$

Using the instantaneous value as the estimate for the gradient we have the following recursion formula for the LMS-Newton algorithm

$$\mathbf{x}_{n+1} = \mathbf{x}_n + 2\mu\lambda_{ave}\mathbf{R}^{-1}e_n\mathbf{x}_n .$$

It is a well known fact that the LMS algorithm is sensitive to the spread in the eigenvalue of the correlation matrix. This property is inherited from the method of steepest descent where several modes of convergence exist which corresponds to the number of distinct eigenvalues in $\mathbf{R}$. Because prior knowledge of $\mathbf{R}$ is rarely known, this property makes the rate of convergence of the LMS algorithm unpredictable. The LMS-Newton algorithm is considered as an improved version where it only has one mode of convergence [9]. However, because the LMS-Newton algorithm requires knowledge of $\mathbf{R}^{-1}$, this algorithm cannot be implemented in practice.

3.2 The Direction Set Based Algorithm

The direction set method is inherited from the Powell and Zangwill method for optimizing unconstrained minimization problem [8, 13]. Given a starting estimate $\mathbf{x}$ and a set of N linearly independent direction $\{\mathbf{d}_1, \ldots, \mathbf{d}_N\}$, the direction set method searches along each direction sequentially for a better estimate. The search through N directions is called one cycle. Before the next search cycle, directions may or may not be modified (depending on the linear independence of the new set of directions), and a new starting estimate maybe chosen. The iterative algorithm for updating the coefficient vector within each cycle takes the form

$$\mathbf{x}_{i+1}^{(n)} = \mathbf{x}_i^n + \alpha_i^{(n)}\mathbf{d}_i^{(n)}, \qquad i = 1, \ldots N$$

where $\alpha_i^{(n)}$ is the stepsize and $\mathbf{d}_i^{(n)}$ is the search direction. The optimal stepsize can be obtained by setting $\nabla_\alpha J(\mathbf{x} + \alpha\mathbf{d}) = 0$ which gives

$$\alpha_i^{(n)} = -\frac{\mathbf{d}_i^{(n)T}\left(\mathbf{R}(n)\mathbf{x}_i^{(n)} - \mathbf{p}(n)\right)}{\mathbf{d}_i^{(n)T}\mathbf{R}(n)\mathbf{d}_i^{(n)}}. \quad (5)$$

The simplest form of the direction set method is obtained by choosing the Euclidean directions as the search direction at each cycle, i.e. $\mathbf{d}_i^{(n)} = [0 \ldots 0\,1\,0 \ldots 0]^T$ where the 1 appears in the i th position. This gives rise to the Euclidean Direction Search (EDS) algorithm. A further modification of the EDS algorithm can be found in [11, 12] (Fast EDS algorithm) and [2] (Scaled EDS algorithm).

3.3 The Conjugate Gradient Based Algorithm

The Conjugate Gradient method (CG) looks for a set of linearly independent direction vectors $\{\mathbf{d}_1, \ldots, \mathbf{d}_N\}$ which are conjugate with respect to $\mathbf{R}$ so that the solution vector $\mathbf{x}^*$ can be expressed as

$$\mathbf{x}^* = \alpha_1\mathbf{d}_1 + \alpha_2\mathbf{d}_2 + \ldots + \alpha_N\mathbf{d}_N .$$

Minimization of the cost function J gives rise to the same formula for the stepsize α as given in (5) although conjugacy with respect to $\mathbf{R}$ is not a requirement for the search directions in the direction set method.

In order to maintain conjugacy with respect to $\mathbf{R}$, the i th search direction is updated as follows

$$\mathbf{d}_i = \mathbf{g}_{i-1} + \beta_{i-1}\mathbf{d}_{i-1}$$

where $\mathbf{g}_i$ is the gradient vector at the i th iteration and β_i is a scalar value given as $\beta_i = \frac{\mathbf{g}_i^T\mathbf{g}_i}{\mathbf{g}_{i-1}^T\mathbf{g}_{i-1}}$. Since the gradient vector in the adaptive filtering problem is dependent on the correlation matrix $\mathbf{R}$ and the cross-correlation vector $\mathbf{p}$, estimates of the gradient requires estimates of both $\mathbf{R}$ and $\mathbf{p}$. We will highlight two ways of implementing the CG method in adaptive filtering problem [3, 4] which employs two different technique of estimating $\mathbf{R}$ and $\mathbf{p}$:

i) The cost function is assumed to be the averaged sum of squared error, i.e.

$$J_n(\mathbf{x}) = \frac{1}{n}\sum_{i=0}^{n}\left(\mathbf{a}_i^T\mathbf{x} - s_i\right)^2$$

so that the correlation matrix $\mathbf{R}_n = \frac{1}{n}\mathbf{A}_n\mathbf{A}_n^T$, where $\mathbf{A}_n = [\mathbf{a}_0, \mathbf{a}_1, \ldots, \mathbf{a}_n]^T$ and the crosscorrelation vector becomes $\mathbf{p}_n = \frac{1}{n}\mathbf{A}_n\mathbf{b}_n$ where $\mathbf{b}_n = [s_0, s_1, \ldots, s_n]^T$. In the CG algorithm for adaptive filtering, the finite sliding data windowing is used to estimate $\mathbf{R}_n$ and $\mathbf{p}_n$, where, only data samples inside a window of finite length M are used. Hence we have the following estimates,

$$\mathbf{R}_n = \frac{1}{M}\sum_{j=n-M+1}^{n}\mathbf{a}_j\mathbf{a}_j^T\,, \qquad \mathbf{p}_n = \frac{1}{M}\sum_{j=n-M+1}^{n}s(j)\mathbf{a}_j\,.$$

For every incoming data sample, the conjugate gradient iteration is run $k_{\max}$ times, where $k_{\max} = \min(N, M)$.

ii) The cost function is assumed to be the exponentially weighted sum of squared error

$$J_n(\mathbf{x}) = \sum_{i=0}^{n}\lambda^{n-i}\left(\mathbf{a}_i^T\mathbf{x} - s_i\right)^2$$

Using an exponentially decaying data windowing, the correlation matrix and the cross-correlation vector may be updated recursively as follows,

$$\mathbf{R}_n = \lambda\mathbf{R}_{n-1} + \mathbf{a}_n\mathbf{a}_n^T$$
$$\mathbf{p}_n = \lambda\mathbf{p}_{n-1} + s(n)\mathbf{a}_n$$

In this implementation, the conjugate gradient iteration is performed once for every incoming data sample.

4 Comparative performance

In our discussions here, we will be assessing the performance of adaptive algorithms in section 3 in the framework of adaptive system modeling. The block diagram for adaptive system modeling is in Fig. 2. The input signal is filtered through a colouring filter with the frequency response $H(z) = \frac{\sqrt{1-\alpha^2}}{1-\alpha z^{-1}}$, where $|\alpha| < 1$. The parameter α controls the eigenvalue spread of the input correlation matrix, where $\alpha = 0$ gives uncorrelated sequence (white) with small eigenvalue spread. The aim is to find the parameters of a model $\mathbf{x}$ through an adaptive algorithm so that the difference between the unknown system output, $d(n)$, and the adaptive model output, $y(n)$, is minimized according to some specified cost function. Noise, $\eta(n)$, with a variance of 0.001 is added to the output of the unknown system.

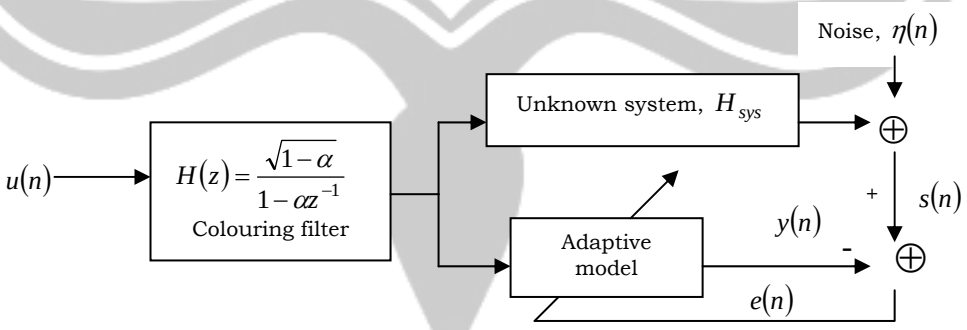


Fig. 2 Adaptive system modeling

4.1 Rate of Convergence and Misadjustment

An efficient adaptive algorithm is one that minimizes usage of data without compromising the quality of solution. In other words we require the algorithm to have a reasonably high rate of convergence and at the same time it keeps the solution as close as possible to the optimum solution. Commonly, in adaptive filtering problems, the rate of convergence is assessed by the number of iterations required to achieve the steady state mean squared error (MSE). In addition to that, the quality of the steady state solution is measured through the quantity called misadjusment which is the ratio of the excess MSE (the difference between the steady state MSE and the MSE corresponds to the optimum solution) to the steady state MSE. In our discussions here, because the exact solution is available, we will be evaluating rate of convergence and misadjustments by looking at the progression of error between the coefficient vector of the unknown system and that of the adaptive filter.

Fig. 3 (i) displays the progression of error (computed in norm-2) as the number of iteration increases. It is clearly shown that both LMS and LMS-Newton algorithms have comparable rate of convergence where a steady state error is achieved after 600 iterations. However the EDS and the Conjugate Gradient based algorithms provide a much superior convergence rate, where steady state error is achieved only after about 40 iterations.

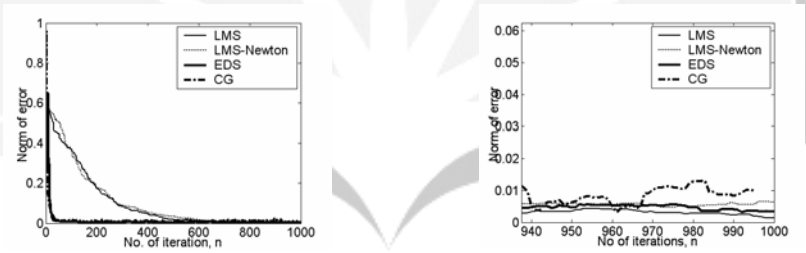


Fig. 3 Progression of error as iteration increases. Error is measured by the norm of the difference between the coefficient vector of the unknown system and the coefficient vector of the adaptive model ($N = 5$)

The steady state errors shown in Fig. 3 (ii) gives a better view of misadjustments produced from the solutions. It can be seen that although conjugate gradient method has a high rate of convergence, it does tend to produce higher misadjustment compared to the other algorithms.

4.2 Computational complexity

Another important assessment criterion is the computational complexity of the algorithm. A higher computational complexity means the algorithm requires a higher storage capacity and computation time. High computational complexity also

makes the algorithm more susceptible to round off errors, hence, reducing the quality of solution.

Table 1 below summarizes the computational complexity of the algorithms discussed as a function of the adaptive filter order, N.

Table 1 Computational complexity

Algorithm	Complexity
LMS	$\mathrm{O}(N)$
EDS	$\mathrm{O}(N^2)$
Fast EDS	$\mathrm{O}(N)$
CG (Implementation (i))	$\mathrm{O}(N^2)$
CG (Implementation (ii))	$\mathrm{O}(N^2)$

To see the effect of computational complexity on convergence and misadjustments, we plot the error profile for a much larger N in Fig. 4.

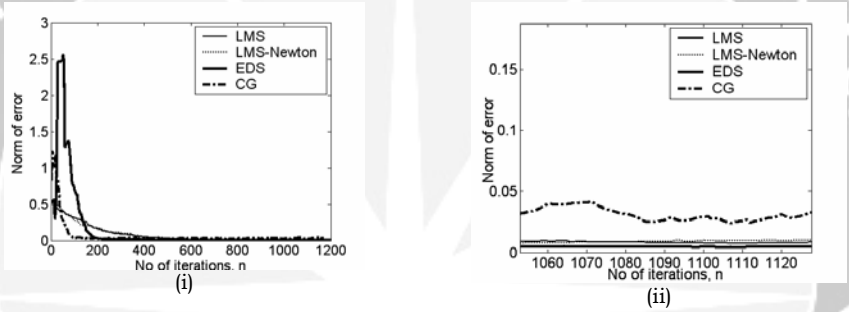
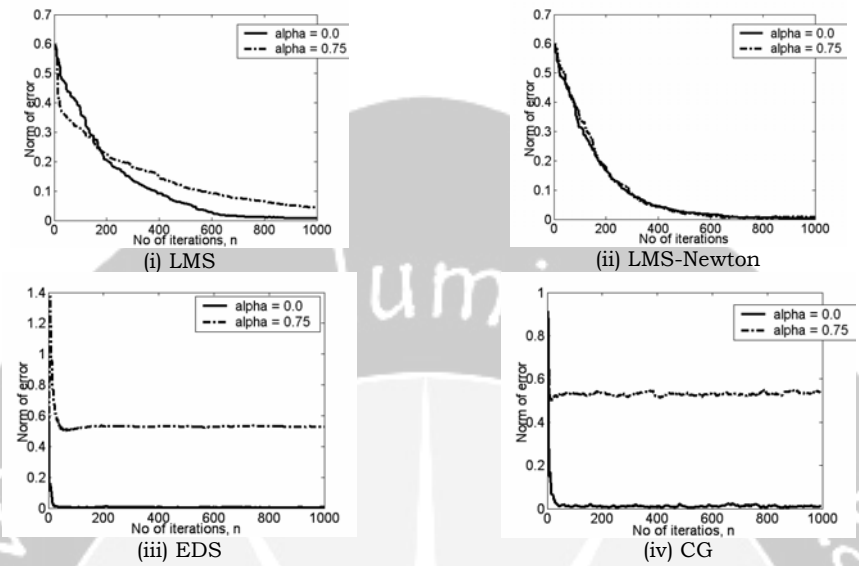


Fig. 4 Rate of convergence and misadjustment for $N = 30$

Increasing the value of N has little effect on the LMS-based algorithms. However, for the EDS algorithm, we see a slower convergence rate than in previously and a significant increase in the initial errors. The misadjustment for the EDS algorithm remains comparable to that of the LMS-based algorithm. For the CG algorithm, although the convergence rate remains unchanged, we see a significant increase in the misadjustment.

4.3 Sensitivity to eigenvalue spread

We now compare the sensitivity of the solutions for our problem towards eigenvalue spread for all the algorithms. Note that eigenvalue spread is defined as $\lambda_{\max}/\lambda_{\min}$, where $\lambda_{\max}$ is the largest eigenvalue of R and $\lambda_{\min}$ is the smallest eigenvalue.



(i) LMS (ii) LMS-Newton

(iii) EDS (iv) CG

Fig.5 Sensitivity towards eigenvalue spread. The value of eigenvalue spread is controlled by the parameter α.

As predicted, the convergence rate of the LMS algorithm is reduced for large eigenvalue spread and almost no effect on the LMS-Newton algorithm is found. For both the EDS and CG algorithm, although eigenvalue spread has little effect on their rate of convergence, the misadjustments suffers terribly from it.

5 Conclusions

We have reviewed four different adaptive algorithms for solving adaptive filtering problem, namely the LMS algorithm, the LMS-Newton algorithm, the direction set based method and the conjugate gradient based method. These algorithms are treated as iterative search method for solving a time-varying linear systems of equation of the form $\mathbf{Rx} = \mathbf{p}$, where $\mathbf{R}$ corresponds to the time-varying correlation matrix of the adaptive problem and $\mathbf{p}$ corresponds to the time-varying cross-correlation vector.

Performance evaluation of the algorithms are conducted within the framework of an adaptive system modeling problem. The EDS and the CG method proved to have superior rate of convergence compared to the LMS-based method. However, due to the high computational complexity of the CG algorithm, it tends to give higher misadjustment compared to the other algorithms.

The convergence rate of the EDS and the CG algorithm is not affected much by the increase in eigenvalue spread. However, poor misadjustments are obtained. To this end, we note that the eigenvalue spread is the same as the condition number of R. Therefore, preconditioning the algorithms with a suitable preconditioning matrix will help reduce this problem.

Acknowledgment

The author acknowledges support from Universiti Sains Malaysia for the travel grants to attend ICAM05.

References

[1] Ahmad, N.A, Lye, W.K & Ho, Y.Y. (2005), Euclidean Direction Search Algorithm for Adaptive Least Squares Problems, *Proceedings of the 13th National Symposium on Mathematical Sciences,*Universiti Utara Malaysia, **1**, 193-201.

[2] Ahmad, N.A., Lye, W.K. & Ho, Y.Y. (2005), Investigation into Modified Euclidean Direction Search Algorithm for Adaptive Least Squares Problem, *Proceedings of the 2nd International Conference on Research and Education in Mathematics,* Universiti Putra Malaysia (CD Proceedings).

[3] Boray, G. & Srinath, M. (1992), Conjugate Gradient Techniques for Adaptive Filtering, *IEEE Trans. Circuits and Syst. I,* **39**(1), 1-10.

[4] Chang, P.S. & Wilson, A.N. (2000), Analysis of the Conjugate Gradient Algorithms for Adaptive Filtering, *IEEE Trans Signal Processing*, **48**(2), 409-418.

[5] Chen, M.Q. (1998), A Direction set based algorithm for least squares problems in adaptive signal processing, *Linear Algebra and its Applications*, **284**, 73-94.

[6] Chen, M.Q., Bose, T. & Xu, G.F. (1999), A direction set based algorithm for adaptive filtering, *IEEE Trans. Signal Processing* **47**(2), 535 – 539.

[7] Farhang-Boroujeny, B. (2000), *Adaptive Filters:Theory and Applications*, John Wiley & Sons, England.

[8] Powell, M.J.D. (1964), An Efficient Method for Finding the Minimum of a Function of Several Variables Without Calculating Derivatives, *Comput. J.* **7**, 155-162.

[9] Widrow, B. & Hoff, M.E. (1960), Adaptive Switching Circuits, *IREWESCON Conv. Rec.*, **4**, 96-104.

[10] Widrow, B. & Kamenetsky, M. (2003), On the Statistical Efficiency of the LMS Family of Adaptive Algorithms, 2872-2880.

[11] Xu, G.F., Bose, T. & Schroeder, J. (1998). Channel equalization using an euclidean direction search adaptive algorithm, *Global Telecommunications Conference (GLOBECOM 98), The Bridge to Global Integration. IEEE.* **6**, 3479 – 3484.

[12] Xu, G.F., Bose, T. & Schroeder, J. (1999). The Euclidean direction search algorithm for adaptive filtering, *Proceedings of the 1999 IEEE International Symposium on Circuits and Systems.* **3**,146 – 149.

[13] Zangwill, W. (1967), Minimizing a Function Without Calculating Derivatives, *Comput. J.* **10**, 293-296.

AHMAD, N.A.: School of Mathematical Sciences, Universiti Sains Malaysia, 11800 Minden, Penang, Malaysia. Phone: +6 +4 6533656, Fax: +6 +4 6570910.
E-mail: atinah@cs.usm.my

Phasor Estimation Under Non-Stationary Conditions

Praveen Pankajakshan

Texas A&M University

Abstract: This paper describes the design and analysis of a recursive Kalman Filter for phasor estimation of non-stationary signals. To detect the transmission or distribution line disturbances (or faults), it is necessary to track the amplitude and phase of the steady state, post-disturbance (post-fault) fundamental signal from the distorted signal. The model assumes a constant-frequency, rotating phasor for the sampled voltage and current signals. The filter was developed in the Simulink environment of MATLAB, and the M Programming Language was used to model the filter. The parameters of the filter and the model are input through S-functions. The developed filter is an extended form of the discrete Kalman filter and estimates the phasors in the presence of decaying dc offsets, odd or even harmonic distortions, and measurement noise. The model settings and parameters are customizable depending on the input signal and the noise characteristics. The filter assumes a white noise characteristic for the signal variance and the measurement noise variance. However, in a real-time system, the signal and the noise frequencies are band limited and are therefore they are not perfectly white. The model can still be applied but the estimation error will be different from the test bench created signal. The open-system solution and user-friendly nature of our interface makes it a useful educational tool for students and engineers in understanding the mathematical theory and reasoning behind power system protection and voltage restoration approaches. Performance evaluation of the algorithm on estimation is also shown in comparison with other non-recursive techniques. The filter has a faster convergence time and the mean square difference between the output and actual values is a minimum. The results of the sensitivity study on the changes in filter parameters, signal sampling rate, and fundamental frequency drifts reveals the effectiveness of the developed scheme.

Keyword: signal analysis, recursive kalman filter, phasor estimation, state-space model, noise variance, digital relay, voltage restorer

On the Sufficient Condition for Solvability of Singular LQR Problem for Descriptor Systems

Muhafzan[a], Malik Hj. Abu Hassan [b], Fudziah Ismail [b] & Leong Wah June[b]

[a]Department of Mathematics,Universitas Andalas, Padang, Indonesia
[b]Department of Mathematics, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, Malaysia

Abstract. The main aim of this paper is to solve the singular LQR problem for descriptor systems. By using the Weierstrass-Kronecker canonical reprentation to the descriptor system, the singular LQR problem can be transformed into two LQR problems, i.e LQR problem for proper descriptor system and LQR problem for nonproper descriptor system. The sufficient conditions for solvability of both LQR problems are established by making use of the semidefinite programming technique.

Key-words: Descriptor system, singular LQR problem, semidefinite programming

1 Introduction

Consider the following linear quadratic regulator (LQR) problem:

$$\min \; J(u, x_0) = \int_0^\infty y^T(t)\, y(t)\, dt \tag{1}$$

$$\text{s.t.} \begin{cases} E\dot{x}(t) = Ax(t) + Bu(t), \quad Ex(0) = x_0 \in \mathbb{R}^n \\ y(t) = Cx(t) + Du(t) \end{cases} \tag{2}$$

where $E, A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times r}, C \in \mathbb{R}^{q\times n}, D \in \mathbb{R}^{q\times r}$ are time invariant, $rank(E) \equiv p < n$, $\det(sE - A) \neq 0$ for some $s \in \mathbb{R}$ and $D^T D \geq 0$. Here $x(t) \in \mathbb{R}^n$ denotes the state variable, $y(t) \in \mathbb{R}^q$ denotes the output variable, $u(t)$ denotes the control (input) variable with $u(t) \in L^2(\mathbb{R}^r)$, the set of all $\mathbb{R}^r$ - valued, measure functions satisfying $\int_0^\infty \| u(t) \|^2 dt < \infty$, where $\| u(t) \| = \left[\sum_i u_i(t)^2 \right]^{\frac{1}{2}}$. For simplicity, we denote the above LQR problem as Π. Remembering the infinite horizon nature of the problem, we further require the control to be stabilizing such that the corresponding state trajectory converges to zero as time goes to infinity.

A fundamental question needs to be answered is concerning the existence of the optimal control-state pair (x^*, u^*) for the problem Π, i.e.under what conditions does there exist an optimal control-state pair (x^*, u^*)?. For the case in which

$D^T D > 0$ and $\det(sE - A) \neq 0$ for some $s \in \mathbb{R}$, has already been answered exhaustively in [3,8,9]. To the best of the author's knowledge, not much work has been reported for the case $D^T D \geq 0$ (singular case). In general setting $D^T D \geq 0$, the existing LQR theories always involve the impulse distributions [4,8]. Thus it does not provide any answer to a basic question such as when the LQR problem presented above posseses an optimal solution in the form of a conventional control, in particular, one that does not involve impulse distribution. It remains and still open, and this motivates our present work.

Nevertheless, the paper from Zhu *et al.* [11] is of interest to consider. It solves the case in which the descriptor system need not be regular and the output vector does not depend on the input vector. Muhafzan *et al.* [6] generalize the problem in [11] for the case in which the output vector depends on the input vector. Furthermore, Muhafzan *et al.* [7] has solved the singular LQR problem for descriptor system of finite horizon case by transforming the problem Π into two classes, i.e. LQR problem for proper descriptor system and LQR problem for nonproper descriptor system. By a certain condition, proper part can be separated into two cases, i.e. regular LQR and singular LQR. However, for infinite horizon case, the regular LQR for proper descriptor system is a classical problem and its solution can be obtained elegantly via the Riccati equation

$$P(A_1 - B_1 Q_{22}^{-1} Q_{12}^T) + (A_1 - B_1 Q_{22}^{-1} Q_{12}^T)^T P - P B_1 Q_{22}^{-1} B_1^T P + Q_{11} - Q_{12} Q_{22}^{-1} Q_{12}^T = 0. \quad (3)$$

Furthermore, the optimal control (provided it is stabilizing) can be expressed explicitly in a feedback form,

$$u^*(t) = -Q_{22}^{-1}(B_1^T P + Q_{12}^T) x_1^*(t).$$

There are still some open problems that have not been tackled in [7], namely, the singular part of LQR problem for proper and nonproper which ill-defined.

The objective of this paper is to study these open problems and provide the sufficient condition so that the optimal control $u^*(t)$ of the singular LQR problem for descriptor system Π exists. To handle this problem, we use the semidefinite programming (SDP) technique in which, it has been solved successfully recently the singular LQR problem for standard state space system [2,10].

Notation: Throughout this paper, the superscript T denotes the transpose, $\mathbb{R}^n$ denotes the n-dimensional Euclidean space, $\mathbb{R}^{m\times n}$ denotes the set of all $m \times n$ real matrices and I is the identity matrix with appropriate dimension.

2 Transformasi of the problem

To discuss the solution of Π, we use the Weierstrass-Kronecker canonical representation [8,9] for the descriptor system (2). It is well known that under condition $\det(sE - A) \neq 0$, there exist the nonsingular matrices $L, M \in \mathbb{R}^{n\times n}$ so that

$$LEM = \begin{bmatrix} I_p & 0 \\ 0 & N \end{bmatrix},\ LAM = \begin{bmatrix} A_1 & 0 \\ 0 & I_{n-p} \end{bmatrix},\ LB = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix},\ CM = \begin{bmatrix} C_1 & C_2 \end{bmatrix}.$$

where $A_1 \in \mathbb{R}^{p\times p}$, $B_1 \in \mathbb{R}^{p\times r}$, $B_2 \in \mathbb{R}^{(n-p)\times r}$, $C_1 \in \mathbb{R}^{q\times p}$, $C_2 \in \mathbb{R}^{q\times(n-p)}$, N is a nilpotent matrix of index k (i.e. $N^k = 0, N^{k-1} \neq 0$) defining the index of the linear descriptor systems. Accordingly, let

$$M^{-1}x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{4}$$

where $x_1 \in \mathbb{R}^p$ and $x_2 \in \mathbb{R}^{(n-p)}$. By the transformation (4), the performance index (1) can be rewritten as

$$J_1 = \int_0^\infty \begin{bmatrix} x_1(t) \\ x_2(t) \\ u(t) \end{bmatrix}^T \begin{bmatrix} C_1^TC_1 & C_1^TC_2 & C_1^TD \\ C_2^TC_1 & C_2^TC_2 & C_2^TD \\ D^TC_1 & D^TC_2 & D^TD \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ u(t) \end{bmatrix} dt \tag{5}$$

and the descriptor system (2) becomes:

$$\begin{cases} \dot{x}_1(t) = A_1x_1(t) + B_1u(t),\ x_1(0) = x_{10} \in \mathbb{R}^p \\ N\dot{x}_2(t) = x_2(t) + B_2u(t) \\ y(t) = C_1x_1(t) + C_2x_2(t) + Du(t) \end{cases} \tag{6}$$

where $x_{10} = [\, I_p \quad 0\,]Lx_0$. Now we must minimize (5) subject to the system (6). The solution of the first differential equation of the system (6) is easily obtained in the classical manner, meanwhile the solution of the second equation of (6) is given

$$x_2(t) = -\sum_{i=0}^{k-1} N^i B_2 u^{(i)}(t) \tag{7}$$

where $u^{(i)}(t)$ denotes i^{th} derivative of $u(t)$. One can see that x_2 generally depends on the higher order time derivatives of control which is a very unusual behavior and must be regarded very carefully. In the following we distinguish the two cases where the solution depends either only on $u(t)$ but not on its derivatives $\dot{u}, \ddot{u}, ..., u^{(k-1)}$ or on $u(t)$ and its derivatives $\dot{u}, \ddot{u}, ..., u^{(k-1)}$.

Definition 2.1. [9] *System* (1) *is termed as proper if its solution depends only on* $u(t)$ *but not on its derivatives* $\dot{u}, \ddot{u}, \cdots, u^{(k-1)}$. *Otherwise the system is nonproper.*

A criterion for properness is derived immediately from Definition 2.1.

Lemma 2.2. *The descriptor system* (2) *or* (6) *is proper if and only if* $NB_2 = 0$ *holds.*

The optimization of Π has to be performed in accordance to the properness and nonproperness of the descriptor system and so two different optimization problems have to be considered.

If $NB_2 = 0$ then solution of the second equation of (6) is

$$x_2(t) = -B_2 u(t) \tag{8}$$

Creating the transformation

$$\begin{bmatrix} x_1(t) \\ x_2(t) \\ u(t) \end{bmatrix} = \begin{bmatrix} I_p & 0 \\ 0 & -B_2 \\ 0 & I_r \end{bmatrix} \begin{bmatrix} x_1(t) \\ u(t) \end{bmatrix} \tag{9}$$

and replace (9) into the performance index (5) and the system (6), we have the following standard LQR problem:

$$\min\ J(u, x_{10}) = \int_0^\infty \begin{bmatrix} x_1(t) \\ u(t) \end{bmatrix}^T \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{12}^T & Q_{22} \end{bmatrix} \begin{bmatrix} x_1(t) \\ u(t) \end{bmatrix} dt \tag{10}$$

$$\text{s.t.} \begin{cases} \dot{x}_1(t) = A_1 x_1(t) + B_1 u(t),\ x_1(0) = x_{10} \in \mathbb{R}^p \\ y(t) = C_1 x_1(t) + (-C_2 B_2 + D) u(t), \end{cases} \tag{11}$$

where $Q_{11} = C_1^T C_1$, $Q_{12} = C_1^T (D - C_2 B_2)$ and $Q_{22} = (D - C_2 B_2)^T (D - C_2 B_2)$. It is easy to verify that $Q_{11} > 0$ if and only if $rank\,(C_1) = p$.

Lemma 2.3. [7] $Q_{22} > 0$ *if and only if* $rank\,[C \quad D] = r$.

The optimization problem (10) and (11) is termed as LQR for proper descriptor system, and for simplicity we denote as Π_1.

If $NB_2 \neq 0$, we need an extension of state and control variables to deal correctly with the influence of the time derivatives of the control input. Let

$$v_1 = u,\ v_2 = \dot{u},\ v_3 = \ddot{u}, \cdots,\ v_{k-1} = u^{(k-2)},\ w = u^{(k-1)},$$

then (7) can be written as

$$x_2(t) = -B_2 v_1(t) - NB_2 v_2(t) - \cdots - N^{k-2} B_2 v_{k-1}(t) - N^{k-1} B_2 w(t) = -[0 \quad \overline{B}_2] x_e(t) - N^{k-1} B_2 w(t)$$

where

$$x_e = \begin{bmatrix} x_1^T & v_1^T & \cdots & v_{k-1}^T \end{bmatrix}^T, \quad \overline{B}_2 = \begin{bmatrix} B_2 & NB_2 & N^2B_2 & \cdots & N^{k-2}B_2 \end{bmatrix}$$

Here $w(t)$ is considered as a new control variable. Create the transformation

$$\begin{bmatrix} x_1(t) \\ x_2(t) \\ u(t) \end{bmatrix} = \left[\begin{array}{c|c} I_p & 0 \\ \hline 0 & -\overline{B}_2 \\ \hline 0 & I_{r,0} \end{array}\right] \left[\begin{array}{c|c} 0 & \\ \hline -N^{k-1}B_2 & \\ \hline 0 & \end{array}\right] \begin{bmatrix} x_e(t) \\ w(t) \end{bmatrix} \tag{12}$$

where $I_{r,0} = \begin{bmatrix} I_r & 0 & \cdots & 0 \end{bmatrix}$. Substituting (12) into the performance index (5), the first and third equation of (6), we have the following LQR problem:

$$\min \; J(w, x_{e0}) = \int_0^\infty \begin{bmatrix} x_e(t) \\ w(t) \end{bmatrix}^T \begin{bmatrix} \overline{Q}_{11} & \overline{Q}_{12} \\ \overline{Q}_{12}^T & \overline{Q}_{22} \end{bmatrix} \begin{bmatrix} x_e(t) \\ w(t) \end{bmatrix} dt \tag{13}$$

$$\text{s.t.} \begin{cases} \dot{x}_e(t) = A_e x_e(t) + B_e w(t), & x_e(0) = x_{e0} \in \mathbb{R}^{p+(k-1)r} \\ y(t) = \overline{C} x_e(t) + (-C_2 N^{k-1} B_2)\, w(t) \end{cases} \tag{14}$$

where

$$A_e = \begin{bmatrix} A_1 & B_1 & 0 & \cdots & 0 \\ 0 & 0 & I_r & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & I_r \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}, \quad B_e = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ I_r \end{bmatrix}, \quad \overline{C} = \begin{bmatrix} C_1^T \\ (DI_{r,0} - C_2\overline{B}_2)^T \end{bmatrix}^T,$$

$$\overline{Q}_{11} = \begin{bmatrix} C_1^T \\ (DI_{r,0} - C_2\overline{B}_2)^T \end{bmatrix} \begin{bmatrix} C_1^T \\ (DI_{r,0} - C_2\overline{B}_2)^T \end{bmatrix}^T, \quad \overline{Q}_{12} = \begin{bmatrix} -C_1^T C_2 N^{k-1} B_2 \\ (DI_{r,0} - C_2\overline{B}_2)^T C_2 N^{k-1} B_2 \end{bmatrix} \text{ and}$$

$$\overline{Q}_{22} = (C_2 N^{k-1} B_2)^T (C_2 N^{k-1} B_2).$$

It is easy to verify that $\overline{Q}_{11} > 0$ if and only if $rank \begin{bmatrix} C_1^T \\ (DI_{r,0} - C_2\overline{B}_2)^T \end{bmatrix} = p + (k-1)r.$

The optimization problem (13) and (14) is termed as LQR for nonproper descriptor system, and for simplicity we denote as Π_2.

3 Results

3.1 LQR problem for Proper Descriptor System

Now, reconsider Π_1. According to the LQR theory for standard state space system [1], if the matrix Q_{22} is positive definite $(Q_{22} > 0)$ and the LQR Π_1 is stabilizable, then the optimal control is given by

$$u^* = -Q_{22}^{-1}(Q_{12}^T + B_1^T P)x_1^*$$

where x_1^* satisfies

$$\dot{x}_1 = \left(A_1 - B_1 Q_{22}^{-1}(Q_{12}^T + B_1^T P)\right) x_1, \quad x_1(0) = x_{10},$$

with P the unique positive semidefinite solution of Riccati equation (3).

Theorem 3.1.1. [7] If $NB_2 = 0$ and $rank\begin{bmatrix} C & D \end{bmatrix} = r$ *then the* LQR *problem* Π *has a unique optimal control-state pair.*

On the other hand, if $rank\begin{bmatrix} C & D \end{bmatrix} < r$ (Q_{22} to be singular) then the Riccati equation (3) seems to be meaningless. However, to handle this situation we can use SDP to solve the problem.

Consider the following equation which is a generalization of the classical Riccati equation (3) as follows:

$$F(P) \equiv A_1^T P + PA_1 + Q_{11} - (PB_1 + Q_{12})Q_{22}^{\dagger}(B_1^T P + Q_{12}^T) = 0 \tag{15}$$

where $Q_{22}^{\dagger}$ stands for the pseudo inverse of Q_{22}. Next, we introduce an affine transformation of matrix P as follows:

$$\pounds(P) \equiv \begin{bmatrix} Q_{22} & B_1^T P + Q_{12}^T \\ PB_1 + Q_{12} & Q_{11} + A_1^T P + PA_1 \end{bmatrix} \tag{16}$$

Note that if $Q_{12} = 0$, (15) and (16) coincide to the equations (3) and (4) in [10]. Corresponding to the above LQR problem Π_1, we introduce SDP and its associated dual problem as follows:

$$\text{(P)} \qquad \begin{aligned} &\max \ \langle I, P \rangle \\ &\text{s.t.} \begin{cases} \pounds(P) \geq 0 \\ P \in S^{n \times n} \end{cases} \end{aligned}$$

$$\text{(D)} \qquad \begin{aligned} \min \quad & \langle Q_{22}, Z_b \rangle + 2\langle Q_{12}, Z_u \rangle + \langle Q_{11}, Z_n \rangle \\ \text{s.t.} \quad & \begin{cases} B_1 Z_u + Z_u^T B_1^T + A_1 Z_n + A_1^T Z_n + I = 0 \\ Z \equiv \begin{bmatrix} Z_b & Z_u \\ Z_u^T & Z_n \end{bmatrix} \geq 0 \end{cases} \end{aligned}$$

where $S^{n\times n}$ denotes the space of symmetric matrices, $\langle X, Y \rangle = \sum_{i,j} X_{ij} Y_{ij}$ denotes the inner product of two matrices X and Y. In particular, $\langle I, P \rangle$ is equal to the trace of the matrix P. $Z \equiv \begin{bmatrix} Z_b & Z_u \\ Z_u^T & Z_n \end{bmatrix}$ denotes the dual variable associated with the primal constraint $£(P) \geq 0$ with Z_b, Z_u and Z_n being a block partitioning of Z of appropriate dimensions.

For the above SDP, a well developed duality theory exist [2,5]. The key points of the theory can be highlighted as follows:

- (P) and (D) are said to satisfy the *strict feasibility* if there exist primal and dual feasible solutions, P^0 and Z^0, such that $(P^0) > 0$ and $Z^0 > 0$, respectively. On the other hand, a primal optimal solution P^* and a dual optimal solution Z^* are called complementary optimal solutions if $£(P^*)Z^* = 0.$
- The weak duality always holds, i.e. any feasible solution to the primal problem always posseses an objective value that is no less than the objective value of any dual feasible solution.
- In contrast, the strong duality, i.e. the optimal values of the primal and dual problems coincide, holds if there exist a pair of complementary optimal solution P^* and Z^*, namely, they satisfy $£(P^*)Z^* = 0.$

Throughout this section, we assume that the LQR Π_1 is stabilizable. The following lemma is base on Lemma (1) in the [10] which shows that $F(P)$ and $£(P)$ are closely related .

Lemma 3.1.2. $£(P) \geq 0$ *iff* $F(P) \geq 0$ *and* $(I - Q_{22}Q_{22}^{\dagger})(B_1^T P + Q_{12}^T) = 0.$

Theorem 3.1.3. *Suppose* $rank(C_1) = p$ *and* $rank\begin{bmatrix} C & D \end{bmatrix} < r$. *If* (P) *and* (D) *have complementary optimal solutions* P^* *and* Z^*, *respectively, then* P^* *must satisfy the generalized Riccati equation* $F(P^*) = 0$.

Proof . Let the hypothesis hold, then P^* and Z^* satisfy $£(P^*)Z^* = 0.$ Since P^* is the optimal solution of (P) then it satisfies $£(P^*) \geq 0$ as well, and by Lemma 3.1.2 we have $(I - Q_{22}Q_{22}^{\dagger})(B_1^T P + Q_{12}^T) = 0.$ Thus, the following decomposition holds:

$$£(P^*) = \begin{bmatrix} I & 0 \\ (PB_1 + Q_{12})Q_{22}^{\dagger} & I \end{bmatrix} \begin{bmatrix} Q_{22} & 0 \\ 0 & F(P^*) \end{bmatrix} \begin{bmatrix} I & Q_{22}^{\dagger}(B_1^T P^* + Q_{12}^T) \\ 0 & I \end{bmatrix}.$$

From the relation $£(P^*)Z^* = 0$, we have

$$\begin{aligned} £(P^*)Z^* &= \begin{bmatrix} Q_{22} & 0 \\ 0 & F(P^*) \end{bmatrix} \begin{bmatrix} I & Q_{22}^{\dagger}(B_1^T P^* + Q_{12}^T) \\ 0 & I \end{bmatrix} \begin{bmatrix} Z_b^* & Z_u^* \\ (Z_u^*)^T & Z_n^* \end{bmatrix} \\ &= \begin{bmatrix} Q_{22}\left(Z_b^* + Q_{22}^{\dagger}(B_1^T P^* + Q_{12}^T)(Z_u^*)^T\right) & Q_{22}\left(Z_u^* + Q_{22}^{\dagger}(B_1^T P^* + Q_{12}^T)Z_n^*\right) \\ F(P^*)(Z_u^*)^T & F(P^*)Z_n^* \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \end{aligned}$$

Therefore $F(P^*)(Z_u^*)^T = 0$, $F(P^*)Z_n^* = 0$, hence $Z_u^* F(P^*) = 0$ and $Z_n^* F(P^*) = 0.$ Since Z^* is dual feasible, then $(Z_u^*)^T B^T + BZ_u^* + Z_n^* A^T + AZ_n^* + I = 0.$ Multiplying $F(P^*)$ on both sides above yields

$$0 = F(P^*)\left((Z_u^*)^T B^T + BZ_u^* + Z_n^* A^T + AZ_n^* + I\right)F(P^*) = F(P^*)^2$$

which implies $F(P^*) = 0.$ ■

Theorem 3.1.4. *If* (P) *has an optimal solution* P^* *satisfying* $F(P^*) = 0$ *then the* LQR Π_1 *has an optimal feedback control wich is determined by*

$$u^* = -Q_{22}^{\dagger}(Q_{12}^T + B_1^T P^*)x_1^*. \tag{17}$$

Proof. Let P be any primal feasible solution and $u(.)$ be any admissible (therefore stabilizing) control. We have,

$$\begin{aligned} \frac{d}{dt}\left(x_1(t)^T P x_1(t)\right) &= \left(A_1 x_1(t) + B_1 u(t)\right)^T P x_1(t) + x_1(t)^T P\left(A_1 x_1(t) + B_1 u(t)\right) \\ &= x_1(t)^T\left(A_1^T P + PA_1\right)x_1(t) + 2u(t)^T B_1^T P x_1(t). \end{aligned}$$

Integrating over $[0,\infty)$ and making use of the fact that $x_1(t)^T P x_1(t) \to 0$ as $t \to \infty$, we have

$$0 = x_{10}^T P x_{10} + \int_0^{\infty} \left[x_1(t)^T\left(A_1^T P + PA_1\right)x_1(t) + 2u(t)^T B_1^T P x_1(t)\right] dt$$

Therefore

$$J(u(.), x_{10}) = \int_0^\infty \left[x_1(t)^T Q_{11} x_1(t) + 2u(t)^T Q_{12}^T P x_1(t) + u(t)^T Q_{22} u(t) \right] dt$$

$$= x_{10}^T P x_{10} + \int_0^\infty \left[x_1(t)^T (A_1^T P + P A_1 + Q_{11}) x_1(t) + 2u(t)^T (B_1^T + Q_{12}^T) P x_1(t) \right.$$

$$\left. + u(t)^T Q_{22} u(t) \right] dt$$

$$= x_{10}^T P x_{10} + \int_0^\infty \left[\left(u(t) + Q_{22}^\dagger (B_1^T P + Q_{12}^T) x_1(t)\right)^T Q_{22} \left(u(t) + Q_{22}^\dagger (B_1^T P + Q_{12}^T) x_1(t)\right) \right.$$

$$\left. + x_1(t)^T F(P) x_1(t) \right] dt$$

Since P is feasible, we have $F(P) \geq 0$. This means that

$$J\left(u(.),\ x_{10}\right)) \geq x_{10}^T P x_{10} \tag{18}$$

for any P feasible to (P) and for any admissible control $u(.)$. On the other hand, if we take into account $P^* \geq 0$ and under the feedback control (17) then we have

$$0 \leq J\left(u^*(.),\ x_{10}\right) = \int_0^\infty \left[x_1(t)^T Q_{11} x_1(t) + 2u(t)^T Q_{12}^T P x_1(t) + u(t)^T Q_{22} u(t) \right] dt$$

$$= \lim_{t\to\infty} \int_0^t \left[x_1(\tau)^T Q_{11} x_1(\tau) + 2u(\tau)^T Q_{12}^T P x_1(\tau) + u(\tau)^T Q_{22} u(\tau) \right] d\tau$$

$$= \lim_{t\to\infty} \left(x_{10}^T P^* x_{10} - x_1(t)^T P^* x_1(t) + \int_0^t \left[x_1(\tau)^T (A_1^T P^* + P^* A_1 + Q_{11}) x_1(\tau) \right.\right.$$

$$\left.\left. + 2u(\tau)^T (B_1^T P^* + Q_{12}^T) P^* x_1(\tau) + u(\tau)^T Q_{22} u(\tau) \right] d\tau \right)$$

$$\leq x_{10}^T P^* x_{10} + \lim_{t\to\infty} \int_0^t \left[\left(u(\tau) + Q_{22}^\dagger (B_1^T P^* + Q_{12}^T) x_1(\tau)\right)^T Q_{22} \left(u(\tau) + Q_{22}^\dagger (B_1^T P^* + Q_{12}^T) x_1(\tau)\right) \right.$$

$$\left. + x_1(\tau)^T F(P^*) x_1(\tau) \right] d\tau$$

$$= x_{10}^T P^* x_{10} \tag{19}$$

First of all, the above shows that the feedback control $u^*(t)$ incurs a finite cost with respect to any initial state x_{10}, then it must be stabilizing. This is because a finite cost in (10) implies $\lim_{t\to\infty} x_1^*(t)^T Q_{11} x_1(t) = 0$ where $x_1^*(.)$ is the corresponding state trajectory, and since $Q_{11} > 0$ we must have $\lim_{t\to\infty} x_1^*(t) = 0$. On the other hand, (19) yields $J\left(u^*(.), x_{10}\right) \le x_{10}^T P^* x_{10}$. Thus in view of (18) we conclude that $u^*(t)$ is an optimal control for Π_1. ■

In view of (4), (9) and (17), the Theorem 3.1.4 implies that the optimal state control pair $\left(x^*, u^*\right)$ for LQR problem Π is given by

$$\begin{bmatrix} x^* \\ u^* \end{bmatrix} = \begin{bmatrix} M & 0 \\ 0 & I_r \end{bmatrix} \begin{bmatrix} x_1^* \\ x_2^* \\ -- \\ u^* \end{bmatrix} = \begin{bmatrix} M & 0 \\ 0 & I_r \end{bmatrix} \left[\begin{array}{c|c|c} I_p & 0 & \\ 0 & -B_2 & \\ \hline 0 & I_r & \end{array}\right] \begin{bmatrix} I_p \\ \\ -Q_{22}^{\dagger}(Q_{12}^T + B_1^T P^*) \end{bmatrix} x_1^*$$

where x_1^* is solution of differential equation

$$\dot{x}_1(t) = \left(A_1 - B_1 Q_{22}^{\dagger}(Q_{12}^T + B_1^T P^*)\right) x_1(t), \quad x_1(0) = x_{10}.$$

3.2 LQR problem for nonproper descriptor system

Now, we consider Π_2. Since N is a nilpotent matrix then $rank(C_2 N^{k-1} B_2) < r$, and it follows that $\overline{Q}_{22}$ is singular. Thereby it is obvious that we arrive at a standard singular LQR problem with respect to variables x_e and w. We can solve this problem using the SDP approach such as in the section (3.1). Here, we also assume that LQR Π_2 is stabilizable. Corresponding to the LQR problem Π_2, we have the following SDP and its associated dual as follows:

$$(\overline{\mathrm{P}}) \qquad \begin{aligned} &\max \quad \left\langle I, \overline{P} \right\rangle \\ &\text{s.t.} \begin{cases} \pounds(\overline{P}) \equiv \begin{bmatrix} \overline{Q}_{22} & B_e^T \overline{P} + \overline{Q}_{12}^T \\ \overline{P} B_e + \overline{Q}_{12} & \overline{Q}_{11} + A_e^T \overline{P} + \overline{P} A_e \end{bmatrix} \ge 0 \\ \overline{P} \in S^{n\times n} \end{cases} \end{aligned}$$

$$(\overline{\mathrm{D}}) \qquad \begin{aligned} &\min \quad \left\langle \overline{Q}_{22}, \overline{Z}_b \right\rangle + 2\left\langle \overline{Q}_{12}, \overline{Z}_u \right\rangle + \left\langle \overline{Q}_{11}, \overline{Z}_n \right\rangle \\ &\text{s.t.} \begin{cases} B_e \overline{Z}_u + \overline{Z}_u^T B_e^T + A_e \overline{Z}_n + A_e^T \overline{Z}_n + I = 0 \\ \overline{Z} \equiv \begin{bmatrix} \overline{Z}_b & \overline{Z}_u \\ \overline{Z}_u^T & \overline{Z}_n \end{bmatrix} \ge 0 \end{cases} \end{aligned}$$

where $\overline{Z}$ denotes the dual variable associated with the primal constraint $\pounds(\overline{P}) \geq 0$ with $\overline{Z}_b$, $\overline{Z}_u$ and $\overline{Z}_n$ being a block partitioning of $\overline{Z}$ with appropriate dimensions. Mimicing Lemma 3.1.2, Theorem 3.1.3 and Theorem 3.1.4 in section 3.1, we have the optimal control for Π_2 as follow:

$$w^* = -\overline{Q}_{22}^{\dagger}(\overline{Q}_{12}^T + B_e^T \overline{P}^*)x_e^*(t)$$

where $\overline{P}^*$ satisfies the generalized Riccati equation

$$F(\overline{P}) \equiv A_e^T \overline{P} + \overline{P}A_e + \overline{Q}_{11} - (\overline{P}B_e + \overline{Q}_{12})\overline{Q}_{22}^{\dagger}(B_e^T \overline{P} + \overline{Q}_{12}^T) = 0$$

and $x_e^*(t)$ is solution of differential equation

$$\dot{x}_e(t) = A_e x_e(t) + B_e w(t), \;\; x_e(0) = x_{e0}.$$

The optimal state control pair $\left(x^*, u^*\right)$ for LQR problem Π is given by

$$\begin{bmatrix} x^* \\ u^* \end{bmatrix} = \begin{bmatrix} M & 0 \\ 0 & I_r \end{bmatrix} \begin{bmatrix} x_1^* \\ x_2^* \\ -- \\ u^* \end{bmatrix} = \begin{bmatrix} M & 0 \\ 0 & I_r \end{bmatrix} \left[\begin{array}{cc|c} I_p & 0 & 0 \\ 0 & -\overline{B}_2 & -N^{k-1}B_2 \\ \hline 0 & I_{r,0} & 0 \end{array}\right] \begin{bmatrix} x_e(t) \\ w(t) \end{bmatrix}$$

$$= \begin{bmatrix} M & 0 \\ 0 & I_r \end{bmatrix} \left[\begin{array}{cc|c} I_p & 0 & 0 \\ 0 & -\overline{B}_2 & -N^{k-1}B_2 \\ \hline 0 & I_{r,0} & 0 \end{array}\right] \begin{bmatrix} I_{p+(k-1)r} \\ -\overline{Q}_{22}^{\dagger}(\overline{Q}_{12}^T + B_e^T \overline{P}^*) \end{bmatrix} x_e^*$$

4 Conclusion

We have established the sufficient condition for solvability of the singular LQR problem for descriptor system. By these conditions, the complete solutions of the problem, i.e. the explicit form of the optimal control-state pair of the problem, are obtained.

References

[1] Anderson, B. D. O. and J. B. Moore (1990), Optimal Control: Linear Quadratic Methods, Prentice Hall, New Jersey.

[2] Balakrishnan, V. and L. Vandenberghe (2003), Semidefinite Programming Duality and Linear Time-Invariant Systems, IEEE Transaction on Automatic Control, 48, 30-41

[3] Bender, D. J. and A. J. Laub (1987), The Linear Quadratic Optimal regulator for Descriptor Systems, IEEE Transaction on Automatic Control, 32 , 672-688.

[4] Geerts, T. (1994), Linear Quadratic Control with and without Stability Subject to General Implicit Continuous Time Systems: Coordinate-Free Interpretations of the Optimal Cost in Terms of Dissipation Inequality and Linear Matrix Inequality, Linear Algebra and Applications, 203, 607-658.

[5] Helmberg, C. (2002), Semidefinite Programming, European Journal of Operational Research, 137, 462-482.

[6] Muhafzan, Malik, Hj. Abu Hassan, Fudziah Ismail, Leong Wah June (2005), On Sufficient Condition for Solvability of the LQ Problem for Nonregular Descriptor Systems (submitted to ESAIM Journal: Control, Optimization and Calculus of Variations).

[7] Muhafzan, Malik, Hj. Abu Hassan, Fudziah Ismail, Leong Wah June (2005), On the Singular LQ Problem for Descriptor Systems, Proceedings of the 2nd International Conference on Research and Education in Mathematics, Kuala lumpur, 203-208.

[8] Mehrmann, V. (1989), Existence, Uniqueness, and Stability of Solutions to Singular Linear Quadratic Optimal Control Problems, Linear Algebra and ItsApplications, 121, 291-331.

[9] Müller, P. C.(1999), Linear Quadratic Optimal Control of Descriptor Systems, Journal of the Brazilian Society of Mechanical Sciences, 21, 1-13.

[10] Yao, D., S. Zhang and X. Y. Zhou (2001), A Primal-Dual Semidefinite Programming Approach to Linear Quadratic Control, EEE Transaction on Automatic Control, 46, 1442-1447.

[11] Zhu, J., S. Ma and Z. Cheng, (1999), Singular LQ Problem for Descriptor Systems, Proceeding of the 38th Conference on Decision & Control, 4098-4099.

MUHAFZAN: Department of Mathematics, Andalas University, Kampus UNAND Limau Manis, Padang, 25163.
E-mail: muhafzan@gmail.com

MALIK Hj. ABU HASSAN: Department of Mathematics, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, Malaysia.
E-mail: malik@fsas.upm.edu.my

FUDZIAH ISMAIL: Department of Mathematics, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, Malaysia.
E-mail: fudziah@fsas.upm.edu.my

LEONG WAH JUNE: Department of Mathematics, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, Malaysia.
E-mail: lwjune@fsas.upm.edu.my

Chattering Free in the Sliding Mode Control of Class of MIMO Systems

Tedy Setiawan[1], Carmadi Machbub[1], Dimitri Mahayana[2], Iwan Pranoto[2]

[1)] Department of electrical Engineering, Institut Teknologi Bandung, Indonesia
[2)] Department of Mathematics, Institut Teknologi Bandung, Indonesia

Abstract: *Sliding-mode control* is powerful non linear control technique that has been intensively developed during the last 35 years. However, this control produces chattering which can cause instability due to unmodeled dynamics and can also cause damage to actuators or the plant. This paper presents result of a research and development of nonlinear system synthesis using sliding mode method to perfect the current sliding mode technique. Replacing the sign function by tan inverse function may deleting or smoothing chattering effect. It is done by ordering the system parameters by considering stability aspects and uncertainty of the system. Simulation results show that using the function of inverse tan shows that chattering effect can be smoothened and deleted by taking declivity curve inverse tan. Observation and accurate comparison through digital simulation and taking of certain function as replacing the function of switching controller (the sign) gives the better result.

Keywords: sliding control, tracking ability, nonlinear control

THE GINI INDEX AND ITS APPLICATION

Budi Nurani Ruchjana

Department of Mathematics Universitas Padjadjaran-Indonesia

Abstract. Italian statistician *Corrado Gini* developed the Gini coefficient is a measure of inequality. It is usually used to measure income inequality, but can be used to measure any form of uneven distribution. The Gini index is the Gini coefficient expressed in percentage form, and is equal to the Gini coefficient multiplied by 100. The Gini coefficient is calculated as a ratio of areas on the Lorenz curve diagram. If the area between the line of perfect equality and Lorenz curve is A, and the area underneath the Lorenz curve is B, then the Gini coefficient is A/(A+B). This ratio is expressed as a percentage or as the numerical equivalent of that percentage, which is always a number between 0 and 1. In this paper, we apply the Gini index to describe the heterogeinities of oil reservoir's characteristics. For case study, we use the permeability and thickness data from several oil wells at Jatibarang Field.

Key words: Gini index, Lorenz curve, heterogeneity, permeability, thickness

1 Introduction

The Gini coefficient or index is perhaps one of the most use indicators of social and economic conditions. It was proposed as a summary statistics odf dispersion of a distribution. It can be used to indicate how the distribution of observation has changed within a location over a period of time, so it is possible to look whether the inequality is increasing or deccreasing. In this paper we will study the Gini index and apply it to indicate the heterogeneity of permeability and porosity of oil reservoir at volcanic layer.

2 The Formulation of Gini Index

The Gini index has a many formulations. There are generally two different approaches for analyzing theoretical result of Gini index: one based on discrete distribution, the other on continous distribution. It can be expressed as a ratio of two regions defined by a 45 degree line and a Lorenz curve in a unit box.

The Gini coefficient is calculated as a ratio of areas on the Lorenz curve diagram. If the area between the line of perfect equality and Lorenz curve is A, and the area underneath the Lorenz curve is B, then the Gini coefficient is A/(A+B). This ratio is expressed as a percentage or as the numerical equivalent of that percentage, which is always a number between 0 and 1 (Figure 2.1).

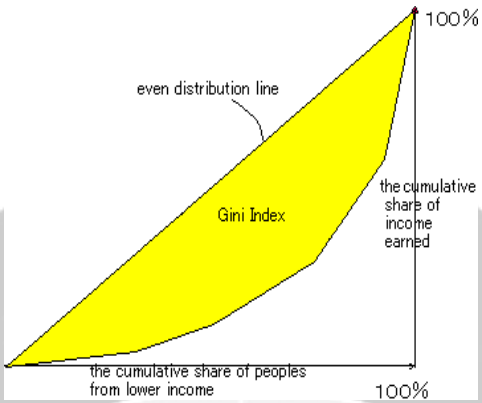
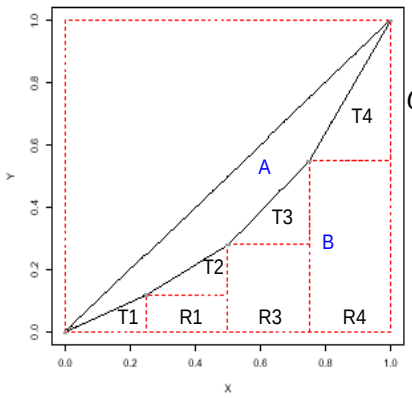


Figure 1. The Gini Index (source: Kuan Xu, 2004)

Furthermore, in this paper we use the geometric approach to get the Gini index. Figure 2, define Gini index geometrically as the ratio of two geometrical areas in the unit box:
(a) the area between the line perfect equality and the Lorenz curve which called A
(b) the area under the 45 degree line, or areas (A+B)

Because areas (A+B) represent the half of the unit box, that is, (A+B) =1/2, so the Gini index , G, can be written as:

$$G = \frac{A}{A+B} = 2A = 1 - 2B \qquad (1)$$



$$G = \frac{A}{1/2} = 2A$$

$$A = \frac{G}{2}$$

Figure 2. The Geometric Approach of Gini Index

The derivation of Gini index are given by Brown (1994) in http://depts.washington.edu /eqhlth/pages /nderby.pdf . The Gini index is defined:

$$G = \left| 1 - \sum_{i=0}^{n-1} (Y_{i+1} + Y_i)(X_{i+1} - X_i) \right| \quad (2)$$

Suppose we have a sample 5-point data set $\{X_0, X_1, X_2, X_3, X_4\}$ and $\{Y_0, Y_1, Y_2, Y_3, Y_4\}$,so in this case k = 5.

Looking at the bottom graph of Figure 2 , we have that G is equal to the area A between the curves divided by the area underneath the top line, as equation (1). We can compute A:

A = 1- areas of rectangles and triangles from Figure 2

$$A = 1 - \frac{1}{2}\sum_{i=1}^{4} area(T_i) - \sum_{i=1}^{3} area(R_i)$$

$$A = \frac{1}{2}\left(1 - \sum_{i=0}^{k=1} (Y_{i+1} + Y_i)(X_{i+1} - X_i)\right)$$

Because A =$G/2$, and the Gini index is an area, so we have the Gini index in absolute value as equation (2).

Zitikis (2002) gave the equation of Gini index as:

$$G = -1 + \frac{1}{n^2 \bar{y}} \sum_{i=1}^{n} (2i - 1) y_i \quad (3)$$

Abdi and Ruchjana (2005) have shown that the equation (2) is equal with the equation (3):

$$1 - \sum_{i=0}^{n-1} (Y_{i+1} + Y_i)(X_{i+1} - X_i) = -1 + \frac{1}{n^2 \bar{y}} \sum_{i=1}^{n} (2i - 1) y_i$$

So, we can use the Brown either Zitikis formula to count the Gini index. In this paper, we transfer the Zitikis formula to S-Plus 2000 software to define the Gini index as in Appendix A.

3 Case Study

We use the porosity and permeability data from 39 oil well at Jatibarang field for case study. The plot of porosity data is shown at Figure 3 below:

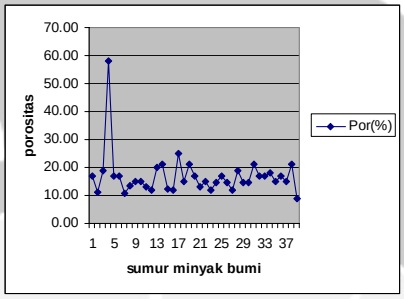


Figure 3. Porosity Data

Using S-Plus 2000, we have the Gini index G_{por} = 0,1697855 = 17 %. It means that the porosity of 39 oil wells is almost homogenous.

Permeability of 39 wells can be plotted as Figure 4:

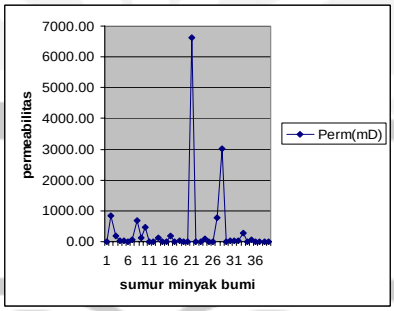


Figure 4. Permeability Data 39 Oil Wells

The Gini index of permeability is G_{perm} = 0,8770102 = 87,8 %. It means that the permeability at 39 wells is heterogenous, because the Gini index is near to 1. We can compare the Gini index with the coefficient of variation: CV_{por} = 50% and CV_{perm} = 327%. CV of permeability is higher than CV porosity, so it makes sense that the permeability more heterogenous that porosity. Using of S-Plus 2000, we can plot the graphic of Gini index of porosity, permeability and the other variables of oil reservoir as shown at Figure 5.

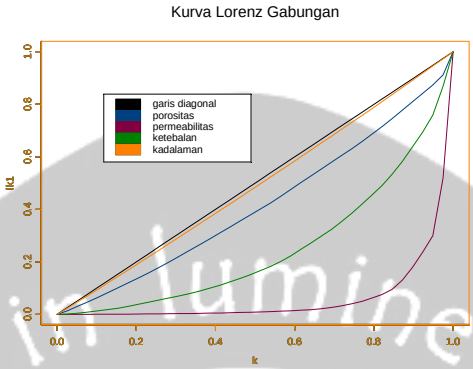


Figure 5. Plot of Gini Index

Acknowledgment

The author would like to thank to Dr. Ricardas Zitikis from Canada for valuable discussion when I attend his workshop at 2002 in ITB and to Ir. Pudjo Tri Raharjo from Pertamina Jatibarang for the data of oil reservoir.

References

[1] Abdi, H. and Ruchjana, B.N. 2005. The Gini Index and Its Applications in Oil Reservoir. Skripsi S1 Jurusan Matematika FMIPA Universitas Padjadjaran.

[2] Gini Coefficient. http://en.wikipedia .org/wiki/Gini_Coefficient , di download tanggal 02/19/2005.

[2] Xu, K., 2004, How HAs the Literature on Gini 's Index Evolved in the PAst 80 Years?, Department of Economics, Canada: Dalhouse University.

[2] Zitikis, R. , 2002, *Statistical inference for Gini indices with applications to economic inequality, Lecture Note of Research Workshop on Computer-Intensive Statistics,* Bandung: P4M ITB, 10-28 Juni 2002.

A. Gini Index using S-Plus or R for Porosity Data

```
x1<-Gini[,1]
x1
k<-0:39/39
k
y1<-c(0,x1)
y1
z1<-sort(y1)
z1
lk1<-cumsum(z1)/sum(z1)
lk1
ord1<-sort(x1)
ord1
ii<-1:39
ii
ci1<-2*ii-1
ci1
cc1<-ci1*ord1
cc1
Giniporosity<--1+sum(cc1)/(mean(x1)*39*39)
Giniporosity
plot(k,k,xlim=c(0,1), ylim=c(0,1), xlab="k",ylab="lk1",type="l",col=1)
par(new=T)
plot(k,lk1,xlim=c(0,1), ylim=c(0,1), xlab="k",ylab="lk1",type="l",col=2)
typ.names <- c("garis diagonal","porositas")
legend(locator(1), legend = typ.names, fill = 1:2)
title("Lorenz Curve for Porosity")
```

Scipt, **Run** and then **F10 to see the result.**

BUDI NURANI RUCHJANA: Department of Mathematics, FMIPA, Universitas Padjadjaran, Jl. Raya Bandung-Sumedang Km 21, Jatinangor-Sumedang 45363, Indonesia. Telp/fax: +62 (0)22 7794696

E-mail: bnurani@yahoo.com

DEVELOPMENT ON SIMPLIFIED MODEL FOR URBAN AIR QUALITY RELATED TO TRANSPORTATION (Case Studi In Jakarta)

Haryo Satriyo Tomo

Environmental Engineering, ITB, Bandung, Indonesia

Abstract. Integrated tool on decision support system is required to analyze urban air quality problems especially those related to transportation. However, the sufficient data related to support an actual mobile emission, emission factor and traffic condition is hardly available. *ISVAQ-2* is an interactive computer system that can be applied to evaluate transportation emission in urban areas. A calculation method is applied to administrative boundaries to select road and volume of vehicle data each peak hour. These data are used to predict emissions that cause ambient conditions. The benefit of this system is to assess the policy scenarios related to transportation and environment impact. Although the method to calculate data is based on independence steady box model, *ISVAQ-2* is yet a complex system due to multitude of possible database selections, because the needs of users on flexibility. The paper explains the overview of system development method as part of decision support system establishment in managing the transport urban air pollution. The result indicates that the model is the appropriate available approach to estimate the mobile emission to emulate the data of idle vehicle emission test as emission factor. Further emission estimation using dynamic mode and the modified local driving condition are vitally important to improve the results of emission estimation as well as actual condition.

Key-words: ISVAQ-2, transportation emission, urban air quality, vehicle problems

1 Introduction

Air pollution in Jakarta is a severe problem. A survey conducted in 1985 exposed that the lead fume emitted in the inner city district was 17 times higher than the WHO warning levels. Ambient levels of particulate matter exceed health standards at least 173 days per year *(WHO, 2000)*. According to Jakarta Land Transport Organization (DLLAJ) explanation, that in the last decade, there was an increase of 15% vehicles per annum adding reports it to the traffic congestion. There are currently more than 3 million vehicles in Jakarta (1999), and their number is predicted to increase every year *(EIA, 1999)*. This increasing number and density of vehicles, is followed with the increasing consumption of oil and other energy sources, such as coal. The fast growing of vehicles is one of the major factors to the declining air quality. Moreover, Transportation sector has huge possibility to emit air pollutant especially CO, NO_x, SO_2 and Total Suspended Particulate (TSP). In Jakarta, according to Bapedal - ITB (1988), showed that 70 - 80 % urban air qualities is contributed by transport sector.

Since August 21st 2000, Badan Pengendalian Dampak Lingkungan Daerah (Bapedalda) Jakarta has extended Inspection and Maintenance (IM) Programs for Private Vehicle, promulgated with SK Gubernur DKI Jakarta No. 95 tahun 2000.

This rule is meant to succeed "Clean Air Program" because reducing emission could be started from efficient vehicle's engine performance.

According to support the implementation of IM Programs, extending a computerized integrated system is required. Development of this system is meant to input data, process data and represent output (post processor) in single user system interface. So, computerizing of the prototype software is a beginning to extent the whole system beside infrastructure preparations. The Prototype of the system is called ***ISVAQ-2***. This prototype is an upgrade version of ISVAQ-1 that was developed in 2001.

2 System Development

ISVAQ-2 has been constructed as a couple systems that join between numerical and declarative entities. Numerical entity is the attribute of every domain, which has non-string value data types. Moreover, the entity could operate mathematically. The couple systems become important because the data is grouping on the domains, which represent logic and actual condition, which consist of two types of entities.

The main goals for the development of ***ISVAQ-2***-system were three folds. First, Bapedalda DKI Jakarta and related institutions should be able to get reliable and updated information on transport emission and also supporting data in each grid. The second one, Bapedalda DKI Jakarta and related institutions should be able to get reliable and updated information on vehicle emission testing data each period. And the last one, Bapedalda DKI should be able to evaluate traffic conditions related to degree of air quality index in each grid from time to time.

Through meteorological mechanisms, emission is converted into ambient air quality. The worst conditions of these would perform in peak hour everyday in each road. Determination of the global urban air quality has to consider the differences about transportation activities of each part of the areas. Moreover, Jakarta is divided into parts related to the homogenous activity area and main road that cross the area. However, to simplify management aspects, such as regional authority, could be changed into closed administrative boundaries (*Thomas et. al., 1989*). For detail information, it could be seen in figure 1.

To attain further information, each grid is specified into road allocation, consist of length and load capacity of main road on it. This also includes traffic fluctuations on predicted peak hours. The counted vehicles are divided into four classifications, motorcycle (MC), light duty gasoline vehicle (LDGV), light duty diesel vehicle (LDDV) and high duty diesel vehicle (HDDV). These classifications related to emission factors that are used to calculate total emission. The calculated air pollution compounds, consist of NOx (Nitrogen oxides), Suspended Particulate Matter, SOx (Sulfur Oxides) and CO (Carbon Monoxide).

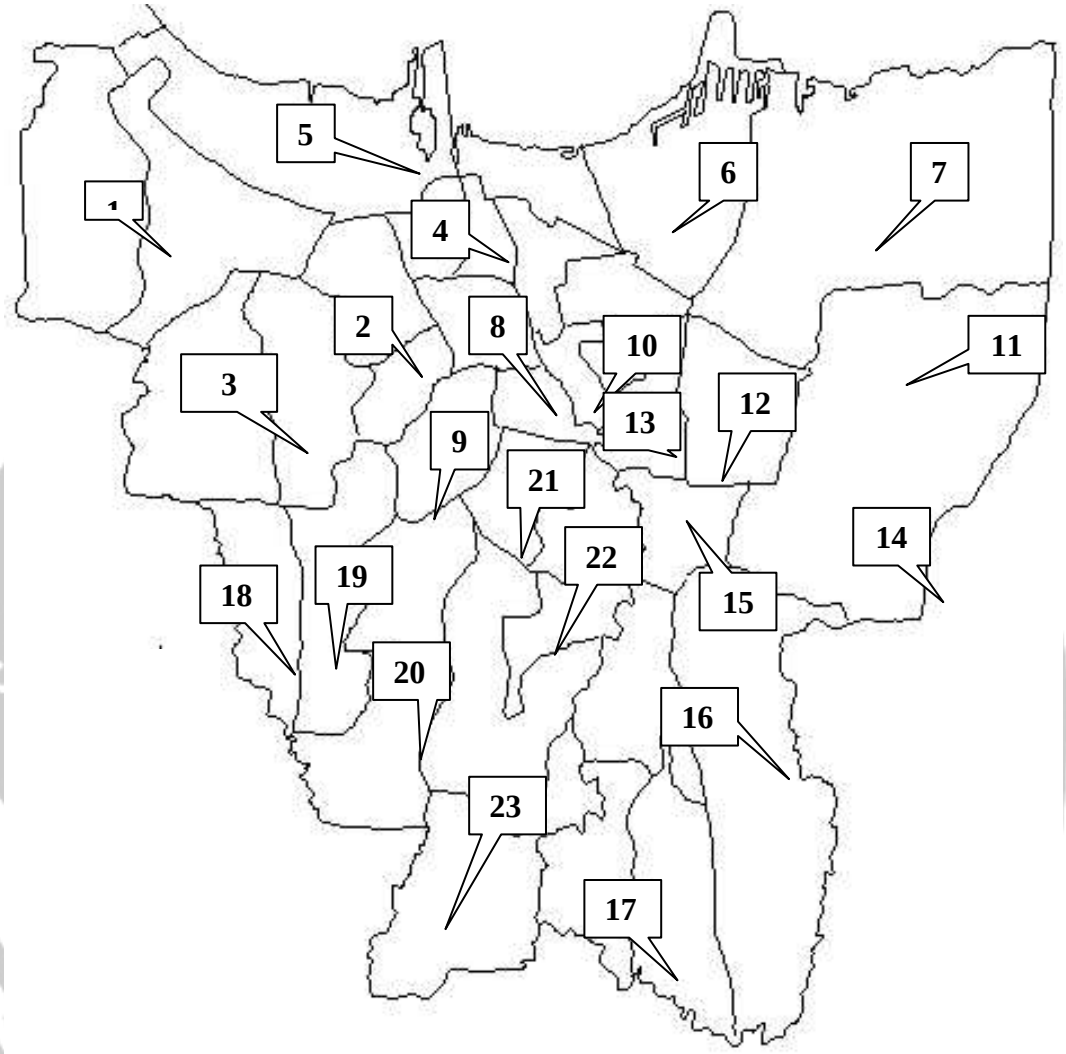


Figure 1. Jakarta in Grids (Segments)

Note :

Cengkareng - Kalideres (Grid 1)
Grogol Petamburan - Palmerah (Grid 2)
Kembangan - Kebon Jeruk (Grid 3)
Tambora - Tamansari - Sawah Besar (Grid 4)
Penjaringan - Pademangan (Grid 5)
Tanjung Priok - Kelapa Gading - Koja (Grid 6)
Cilincing (Grid 7)
Gambir - Menteng (Grid 8)
Tanah Abang (Grid 9)
Kmyn-Cpk Putih-Johar Baru-Senen (Grid 10)
Cakung (Grid 11)
Pulogadung (Grid 12)
Matraman (Grid 13)
Duren Sawit (Grid 14)
Jatinegara (Grid 15)
Kramat Jati - Makasar (Grid 16)
Pasar Rebo - Ciracas - Cipayung (Grid 17)
Pasanggrahan (Grid 18)
Kebayoran Lama (Grid 19)
Cilandak - Kebayoran Baru (Grid 20)
Setiabudi - Mampang Perapatan (Grid 21)
Tebet - Pancoran (Grid 22)
Pasar Minggu - Jagakarsa (Grid 23)

3 Method of Calculation and Implementation

The calculation of air quality condition has many ways. There are two basic concepts that could apply to purpose analytical and numerical solution using Gauss equation (dispersion model) or box model. Dispersion model is commonly used to gain excellent calculation accuracy that is not showed by box model. But in some cases, dispersion model has difficulties in applying because of insufficient data and counting process. These difficulties, especially on insufficient data, are commonly happened in many district authorities (PEMDA), included Jakarta. Therefore, the box model, as a basic concept in calculating air quality is appropriate alternative.

Air quality condition is calculated each grid using equation as follows:

$$E_{(i,j,k,l)} = Veh_{(j,k,l)} * [Lr_{(l)} / Vel_{(k,l)}] * EF_{(i,j)} * k_v \quad \text{.....(1)}$$

$$Vol_{(k,l)} = Ar_{(l)} * MH_{(k,l)} * k_m \quad \text{.....(2)}$$

$$C_{(i,j,k,l)} = E_{(i,j,k,l)} / Vol_{(k,l)} \quad \text{.....(3)}$$

$$k_v = f(Li, Wr, LOS) \quad \text{.....(4)}$$

$$k_m = f(u, \alpha, R) \quad \text{.....(5)}$$

where: E = emission, Veh = vehicle, Lr = length of road, Vel = mean vehicle's velocity, EF = emission factor, Vol = grid volume, Ar = grid area, MH = mixing height, C = concentration, k_v = traffic constant, k_m = meteorological constant, Li = number of line, Wr = width of line, LOS = level of service, u = wind speed, α = wind direction, R = solar radiation, i = type of pollutant, j = type of vehicle, k = type of peak hour, l = grid.

From equation (1) it can be seen, the emission of a compound by vehicles in each road can be computed for each type of peak hour. Emission depends on vehicle traffics and the corresponding emission factor. Computation of ambient air quality based on total emission and meteorological condition, described by mixing height of each grid for each type of peak hour. This can be depicted from equation (2) and (3).

The formulas indicate, there are no relation between air quality from grid to one another. This calls independent multiple-box models. Because of that, it would be necessary to measure all kind of data input in the same time frame condition (weekday, duration, etc). So, The results could be determined closely to actual conditions.

Constructed Database in the system has been transformed into relational database to define connection of each component. All relation of components is connected through knowledge base system. This knowledge is based on logic and structural understanding between components, as depicted in figure 3 and 4.

Air quality model and emission test domains build the database. Air quality domain is meant to count concentration of each compound by grid. There are four sub-domains related in air quality domain: Grid Information, Street Information, Traffic Information, and Local Meteorology. System query is needed to joins all data in each domain to develop data communication.

The other functions are to determine method of data searching and knowledge rules to meet user information about air quality. Data searching method is related to system query that communicate the joined domain. The higher level of field data in domain is priority to lookup as keyword variable. This meant field data is prior if spotted on every domain, as example (Figure 3) No_grid.

Knowledge rules are adjusted from calculation methods that aggregate query system and searching system. Mechanism of knowledge rules could be described in figure 2. Meanwhile, emission test domain mechanism is only to manipulate data and search logic solution. Data manipulation is meant to add, delete, and view data in the domain as depicted in figure 4. Logic solution is based on common 'bengkel' lookup components related to emission test fail. Failing test is indicated by over or lower compound from exhaust systems.

At first, the system would be defined in three main sub-program, input data, analyzing data and output data. Those definitions are accorded to user interface requirements. Because of this, the system is suited to become more complex chaining. User interface is designed to support Bapedalda Jakarta and related institutions that are involved in urban air quality management especially from transportation. So, all components and instructions must be familiar to them.

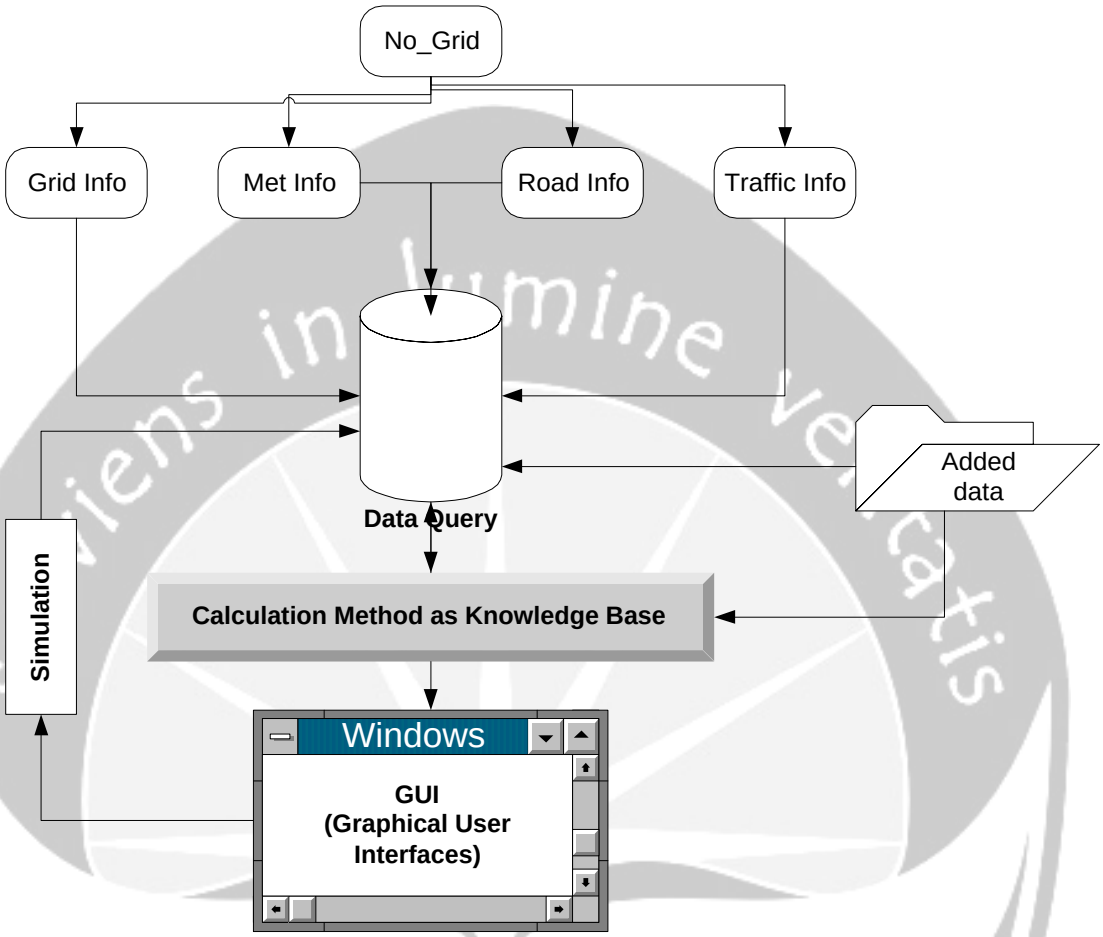


Figure 2. System Query, Searching and Calculations in Air Quality Domain

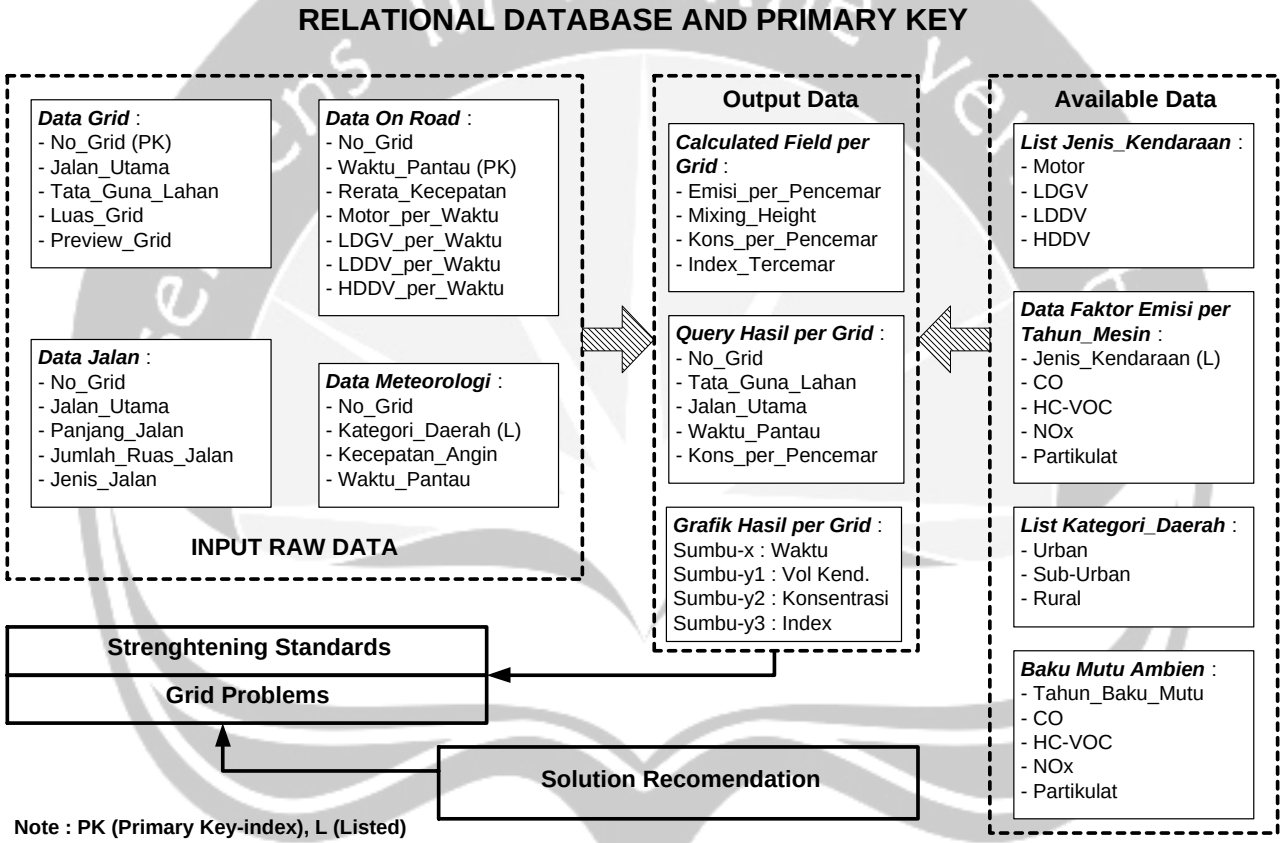


Figure 3. Database Relation to Represent Urban Air Quality

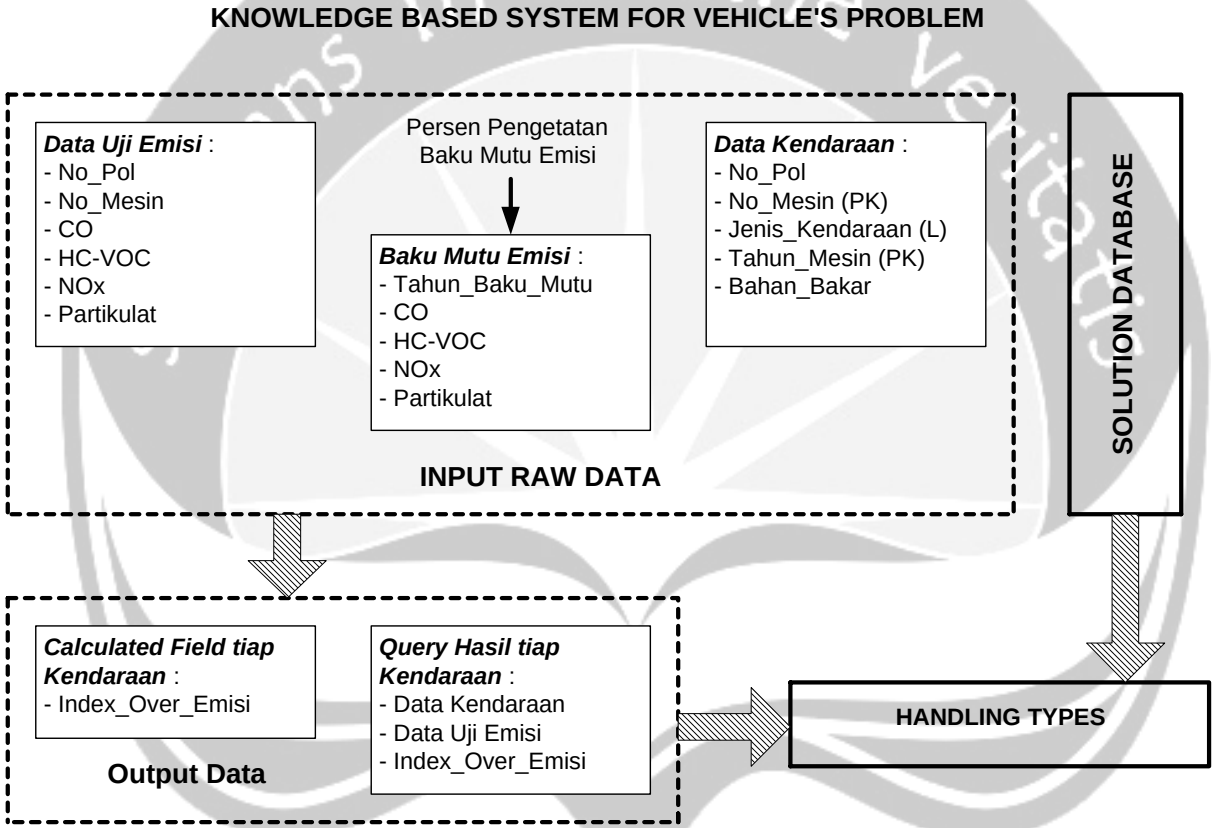


Figure 4. Database Relation to Represent Vehicle's Problem Handling

Most important of the system development is extending into computer system. This is remarkable to select the kind of format data and language that is used. ***ISVAQ-2*** is a prototype which is developed in Paradox-4 data type, using Borland Delphi 5 (base on Pascal) as Integrator and to be communicated with Fortran90 and SQL by dynamic library link (DLL). The system is applied on a common personal computer now but only in stand-alone mode. The future requirement is to integrate the whole system into networks mode according to multi user, multi tasking and multi transaction data.

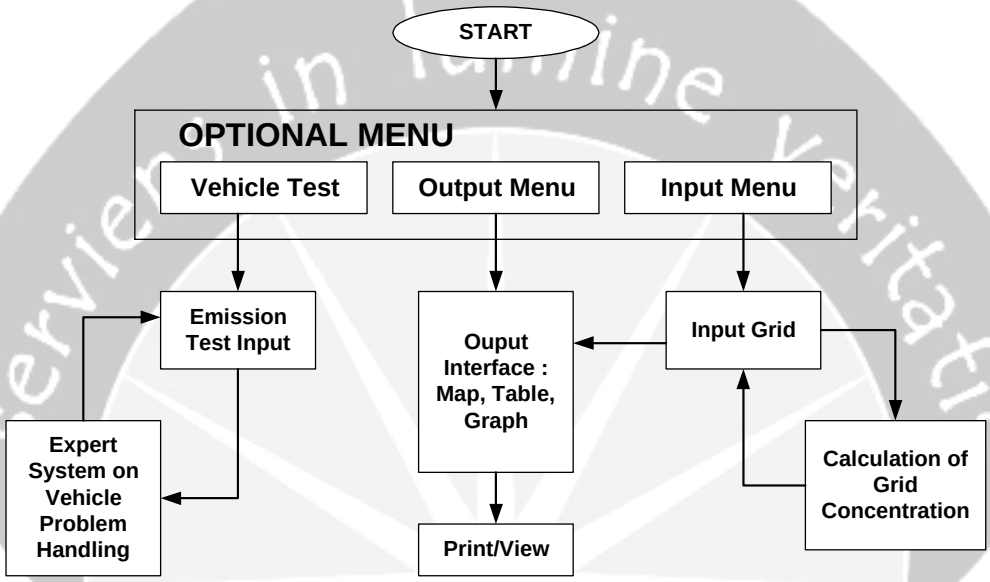


Figure 5. The System (ISVAQ-1) Interface Flow

It could be seen in SK Gubernur Jakarta No 95 tahun 2000, that one of important things to succeed "Clean Air Program" is vehicle emission test. This is meant to regulate emission standards on the source. The point of this program is to earn data from vehicle emission test periodically. Those data are used to recognize machine problems and solving (Bengkel Level) and also to review transportation and automotive policies (Government Sector). Additional Interface in the system is required to input data that would accommodate it. Relation data to represent vehicle's problem handling is seen in figure 4. Beside that, the system also supports to acknowledge and explore about SK Gubernur Jakarta No 95 tahun 2000. This part of the system is to help users whenever they want to search components of the regulation by keyword method or content method.

4 Conclusion

ISVAQ-2 concerns on database management system (DBMS). The output of the system depends on the reliability of input data from surveys or laboratory tests.

Beside that, to attain all requirement, data must be worked the prerequisite analysis to meet the suitable format data. The main benefit of this system is to assess the policy scenarios related to transportation and environment impacts.

ISVAQ-2 is a couple systems that would be able to integrate every component of domains and their relations. The characterization of relations between numerical and declarative entities becomes important in the systems. The relation has to facilitate an identification of data type and possibility of processing among domains. Furthermore, their modules will stay simplifications of the more sophisticated systems used in Bapedalda Jakarta. According to detailed model and calculation method, which is used in the systems, the latter research is needed to propose k_m and k_v as an appropriate constant.

Moreover this system embodies much of the knowledge of the decision-makers. It is their Indonesian first tools to present the proper picture of their compartments and to gain necessary insights into their field. A good cooperation between research institute and government as a user must therefore be built to cover urban air quality problem and solving entirely.

Acknowledgment

The author would like to thank to Dr. Priana Sudjono as the head of Computation of Environmental Management Laboratory TL ITB who supervise and assist me during the research and the paper completion.

References

Energy Information Administration (EIA), December 1999, Indonesia: Environmental Issues, US-Energy Information Administration, http://www.eia.doe.gov/emeu/cabs/indoe.html

Olsthoorn, T.N., et.al., ***Integrated Modeling in Netherlands***, Riso International Conference, Elsevier, 1990, Amsterdam

Rubin E.S., ***Characterizing Uncertainty in Integrated Environmental Models***, Riso International Conference, Elsevier, 1990, Amsterdam

Shaw, R.W., Amann, M., ***Effect of Uncertainty in Source-Receptor Relationships on Transboundary Air Pollution Control Strategies***, Riso International Conference, Elsevier, 1990, Amsterdam

Thomas, R., et.al., ***Environmental Information and Planning Model***, Riso International Conference, Elsevier, 1990, Amsterdam

World Health Organization (WHO), September 2000, Air Pollution, Fact Sheets No. 187 - WHO Information, http://www.who.int/inf-fs/en/fact187.html

World Resource Institute (WRI), 2000, Problems and Priorities in Jakarta and Detroit: A World of Differences, http://www.igc.org/wir/enved/suscom-jakarta-detroit.html

HARYO SATRIYO TOMO: Department of Environmental Engineering, Air Quality Laboratory, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia. Phone/Fax: +62 +22 2534189
E-mail: haryo@tl.itb.ac.id , haryost@gmx.net

Chaotic Motion Simulation Of Water in a Reservoir Heated From Below

Marjono, E. Siswanto, ING Wardana

Dept. of Mathematics, Brawijaya University, Indonesia

Abstract: Heat transfer by using water as a media has a very wide application in industries. Usually the efficiency for heating process is low. This is caused by the fact that the process goes on the stable condition. Based on the problem, it is important to create unstable condition, namely chaotic motion, during water heating process. Here was investigated the critical chaotic motion in water inside container heated from below. The width of the container is 1.3 unit length width varying the Rayleigh number (R). The indicators used are attractor shape, the Lyapunov Exponent Equivalent (LEE) and the state of the system. The results are that critical chaos occur at R=50.750 to R=80.000 and that a quasiperiodic condition was held on R=80.000 to R=200.000.

Keywords: chaos, heat transfer

THE EFFECTIVE PREDICTION MODELING IN OIL PALM INDUSTRY USING DATA MINING

Mohd.Najib B. Mohd.Salleh, Saifulah Bin Rusiman

Kolej Universiti Teknologi Tun Hussein Onn
86100 Batu Pahat, Johor, Malaysia.

Abstract. As Malaysia becomes the largest exporter of palm oil in the world, there is a need to generate information, increase production, processing efficiency and expand uses of palm oil through the combination of some existing technologies such as data mining application. This paper proposes the use of data mining technique incorporating induction decision tree algorithms in term of their performance as classifier to predict the profile of elite genotype of planting materials behavior in oil palm industry. The data sets are studied and validated from previous physiological traits of germplasm planting material. The clustering and classification performance of the induction decision tree is analyzed. The results indicate that data mining algorithms are implemented as classifier for the oil palm germaplasm behavior-based problem. In addition, the uncertainty inherent such classification and clustering decisions were examined with a limitation of sampling data set. The output of this intelligent system can be highly beneficial to breeders in designing effective policies and decision making.

1. INTRODUCTION

The prediction of elite genotype oil palm in planting material is a difficult and complex problem since it depends on a large number of physiological traits: genotype x environment, vegetative traits and fresh fruit bunch index. To provide the Malaysian oil palm industry novel planting material[1], we presents an application of data mining techniques[2] to improve and exploit the productivity of the oil palm from the production of high value products. The main purpose of this paper is to establish an optimum data mining technique and use well-suited training algorithms for planting materials in oil palm industry in Malaysia. Data mining [3] has widely been used, which is the process of extracting and discovering interesting knowledge in the form of pattern, association and significant structures from large amounts of data stored in databases, data warehouses or other information repositories. Our research study is based on an explorative review on the concept of manual process and the scales for measuring features of palm tree and oil extraction in the selected plantation in Malaysia. We want to outline a methodology for extracting feature selection algorithm and demonstrate on planting material, where it has been used to improve the comprehensibility and accuracy. The large genetic variability within collection is being evaluated and selected for novel traits are focused on oil yield, palm height, bunch index and vegetative traits. The data collections are evaluated using decision tree induction to elicit useful information and the results of the research for rule extraction are analyzed on individual palm basis. This is followed by a discussion of typical

issues during data mining process and some of the more successful algorithms in research work. We expected the experiment results reveal the production rules to show the most important feature in oil production.

2. DATA MINING TECHNIQUE

In this section, we describe the required existing features or parameters involved in predicting the elite genotype in oil palm. We also discuss some method of clustering and classification technique to our research study before we build the model. Data Mining is still based on the conceptual principles of statistics including the traditional Exploratory Data Analysis (EDA) and modeling and it shares with them both some components of its general approaches and specific techniques. Decision Tree algorithms and associated techniques promise to resolve these problems by giving more transparent solutions with similar non-linear capabilities to those of Artificial Neural networks[4]. Here we will present decision tree as an appropriate machine intelligence technology for the oil palm industry applications. Roiger[5] has shown that the best performing model in classification was decision tree techniques. A successful application has been developed using statistical concepts and machine learning logic algorithms [6][7]. In our study, we found the data are not predefined as unsupervised learning. With this approaches, a training set can be applied based on similarity scheme defined by the clustering approach. This is done to partition or segment the database into components that make more general view of the data. We add another attribute to the data set to identify the class. The algorithm used to generate this new attribute is also called labeling algorithm. In our case, the rule production created to partition our database into several clusters.

2.1 DECISION TREE-BASED CONCEPT

The decision tree approach[8] is most useful of presenting a series of rules that lead to clustering problems. Decision tree is easy to understand and successfully applied in real problems. In addition, it is able to build models with datasets belong to numerical as well as categorical. Decision Tree can be mapped to a set of induction rules. Rule extraction is a method for deriving a set of rules to classify cases. Although decision trees produce a set of rules and generate a set of independent rules that do not forcing splits at each level, it may be able to find different and sometimes better patterns for classification. The algorithm for building a decision tree uses the divide and conquers strategy to recursively partition the data to produce the tree. Each successive step greedily chooses the best cut to partition the space into two parts in order to obtain purer regions. A commonly used criterion for choosing the best cut is the information gain. It is a standard method to select a test attribute in classical decision tree induction by choosing the attributes that yields the highest information gain. In this subsection, we apply this information measure in the decision tree induction.

1 for each attribute $A_i \in \{ A_1, A_2, \ldots\ldots\ldots A_d \}$ of the data set D do
2 for each value x of A_i in D do
3 Compute the **information gain at** x
4 end
5 Select the test or cut that gives the best information gain to partition the space.

We also present induction tree algorithm that generate a rule base from a set of input-output. Pruning is the process of removing leaves and branches to improve the performance of the decision tree when it moves from the training data. The tree-building algorithm makes the best split at the root node where there are the largest number of records and, hence, a lot of information. Each subsequent split has a smaller and less representative population with which to work. A systematic way of determining the rules that associate the inputs to the outputs is needed. The information available to the extracting the rules include empirical input-output from the physical plant site and heuristic information from the experts. Both types of information are useful and should complement one another in determining the rules.

2.2 INFORMATION MEASURE

The core problem of how to compute the information measure or entropy used in the induction trees is essential to construct the decision tree[9]. In addition, we consider rule bases derived from decision trees and present some heuristic strategies to prune them. Before the tree induction partition has to be created for each attribute. The partitions used as tests in the nodes of the induction tree. To initialize these partitions, we create completely automatically based on a given data set, calculating the highest information gain for the given data set. The definition of information gain is based on probability theory.

$$Entropy = \sum - p_i log_2 p_i$$

The order in which attributes are chosen determines how complicated the tree is. The entropy is used to determine the most informative attribute. A measure of the information content of a message is the inverse of the probability of receiving the message. The information content of a message should be related to the degree of surprise in receiving the message. Messages with a high probability of arrival are not as informative as messages with low probability. Probabilities are multiplied to get the probability of two or more things both or all happening.

Given entropy as a measure of the impurity in a collection of training examples, we are able to measure of the effectiveness of an attribute in classifying the training data using information gain.

$$Gain(S,A) = Entropy(S) - \sum \frac{|S_v|}{|S|} Entropy(S_v)$$

By learning as building many-to-one mapping input and output. Learning tries to reduce the information content of the inputs by mapping them to fewer outputs. Hence we try to minimize entropy. The simplest mapping is to map everything to one output. We can seek a trade-off between accuracy and simplicity. These patterns can become meaningless and sometimes harmful for prediction if you try to extend rules based on them to larger populations. Since the tree is grown from the training data set, when it has reached full structure it usually suffers from over-fitting (i.e. it is "explaining" random elements of the training data that are not likely to be features of the larger population of data). This results in poor performance on real life data. Therefore, it has to be pruned using the validation data set and the user-specified cost complexity factor. The tree is pruned to minimize the sum of the output variable variance in the validation data, taken a terminal node at a time, and the product of the cost complexity factor and the number of terminal nodes. If the cost complexity factor is specified as zero then pruning is simply finding the tree that performs best on validation data in terms of total terminal node variance. Larger values of the cost complexity factor result in smaller trees. The pruning is done such that the last grown node is chopped off first and so on. Expected error EE_S pruning is considered when approximate expected errors assuming that we prune at a particular node. Approximate backed-up error from child assuming we did not prune. If we consider expected error is less than backed-up error, a particular node is pruned. If we prune a node, it becomes a leaf labeled as C. We used the Laplace error estimate, based on the assumption that the distribution of probabilities that examples will belong to different classes is uniform.

$$EE_S = \frac{(N - n + k - 1)}{(N + k)}$$

S is the set of examples in a node, k is the number of classes N examples in S, n out of N examples in S belong to Class. For a non-leaf node, we calculate the backed up error based on let the children of $Node_j$,

$$\mathbf{BackedUpError(Node) = \Sigma P_i \, x \; Error(\; Node_i)}$$

The probabilities can be estimated by relative frequencies of attribute values in sets of examples that fall into child nodes. Finally we are able to calculate the error to prune by selecting the minimum value.

$$\mathbf{Error(Node) = min \; (E(Node) \; , \; BackedUpError(\; Node_i))}$$

2.4 RULE EXTRACTION

We have to select the relevant inputs from a set of candidate inputs. It is important to extract rules indicate output depends on a particular input. Often it is also important to know whether including an input improves the accuracy of the rules significantly.

There have been several methods proposed for the identification of a rule-based model from input-output data. In this research, we used induction tree based to form clustering and classification[10] task to extract rules. The induction tree approach provides framework for combining empirical knowledge in the form of input-output data with qualitative knowledge in the form of IF-THEN rules provided by a human expert. The iterative algorithm attempts to find the subset from the candidate set that minimizes the prediction error. The algorithm also includes pruning condition that assist in screening output irrelevant attributes.

All unsupervised clustering techniques compute some measure of cluster quality. A common technique is to calculate the summation of squared error differences between the instances of each cluster and their corresponding cluster centre. Smaller values for sums of squared error differences indicate clusters of higher quality. The evaluation method has at least two advantages. Euclidean method is used to measure the dissimilarities or distances between objects when forming the clusters. These distances can be based on a single dimension or multiple dimensions.

$$\text{distance}(x,y) = \{\sum_i (x_i - y_i)^2\}^{1/2}$$

This method distance is suitable to our case whereby it usually computed from raw data, and not from standardized data. The distance between any two objects is not affected by the addition of new objects to the analysis, which may be outliers. However, the distances can be greatly affected by differences in scale among the dimensions from which the distances are computed. By implementing square the standard Euclidean distance in order to place progressively greater weight on objects that are further apart.

We also address of implementing k-means cluster analysis or segmentation in this research. It is suitable for use with very large datasets such as arise in data mining and survey analysis. An exact assignment test assures that the algorithm must converge and mapping of complex cluster. Clustering problems routinely occur in survey analysis and data mining where incomplete data and different types of variables can be presented. A popular method of classification is k-means analysis, which partitions a set of cases into k clusters so as to minimize the “error” or sum of squared distances of the cases about the cluster means. However, k-means analysis is usually only implemented with quantitative variables.

3. EXPERIMENT

The aim of the rule extraction process described to generate a valid set of prediction rules for elite genotype during progeny test in oil palm planting materials. These rules will have accurately recognized particular patterns in the data that indicate an upcoming feature of the planting material. The rule inference process starts by the

selection of the data related to the feature of interest. We retrieve data sets from the previous recorded sheets and annual oil palm statistical reports. The information retained physiological traits for almost 6 months period of time with 3340 records to create the data set to build predictive model.

3.1 EXPERIMENT RESULT

In the first experiment, we used k-means cluster analysis in SPSS software and ESX data miner[11] to generate cluster score using 7 attributes of physiological trait in the planting materials, such as: fruit type, petiole cross section, trunk diameter, height, frond production, total economic product as predictors in ESX data mining process. We test the same instances to find the significance value using SPSS software. As a result, we select the attribute with larger significance score.

Attribute Name	Clustering By SPSS			Cluster By ESX		
	1	2	3	1	2	3
Frond Production	24	24	24	23	23	33
Petiole Cross	20	21	19	17	23	34
Height	1	1	1	1	1	2
Leaf Area	7	7	7	6	8	8
Trunk Diameter	1	1	1	1	1	1
Total Economic Product	28	43	15	25	32	43
Fruit Type	D	T	D	D	T	T
Instances	1482	606	1252	2052	760	528

table 1 : statistical result of physiological trait using cluster analysis of SPSS and ESX data miner

In table 1, we produce the result of physiological trait using cluster analysis of SPSS and ESX with 3 clusters generated. Concerning the output results, after each training session, we observed homogeneity of the difference clustering between SPSS and ESX. There is no reason to compare the accuracy of the results among the mentioned

method. The only parameters that can establish which algorithm is more suitable remain the higher score. The typicality values defined as the average similarity of an instance to all other members of its cluster. Concerning the prediction further into the cluster analysis, we have obtained satisfactory results with extension of cluster 3.

We repeat the process by adding extra 3 attributes to the existing physiological attributes to run on ESX data miner and SPSS. The result of unsupervised mining ESX produces in table 2, which incorporates with statistical summary of the instances in the cluster. The results obviously shown the different number of clusters generated by both tools. In this section we report the results obtained by same induction decision tree approach to generate 2 clusters by the ESX. The class resemblance score for both clusters higher than domain with 0.645 and 0.585 respectively, which is more than the domain resemblance score.

Attribute Name	Cluster By SPSS			Cluster By ESX	
	1	2	3	1	2
Frond Production	24	24	24	24	24
Petiole Cross	20	20	21	19	18
Height	1	1	1	1	1
Leaf Area	7	7	7	7	6
Trunk Diameter	1	1	1	1	1
Mesocarp To Fruit	47	70	71	45	71
Shell To Fruit	39	17	16	42	17
Kernel/Palm/Year	10	8	13	16	9
Total Economic Product	22	24	43	21	29
Fruit Type	1	2	2	1	2
Instances	2052	760	528	2130	1210

Table 2: statistical summary of the instances in the cluster using cluster analysis and ESX data miner

But cluster quality score only shown 9% better than the other. Comparing the result in the first experiment does not show much different in cluster generation, while in the second experiment, we are able to conclude that feature selection of vegetative trait performed better cluster interpretation. The induction decision tree constructs the most relevant information content to grow and prune the unnecessary feature in the experiment. Production rule generate by ESX data miner explained the result and can be easily understand by users.

3.2 DISCUSSION AND FUTURE WORK

In the research work, we attempt to apply decision tree approach to generate meaningful rules. A number of related issues are to be further studied. When clustering is applied to real problem, we found some issues that need to be discussed, such there are some instances do not fall into any cluster. Most of the experiments show that adding useless redundant attributes causes the performance of learning schemes to deteriorate. To learn the whole process of fresh fruit bunch analysis is essential to recognize the important variables in order to construct models which describe patterns and relationships presented in data. Collecting some useful data during the whole process would be contributed at effective prediction of genotype in oil palm planting material. The production rule in decision tree techniques work well when new data are provided. By selecting suitable parameters for the model, the data are pre-processed by scaling and targets so they always fall with the mentioned range. In the studied cases, we found that the rules can be performed and some hidden instances able to elicit to explain the significance influence of the instances over the accuracy of the result. Data imperfection might have been the result of noise, imprecise measurements, subjective evaluations, inadequate descriptive language or simply missing data. As a result, the knowledge generally exhibits lower comprehensibility. This outlier sometimes is meaningful to be used to interpret abnormal situation. A domain expert still needed to interpret the exact meaning of some cluster as we realize that to determine the exact number of cluster is not an easy task.

4. CONCLUSION

Decision tree algorithms provide most popular methodologies for symbolic knowledge acquisition. In the process of extracting patterns into cluster as well as predicting some unseen useful knowledge from large quantities of data, elicit some basic application of variables and describe behavior captured in the data are most important issues. The use of decision tree induction methods turns to be useful for improving the insight into complex problem of genotype selection. The nature of planting material where different factors influence the physiological and vegetative trait makes the problem very difficult.

REFERENCES

[1] A.Khusairi, A. Rajanaidu, Mohd Din, A. " Mining The Germaplasm", International Seminar On The Progress of Oil Palm Breeding And Selection, 6-9October 2003
[2] Micheal J.A. Berry, Gordon Linoff, "Data Mining Techniques For marketing, Sales and Customer Support", John Wiley & Sons Inc, 1997.
[3] Paolo Giudici, "Applied Data Mining", John Wiley & Sons, Ltd, 2003
[4] Hongjun Lu, Rudy Setiono, Huan Liu. "Effective Data Ming Using Neural Networks". IEEE Transactions on Knowledge and data Engineering, Vol 8, No.6, December 1996.
[5] Richard J Roiger, Cyrus Azarbod, Rajiv R. Sant. " A Majority Approach To Data Mining". IEEE Reports 1997.
[6] Tom M.Mitchell. " Machine Learning". McGraw-Hill International Editions. 1997.
[7] Mehmed Kantardzic. " Data Mining : Concepts, Models and Algorithms". IEEE Press, 2003.
[8] J.Quinlan, "Introduction to Decision Tree", Morgan Kaufman, 1993
[9] Ian H.Witten, Eibe Frank. "Data Mining – Practical Machine Learning Tools and techniques with Java Implementations". Morgan Kaufmann Publishers Inc, 2000.
[10] Breiman, L., J. Friedman, R. Olshen and C. Stone, "Classification and Regression Trees", Pacific Grove: Wadsworth, 1984.
[11] Richard J.Roiger, Michael W.Geatz. " Data Mining –A Tutorial-Based Primer", Pearson Education. 2003.

MOHD.NAJIB B. MOHD.SALLEH: Kolej Universiti Teknologi Tun Hussein Onn
86100 Batu Pahat, Johor, Malaysia.
E-mail: najib@kuittho.edu.my

SAIFULAH BIN RUSIMAN: Kolej Universiti Teknologi Tun Hussein Onn
86100 Batu Pahat, Johor, Malaysia.
E-mail: saifulah@kuittho.edu.my

MODELING OF PRODUCTION PROCESS: A STUDY CASE IN A MIRROR INDUSTRY

E. Cahyono[a], R. Raya[a], L. D. Saidi[a] & T Masriyati[a]

[a] Universitas Haluoleo, Kendari, Indonesia

Abstract. In this paper we discuss a glass velocity problem that appears in a mirror industry. There are several steps to process glasses into mirrors; such as cleaning, silvering and painting. For those processes the glasses move in a compact machine at the velocity of 4 m/minute, except for the painting process. In the painting area the glasses are accelerated until reaching the velocity about 80 m/minute in the middle of the area, then decelerated to the velocity of 4 m/minute when leaving the painting area. In the middle of the painting area it is observed there is a small velocity variation; the glass velocities are in the interval of 78 to 81 m/minute. This may affect a crash of two adjacent glasses at the end of the process, especially if the distance between the glasses is too small. In this paper we develop a mathematical model to obtain an optimum distance of two adjacent glasses before entering the painting area to avoid the crash. This leads to a non-linear ordinary differential equation where the existence and uniqueness solution is guaranteed. The solution is computed numerically by using Runge-Kutta method.
Key-words: Differential equation, modeling of production process, Runge-Kutta method

1 Introduction

This paper is motivated by a problem appearing in a mirror industry. There are several steps to process glasses into mirrors such as cleaning, silvering and painting. Cleaning process which uses water is the first process to clean the glasses before silvering. Silvering process is intended to give a thin layer of silver on the glasses. This layer creates an image of an object located before the mirror. The thin layer of paint outside is mainly to protect the layer of silver.

All processes are carried out in a compact machine which consists of several part (areas) for specific processes. The glasses move at the velocity of 4 m/minute in all areas, except in the painting area. Observation shows that in the middle of the painting area the glass velocities may reach 80 m/minute. Moreover, it is also observed that there is a small velocity variation; the velocities are in the interval of 78 to 81 m/minute. This velocity variation may create a crash of two adjacent glasses during and after the painting process if the distance between the glasses before entering the painting area is 'too small'. If the distance is 'too large', however, it creates inefficiency. The industry is interested in obtaining the optimal distance.

We focus on the position of two adjacent glasses while moving inside the machine as illustrated in Figure 1. We call glass 1 for the glass that moves in the front, and glass 2 for the other. Note that since their velocities depend on their positions, their distance may not the same all the time. The aim of the present paper is to develop a mathematical model in order to understand the process better.

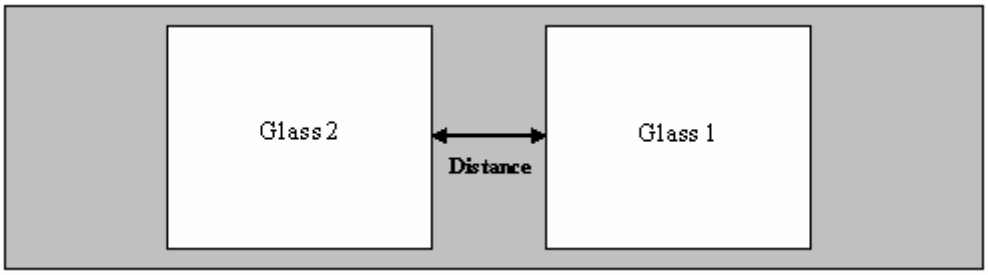


Figure 1. Position of two adjacent glasses.

2 Mathematical model, existence and uniqueness of the solution

The position of a glass is defined by the position of its front part, and at time *t* it is denoted by $x = x(t)$. While the glass is moving in the machine, its velocity is determined by its position, and in the area of painting process it also depends on the time. Hence, x satisfies

$$\overset{\circ}{x} = v(t, x) \tag{1}$$

where $\overset{\circ}{x} = \dfrac{dx}{dt}$ and the velocity function $v(t, x)$ is given by the velocity setting of the machine.

We write the velocity $v(t, x) = f(x) + \varepsilon(x, t)$, where the function f is the deterministic term refers to the average velocity at the position x. The function ε is the small velocity variation in the painting area. It cannot be predicted, but its value vanishes outside the painting area. The function ε is caused by the inaccuracy of the machine which is responsible for the crash of two adjacent glass in and after the painting area if their distance before entering the area is 'too small'.

Observation in industry shows that there is no immediate change of velocity before, inside and after the painting process, the velocity change is smooth. Hence, we may assume that we have a smooth function $v(t, x)$. This guarantees the existence and uniqueness of the solution for a given initial condition $x(t_0) = x_0$.

Theorem 2.1. *Let* $v(t, x(t))$ *be continuous in a domain* D *of the* tx*-plane. Let* $v_x(t, x(t))$ *be continuous in* D *and let* (t_0, x_0) *be a point of* D*. Then a solution* $x = g(t)$, $(|t - t_0| < h)$ *of the differential equation*

$$\overset{\circ}{x} = v(t, x). \tag{1}$$

exists such that $g(t_0) = x_0$ *and, for* x, $[t, g(t)]$ *each is in* D.

Theorem 2.2. *Let* $v(t, x(t))$ *be continuous in a domain* D *of the* xy*-plane. Let* $v_x(t, x(t))$ *be continuous in* D *and let* (t_0, x_0) *be a point of* D. *Let* $g_1(t)$ *and* $g_2(t)$ *both be solutions of the differential equation* $\overset{\circ}{x} = F(t, x)$ *for* $|t - t_0| < h$, *with* $g_1(t_0) = g_2(t_0) = x_0$. *Then* $g_1(t) \equiv g_2(t)$ *for* $|t - t_0| < h$.

The proof of theorem 2.1 and 2.2 can easily be found in some standard books such as [3:473-483].

There is no complete information yet from industry about the velocity, except it is 4 m/minute outside the painting area. At the middle of the painting area it varies from 77 to 81 m/minute, the average is 79 m/minute. One may consider the middle of the painting area as a reference of point, and approximates the velocity change linearly as reported by Cahyono & Kartono [2]. Suppose the length of the painting area is $2a$, then $f(x)$ is given by

$$f(x) = \begin{cases} 4, & x < -a \\ \dfrac{79-4}{a}x + 79, & -a \le x < 0 \\ -\dfrac{79-4}{a}x + 79, & 0 \le x < a \\ 4, & x \ge a. \end{cases} \tag{2}$$

Writing the velocity deviation at the middle of painting area $\varepsilon_{mid} = \varepsilon_{mid}(t)$ the term $\varepsilon(x,t)$ satisfies

$$\varepsilon(x,t) = \begin{cases} 0, & x < -a \\ \dfrac{\varepsilon_{mid}}{a}x + \varepsilon_{mid}, & -a \le x < 0 \\ -\dfrac{\varepsilon_{mid}}{a}x + \varepsilon_{mid}, & 0 \le x < a \\ 0, & x \ge 0. \end{cases} \tag{3}$$

An illustrative plot of this approximation is given in Figure 2. A glass experiences velocity as given by curve 1, while another glass does 2 and other glasses follow other curves.

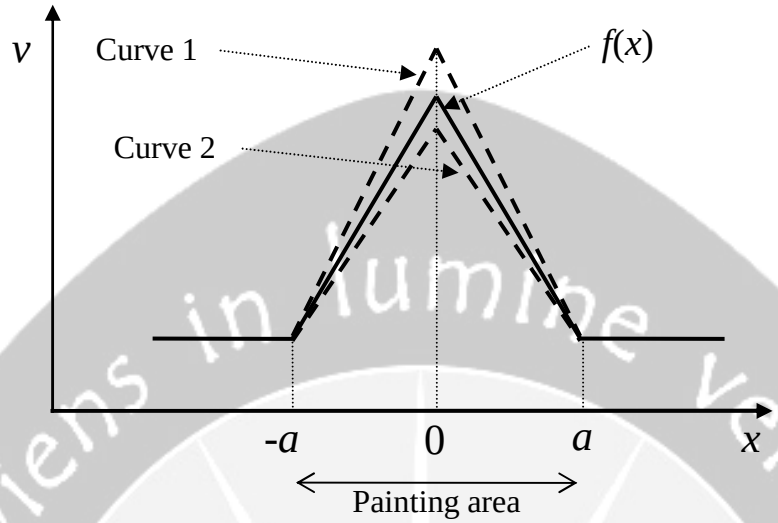


Figure 2. Illustrative plot of a linear approximation to the velocity function.

On substituting (2) and (3) into (1), the solution is given as follows. Assume that at $t = 0$ the position of the glass at $x = -2a - b$, where b is a constant. We write the solution in the form of

$$x(t) = \begin{cases} x_1(t), & t_0 < t \le t_1 \\ x_2(t), & t_1 < t \le t_2 \\ x_3(t), & t_2 < t \le t_3 \\ x_4(t), & t > t_3 \end{cases}$$

where

$$x_1(t) = 4t - 2a - b$$

$$x_2(t) = \frac{4a \exp\left(\dfrac{(75+\varepsilon_{mid})t + b}{a}\right)}{(75+\varepsilon_{mid})\exp\left(\dfrac{79a + \varepsilon_{mid} a + 79b + \varepsilon_{mid} b}{4a} - 1\right)} - \frac{a(79+\varepsilon_{mid})}{75+\varepsilon_{mid}}$$

$$x_3(t) = \frac{-a(79+\varepsilon_{mid})^2 \exp\left(-\dfrac{(75+\varepsilon_{mid})t}{a}\right)}{\left(4(75+\varepsilon_{mid})\exp\left(-\dfrac{79(a+b) - \varepsilon_{mid}(a+b) - 4b}{4a}\right)\right)} + \frac{a(79+\varepsilon_{mid})}{75+\varepsilon_{mid}}$$

$$x_4(t) = 4t + a + \frac{3}{4}\frac{4a\ln\left(\frac{16a}{a(79+\varepsilon_{mid})^2}\right) - (a+b)(75+\varepsilon_{mid})}{75+\varepsilon_{mid}}$$

and

$$t_1 = \frac{a+b}{4}$$

$$t_2 = \frac{1}{4}\frac{4a\ln\left(\frac{79+\varepsilon_{mid}}{4}\right) + (a+b)(75+\varepsilon_{mid})}{75+\varepsilon_{mid}}$$

$$t_3 = \frac{1}{4}\frac{-4a\ln\left(\frac{4a}{a(79+\varepsilon_{mid})^2}\right) + (a+b)(75+\varepsilon_{mid})}{75+\varepsilon_{mid}}.$$

Note that this approximation give un-smooth velocity at the end points and the middle of the painting area.

3 Approximation and simulation of the solution

We have discussed a linear approximation for the velocity function in the previous section. In this section we approximate the velocity function by using a smooth function given by

$$f(x) = \begin{cases} 4, & x < -a \\ \frac{75}{a^4}(x-a)^2(x+a)^2 + 4, & -a \le x \le a \\ 4, & x > a \end{cases}. \quad (4)$$

The maximum velocity variation is given by

$$\max|\varepsilon(x,t)| = \begin{cases} 0, & x < -a \\ \frac{\varepsilon_0}{a^4}(x-a)^2(x+a)^2, & -a \le x \le a \\ 0, & x > a \end{cases} \quad (5)$$

which is also a smooth function. An illustrative plot of this velocity function is given in Figure 3. The worst cases of the velocity variation result in the upper and lower dotted curves.

Solving (1) for f and ε given by (4) and (5) analytically is not easy. Hence, we solve them numerically using fourth order Runge-Kutta method. This method can easily be found in many elementary textbooks such as [1:314-323]. Consider set of

nodes of $t\ 0 = t_0 < t_1 < t_2 < \cdots < t_N$, which is evenly spaced: $t_n = t_0 + nh$ for $n = 0,1,2,\cdots,N$, and let $x_n = x(t_n)$. Runge-Kutta method gives the following

$$k_1 = v(t_n, x_n)$$
$$k_2 = v(t_n + \frac{h}{2}, x_n + \frac{h}{2}k_1)$$
$$k_3 = v(t_n + \frac{h}{2}, x_n + \frac{h}{2}k_2)$$
$$k_4 = v(t_n + h, x_n + hk_3)$$
$$x_{n+1} = x_n + \frac{h}{6}[k_1 + 2k_2 + 2k_3 + k_4].$$

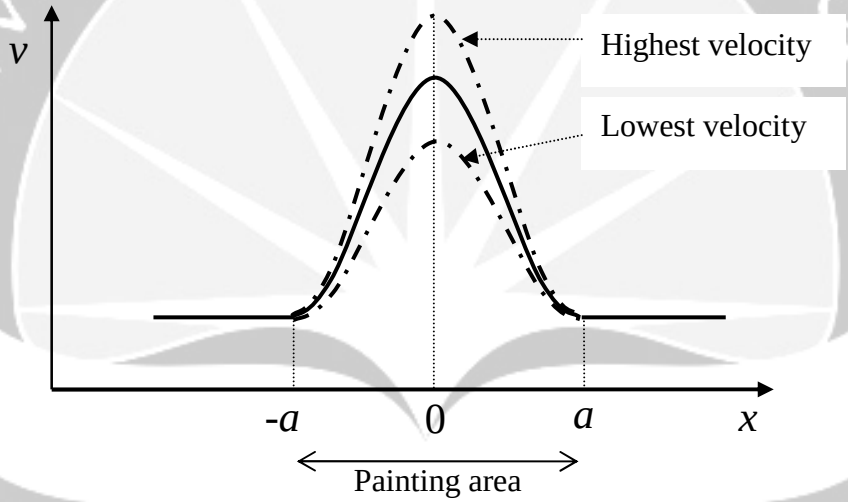


Figure 3. Illustrative plot of a smooth approximation to the velocity function.

We are interested in avoiding the worst case to happen. Hence, we consider glass 1 moves at the lowest velocity, then followed by glass 2 which moves at the highest velocity, see Figure 3. We solve (1) numerically for the two cases:

$$v(t,x) = f(x) - \max(|\varepsilon(x,t)|) \text{ for } x(t_0) = x_0$$

then

$$v(t,x) = f(x) + \max(|\varepsilon(x,t)|) \text{ for } x(t_0) = x_0 - (l+d)$$

where l is the length of the glasses and d is the distance of the glasses, i.e. the distance between the back part of the glass 1 and the front part of the glass 2. For the need of numerical simulation we use the parameters given in Table 1. Note that the length of the painting area is 2m.

Parameter	Value
a	1 m
d	0.05 m
h	0.03 minute
l	1.5 m
x_0	-2
ε_0	2 m/minute

Table 1. Parameters applied for numerical simulation.

Figure 4 shows the distance between the glasses as a function of time. Before entering the painting area the distance is 5 cm. When the glass 1 already enters the painting area and leaving the glass still outside, the distance grows up to 1 m. However, when the glass 2 already enters the painting area, the distance become smaller and reaching the minimum at about 3 cm after both glasses leaving the painting area.

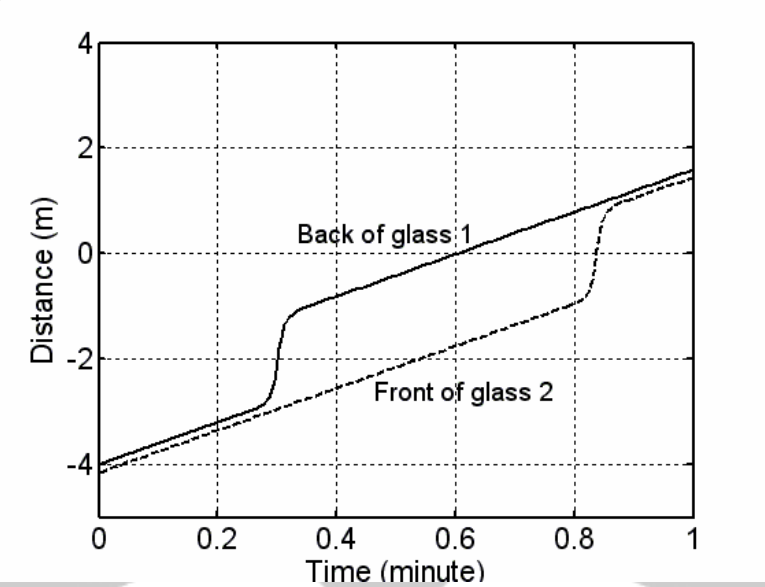


Figure 4. The distance of two adjacent glasses for the worst case.

4 Conclusion and recommendation

We have developed a mathematical model for a production process in a mirror industry, which is a velocity problem in the painting area where often causes crashes of two adjacent glasses or inefficiency. The process leads to a non-linear ordinary differential equation, where the existence and uniqueness of the solution is guaranteed.

The main interest is on the worst case, and to avoid the crash of two adjacent glasses. Based on the information of the velocity outside and at the middle of the painting area, the glass velocity is approximated by a polynomial function of order four. We have solved the model numerically using fourth order Runge-Kutta method, and found that the distance between the glasses after leaving the painting area is smaller than before entering this area. The numerical simulation shows that the distance of 5 cm before entering the painting area does not result in the crash of two adjacent glasses. Applying in the real production process, however, we need more information about the glass velocity in the whole painting area.

Acknowledgment

The first author thanks to Mr. Budiono Wijaya of PT Matahari Silverindo Jaya for the help to make the study of the industrial process possible.

References

[1] Atkinson, K (1985), *Elementary numerical analysis*, John Wiley & Sons, New York.

[2] Cahyono, E. & Kartono (2004), A glass velocity problem in a mirror industries, *Paradigma: Majalah Matematika dan Sains, Unhalu* **8**,1-10.

[3] Kaplan, W (1958), *Ordinary differential equations*, Addison-Wesley, London.

E. CAHYONO: Department of Mathematics, FMIPA, Universitas Haluoleo, Kampus Bumi Tridharma Anduonohu Kendari 93232, Indonesia. Phone
E-mail: edi_cahyono@unhalu.ac.id

R. RAYA: Department of Mathematics, FMIPA, Universitas Haluoleo, Kampus Bumi Tridharma Anduonohu Kendari 93232, Indonesia. Phone
E-mail: rasas_raya@unhalu.ac.id

L. D. SAIDI: Department of Mathematics, FMIPA, Universitas Haluoleo, Kampus Bumi Tridharma Anduonohu Kendari 93232, Indonesia. Phone
E-mail: ld_saidi@unhalu.ac.id

T. MASRIYATI: Department of Mathematics, FMIPA, Universitas Haluoleo, Kampus Bumi Tridharma Anduonohu Kendari 93232, Indonesia. Phone
E-mail: tati_masriyati@yahoo.com

Inverse Problems of Coal Gasification

Sapna Somani , Patel Dhaneshkumar

Department of Applied Mathematics, Faculty of Technology and Engineering, The M. S. University of Baroda, Vadodara, Gujarat, India.

Abstract: UCG is an alternative to conventional mining of coal. Recent CRIP (Controlled Retractable Injection Point) technology utilizes injection/production well pair in the selected site. Coal energy is recovered by igniting the coal remotely, partially combusting and gasifying coal by means of injected oxygen-steam-air mixtures through horizontal injection well. Product gases are evolved and produced through production well. The gases are cleaned and processed for a variety of end uses. The combustion and gasification of coal develops a cavity during UCG. The modeling of the geometrical evolution of a UCG cavity and the factors controlling its ultimate dimensions is a significant step towards understanding UCG process.
Mathematical modeling of UCG processes at various stages is essential. For example, analysis of coal pyrolysis and drying history for various coal geometries, prediction of cavity growth and product composition, etc. Post UCG requires modeling of subsidence using the overburden and mechanical properties of overlying rocks.
In this paper, we shall consider the problem of cavity growth process. Cavity consists of rubble, ash and void space left by the combustion/ gasification of coal. We have developed mathematical models for growth of cavity, which turns out to be **Inverse problem** in nature. We have also made an attempt to solve this models.

References:

[1]. Hinch E. J.: *Perturbation methods*, Cambridge University Press (1992). Ockendon John, Howison Sam, Andrew Lacy, and Movchan Alexander: *Applied Partial Differential Equations*, Oxford University Press (1999).

The Application of Cobb Douglas Function for Solving Linear Programming to analyze its Optimum

Aidawayati Rangkuti

Department of Mathematics, Faculty of Mathematics and Natural Sciences,
Hasanuddin University, Makassar

Abstract. This study attempts to analyze Linear Programming with Cobb Douglas function in solve the use of economic resources owned by Local Transmigrates in South Sulawesi, formulate the optimum use of the resources producing crops. The level of resources used and the economic scale is analyzed by using Cobb Douglas function. Optimization of the use of resources by ratio $(\alpha_j(y^*) \quad p/x_{mi}c_j)$, optimization of crops by Linear Programming. The estimation result of the use of resources indicates a positive and very significant role on production in which the production scale is at the decreasing returns to scale. The optimum profit increases to 801.95 %; 251.96 %; 455.84 %; and 346.67 % at the Transmigration Settlement Units of Lombok I; II; III; Bulukatoang; Timusu; and Pencong respectively.
Key-words: Linear Programming, Cobb Douglas Function

1. INTRODUCTION

The application of this research in mathematics is to develop a solution, and if it's possible, it can improve an optimum result from the level of a system tendency. Therefore, this operational research is considered that it can find an optimal and the best solution in an area of this operational research namely Linear Programming (LP) [3].

George B. Dantzig as a pioneer who develop Linear programming, to determine a method and to find the solution of the linear programming problems using many variables. The problem of the general application of Linear Programming is how to allocate the limited resources on the competitive activities in a good way [4].

Cobb Douglas function is a non-linier function which is widely used in determining the optimum level of production, the income level, the efficiency, etc. in this research Cobb Douglas function is used to determine the efficiency of the resource and to find out its correlation with Linear Programming.

In this research, the Linear Programming is used to determine the optimum level of income of the local transmigrates in South Sulawesi namely the transmigration of settlement units (TSU) of Lombok I, II, III, Timusu, Bulu katoang, and Pencong, where

each of them located in Sidrap, Maros, Soppeng, and Gowa regency. The income of local transmigrates in south sulawesi decreased because the resources owned by the local transmigrates is not optimum yet [1]. The second function is determining the efficient level of the usage of resources, such as : land, fertilizer, meds, seeds, tools, and labour by using The Cobb Douglas analysis to increase the income of local transmigrates on South Sulawesi, in this case planting pattern will be done by combining between crops (rice plants, corns, beans) and garden plants (cashew nuts, cocoas).

The optimum income is the optimum results which is giving a maximum positive impact, furthermore sensitivity analysis (post optimal) is done if any changes occurs on the resource cost.

2. LITERATUR REVIEW

The correlation between input and output

The technical correlation between input and output stated in a production function, this correlation is formulated with mathematics equation, namely y = $f(x_1, x_2, ..., x_n)$ with y as the output which is resulted from the input usage $(x_1, x_2, ..., x_n)$. One of the algebra form of production function is The Cobb Douglas form. Production function is a requirement that an equal input (x) and output (y) correlation, so that we can find the first descendant $\frac{\partial y}{\partial x}$, with several condition where adding an output cause the decrease of a result ($\frac{\partial y}{\partial x}$, negative) or the second descendant of production function gives a negative grade $\left(\frac{\partial^2 x}{\partial {x_i}^2} < 0\right)$. So that a condition happen namely the increment input reasonably causing the decrease of the increasing output, which means :

$$\left(\frac{\partial y}{\partial x_i}\right)\left(\frac{x_i}{y}\right) = 1, \text{ [9].}$$

The Cobb Douglas function stated as :

$$y_j = \prod_{j=1}^{m} a_0 \, {x_{ij}}^{\alpha j} \, {e_j}^{u} \qquad (1)$$

The marginal production of production factor * is :

$$\frac{\partial y_j}{\partial x_{1i}} = a_0 \alpha_1 x_{1i}^{\alpha_1 - 1} x_{2i}^{\alpha_2} \ldots x_{ni}^{\alpha_{nm}} \quad (2)$$

$$\frac{\partial y_j}{\partial x_{mi}} = \frac{\alpha_j (y^*)}{x^*_{mi}} \quad (3)$$

y^* is a geometric average production and x^*_{mi} is a geometric average from the total of the production factor j. The using of this production factor would be efficient if the marginal production value equal with the production cost, then mathematic form can be written :

$$p\left(\frac{\alpha_j (y^*)}{x^*_{mi}}\right) = c_j \quad (4)$$

With P is production cost per unit and c_j is production factor cost per unit [2].

From the (4) equation can be re-written :

$$\left(\frac{\alpha_j (y^*)}{x^*_{mi}}\right)\frac{p}{c_j} = 1 \quad (5)$$

The Cobb Douglas Model based on the assumption of farming company scale (elasticity) with categories :

a. if $\sum \alpha_j > 1$, production scale will be on the increasing return to scale position
b. if $\sum \alpha_j = 1$, production scale will be on the constant return to scale position
c. if $\sum \alpha_j < 1$, production scale will be on the decreasing return to scale position

The correlation of Cobb Douglas function with Linear Programming

Frontier production function (FPF) is a production function which is used to measure how the production function compare with the frontier position located on the isoquant line [4].

The Cobb Douglas Function model :

$$y_i = a_0 \sum_{j=1}^{m} x_{ij}^{\alpha_j} e_i^{u} \quad (6)$$

The value of y_i can be found by logarithming the equation :

$$\ln y_i = \ln a_0 + \sum_{j=1}^{m} \alpha_j \ln x_{ij} + \ln e_i^{\,u}$$

$$\text{or } y_i = \sum_{j=1}^{m} \alpha_j x_{ij} + u \qquad (7)$$

If the 7 (seventh) equation estimated with the frontier so u have to be minimized, so that minimizing u on condition that $\hat{y}_i = \sum_{j=1}^{m} a_j \times_{ij}$, thereby $\hat{y}_i \geq y_i$, and $a_j \geq 0$,, therefore this problems is a Linear programming problem with a_j can be counted.

By summing the perceived sample from the 7 (seventh) equation can also be re-written :

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_j x_{ij} - \sum_{i=1}^{n} u \quad \text{and} \quad \sum_{i=1}^{n} u = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_j x_{ij} - \sum_{i=1}^{n} y_i \qquad (8)$$

Or minimizing $\sum_{j=1}^{m} a_j \times_{ij}$ with note $\sum_{j=1}^{m} a_j \times_{ij} \geq y_i$, [7]

Thereby, generally the equation become :

Minimizing : $a_1 \times_1 + a_2 \times_2 + \ldots + a_m \times_n$

Condition : $a_1 \times_{11} + a_2 \times_{12} + \ldots + a_m \times_{1m} \geq y_1$

$$a_1 \times_{n1} + a_2 \times_{n2} + \ldots + a_m \times_{nm} \geq y_n$$

The Linear Programming Analysis

The general form of Linear programming analysis

Maximizing/minimizing Z = $\sum_{j=1}^{m} c_j \times_j$

With boundary condition $\sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij} \times_j \leq = \geq b_i$

$$\times_j \geq 0$$

i = 1, 2, 3, 4, ... , 8
j = 1, 2, 3, 4, 5

$(c_1 - c_5)$	=	net earning from farming, rice plants, corns, cashew nuts, beans, cocoas.
$(x_1 - x_2)$	=	the wide of pattern plant which wanted
$(a_{1j} - a_{8j})$	=	the coefficient of resource that had been used
$(b_1 - b_8)$	=	the limited economic resource (land, fertilizer, seeds, pesticide, tools, labour)
Z	=	the optimum opinion of local transmigrates

The food crops x_1 = rice plants (kg/Ha); x_2 = Beans (kg/Ha); x_3 = corns (kg/Ha);

The plantation crops x_4 = cashew nuts (kg/Ha); x_5 = cocoas (kg/Ha)

Compare with the other method, the using of linear programming is more efficient in the purpose of cost, capital, and the ability to analyze the result and using of data.

The Linear programming analysis supported by five basic assumption which become the power of this analysis , namely : (1) Linearity; (2) proporsionality; (3) addictive; (4) divisibility; (5) determinism. [8]

The utilized variable in this research is the labour who is measured by people days work (PDW), the wide of the land, fertilizer, seeds in Kg while equipments pesticide measured by nominal rupiahs value.

Primer data through the field observation with the interviews at the local transmigrates farmer using a questionnaire. The election of responder is conducted proporsionally from 570 questionnaire spread over the transmigration settlements unit of local transmigration and obtained 268 questionnaire (268kk) namely 20 % from populations amount (1.348) which assumed representative.

3. SOLUTION

The advantage from each farming like rice plants, cashew nuts, corns, beans, cocoas, is the target function coefficient and problem coefficient function, is obtained from the optimum usage of the resource which had been processed by the Cobb Douglas function, also the constraint boundary is also obtained from mean mount usage of rice plants resource, corns, beans, cashew nuts, and cocoas at the transmigration settlement units (TSU) in Lombok I, II, III.

The analysis result of Linear Programming (LP) by using the resource, such as : lands, urea fertilizer, TSP fertilizer, KCl fertilizer, pesticide, seeds, and labour in Lombok I, II, III.is obtained an advantage per Ha equal to Rp. 396.878,80. The (1) optimum result, a

profit was got to Rp. 1.369.009 from wide of corn farm 0,864 Ha and cashew nut farm 0,136 Ha. The optimum profit at (1) condition increase to 244,95 % from the actual condition. Next, on the (2) optimal by paying attention the allowable increase object coefficient range (Al-OCR) namely the function coefficient of target changing from the quantifying of coefficients early with current increase as a maximum increase boundary, so that the maximum profit on (2) optimum is Rp. 3.579.642 which is obtained from the wide of corn farm 0,864 Ha and corns for the width of 0,136 Ha, this condition increase to 801,95 %.

With the existence of the increasing input and output price, each of them 15 % on the (3) optimum condition showing a maximum profit equal to Rp. 1.574.361 obtained from the wide of corn farm 0,864 Ha and corns for the width of 0,136 Ha increase to 296,9 %. From the three condition, the (2) optimum condition was the maximum profit of all. The same way can be done for the transmigration settlement unit Timusu, Bulu katoang, and Pencong, like shown on the table 1.

From table 1, it can be seen for : (1) the transmigration settlement unit Timusu; the analysis of Linear programming on actual condition, getting a maximum profit equal to Rp.290.884, on the (1) optimum condition the maximum profit is Rp. 1.398.166; the (2) optimum condition equal to Rp. 1.616.836 and the (3) optimum condition equal to Rp. 1.607.89. This maximum profit is increase, each of (360,66%; 455,84%; 452,76%) from the actual condition. And so do to the transmigration settlement unit of Bulu Katoang and Pencong.

Table 1 : The recapitulation of analysis in order to make the farming of rice plants, corns, beans, cashew nuts, and cocoas in Local TSU optimum.

Local TSU	Optimum Condition	Changes	Optimum Solution	*) OFV (Rp)	**) PIL (%)
Lombok I,II,III TSU		Actual condition		396877	
	(1)	Present condition	$0,864X_2 + 0,136\ X_4$	1369009	244,95
	(2)	Price increase on interval Al-OCR	$0,864X_2 + 0,136\ X_4$	3579642	801,95
	(3)	Input and output price increase 15 %	$0,864X_2 + 0,136\ X_4$	1574361	296,69
Bulu Katoang TSU		Actual condition		483026	
	(1)	Present condition	$0,244X_4+0,542X_2+0,214X_5$	1210836	167,42
	(2)	Price increase on interval Al-OCR	$0,244X_4+0,542X_2+0,214X_5$	1700052	251,96
	(3)	Input and output price increase 15 %	$0,244X_4+0,542X_2+0,214X_5$	1392461	188,28
Timusu TSU		Actual condition		290884	
	(1)	Present condition	$0,089X_1+0,041X_3+0,214X_4$	1398166	360,66
	(2)	Price increase on interval Al-OCR	$0,089X_1+0,041X_3+0,214X_4$	1616836	455,84
	(3)	Input and output price increase 15 %	$0,089X_1+0,041X_3+0,214X_4$	1607891	452,76

Pencong TSU	(1)	Actual condition Present condition	$0{,}778X_1+0{,}055X_2+0{,}294X_3+0{,}138X_5$	463140 1220593	163,55
	(2)	Price increase on interval Al-OCR	$0{,}778X_1+0{,}055X_2+0{,}294X_3+0{,}138X_5$	2068687	346,67
	(3)	Input and output price increase 15 %	$0{,}778X_1+0{,}055X_2+0{,}294X_3+0{,}138X_5$	1403682	203,08

Source : Data Analysis
OFV : Objective Function Value
PIL : The percentage of Increasing Level

CONCLUSION

The owned resource of the local transmigrates in South Sulawesi, hasn't been optimally exploited, the governance area on each regency need to pay attention on the local transmigrates, because since the local transmigrates delivered to the governance area, seems like less pay attention, beside that we need to have some help from the governance in the case of fertilizer and medicines also the presentation of credit so that the local transmigrates can't go out from the location.

The Cobb Douglas function can measure the real production function to the frontier, located on the isoquant line, and the income function is the Cobb Douglas function which is can be found by using the Linear Program.

Bibliography

[1] Aidawayati R. 2004. *Optimalisasi Pemanfaatan Sumberdaya EkonomiTransmigran Lokal di Sulawesi Selatan***.** Proyek Peningkatan Unhas dalam Hibah Penelitian dan Hibah Pengajaran (SP-4 Tahun 2004). Makassar.
[2] Chermak and Patrick. 1995. *A well Based Cost Function and Economics of exhaustible Resources*. American Journal of Environmental Economics and Management. Volume 28. Number 2.
[3] Frederick S. H., Gerald J. L. 1990. *Pengantar Riset Operasi. Edisi Kelima.* Erlangga, Surabaya.
[4] Farrel. M.J. 1962. *Estimating Efficient Production Funciton Under Increasing Return for Scale*. Journal of the Royal Series A part 2 (125) page 252-267
[5] Hamdy A. Taha. 1996. *Riset Operasi Suatu Pengantar. Edisi Kelima***,** Jilid I. Binarupa Aksara, Jakarta.
[6] Karyono F. 1997. *Analisis Linear Programming Sektor Pertanian di Indonesia***.**Agro Ekonomi, No II Tahun X, PERHEPI.
[7] Soekartawi. 2002. *Teori Ekonomi Produksi dengan Pokok Bahasan Analisis Fungsi Cobb-Douglas*. PT. RajaGrafindo Persada, Jakarta.

[8] Taylor III, 2000. *Sains Manajemen. Pendekatan Matematika Untuk Bisnis.* Buku Satu, Edisi Kelima, Indonesia Jakarta.
[9] Wesly D. S.Grald, Harold. 1994. *Economics of Resources Agriculture and Foot.* MC, Graw Hill Inc, Singapore.

AIDAWAYATI RANGKUTI: Department of Mathematics, Faculty of Mathematics and Natural Sciences, Hasanuddin University, Makassar
Email: aidarangkuti05@Yahoo.com

PARALLEL PERFORMANCE OF EXPLICIT GROUP ITERATIVE ALGORITHMS ON SMP MULTIPROCESSORS

M. Othman[a], A.R. Abdullah[b]

[a] University Putra Malaysia, Malaysia
[b] University Kebangsaan Malaysia, Malaysia

Abstract. Recently, the modified explicit group method for solving 2D Poisson problem was introduced and it was shown to be the most superior as compared to the explicit decoupled group and standard explicit group methods. While the parallel version of standard explicit group, explicit decoupled group and modified explicit group iterative algorithms were implemented successfully on Shared Memory Multiprocessors computer system. In this paper, we will discusses the performance of parallel explicit group algorithms and the results were compared among the them in order to show their outstanding performances.

Key-words: Parallel Explicit Group Algorithms, Shared Memory Multiprocessors, Parallel Performance Evaluation

1 Introduction

The parallel point iterative algorithm which incorporates the full-sweep approach for solving a large and sparse linear system has been implemented successfully, see [2] and [3]. While the half-sweep approach was introduced by Abdullah [1] for the derivation of the Explicit Decoupled Group (EDG) method. Since the EDG method is explicit, it is suitable to be implemented in parallel on any parallel computer. Consequently, the parallel standard Explicit Group (EG) and EDG algorithms have been developed extensively by Evans *et al.* [5] and Yousif, *et al.*, [10], respectively for solving 2D Poisson problem. In [7], the four points Modified Explicit Group (MEG) method was proposed and the method is shown to be the most superior as compared to the four points- EDG and EG algorithms. The parallel version of four points MEG iterative algorithm was developed by Othman *et al.*, [9] for solving the same problem. All the parallel either point or group iterative algorithms were implemented on shared memory multiprocessor (SMP).

2 The Shared Memory Multiprocessors

Based on Flynn taxonomy [6], the shared memory multiprocessor is in a class of Multiple Instruction Multiple Data (MIMD) parallel computer and the basic architecture of SMP parallel computer is shown in Figure 1. In this research, the parallel computer consists of six tightly coupled processors, each a 32-bit Intel Pentium processor with 64Kb cache. The main memory of 64 MB is shared by all processors. The machine runs under the Dynix/ptx v2.1.0 operating system which is UNIX system. In addition to all the usual UNIX facilities, Dynix/ptx provides

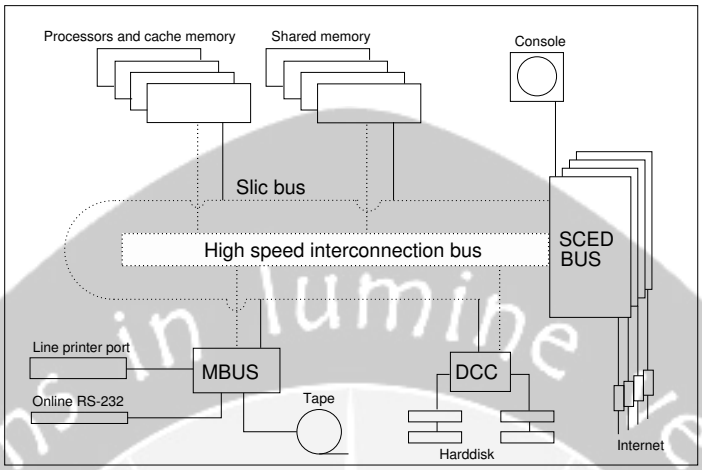


Figure 1: The basic architecture of SMP parallel computer.

capability of multiprocessing, employing all the CPUs available to support parallel processing. The processors are shared by all the processes including operating system and user processes. The Dynix/ptx C language is extended with parallel features and augmented by parallel library routines. It is possible to reserve a certain number of processors to be dedicated to a certain parallel program for the duration of its execution. This makes performing parallel experiments easy. Programming primitives for allocating shared memory, synchronization and timing of parallel processes are provided. Parallel activities are initiated in a program by creating child processes or tasks that execute independently but simultaneously with the parent process. Each task is essentially a separately running program. Communication among a group of cooperating parallel tasks was achieved through shared memory.

3 Design and Implementation of Parallel Explicit Group Algorithms

Assuming that the solution domain Ω is large with the mesh size n is an even number. Let N is a number of four points group or task T_i for $i = 1, 2, \ldots, N$ (in Figure 2, $N = 16$) which is greater than the number of processors, p_j for $j = 1, 2, \ldots, 6$ i.e. $N \gg p_j$. Since all groups are identical, the data partitioning approach is suitable in the implementation of the method and all the identical tasks can be executed in parallel. Again, the static scheduling is employed in this implementation.

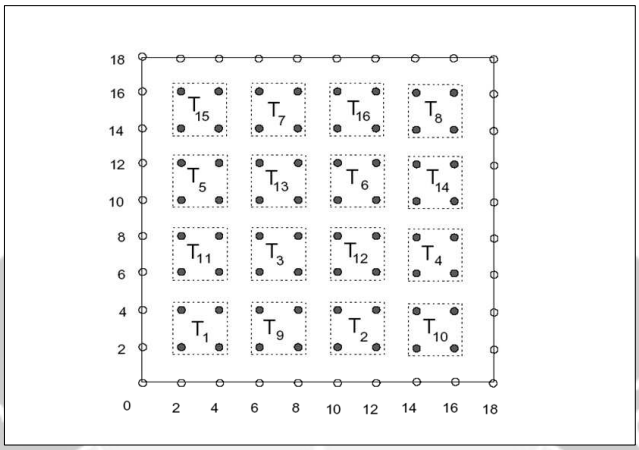


Figure 2: shows red black (RB) strategy for $n = 18$ and $T_i \ \forall i = 1, 2, \ldots, (\lfloor \frac{n-1}{4} \rfloor)^2$.

3.1 The Parallel MEG Algorithm

Let we consider all the tasks T_i, for all $i = 1, 2, \ldots, N$ with $N = (\lfloor \frac{n-1}{4} \rfloor)^2$ are assigned to available processors p_j in RB (i.e. red (r) and black (b)) ordering strategy, see Figure 2. By applying the following Eq. (see [7])

$$\begin{bmatrix} v_{i,j} \\ v_{i+2,j} \\ v_{i+2,j+2} \\ v_{i,j+2} \end{bmatrix} = \frac{1}{24} \begin{bmatrix} 7 & 2 & 1 & 2 \\ 2 & 7 & 2 & 1 \\ 1 & 2 & 7 & 2 \\ 2 & 1 & 2 & 7 \end{bmatrix} \begin{bmatrix} P_{i,j} \\ P_{i+2,j} \\ P_{i+2,j+2} \\ P_{i,j+2} \end{bmatrix} \tag{1}$$

where $P_{i,j} = v_{i-2,j} + v_{i,j-2} - 4h^2 f_{i,j}$, $P_{i+2,j} = v_{i+4,j} + v_{i+2,j-2} - 4h^2 f_{i+2,j}$, $P_{i+2,j+2} = v_{i+4,j+2} + v_{i+2,j+4} - 4h^2 f_{i+2,j+2}$ and $P_{i,j+2} = v_{i-2,j+2} + v_{i,j+2} - 4h^2 f_{i,j+2}$ to all tasks and thus will leads to a (2×2) system of equations,

$$\begin{bmatrix} D & U \\ U^T & D \end{bmatrix} \begin{bmatrix} u_r \\ u_b \end{bmatrix} = \begin{bmatrix} f_r \\ f_b \end{bmatrix} \tag{2}$$

with the diagonal sub matric D in which their diagonal elements consist of sub matric R_0. While the diagonal sub matric U consists of the following sub matric

$$\begin{bmatrix} R_1 & & R_2 & & & & & \\ R_1^T & R_1 & & R_2 & & & & \\ R_2^T & & R_1^T & R_1 & R_2 & & & \\ & R_2^T & & R_1^T & & R_2 & & \\ & & R_2^T & & R_1 & & R_2 & \\ & & & R_2^T & R_1^T & R_1 & & R_2 \\ & & & & R_2^T & & R_1^T & R_1 \\ & & & & & R_2^T & & R_1^T \end{bmatrix}$$

with sub matrices R_0, R_1 and R_2 consist of the following matrices

$$\begin{bmatrix} 4 & -1 & & -1 \\ -1 & 4 & -1 & \\ & -1 & 4 & -1 \\ -1 & & -1 & 4 \end{bmatrix}, \begin{bmatrix} 0 & & & \\ -1 & 0 & & \\ & & 0 & -1 \\ & & & 0 \end{bmatrix} \text{ and } \begin{bmatrix} 0 & & & 0 \\ & 0 & 0 & \\ & -1 & 0 & \\ -1 & & & 0 \end{bmatrix},$$

respectively. By simplifying Eq. (2), we will have

$$\begin{bmatrix} D & 0 \\ U^T & D \end{bmatrix} \begin{bmatrix} u_r \\ u_b \end{bmatrix}^{(k+1)} = \begin{bmatrix} f_r \\ f_b \end{bmatrix} - \begin{bmatrix} 0 & U \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_r \\ u_b \end{bmatrix}^{(k)}. \tag{3}$$

If the diagonal sub matric D^{-1} exist, we can evaluate Eq. (3) by first calculating

$$u_r^{(k+1)} = (1-\omega_e)u_r^{(k)} + \omega_e D^{-1}\left[f_r - U u_b^{(k)}\right] \tag{4}$$

followed by

$$u_b^{(k+1)} = (1-\omega_e)u_b^{(k)} + \omega_e D^{-1}\left[f_r - U^T u_r^{(k+1)}\right] \tag{5}$$

with the generated relaxation factor, ω_e.

From Eq. (4) and Figure 2, it is clear that all the red tasks $u_r^{(k+1)}$ (i.e. consists of tasks T_1, T_2, $\ldots, T_{\frac{N}{2}}$) are independent of each other and can be computed in parallel. After $u_r^{(k+1)}$ has been completed, all the black tasks (i.e. $T_{\frac{N}{2}+1}, \ldots, T_N$) which represented as $u_b^{(k+1)}$ or Eq. (5) can be calculated simultaneously using the updated values of $u_r^{(k+1)}$ since these calculations are independent. At the end of each stage, a synchronization call $m\_sync()$ is executed to ensure the updated values are used in the subsequent iteration. Each processor independently iterate on its own task T_i and check for its own local convergence.

If converge globally then the solution of the remaining points in Ω are evaluated directly at once using the rotated and standard stencils with the width of $\sqrt{2}h$ and h, respectively. The direct evaluations are executed in parallel. Otherwise, increased the number of iteration and repeat the iteration cycle.

3.2 The Parallel EDG Algorithm

Let we assigned all the tasks T_{l_i}, for all $i = 1, 2, \ldots, N$ with $N = (\lfloor \frac{n-1}{4} \rfloor)^2$ to available processors p_j in horizontal zebra line (HZL) ordering strategy, see Figure 3. In this strategy, it consists of two stages, the first stage is l_1 and l_2, and the second stage is l_3 and l_4. By applying the following Eq.

$$\begin{bmatrix} v_{i,j} \\ v_{i+1,j+1} \end{bmatrix} = \frac{1}{15} \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} v_{i-1,j-1} + v_{i+1,j-1} + v_{i-1,j+1} - 2h^2 f_{i,j}, \\ v_{i,j+2} + v_{i+2,j} + v_{i+2,j+2} - 2h^2 f_{i+1,j+1}, \end{bmatrix} \tag{6}$$

or

$$\begin{bmatrix} v_{i+1,j} \\ v_{i,j+1} \end{bmatrix} = \frac{1}{15} \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} v_{i,j-1} + v_{i+2,j-1} + v_{i+2,j+1} - 2h^2 f_{i+1,j}, \\ v_{i-1,j+2} + v_{i-1,j} + v_{i+1,j+2} - 2h^2 f_{i,j+1} \end{bmatrix} \tag{7}$$

in turn to each stage of tasks with such strategy, and leads the following equation

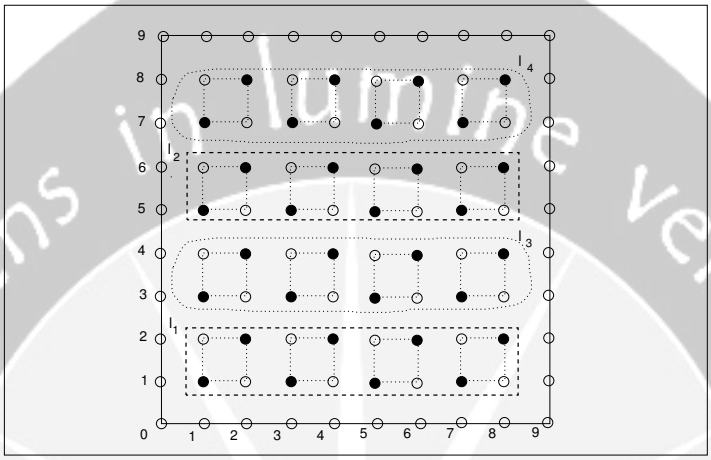


Figure 3: shows horizontal zebra line (HZL) strategy for $n = 9$ and T_{l_i} $\forall i = 1, 2, \ldots, (\lfloor \frac{n-1}{4} \rfloor)^2$.

$$\begin{bmatrix} D_{l_1} & & U & \\ & D_{l_2} & V & U \\ U^T & V^T & D_{l_3} & \\ & U^T & & D_{l_4} \end{bmatrix} \begin{bmatrix} u_{l_1} \\ u_{l_2} \\ u_{l_3} \\ u_{l_4} \end{bmatrix} = \begin{bmatrix} f_{l_1} \\ f_{l_2} \\ f_{l_3} \\ f_{l_4} \end{bmatrix} \tag{8}$$

with sub matrices U, V and D_{li} for $i = 1, 2, \ldots, 4$ consist of the following diagonal sub matrices

$$\begin{bmatrix} R_1 & R_1 & & \\ R_1 & R_1 & R_1 & \\ & R_1 & R_1 & R_1 \\ & & R_1 & \end{bmatrix}, \begin{bmatrix} R_1 & & & \\ R_1 & R_1 & & \\ & R_1 & R_1 & \\ & & R_1 & R_1 \end{bmatrix} \text{ and } \begin{bmatrix} R_0 & R_1 & & \\ R_1^T & R_0 & R_1 & \\ & R_1^T & R_0 & R_1 \\ & & R_1^T & R_0 \end{bmatrix},$$

respectively. While the sub matrices

$$R_0 = \begin{bmatrix} 4 & -1 \\ -1 & 4 \end{bmatrix} \text{ and } R_1 = \begin{bmatrix} 0 & \\ -1 & 0 \end{bmatrix}.$$

Since there are two stages, Eq. (8) can be rewrite as the following form

$$\begin{bmatrix} \hat{D}_1 & C \\ C^T & \hat{D}_2 \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}. \tag{9}$$

By simplifying Eq. (9), this will leads to

$$\begin{bmatrix} \hat{D}_1 & 0 \\ C^T & \hat{D}_2 \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \end{bmatrix}^{(k+1)} = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} - \begin{bmatrix} 0 & C \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \end{bmatrix}^{(k)}. \quad (10)$$

The explicit solution of Eq. (10) can be de-coupled into the following equations

$$U_1^{(k+1)} = (1-\omega_e)U_1^{(k)} + \omega_e \hat{D}_1^{-1}\left[B_1 - CU_2^{(k)}\right] \quad (11)$$

and

$$U_2^{(k+1)} = (1-\omega_e)U_2^{(k)} + \omega_e \hat{D}_2^{-1}\left[B_2 - C^T U_1^{(k+1)}\right] \quad (12)$$

with the diagonal sub matrices $\hat{D}_1^{-1}$ and $\hat{D}_2^{-1}$ exist.

Clearly, we can see that all the tasks in $U_1^{(k+1)}$ are independent of each other and can be computed in parallel. After $U_1^{(k+1)}$ has been calculated, $U_2^{(k+1)}$ can be calculated simultaneously using the updated values of $U_1^{(k+1)}$ since this calculation is independent. However, the most recent values of $U_1^{(k+1)}$ are to be used in Eq. (12), a synchronizing call $m\_sync()$ has to be made before the calculation of $U_2^{(k+1)}$ starts. Each processor then checks for its local and global convergence criteria, the same way as described in the previous algorithm. Once the global convergence is achieved, the solution at the remaining tasks will be evaluated directly in parallel.

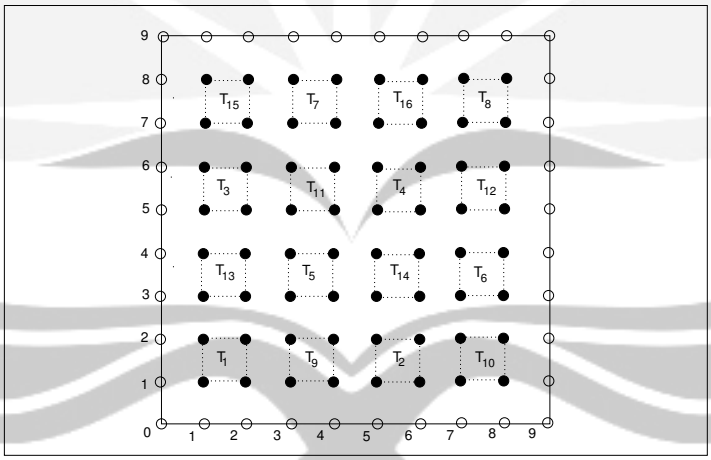


Figure 4: shows four color (4C) strategy for $n = 9$ and T_i $\forall i = 1, 2, \ldots, (\lfloor \frac{n-1}{2} \rfloor)^2$.

3.3 The Parallel Standard EG Algorithm

All the tasks T_i, for all $i = 1, 2, 3, \ldots, N$ with $N = (\lfloor \frac{n-1}{2} \rfloor)^2$ are allocated to the available processors p_j in four colors strategy, see Figure 4. The 4C strat-

egy, i.e. white (w), yellow (y), green (g) and red (r) consist of tasks groups $(T_1, \ldots, T_{\frac{N}{4}})$, $(T_{\frac{N}{4}+1}, \ldots, T_{\frac{N}{2}})$,$(T_{\frac{N}{2}+1}, \ldots, T_{\frac{3N}{4}})$ and $(T_{\frac{3N}{4}+1}, T_{\frac{3N}{4}+2}, \ldots, T_N)$, respectively. Then, iterate all the tasks by using the following Eq.

$$\begin{bmatrix} 4 & -1 & 0 & -1 \\ -1 & 4 & -1 & 0 \\ 0 & -1 & 4 & -1 \\ -1 & 0 & -1 & 4 \end{bmatrix} \begin{bmatrix} v_{i,j} \\ v_{i+1,j} \\ v_{i+1,j+1} \\ v_{i,j+1} \end{bmatrix} = \begin{bmatrix} Q_{i,j} \\ Q_{i+1,j} \\ Q_{i+1,j+1} \\ Q_{i,j+1} \end{bmatrix} \tag{13}$$

where $Q_{i,j}, Q_{i+1,j+1}, Q_{i+1,j}$ and $Q_{i,j+1}$ are equal to $v_{i-1,j}+v_{i,j-1}-h^2 f_{i,j}$, $v_{i+2,j}+v_{i+1,j-1}-h^2 f_{i+1,j}$, $v_{i+2,j+1}+v_{i+1,j+2}-h^2 f_{i+1,j+1}$ and $v_{i-1,j+1}+v_{i,j+1}-h^2 f_{i,j+1}$, respectively. Thus will leads to the following equation,

$$\begin{bmatrix} D & & E & F \\ & D & F^T & E^T \\ E^T & F & D & \\ F^T & E & & D \end{bmatrix} \begin{bmatrix} u_w \\ u_y \\ u_g \\ u_r \end{bmatrix} = \begin{bmatrix} f_w \\ f_y \\ f_g \\ f_r \end{bmatrix}. \tag{14}$$

with sub matrices D, E and F consist of the following sub matrices

$$\begin{bmatrix} R_0 & & & \\ & R_0 & & \\ & & R_0 & \\ & & & R_0 \end{bmatrix}, \begin{bmatrix} R_1 & & & \\ R_1^T & R_1 & & \\ & & R_1 & \\ & & R_1^T & R_1 \end{bmatrix} \text{ and } \begin{bmatrix} R_2 & & & \\ & R_2 & & \\ R_2 & & R_2 & \\ & R_2 & & R_2 \end{bmatrix},$$

respectively. The element of sub matrices R_0, R_1 and R_2 are the same as the element of sub matrices in Eq. (2). If the diagonal sub matric D^{-1} exist, we can evaluate Eq. (14) by first calculating

$$u_w^{(k+1)} = (1-\omega_e)u_w^{(k)} + \omega_e D^{-1}\left[f_w - Eu_g^{(k)} - Fu_r^{(k)}\right], \tag{15}$$

followed by

$$u_y^{(k+1)} = (1-\omega_e)u_y^{(k)} + \omega_e D^{-1}\left[f_y - F^T u_g^{(k)} - E^T u_r^{(k)}\right], \tag{16}$$

$$u_g^{(k+1)} = (1-\omega_e)u_g^{(k)} + \omega_e D^{-1}\left[f_g - E^T u_w^{(k+1)} - Fu_y^{(k+1)}\right], \tag{17}$$

and

$$u_r^{(k+1)} = (1-\omega_e)u_r^{(k)} + \omega_e D^{-1}\left[f_r - F^T u_w^{(k+1)} - Eu_y^{(k+1)}\right]. \tag{18}$$

From Eqs. (15), (16), (17), (18) and Figure 4, it is clear that all the tasks are independent of each other and can be computed in parallel. A group of tasks $u_w^{(k+1)}$ is allocated first to the available processors p_j, then after all the calculations in a group are completed, the synchronization call $m\_sync()$ will take place to ensure that the updated values of each points in the group are used in the subsequent iteration. Then the second groups of tasks $u_y^{(k+1)}$ is allocated to the available processors p_j followed by the third group $u_g^{(k+1)}$ and finally the group $u_r^{(k+1)}$. Each processor will checks for its local and global convergence, the same way as described in the previous algorithm.

4 Experimental Results

All the parallel algorithms described above were applied to a unit solution, Ω and the model used was $u_{xx} + u_{yy} = (x^2 + y^2)e^{xy}$ subject to the Dirichlet boundary conditions and satisfying the exact solution $u(x, y) = e^{xy},\ (x, y) \in \partial\Omega$. Throughout the experiments, a tolerance $\varepsilon = 10^{-10}$ in the local convergence test was used. The optimal relaxation factor ω was used and the experiments were carried out on the different sizes, $n = 26, 50, 74$ and 98.

Table 1: The iteration numbers and maximum errors of the parallel EG, EDG and MEG algorithms when number of processor is one.

n	Parallel Algo.	Strategies	ω	Ite no.	Max. error
	EG	4C	1.72	72	4.63×10^{-6}
26	EDG	HZL	1.69	69	2.46×10^{-4}
	MEG	RB	1.51	38	2.21×10^{-5}
	EG	4C	1.84	135	1.25×10^{-6}
50	EDG	HZL	1.83	129	6.64×10^{-5}
	MEG	RB	1.71	72	5.28×10^{-6}
	EG	4C	1.89	201	5.75×10^{-7}
74	EDG	HZL	1.88	198	3.03×10^{-5}
	MEG	RB	1.79	103	2.35×10^{-6}
	EG	4C	1.92	280	3.27×10^{-7}
98	EDG	HZL	1.91	265	1.72×10^{-5}
	MEG	RB	1.84	139	1.32×10^{-6}

Table 1 lists the strategies, ω, iteration numbers and maximum errors for all the parallel algorithms. While results in Table 2 show the total execution time, speedup and efficiency of all the parallel algorithms. The execution time and temporal performance were plotted in Figures 5 and 6, respectively.

5 Conclusion

The results obtained in Table 2 and Figure 5 have shown that the parallel performance of all the explicit group algorithms with their respective strategy produced a very good performance. However, the parallel performance of MEG algorithm with RB strategy regardless of the number of processors is the fastest among them. In addition, the temporal performance of the parallel MEG algorithm is the best as compared to the other two algorithms which showed the highest values, see Figure 6. While the results in Table 1 show the correctness of all the parallel explicit group algorithms. In conclusion, the parallel MEG algorithm with the RB strategy is the most superior among the family of parallel explicit group algorithms as the size of mesh points getting larger.

Table 2: The total execution time (Tet), speedup (Spe) and efficiency (Eff) for all the parallel EG, EDG and MEG algorithms. Note: #p indicates no. processors

n	#p	Parallel EG Algo.			Parallel EDG Algo.			Parallel MEG Algo.		
		Tet	Spe	Eff	Tet	Spe	Eff.	Tet	Spe	Eff
	1	3.35	1.00	1.00	1.60	1.00	1.00	0.48	1.00	1.00
	2	1.88	1.78	0.89	0.91	1.76	0.88	0.32	1.51	0.75
26	3	1.69	2.43	0.81	0.72	2.23	0.74	0.24	2.02	0.67
	4	1.06	3.14	0.78	0.55	2.91	0.72	0.22	2.13	0.53
	5	0.95	3.51	0.70	0.48	3.31	0.66	0.20	2.39	0.47
	1	24.71	1.00	1.00	11.97	1.00	1.00	3.35	1.00	1.00
	2	13.14	1.88	0.94	6.49	1.84	0.92	1.94	1.72	0.86
50	3	9.90	2.49	0.83	5.06	2.36	0.78	1.50	2.22	0.74
	4	8.52	3.19	0.79	3.71	3.22	0.80	1.24	2.69	0.67
	5	6.29	3.92	0.78	3.09	3.86	0.77	1.02	3.26	0.65
	1	80.94	1.00	1.00	38.05	1.00	1.00	10.22	1.00	1.00
	2	43.02	1.88	0.94	20.44	1.86	0.93	5.67	1.80	0.90
74	3	30.90	2.61	0.87	15.19	2.50	0.83	4.57	2.23	0.74
	4	23.61	3.42	0.85	11.82	3.21	0.80	3.35	3.04	0.76
	5	19.84	4.07	0.81	9.73	3.91	0.78	2.80	3.64	0.72
	1	205.05	1.00	1.00	97.16	1.00	1.00	25.40	1.00	1.00
	2	104.43	1.95	0.79	49.98	1.89	0.94	13.65	1.86	0.93
98	3	73.75	2.78	0.92	37.01	2.62	0.87	9.82	2.58	0.86
	4	60.59	3.38	0.84	28.65	3.39	0.84	7.67	3.31	0.82
	5	49.87	4.11	0.82	24.16	4.02	0.80	6.54	3.88	0.77

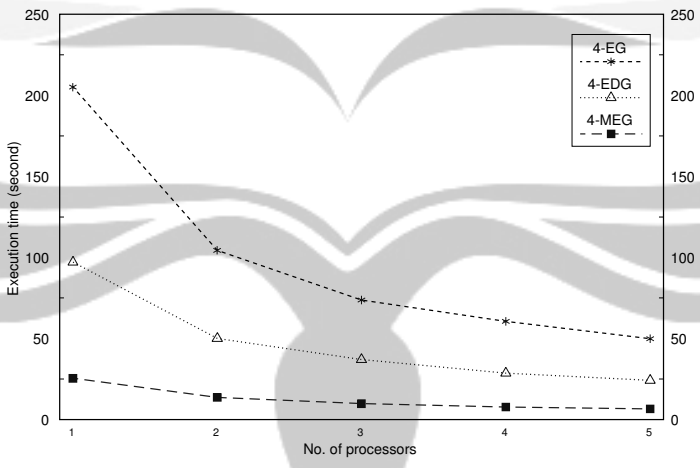


Figure 5: Total execution time versus no. of processors when $n = 98$.

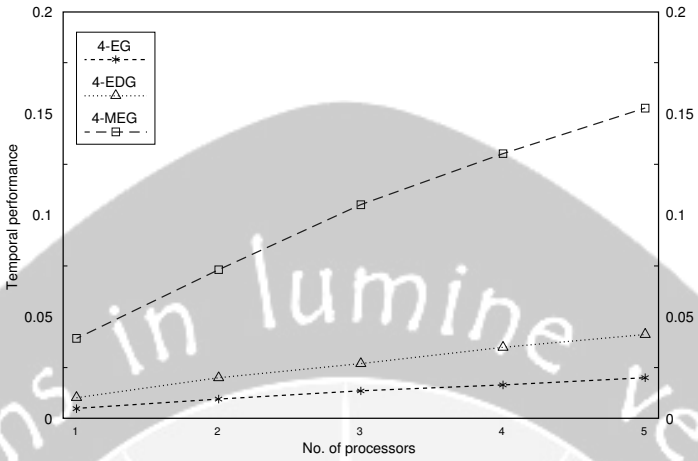


Figure 6: Temporal performances of all the parallel algorithms when $n = 98$.

References

[1] Abdullah, A.R. (1991), The Four Explicit Decoupled Group ($\mathcal{EDG}$) Method: A Fast Poisson Solver, *Intern. J. Comput. Math.*, **38**, 60–70.

[2] Barlow R.H. and Evans, D.J. (1982) Parallel Algorithms for the Iterative Solution to Linear System", *Computer Journal*, **25(1)**, 56–60.

[3] Evans, D.J. (1984) Parallel S.O.R. Iterative Methods, *Parallel Computing*, **1**, 3–18.

[4] Evans D.J. and Biggins, M.J. (1982), The Solution of Elliptic Partial Differential Equations by A New Block Over-Relaxation Technique, *Intern. J. Comput. Math.*, **10**, 269–282.

[5] Evans D.J. and Yousif, W.S. (1990) The Implementation of the Explicit Block Iterative Methods on the Balance 8000 Parallel Computer, *Parallel Computing*, **16**, 81–97.

[6] Flynn, M.J. (1972) Some Computer Organizations and Their Effectiveness, *IEEE Transactions on Computers*, **C-21(9)**, 948–960.

[7] Othman M. and Abdullah, A.R. (2000), An Efficient Four Points Modified Explicit Group Poisson Solver, *Intern. J. Comput. Math.*, **76**, 203–217.

[8] Othman, M., Abdullah, A.R. and Evans D.J., (2004), A Parallel Four Point Modified Explicit Group Iterative Algorithm on Shared Memory Multiprocessors, *Parallel Algorithm and Application*, **19(1)**, 1–9.

[9] Othman, M., Abdullah, A.R. and Evans D.J., (2004), A Parallel Four Point Modified Explicit Group Iterative Algorithm on Shared Memory Multiprocessors, *Parallel Algorithm and Application*, **19(1)**, 1–9.

[10] Yousif W.S. and Evans, D.J. (1995) Explicit De-coupled Group Iterative Methods and Their Parallel Implementations, *Parallel Algorithms and Applications*, **7**, 53–71.

A Authors Full Addresses

M. Othman: Department of Communication Technology and Network, University Putra Malaysia, 43400 UPM Serdang, Selangor D.E., MALAYSIA.
http://www.fsktm.upm.edu.my/˜ mothman/
e-mail: mothman@fsktm.upm.edu.my; mothmanupm@yahoo.com

A.R. Abdullah: Department of Industrial Computing, University Kebangsaan Malaysia, 43600 UKM Bangi, Selangor D.E., MALAYSIA.

SELF-CONSISTENT MODELING OF NANOMETER-WIDTH SILICON SUBSURFACE POTENTIAL WELL IN CALCULATING FOWLER-NORDHEIM EMISSION

Adi Bagus Suryamas, Khairurrijal, and Mikrajuddin

Institut Teknologi Bandung, Bandung, Indonesia

Abstract. The Fowler-Nordheim emission from the n-type silicon with (100) orientation has received a lot of attention in recent years due to their promise as high frequency devices, sensors, and flat-panel displays. A theoretical study on the Fowler-Nordheim emission has been performed by considering the quantum effect of the nanometer-width quantum well in the silicon accumulation layer. The coupled Schrödinger-Poisson equation has been self-consistently solved in obtaining the potential profile of the quantum well, the electron-energy levels and wave functions in the quantum well. It has been found that there are five lowest sub-band states. The lifetimes, occupancies, and Fowler-Nordheim emission currents of the five lowest quasi-bound states have been calculated.
Key-words: Accumulation layer, bound states, field emission, Fowler-Nordheim emission, quantum well, self-consistent solution, subsurface

1 Introduction

Field emission from silicon has received a lot of attention in recent years due to their promise as high frequency devices, sensors, and flat-panel displays [1]. Field emission was first introduced by Fowler and Nordheim [2] to explain electron emission from a metal surface into vacuum because of high electric field. Next, Stratton analyzed the field emission in a semiconductor-vacuum structure by considering the energy bands bending in the semiconductor, which does not exist in the metal-vacuum structure [3]. However, his work did not consider the quantum effect in the semiconductor. In the present article, we report a theoretical study on the Fowler-Nordheim emission from an n-type silicon (Si)-vacuum structure by considering the quantum effect that occurs in the silicon region. When a negative voltage is applied to the silicon, the energy bands of the silicon near the silicon-vacuum interface will be bent downward and an accumulation layer will be formed [4]. This accumulation layer then establishes a nanometer-width quantum well in which sub-band states are created. The electron lifetime, occupancy, and current density are then calculated.

2 Theoretical model

The potential profile of the n-type Si with (100) orientation-vacuum structure when a negative bias is applied to the silicon side (the accumulation case) is shown in Fig. 1. In the presence of an external uniform electric field F_0 along z in vacuum region, the potential profile is written as

$$V(z) = \begin{cases} q\psi(z) & \text{for} \quad -\infty < z \le 0 \\ \chi - qF_0 z & \text{for} \quad 0 < z < +\infty, \end{cases} \tag{1}$$

where $\psi(z)$ is the potential to be obtained by solving the Schrödinger and Poisson equations self-consistently, q the electronic charge, and χ the electron affinity of silicon.

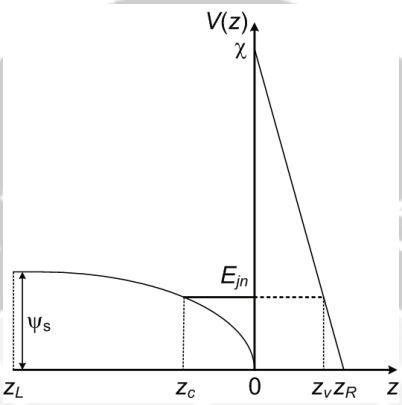


Fig 1. Potential profile of the n-type Si(100)-vacuum structure

The behavior of an electron in the system indicated in Fig. 1 is described by the time-independent Schrödinger equation

$$-\frac{\hbar^2}{2m^*}\frac{d^2\Phi(z)}{dz^2}+V(z)\Phi(z)=E\Phi(z) \quad . \tag{2}$$

Here, $\hbar$ is the reduced Planck constant, m^* the electron effective mass normal to the interface, E the electron energy, and $\Phi(z)$ the electron wave function. The Schrödinger equation is then solved by employing the finite difference method with the boundary conditions of the wave function $\Phi(z_L)=\Phi(z_R)=0$ [5] and the continuity condition of the wave function $m_s{}^{-1}(d\Phi/dz)_{z=0^-}=m_0{}^{-1}(d\Phi/dz)_{z=0^+}$ [6], where m_s and m_0 are the electron effective masses in silicon and vacuum regions, respectively.

The Poisson equation to be solved in obtaining the potential $V(z)$ is

$$\frac{d^2V}{dz^2}=-\frac{\rho(z)}{\varepsilon}, \tag{3}$$

where ε is the permittivity, in which ε_0 in vacuum region and ε_s in silicon region. The charge density is given by $\rho(z)=qN_D-q\sum_{j,n}N_{jn}|\Phi(z)|^2$ [7], where N_D is the donor concentration and N_{jn} the electron density of the level n and the valley j per unit area expressed as

$$N_{jn}=n_{vj}\frac{m_{dj}}{\pi\hbar^2}kT\ln\left[1+\exp\left(\frac{E_F-E_{jn}}{kT}\right)\right]. \tag{4}$$

Here n_{vj} is the valley degeneracy, m_{dj} the density of state mass per valley parallel to the silicon surface, k the Boltzmann constant, and T the temperature. The Fermi level E_F for a set of the sub-band energy levels E_{jn} is obtained from the surface electron density $N_s=\varepsilon_0F_0/q=\sum N_{jn}$ [8].

The finite difference method is also used to solve the Poisson equation with boundary conditions $V(z_L) = q\psi_s$ and $V(z_R) = 0$ [7], where ψ_S is the surface potential and the continuity condition given by $\varepsilon_s (dV/dz)_{z=0^-} = \varepsilon_0 (dV/dz)_{z=0^+}$.

The self-consistent calculation starts by solving Eq. (2) with the guess potential [8]

$$V(z) = q\psi(z) = \begin{cases} \psi_s - [\psi_s - \psi_1(z)]\exp\{(z - z_i)/L_D\} & for \quad z_i \le z \le 0 \\ 2kT \ln\{1 - [qE_s z/(2kT)]\} & for \quad -\infty \le z \le z_i \end{cases}. \quad (5)$$

Here, $E_s = F_0/K_s$ is the electric field at the silicon surface, K_s the dielectric constant of silicon, $L_D = \{\varepsilon_s kT/(q^2 N_D)\}^{1/2}$ the Debye length for the electron, $\psi_s = (2kT/q)\ln[qL_D E_s/(\sqrt{2}kT)]$ the surface potential, and z_i the intersection point determined from $[\sqrt{2}kT/(q\psi_s)] = \exp(z_i/L_D)\{\exp[-q\psi_s/(2kT)] - (z_i/L_D)\}$.

The electron energy E and wave function $\Phi(z)$ obtained from Eq. (2) are then substituted into the Poisson equation given by Eq. (3) to get a new potential $V(z)$. The new potential is compared with the previous one, which is used by the Schrödinger equation in Eq. (2). This process is done repeatedly until convergent.

A quasi-bound electron existing in the quantum well has a finite lifetime before leaving the well and tunneling through the triangular barrier of the vacuum region. The lifetime of the quasi-bound states is expressed as [7]

$$\tau_{jn} = \frac{\int_{z_c}^{0} \left[2m_{3j}/(E_{jn} - V(z))\right]^{1/2} dz}{T(E_{jn})}, \quad (6)$$

where m_{3j} is the electron mass in the silicon region normal to the interface. The probability of an electron tunneling through the triangular barrier is given by $T(E_{jn}) = \exp\left\{-2(2m_0/\hbar^2)^{1/2}\int_0^{z_v} (\chi - E_{jn} - qF_0 z)^{1/2} dz\right\}$ [8], where and z_v and z_c are the two positions shown in Fig. 1.

The Fowler-Nordheim emission current due to the quasi-bound states $[j,n]$ is [10]

$$J_{jn} = q N_{jn}/\tau_{jn}, \quad (7)$$

with the total current is $J = \sum_{j,n} J_{jn}$.

3 Calculated results and discussion

In order to obtain the electron lifetime, occupancy level, and current density for the n-type Si(100) electron field emitter, the following parameters were used: T = 300 K, χ= 4.05 eV, N_D= 10^{23} m^{-3}. For silicon with the (100) orientation, m_{si} has the different values depend on the valley and the number of degeneracy as summarized in Table I [9].

Five lowest sub-band states (j,n), where j = L or H stands for lower L or twofold degenerated and higher H or fourfold degenerated valleys and n = 0,1 or 2 stands for energy level of each valley, were resulted from this simulation. It is found that

the order of the five sub-band energy levels is as follows: $E_{(L,0)}$, $E_{(H,0)}$, $E_{(L,1)}$, $E_{(L,2)}$, and $E_{(H,1)}$, in which $E_{(L,0)}$ is the lowest one.

Table I. Parameters of Si(100) used in calculation
(where m_0 is the free electron mass)

Valley *j*		**Lower (*L*)**	**Higher (*H*)**
Degeneracy	n_v	2	4
Normal mass	m_3	0.916 m_0	0.190 m_0
Longitudinal mass	m_1	0.190 m_0	0.190 m_0
	m_2	0.190 m_0	0.916 m_0
Density-of-states mass	m_d	0.190 m_0	0.417 m_0

The occupancy of each sub-band state, which is defined as N_{jn}/N_s and expressed in percent, for the five lowest sub-band states is described in Fig. 2. The occupancy is nearly dependent of the electric field. It is also found that most electrons (higher than 80%) occupy the lowest states of each valley, i.e. the states (L,0) and (H,0). The occupancy of the state (H,0) is higher than that of (L,0) because the degeneracy and density of states mass of the state (H,0) are higher than those of the state (L,0) as explained by Eq. (4) and Table I.

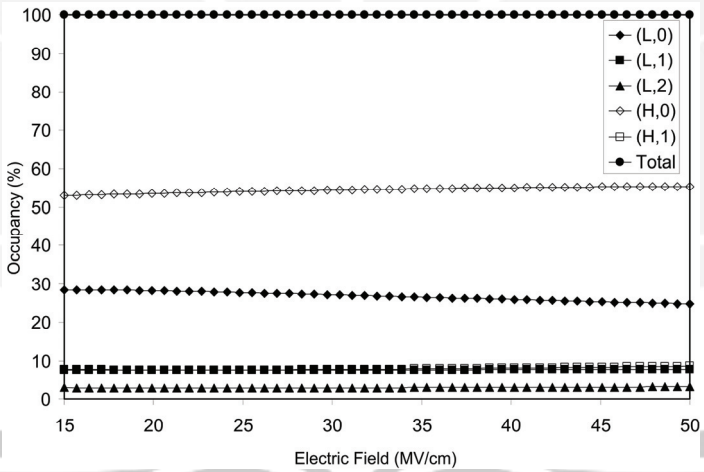


Fig. 2. Occupancy of five lowest sub-band states.

Figure 3 gives the lifetimes of the five quasi-bound states as a function of the external electric field. It is shown that the lifetimes of all states decreases as the electric field is increased. The state (*L*,0) with the lowest energy has the longest lifetime (around 10,000 times longer than the lifetimes of other states).

The Fowler-Nordheim emission currents from each quasi-bound state was calculated by using Eq. (7) and the results plotted as a function of the electric field is shown in Fig. 4. The total current was mainly contributed by the current originating from the state (H,0). This can be easily explained by inspecting Eq. (7) along with Figs. 2 and 3. It is clearly seen that the emission current of each state is significantly determined by the lifetime. Since the lifetime of the state (L,0) is about

ten thousands times longer than that of the state (H,0), the current from the state (H,0) is about four orders in magnitude higher than that from the state (L,0).

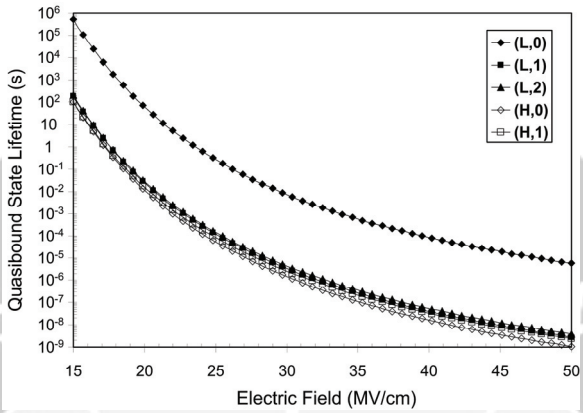


Fig. 3. Lifetimes of five lowest quasi-bound states.

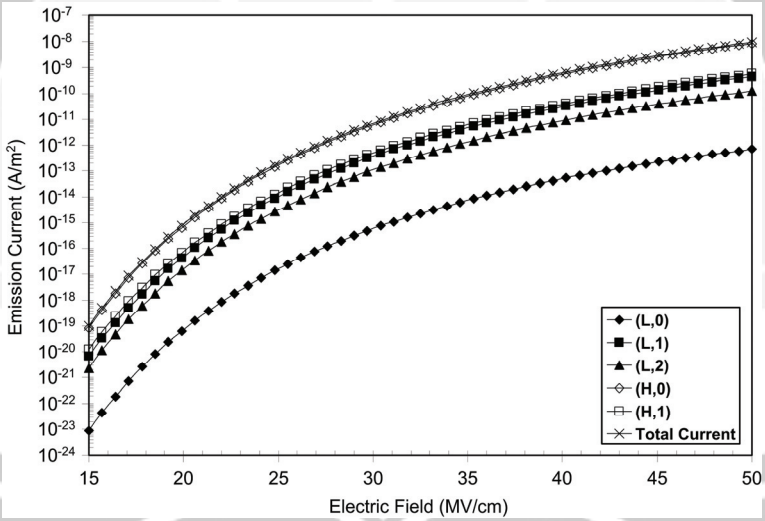


Fig. 4. Emission currents due to each state.

4 Conclusion

We studied theoretically the Fowler-Nordheim emission current of the n-type Si(100)-vacuum structure by considering the quantum effect in the Si accumulation layer. The potential profile of the nanometer-width quantum well in the Si accumulation layer was obtained by solving the coupled Schrödinger-Poisson equation self-consistently. Five lowest sub-band states were found in the quantum well. We calculated the lifetimes, occupancies, and Fowler-Nordheim emission currents of the quasi-bound states. It was found that the occupancy of the state (H,0) is higher than that of the state (L,0). The lifetime of the state (L,0) is the

longest as compared to those of other states. The current from the state (H,0) is the main contribution to the total Fowler-Nordheim emission current because the lifetime of the state (H,0) is shorter than that of the state (L,0).

References

[1] Yang, Y.J. (1999), *Numerical analysis and design strategy for field emission devices,* Doctoral Dissertation, Massachusset Institute of Technology.
[2] Fowler, R.H. and L. Nordheim (1928), Electron emission in intense electric fields, *Proc. R. Soc. London A*, **119**, 173-181.
[3] Stratton, R. (1962), Theory of field emission from semiconductors, *Phys. Rev.*, **125,** 67-83.
[4] Sze, S.M. (1981), *Physics of semiconductor devices 2nd ed.*, John Wiley & Sons, New York.
[5] Morrison, M.A. (1990), *Understanding quantum physics*: *A user's manual*, Prentice Hall, USA.
[6] BenDaniel, D.J. and C.B. Duke (1966), Space-charge effects on electron tunneling. *Phys Rev.*, **152**, 683-692.
[7] Rana, F., S. Tiwari, and D.A. Buchanan [1996], Self-consistent modeling of accumulation layers and tunneling currents through very thin oxides, *Appl. Phys. Lett.*, **69**, 1104-1106.
[8] Khairurrijal, S. Miyazaki, and M. Hirose (1999), Electron field emission from a silicon subsurface based on generalized Airy function approach, *J. Vac. Sci. Technol. B*, **17,** 306-310.
[9] Stern, F. (1972), Self-consistent results for n-type Si inversion layers, *Phys. Rev. B* **5**, 4891-4899.

ADI BAGUS SURYAMAS: Master student, Department of Physics, Institut Teknologi Bandung, Jalan Ganesa 10, Bandung 40132, Indonesia. Phone: +62-22-250 0834. Fax: +62-22-250 6452.

KHAIRURRIJAL: Department of Physics, Institut Teknologi Bandung, Jalan Ganesa 10, Bandung 40132, Indonesia. Phone: +62-22-250 0834. Fax: +62-22-250 6452. E-mail: krijal@fi.itb.ac.id

MIKRAJUDDIN: Department of Physics, Institut Teknologi Bandung, Jalan Ganesa 10, Bandung 40132, Indonesia. Phone: +62-22-250 0834. Fax: +62-22-250 6452. E-mail: din@fi.itb.ac.id

Non-Relativistic Stochastic Quantum Mechanics of a Particle Subjected to a Coulumb Force

W.S.B. Dwandaru

Computational and Theoretical Physics Laboratory, Physics Education Department, Yogyakarta State University, Indonesia

Abstract: Stochastic processes has been applied in quantum mechanics, which originated by E. Nelson. It is fascinating to acknowledge that a conservative diffusion process based on Ito stochastic differential equation turns out to be identical to non-relativistic quantum mechanics. Moreover, stochastic quantum mechanics has been applied mostly for harmonic oscillator and quantum fields, but not for other conservative forces. Therefore this paper will discuss the role of stochastic quantum mechanics in explaining a stochastic particle which is subjected to a Coulumb force. This will obviously give interesting benefit in the understanding of stochastic processes in quantum mechanics.

Keywords: Mathematical Physics, Quantum Mechanics

INFLUENCE OF SURFACE TREATEMENTS ON FATIGUE LIFE OF A FREE PISTON LINEAR GENERATOR ENGINE COMPONENTS USING RANDOM LOADING

M. M. Rahman, A. K. Ariffin and Tulus

Computational and Experimental Mechanics Group
Department of Mechanical and Materials Engineering
Universiti Kebangsaan Malaysia (UKM)
43600, Bangi, Selangor DE, Malaysia.
Phone:+ (6)03-89216012; Fax: + (6)-03-89216040
E-mail: mustafiz@eng.ukm.my, kamal@eng.ukm.my

Abstract. This paper describes finite element based vibration fatigue analysis techniques that can be used to predict fatigue life using the narrow band frequency response approach. Such life prediction results are useful for improving the component design at very early stage. This approach is satisfactory for periodic loading but requires very large time records to accurately describe random loading processes. The focus of this paper is to investigate the effects of nitrided and shot peening on the fatigue life of the components of free piston engine. The finite element modeling and frequency response analysis have been performed using finite element analysis software package MSC.PATRAN/ MSC.NASTRAN and the fatigue life prediction was carried out using MSC.FATIGUE software. Results indicate the great effects for all surface treatment. It is concluded that nitrided treatment condition has been found the highest lives. This significantly reduces time to market, improve product reliability and customer confidence consequences of premature produce failure.
Keywords: Vibration fatigue, finite element, power spectral density function, frequency response, surface treatment.

1 Introduction

The principal surface treatments such as carburizing or carbonitriding, carried out on many mechanical components before their delivery, are aimed to differentiate the response of surface and core to external loading by changing the surface material properties and by introducing appropriate residual stress distribution in order to improve their fatigue and wear behaviour [1]. Among the different treatments that can be carried out to locally improve the material response and to modify the stress field, a combination of case hardening followed compared with other surface treatments, mainly because of the nitriding, shot peening improvement of the residual stress profile introduced by case hardening [2]-[3]. Shot peening after case hardening contributes to an improvement both of the microstructure and of the residual stress distribution. Usually residual stresses are introduced by shot peening because of the intense plastic deformation in the surface region [4], which distinguishes between plastic deformation induced by the Hertzian pressure responsible for the subsurface peak, and plastic deformation due to surface hammering which tends to localize the peak on the surface. Depending

on whether the plastic deformation takes place on or below the surface, a shift of the residual stress peaks can be observed with respect to the surface.

Fatigue is an important parameter to be considered in the behaviour of components subjected to constant and variable amplitude loading [5]. Fatigue is of great concern for components subject to cyclic stresses, particularly where safety is paramount, for examples free piston linear generator engine components. It has long been recognized that fatigue cracks generally initiate from free surfaces and that performance is therefore reliant on the surface topology/integrity produced by surface finishing. It was well known that, in service, many more components and structures fail by cyclic than by static loading. The failure by fracture depends on a large number of parameters and vary often develops from particular surface areas of engineering parts. Therefore, it is possible to improve the fatigue strength of fatigue parts by the application of suitable surface treatments. In order to enhance the surface properties of today's materials, producers of components are turning to different surface treatments. There are various methods have so far been employed in order to improve fatigue strength, including optimization of geometric design, stronger, materials and surface processing such as Nitriding, cold Rolled, shot peening etc [6].

The surface treatments have been the most effective and widely used method of introducing compressive residual stresses into the surface of metals to improve fatigue performance [7]. The significance of nitrided, cold rolled and shot peened as on surface treatments have risen to even greater importance with the advancement of new analysis methods in metal physics and materials science over the past decades. Typical characteristics of nitrided and shot peened surfaces are compressive residual stresses and extremely high dislocation densities in near surface layers resulting from inhomogeneous plastic deformations. In some cases phase transformations occur, leading to additional surface hardening. These microstructural features are generally considered as the reason for inhibited crack initiation and propagation in components which are cyclically loaded [8].

The objective of this paper was to study the influence of surface treatments on the high cycle fatigue of aluminum alloys vibrating cylinder block of a two-stroke free piston engine. However, these investigations are essential in order to understand the involved microstructural mechanisms of hardening or softening in the wake of service load. Numerical investigates ere performed to characterize completely the different induced effects before and after surface treatments. The numerical results were discussed and analysed.

2 Theoretical Basis

The equation of motion of a linear structural system is expressed in matrix format in equation (1). The system of time domain differential equations can be solved directly in the physical coordinate system.

$$[M]\{\ddot{x}(t)\}+[C]\{\dot{x}(t)\}+[K]\{x(t)\}=\{p(t)\} \quad (1)$$

where $\{x(t)\}$ is a system displacement vector, $[M]$, $[C]$ and $[K]$ are mass, damping and stiffness matrices, respectively, {p(t)}is an applied load vector.

When loads are in random in nature, a matrix of the loading power spectral density (PSD) functions $[S_p(\omega)]$ can be generated by employing Fourier transform of load vector $\{p(t)\}$.

$$\left[S_p(\omega)\right]_{m\times m} = \begin{bmatrix} S_{11}(\omega) & \Lambda & S_{1i}(\omega) & \Lambda & S_{1m}(\omega) \\ M & 0 & & \Lambda & M \\ S_{i1}(\omega) & & S_{ii}(\omega) & & S_{im}(\omega) \\ M & \Lambda & & 0 & M \\ S_{m1}(\omega) & \Lambda & S_{mi}(\omega) & \Lambda & S_{mm}(\omega) \end{bmatrix} \tag{2}$$

where m is the number of input loads. The diagonal term $S_{ii}(\omega)$ is the auto-correlation function of load $p_i(t)$, and the off-diagonal term $S_{ij}(\omega)$ is the cross-correlation function between loads $p_i(t)$ and $p_j(t)$. from the properties of the cross PSDs, it can be shown that the multiple input PSD matrix $[S_p(\omega)]$ is a Hermitian matrix.

The system of time domain differential equation of motion of the structure in equation (1), is then reduced to a system of frequency domain algebra equations

$$\left[S_x(\omega)\right]_{n\times n} = \left[H(\omega)\right]_{n\times m} \left[S_p(\omega)\right]_{m\times m} \left[H(\omega)\right]^T_{m\times n} \tag{3}$$

where n is the number of output response variables. The T denotes the transpose of a matrix. $[H(\omega)]$ is the transfer function matrix between the input loadings and output response variables.

$$\left[H(\omega)\right] = \left(-\left[M\right]\omega^2 + i\left[C\right]\omega + \left[K\right]\right)^{-1} \tag{4}$$

The response variables $[S_p(\omega)]$ such as displacement, acceleration and stress response in terms of PSD functions are obtained by solving the system of the linear algebra equations in equation (3).

3 The Spectral Moments from PSD

The stress power spectra density represents the frequency domain approach input into the fatigue [9]-[10]. This is a scalar function that describes how the power of the time signal is distributed among frequencies [11]. Mathematically this function can be obtained by using a Fourier Transform of the stress time history's auto-correlation function, and its area represents the signal's standard deviation. It is clear that PSD is the most complete and concise representation of a random process. The statistical properties of a stationary ergodic process [12]-[15] can be computed from a single time history of sufficiently long period. The time average of a random variable x(t) is equal to the expected value of x(t), as defined as

$$E[x(t)] = \int_{-\infty}^{\infty} x(t)\, dt \tag{5}$$

The mean square value of x(t) is $E[x^2(t)] = \int_{-\infty}^{\infty} x^2(t)\, dt$

Correlation function is a measure of the similarity between two random quantities in a time domain τ .For a single record x(t), the autocorrelation R(τ) of x(t) is the expected value of the product x(t) x(t+τ):

$$R(\tau) = E[x(t)x(t+\tau)] = \int_{-\infty}^{\infty} x(t)x(t+\tau)\, dt \tag{6}$$

When τ = 0, the equation (6) definition reduces to the mean square value.

$$R(0) = E[x^2(t)]$$

For two random quantities *x(t)* and *y(t)*, the cross correlation function is defined as

$$R_{xx}(\tau) = \int_{-\infty}^{\infty} x(t)x(t+\tau)dt = E[x(t)x(t+\tau)] \tag{7}$$

The autocorrelation and PSD functions are related by the Fourier Transform pair

$$S_{xx}(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} R_{xx}(\tau)e^{-j\omega\tau} d\tau \tag{8}$$

$$R_{xx}(\tau) = \int_{-\infty}^{\infty} S_{xx}(\omega)e^{j\omega\tau} d\omega \tag{9}$$

As S_{xx} is a real even valued function

$$R_{xx}(\tau) = \int_{-\infty}^{\infty} S_{xx}(\omega)\cos\omega\tau\, d\omega \tag{10}$$

By differentiating $R_{xx}(\tau)$ several times with respect to τ

$$R'_{xx}(\tau) = R_{x\dot{x}}(\tau) = -\int_{-\infty}^{\infty} \omega S_{xx}(f)\sin\omega\tau\, df \tag{11}$$

$$R''_{xx}(\tau) = -R_{\dot{x}\dot{x}}(\tau) = -\int_{-\infty}^{\infty} \omega^2 S_{xx}(f)\cos\omega\tau\, df \tag{12}$$

$$R'''_{xx}(\tau) = -R_{\dot{x}\ddot{x}}(\tau) = \int_{-\infty}^{\infty} \omega^3 S_{xx}(f)\sin\omega\tau\, df \tag{13}$$

$$R''''_{xx}(\tau) = R_{\ddot{x}\ddot{x}}(\tau) = \int_{-\infty}^{\infty} \omega^4 S_{xx}(f)\cos\omega\tau\, df \tag{14}$$

The moments therefore, define how each of the processes x, x', x'', etc are elated to the other processes when τ =0,

$$\mu_n = \frac{d^n}{d\tau^n} R_{xx}(0) = \frac{d^n}{dt^n} R_{xx}(0) = \int_{-\infty}^{\infty} \omega^n S_{xx}(f)\, df \tag{15}$$

or in terms of the one sided PSD *G(f)*

$$\mu_n = \int_0^{\infty} (2\pi f)^n\, 2S_{xx}(f)\, df = (2\pi)^n \int_0^{\infty} f^n G_{xx}(f)\, df = m_n (2\pi)^n \tag{16}$$

where, $m_n = \int_0^{\infty} f^n G_{xx}(f)\, df$

A method for computing these moments is shown in Figure 1.

It is important to note that μ_1 and μ_3 are zero, but m_1 and m_3 are not. Remember that μ_n is produced by integrating from -∞ and +∞, and m_n is produced by integrating from 0 and +∞. Typically, we calculate m_0, m_1, m_2, and m_4.
The most common spectral moment is μ_0, which determine the variance of a PSD

$$\mu_0 = \sigma_x^2 = \int_{-\infty}^{\infty} S_{xx}(\omega)\, d\omega = 2\int_{0}^{\infty} S_{xx}(\omega)\, d\omega = \int_{0}^{\infty} G(f)\, df = m_0 \tag{17}$$

In this case μ_0 and m_0 are equal. The root mean square (rms) value of the zero mean process is given by $\sqrt{m_0}$.

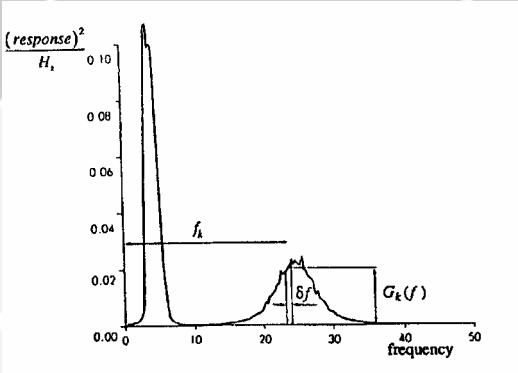


Figure 1. Calculating moments from a PSD

A more complicated example of the use of these moments considers the number of zero crossings in a stationary random and Gaussian (normal) process. Consider the 2D probability function $p(\alpha, \beta)$ of x and $\dot{x}$

$$p(\alpha, \beta)\, \Delta\alpha\, \Delta\beta \approx \text{Prob}[\alpha \le x(t) \le \alpha + \Delta\alpha \text{ and } \beta \le \dot{x}(t) \le \beta + \Delta\beta] \tag{18}$$

This probability represents the fraction of time that x is between α and $\Delta\alpha$, when the velocity $\dot{x}$
is between $\beta+\Delta\beta$. If we define the time to cross one interval as Δt.

$$\Delta t = \frac{\Delta\alpha}{|\beta|} \tag{19}$$

From which we can obtain the expected total number of positive crossings of level

$$\frac{p(\alpha,\beta)\Delta\alpha\Delta\beta}{\Delta t} \approx |\beta|\, p(\alpha,\beta)\Delta\beta \tag{20}$$

As $\Delta\beta \to 0$, the total expected number of passages per unit time through $x(t) = \alpha$ for all possible values of β is given by

$$E[\alpha] = \int_{0}^{\infty} |\beta|\, p(\alpha,\beta)\, d\beta \tag{21}$$

By setting $\alpha = 0$, we get the required number of zero crossings per unit time

$$E[0] = \int_0^{\infty} |\beta| p(0,\beta)\, d\beta \tag{22}$$

The 2D normal density function of x and $\tilde{x}$ is given by

$$p(\alpha,\beta) = (2\pi)^{-1} |A|^{-0.5} e^{\left[\frac{-1}{2|A|}\left(A_{11}\alpha^2 + 2A_{12}\alpha\beta + A_{22}\beta^2\right)\right]} \tag{23}$$

where, $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ and $a_{ij} = E\left[x_i x_j\right] = a_{ji}$ (24)

The a_{ij} terms are the covariances or second moments of x_i and x_j. The a_{ii} terms are the variances of x_i and x_j. $|A|$ is the determinant of A and A_{ij} is the cofactor of a_{ij}

With a little effort, we get, $a_{11} = R(0) = \mu_0$; $a_{12} = a_{21} = \mu_1 = 0$; $a_{22} = \mu_2$ (25)

Therefore, $A = \begin{bmatrix} \mu_0 & 0 \\ 0 & \mu_2 \end{bmatrix}$

From which we get, $E[\alpha] = \frac{1}{2\pi}\left[\frac{\mu_2}{\mu_0}\right]^{\frac{1}{2}} e^{\frac{\alpha^2}{2\mu_0}}$ (26)

If we set α = 0, then $E[0] = \left[\frac{m_2}{m_0}\right]^{\frac{1}{2}}$

In a similar way, we can derive results for the number of peaks per unit time is

$$E[P] = \left[\frac{m_4}{m_2}\right]^{\frac{1}{2}}$$

The irregularity factor is $\gamma = \frac{E[0]}{E[P]} = \left[\frac{m_2}{\sqrt{m_0 m_4}}\right]$

The irregularity factor γ is an important parameter that can be used to evaluate how concentrated near a central frequency the process is. So it can be used to determinate whether or not the process is narrow band or wide band. A narrow band process ($\gamma \to 1$) is characterized by only one predominant central frequency meaning that the number of peaks per second is very similar to the number of zero crossings of the signal. This assumption leads to the fact that the pdf of the fatigue cycles range is the same as the pdf of the peaks in the signal (Bendat theory). In this case fatigue life is easy to estimate. In contrast, the same property is not true for wide bend process ($\gamma \to 0$). Figure 2(a) shows different type of time histories and its corresponding PSD function.

4 Probability Density Functions (pdf's)

The most convenient way, mathematically, of storing stress range histogram information is in the form of a probability density function (pdf) of stress ranges [11]-[12]. A typical representation of this function is shown in Figure 2(b). It is very

easy to transform from a stress range histogram to a pdf, or back. The bin widths used, and the total number of cycles recorded in the histogram are the only additional pieces of information required. To get a pdf from a rainflow histogram each bin in the rainflow count has to be multiplied by $S_t \times dS$. where S_t is the total number of cycles in histogram; dS is the interval width.

The probability of the stress range occurring between $S_i - dS/2$ and $S_i + dS/2$ is given by $p(S_i)dS$.

The actual counted number of cycles, $n_i = p(s)\,dS\,S_t$

The allowable number of cycles, $N(S_i) = \frac{k}{S^b}$

Damage, $E[D] = \sum_i \frac{n_i}{N(S_i)} = \frac{S_t}{k} \int S^b p(s) dS$ (27)

Failure occurs, $D \geq 1.0$.

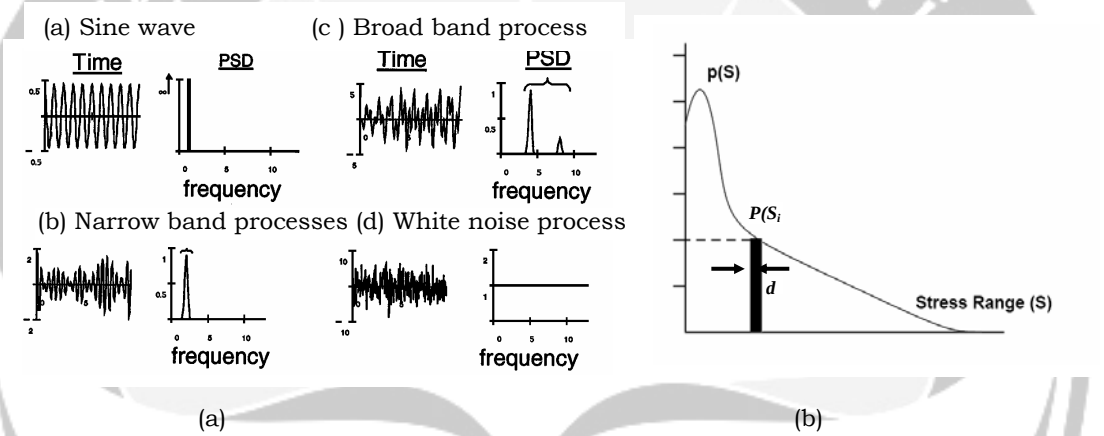


Figure 2. (a) Equivalent time histories and PSDs; (b) Probability density functions.

In order to compute fatigue damage over the lifetime of the structure in seconds the form of materials S-N data must also be defined using the parameters k and b. In addition, the total number of cycles in time T must be determined from the number of peaks per seconds $E[P]$. If the damage caused in time T is greater than 1.0 then the structure is assumed to have failed or alternatively the fatigue life can be obtained by setting $E[D]$ =1.0 and then finding the fatigue life T in seconds from the fatigue damage is given by equation (27) .

5 Narrow Band Solution

Bendat [11] presented the theoretical basis for the first of these of these frequency domain fatigue models, so called Narrow band solution. This expression was defined solely in terms of the spectral moments up to m_4. However, the fact that this solution was suitable only for a specific class of response conditions was an unhelpful limitations for the practical engineer. The narrow band formula [9],[12] is given by the equation (28).

$$E[D] = \sum_i \frac{n_i}{N(S_i)} = \frac{S_t}{K} \int S^b p(S) dS \; N(S) = \frac{E[P]T}{K} \int S^b [\frac{S}{4.m_0} e^{\frac{s^2}{8m_0}}] \, dS = E[P]T \left\{ \frac{S}{4m_0} e^{\frac{s^2}{8m_0}} \right\} \quad (28)$$

This is the first frequency domain method for predicting fatigue damage from PSDs and it assumes that the pdf of peaks is equal to the pdf of stress amplitudes. The narrow band solution was then obtained by substitutions the Rayleigh pdf of peaks with the pdf of stress ranges. The full equation is obtained by noting that S_t is equal to $E[P].T$, where T is the life of the structure in seconds. The basis of the narrow band solution is shown in Figure 3.

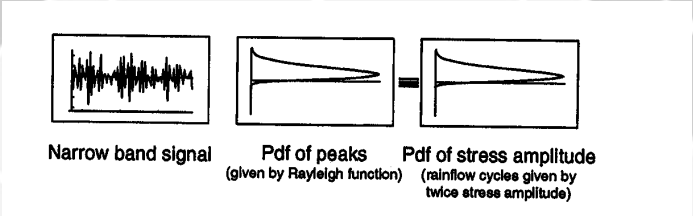
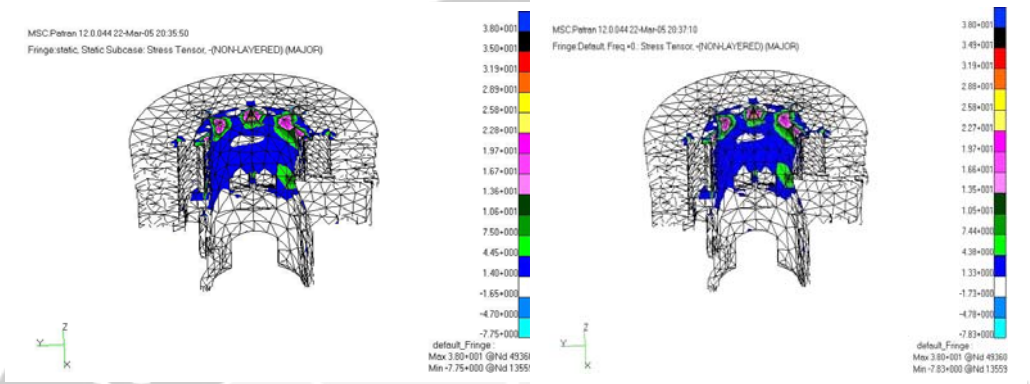


Figure 3. The Basis of the narrow band solution

6 Application of Linear Generator Engine Component

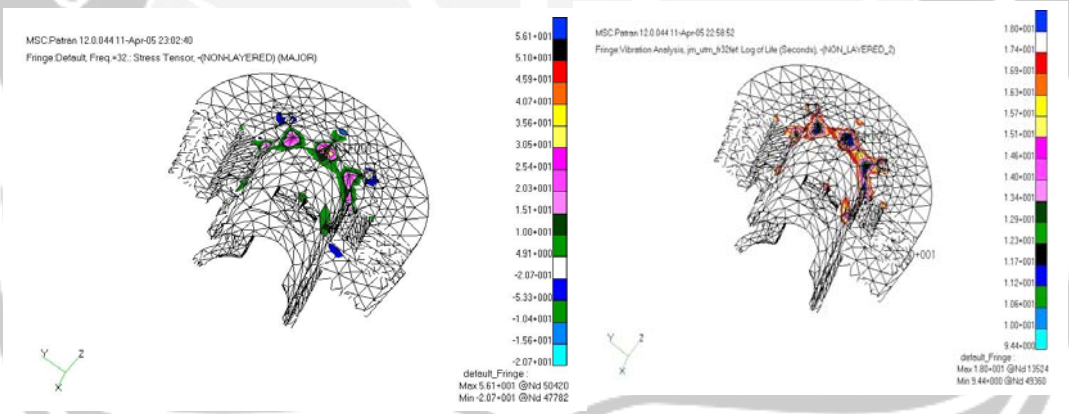
A geometric model of the cylinder block of the free piston engine is considered in this study. First model is imported to finite element software and has created fine mesh using Tetra10 elements. The Pseudo-static and frequency response analyses are performed using MSC.NASTRAN finite element software. The frequency response analysis used a damping ratio of 5% of critical. The results of Pseudo-static and frequency response finite element analysis at zero Hz i.e. the maximum principal stresses distribution of cylinder block are presented in Figures 4(a) and 4(b) respectively. These two are almost identical. When plot higher frequencies, it will be seen a small divergence from the static cases. This is due to dynamic influences of the first mode shape. The maximum principal stresses of the cylinder block for 32 Hz is presented in Figures 5(a). From the results, maximum principal stresses of 56.1 MPa were obtained at node 50420 for 32 Hz. The fatigue life contour result for the most critical locations for 32 Hz is shown in Figures 5(b) using the SAETRN loading histories [12]. The minimum life prediction is $10^{9.44}$ seconds for 32 Hz. Table 1 shows that the comparison between Pseudo-static and vibration fatigue analysis using narrow band frequency response method for different loading conditions and material AA6061-T6 is considered in this comparison. This can be seen that the good agreement between the two approaches. The full set of comparison results for untreated polished cylinder block at critical location (node 49360) is given in Table 2 at different loading conditions. Narrow band solution is considered in this study. It can be seen that from the

Table 2 SAESUS loading condition has been found to give the highest lives for all materials while ASTM A-G loading conditions is given the lowest lives for all materials.



(a) (b)

Figure 4. Maximum principal stresses distribution (a) linear static analysis; (b) frequency response analysis



(a) (b)

Figure 5. (a) Maximum Principal stresses contour for 32 Hz; (b) Vibration fatigue life in log contour plotted for 32 Hz

There are several types of loading histories were selected for the simulation from the SAE and ASTM profiles. Raw time loading histories are shown in Figure 6 and the corresponding PSD plot are also shown in Figure 7. The SAETRN, SAESUS, and SAEBKT in the figure mean the SAE's load-time history obtained from the transmission, suspension, and bracket respectively. I-N, A-A, A-G, R-C, and TRANSP are the ASTM instrumentation & navigation typical fighter, ASTM air-to-air typical fighter, ASTM air to ground typical fighter, ASTM composite mission typical fighter, and ASTM composite mission typical transport loading history, respectively [12].

Table 1 Predicted life in seconds between two approaches at critical location.

Loading Conditions	Pseudo-static	Vibration
SAETRN	1.14E8	2.10E7
SAESUS	6.34E9	8.74E10
SAEBKT	7.56E7	4.06E8
ASTM I-N	3.02E9	2.30E8
ASTM A-A	5.39E8	3.93E7
ASTM A-G	2.72E9	8.23E6
ASTM R-C	1.27E6	6.02E7
ASTM TRANSP	1.15E7	2.27E9

TABLE 2 Predicted life in seconds at weakest location (at node 49360)

Loading conditions	Predicted Life in seconds at critical location (node 49360)							
	2014-T6	2024-T86	2219-87	5083-87	5454-CF	6061-T6	7075-T6	7175-T73
SAETRN	4.27E7	1.25E9	8.48E8	3.56E7	1.30E7	2.10E7	1.08E10	2.33E9
SAESUS	5.01E10	3.78E12	3.36E11	2.51E11	9.24E10	8.74E10	9.51E12	1.10E15
SAEBKT	4.96E8	1.53E10	9.07E9	7.78E8	3.18E8	4.06E8	1.59E11	1.55E11
ASTM I-N	2.64E8	8.02E9	4.66E9	4.52E8	1.86E8	2.30E8	8.52E10	1.13E11
ASTM A-A	5.99E7	1.85E9	1.17E9	7.08E7	2.78E7	3.93E7	1.76E10	7.71E9
ASTM A-G	1.66E7	4.87E8	3.29E8	1.40E7	5.11E6	8.23E6	4.21E9	9.26E8
ASTM R-C	3.19E7	9.66E8	6.29E8	3.19E7	1.21E7	6.02E7	8.78E9	2.65E9
ASTM TRANSP	1.95E9	4.99E10	2.67E10	5.07E9	2.07E9	2.27E9	5.92E11	4.20E12

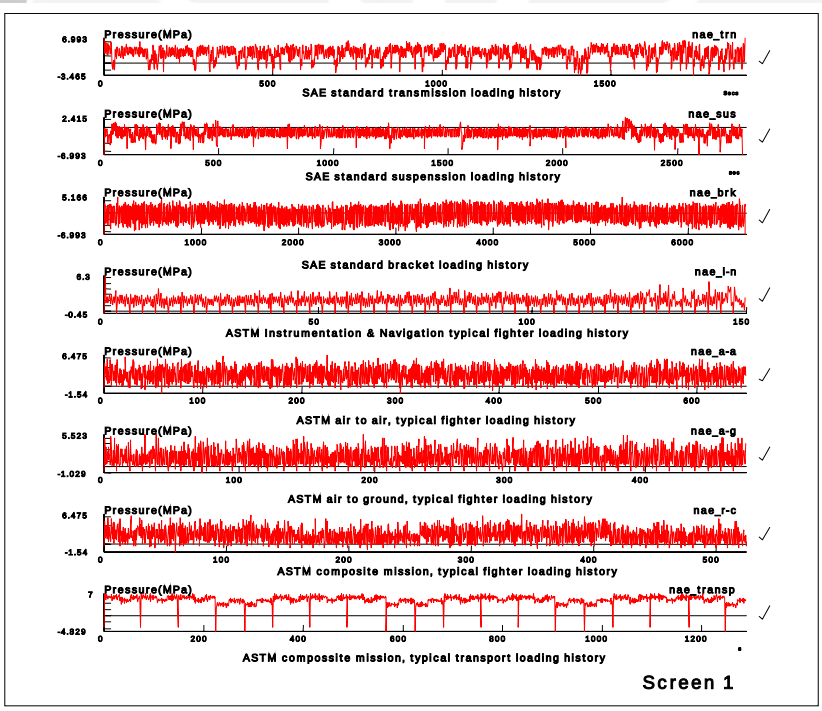

Pressure(MPa)
nae_trn
6.993
-3.465
SAE standard transmission loading history
Pressure(MPa)
nae_sus
2.415
-6.993
SAE standard suspenssion loading history
Pressure(MPa)
nae_brk
5.166
-6.993
SAE standard bracket loading history
Pressure(MPa)
nae_l-n
6.3
-0.45
ASTM Instrumentation & Navigation typical fighter loading history
Pressure(MPa)
nae_a-a
6.475
-1.54
ASTM air to air, typical fighter loading history
Pressure(MPa)
nae_a-g
5.523
-1.029
ASTM air to ground, typical fighter loading history
Pressure(MPa)
nae_r-c
6.475
-1.54
ASTM composite mission, typical fighter loading history
Pressure(MPa)
nae_transp
7
-4.829
ASTM compossite mission, typical transport loading history
Screen 1

Figure 6. Different Time loading histories

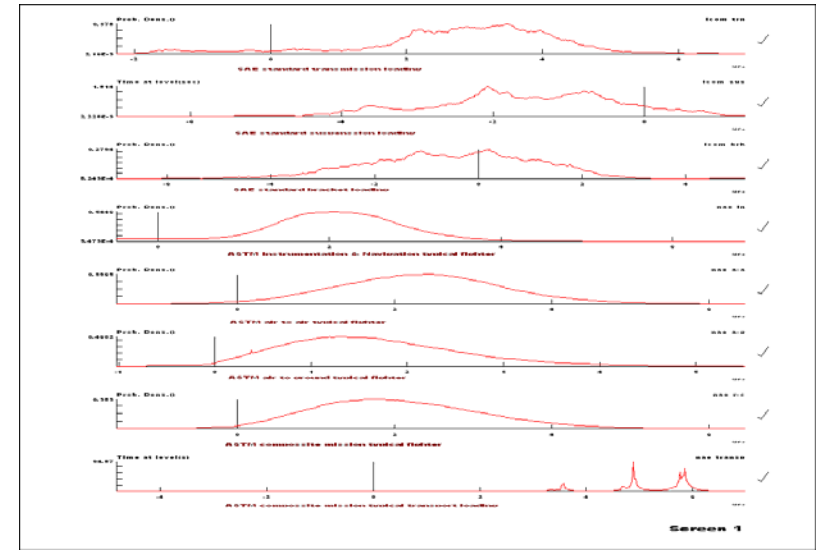


Figure 7. Power Spectral Densities (PSD) responses

Now investigated the effect of surface treatments upon the fatigue life of component subjected to variable amplitude loading conditions. The material used in this study is AA6061-T6. Polished surface finish condition are also used in this study. The contributions of surface treatments on the fatigue lives at different loading conditions using narrow band frequency response method at critical location are summarized in Table 3. Surface treatments (nitriding, cold rolled, shot peening) that produce compressive residual surface stresses are useful. These treatments cause the maximum tensile stress to occur below the surface. The reverse is also true and tensile residual surface stresses are very detrimental and promotes corrosion fatigue. Surface treatments are also increase the endurance limit. Diffusion process such as nitriding is very beneficial for fatigue strength. This process has the combined effect of producing a higher strength material on the surface as well as causing volumetric changes which produce residual compressive surface stresses. There are several methods used to cold work the surface of a component to produce a residual compressive stress. The two most important are cold rolled and shot peening. Along with producing compressive residual stresses, these methods also work-harden the surface material. The great improvement in fatigue life is due primarily to the residual compressive stresses. In shot peening process, the surface of the component undergoes plastic deformation due to the hit of many hard shots. The fatigue life of the component is improved due to the development of compressive residual stresses and the increase of hardness near the surface. The effect of surface treatment at different loading conditions for polished vibrating cylinder block is summarized in Table 3. It can be seen that the fatigue life for nitriding surface treatments is surprisingly increases than those other surface treatment processes. Figure 8 shows that the effect of different surface treatment processes for ASTM A-G loading conditions and polished

specimen. It is clearly shown that nitrided processes is surprisingly increases the fatigue life at critical location than other processes.

Table 3 Effect of surface treatments at different loading conditions for polished components

Loading Conditions	Predicted life in seconds for different surface treatment processes			
	Nitrided	Cold Rolled	Shot Peened	Untreated
SAETRN	4.52E10	8.81E8	6.40E7	2.10E7
SAESUS	3.41E16	1.21E14	8.31E11	8.74E10
SAEBKT	8.08E12	5.31E10	1.68E9	4.06E8
ASTM I-N	6.35E12	3.68E10	1.01E9	2.30E8
ASTM A-A	2.77E11	2.92E9	1.40E8	3.93E7
ASTM A-G	1.84E10	3.51E8	2.52E7	8.23E6
ASTM R-C	7.21E10	1.02E9	6.01E7	1.83E7
ASTM TRANSP	2.40E14	9.55E11	1.33E10	2.27E9

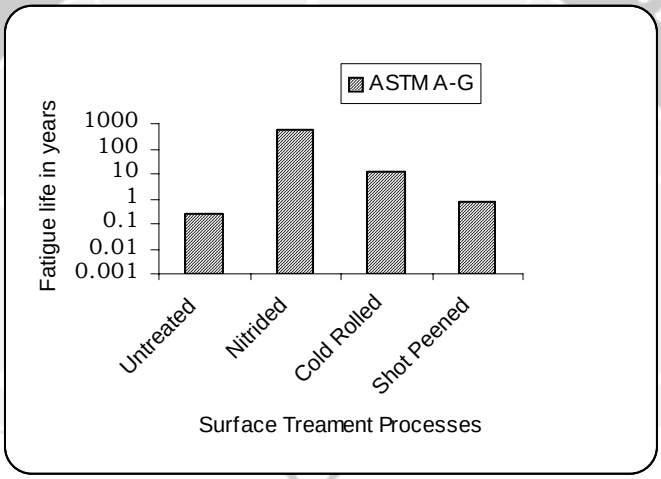


Figure 8. Effect of different surface treatment processes for polished and ASTM A-G loading conditions.

7 Conclusions

The concept of spectral moments has been presented. A state of art of vibration fatigue techniques has been presented where the random loading and response are categorized using PSD functions. Narrow band frequency domain fatigue analysis has been applied to a typical cylinder block of new two-stroke free piston engine. From the results, it can be concluded that compressive mean stress loading conditions has been found to give the highest lives for all materials. According to the results, all surface treatment processes can be applied to increase the fatigue life of the aluminum alloys component. The surface residual compressive stress has the greatest effect on the fatigue life. It can also be concluded that the polished and nitriding combinations have been found the highest lives of the cylinder block. Surface treatment to produce compressive forces in the outer layers of the

component which will be cyclically loaded at stress raising locations. In addition, the vibration fatigue analysis can improve understanding of the system behaviors in terms of frequency characteristics of both structures and loads and their couplings.

Acknowledgments

The authors are grateful to Malaysia Government especially Ministry of Science, Technology and Environment under IRPA project (IRPA project no: 03-02-02-0056 PR0025/04-03) for providing financial support.

References

[1] Benedetti, M., Fortanari, V., Hohn, B.R., Oster, P. & Tobie, T. (2002), Influence of shot peening on bending tooth fatigue limit of case hardened gears, *International Journal of Fatigue*, **24,** 1127-1136.

[2] Hirsch, T., Wohlfahrt, H. & Macherauch, E. (1987), Fatigue Strength of case Hardened shot peening gears. *Proc. of 3rd International Conference on Shot Peening (ICSP-3),* Garmish-Partenkichen, 547-560.

[3] Inoue, K., Maehara, T. & Yamanaka, M. (1989), The effect of Shot peening on the bending strength of carburized gear teeth, *JSME Inter. Journal Series III,* **32**(3), 448-454.

[4] Kobayashi, M., Matsui, T. & Murakami, Y. (1998), Mechanism of creation of compressive residual stress by shot peening, *International Journal of Fatigue,* **20**(5), 351-357.

[5] Torres, M.A.S. & Voorwald, H.J.C. (2002), An evaluation of shot peening, residual stress and stress relaxation on the fatigue life of AISI 4340 steel, *International Journal of Fatigue*, **24**, 877-886.

[6] Rodopoulos, C.A., Curtis, S.A., Rios, E.R. de Los & SolisRomero, J. (2004), Optimisation of the fatigue resistance of 2024-T351 aluminum alloys by controlled shot peening – methodology, results and analysis, *International Journal of Fatigue*, **26**, 849-856.

[7] Novovic, D., Dewes, R.C., Aspinwall, D.K., Voice, W. & Bowen, P. (2004), The effect of machined topology and integrity on fatigue life, *International Journal of machine Tools & Manufacture*, **44**, 125-134.

[8] Martin, U., Altenberger, I., Scholtes, B., Kremmer, K. & Oettel, H. (1998), Cyclic deformation and near microstructures of normalized shot peened steel SAE 1045, *Materials Science and Engineering*, **A246**, 69-80.

[9] Bishop, N.W.M. & Sherratt, F. (2000), *Finite element based fatigue calculations*, NAFEMS Ltd. UK.

[10] MSC/FATIGUE user's guide, Vol. 1& 2. (2004), MSC/Corporation. CANADA.

[11] Bendat, J.S. (1964), Probability Functions for Random Responses, *NASA report on Contract NASA-5-4590.*

[12] Rahman, M.M., Ariffin, A. K., Jamaludin, N. & Haron, C.H.C. (2005), Analytical and finite element based fatigue life assessment of vibration induced fatigue damage, *The second International Conference on Research and Education in Mathematics (ICREM 2)*, Residence Hotel, Putrajaya, Malaysia, 331-345.

[13] Crandell, S.H. & Mark, W.D. (1973), *Random vibration in mechanical systems,* Academic Press, New York.

[14] Newland, D.E. (1993), *An Introduction to random vibrations, spectral and wavelet analysis*, Longman Scientific and Technical, Essex, UK.
[15] Wirsching, P.H., Paez, T.L. & Oritz, K. (1995), *Random vibration, theory and practice.* John Wiley and Sons, Inc. USA.

M. M. Rahman: Computational and Experimental Mechanics Group, Department of Mechanical and Materials Engineering, Universiti Kebangsaan Malaysia (UKM), 43600, Bangi, Selangor DE, Malaysia. Phone:+ (6)03-89216012;
Fax: + (6)-03-89216040
E-mail: mustafiz@eng.ukm.my

A. K. Ariffin: Computational and Experimental Mechanics Group, Department of Mechanical and Materials Engineering, Universiti Kebangsaan Malaysia (UKM), 43600, Bangi, Selangor DE, Malaysia. Phone:+ (6)03-89216012;
Fax: + (6)-03-89216040
E-mail: kamal@eng.ukm.my

Tulus: Computational and Experimental Mechanics Group, Department of Mechanical and Materials Engineering, Universiti Kebangsaan Malaysia (UKM), 43600, Bangi, Selangor DE, Malaysia. Phone:+ (6)03-89216012;
Fax: + (6)-03-89216040

An Algorithm For Timetable Scheduling

Muhammad Rafiullah Arain

NED University of Engineering & Technology Karachi – Pakistan.

Abstract: It is a fact that the timetable designing is a difficult process in educational institutions and it is too hard and so complex where there is the faculty is not working at the basis of fixed timings, courses are offered at different times periods in different days. I have developed an algorithm[i] to design timetable using 6 variables as a(s, c, r, t, d). The proposition a(s, c, r, t, d) stats that Teacher a, teaches Subject s in class c in room r at time t on the day d.
This algorithm has three phases.

1. Build dynamic structure (matrix/table) for a class and mark Title and Captions.
2. Select a perfect available teacher for the particular subject according to teacher's expertise in the subject and according to his/her available time.
3. The selected data arrange into the matrix according to time priority of teachers and availability of room.

Key-words: Timetable designing, timetable scheduling, time management, an algorithm for timetable, optimization, event management.

Introduction:

Since the beginning of human civilization, time management and scheduling have always been an integral part of our society. In order to interact with people in a society, one has to meet at the same time and space. As the numbers of individuals and society becomes larger, in order to achieve some common goal, it becomes very important that individuals follow fixed and well define path (pattern) for maximum benefit. In educational institutions the timetable has a great importance and still done by using the manual process. The timetable designing is time consuming process and it takes effort. Due to clashes (limitations, constraints) the whole system is disturbed. It requires a lot of changes to obtain the desired timetable.
Now the best way for arranging timetable is interactive computer programming that helps to manage by using a well-defined method.

Problem Definition:

1. Dynamic structure of Timetable with 6 variable constraints, which are teacher, subject, class, time, day and room.
2. Select best teacher for the subject.
3. More than n subject can not assign to any teacher.
4. More than m periods (lecturers) can not assign to any teacher.
5. Teacher can not give more than p lectures consecutively.
6. Teacher can not give more than q lectures in a day.
7. More than 1 subject can not assign to any class at the same time.
8. Room must be sufficient for the number of students.
9. Room assign for lectures and Lab assign for practical.
10. Only one class conducted in a room at the same time.

Procedure:

In manual work we perform our task according to set priorities and following rule and regulation. There are different ways to perform a task. Timetable designing in educational institutes is very difficult process, when the faculty members are working as a visiting faculty members, when different courses are running at the same time, when there are different shifts are going on in different time frames. So it is very hard to manage manually because there are many clashes, restrictions and limitations (constraints). I have considered many aspect of timetable scheduling. The development of algorithm is based on: First of all the organization of data. Secondly set the priorities and expertise. And in last some rules and steps are defined to complete the task.

Algorithm has three phases.

Figure No. 1

1. Build dynamic structure (matrix) for a class and mark Title and Captions.
2. Select a perfect available teacher for the specific subject according to teacher's expertise in the subject.
4. Arrange into the matrix according to time priority of teachers and availability of room.

By using the logical equivalencies and mathematical expression algorithms, extracts required data and put it into a new matrix [table] for further use and update others tables.

Algorithm:

1. Variables: define here
2. Select a class.
3. Read: days, number of periods on a day, duration of the period, shift start & end timing, numbers of breaks.
4. tab = days x no_of _periods. Construct a matrix timetable = day x pod and assign captions/titles.
5. Read: subject in a class and credit hours.
6. tch = Σ credit_hours_of_each_subject
7. If tch > tab than stop [matrix timetable is not comfortable for timetable] and go to step 2 to select another class.
8. Read: Shift timing, available teachers' timings and teacher expertise in the subject. Select teachers for the class.
9. Assign subjects to respective teachers according to their expertise in subject,
    a. not more than 4 subjects for a teacher, and
    b. not more than 1 subject in a class.
        i. If yes than go to Step 8 to select another teacher.

Construct a matrix available_teacher_for_class for selected the class. Count time hours and sort it in ascending order by time field.

10. Select a valid and avail room for class or Lab.
11. Arrange subjects in matrix timetable according to available teacher, giving the preferences to the low time.
    a. Does not a teacher teach more than 2 classes consecutively and
    b. Does not conduct more than 3 classes in a day?
        i. If yes change the day. Repeat step 10

    And update the matrix teachers_timing and matrix Room.
12. Repeat step 10 until the final form of timetable (Matrix timetable is filled according to credit hours).
13. If the numbers of classes are completed then stop else go to step 2.

Sequence flow [flowchart]:

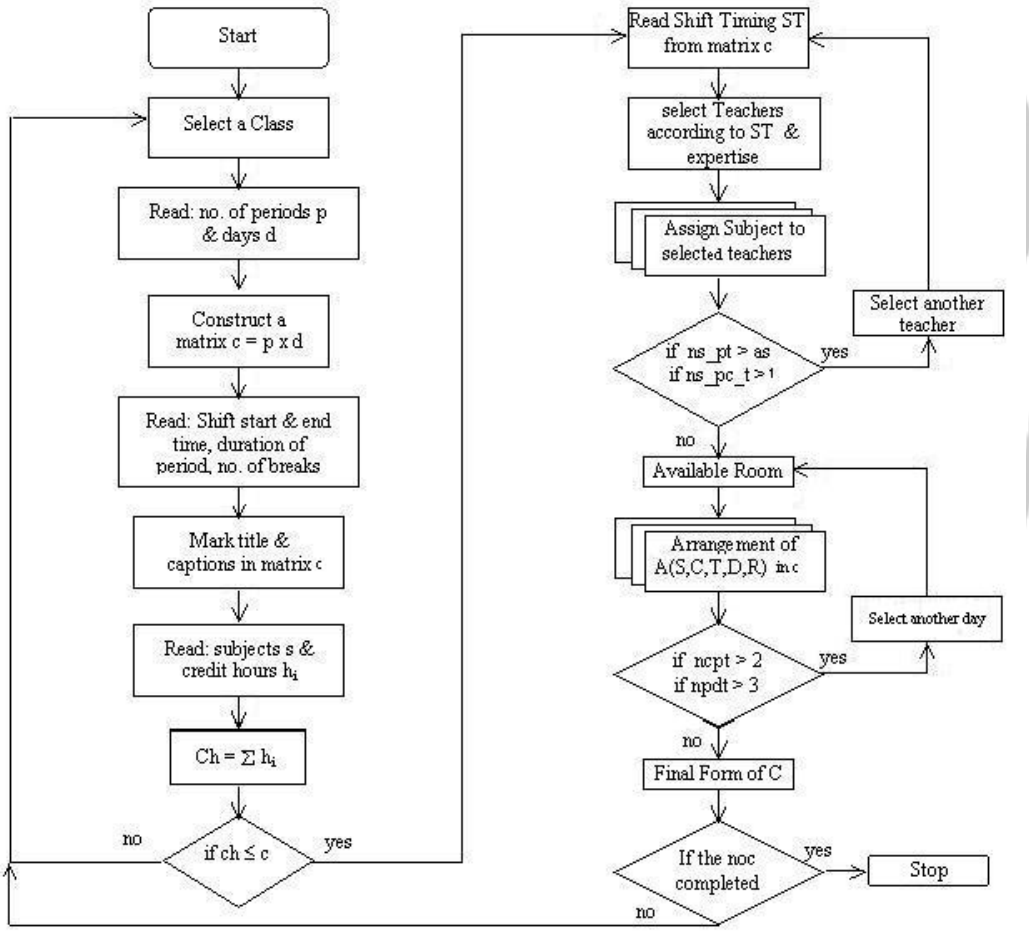


Figure No. 2

Data Structure:

Number of periods and time duration in a day, shift time and number of students.

ID	Course	No. of periods	No. of days	Duration	Shift-start	Shift-End	Number of Student
	BCIT-1	5	5	45	9	12:30	50
	BCIT-2	5	5	45	9	12:30	40
	BCIT-3	5	5	45	9	12:30	25
	BCIT-4	5	5	45	1	5	30
	BBA-1	5	5	45	9	12:30	60
	BBA-2	5	5	45	9	12:30	40
	MBA-1	4	5	45	5	8:30	20
	MBA-2	4	5	45	5	8:30	25
	MBAE-3	4	3	90	1	5	20
	MBAE-4	4	3	90	1	5	15

Table No. 1

Mathematical Presentation of Table no. 1.

Matrix A a x b
A a, 1 = ID
A a, 2 = Class Name
A a, 3 = Numbers of Periods in a Day
A a, 4 = Number of Days
A a, 5 = Period Duration
A a, 6 = Shift Start
A a, 7 = Shift End
A a, 8 = Number of Student in a Class

Days for class

ID	Course	Days
	BCIT-1	Mod
	BCIT-1	Tue
	BCIT-1	Wed
	BCIT-1	Thu
	BCIT-1	Fri
	BCIT-2	Mod
	BCIT-2	Tue
	BCIT-2	Wed
	BCIT-2	Thu

Table No. 2

Subjects of courses

ID	Course	Subject	Subject Type	Credit Hour
	BCIT-1	English-I	0	3
	BCIT-1	C language	0	3
	BCIT-1	Electronics	0	3
	BCIT-1	Calculus	0	3
	BCIT-1	Lab1	1	2
	BCIT-1	Islamyat	0	2
	BCIT-2	English-II	0	3
	BCIT-2	OOP	0	3
	BCIT-2	Oracle	0	3
	BCIT-2	Diff: Equ:	0	3
	BCIT-2	Pak Studies	0	2
	BCIT-3	JAVA	0	3
	BCIT-3	Math I	0	3
	BCIT-3	Probability	0	3
	BCIT-3	Digital Logic	0	3
	BCIT-3	Visual Basic	0	3

Table No. 3

Subject Type:
0 for Lecture/Room.
1 for Lab

Teachers' Name and Expertise in Subject

ID	Teacher Name	Subject	Expertise
	Farooq Ahmed	Oracle	1
	Fazal Imran	Multimedia	1
	Ahmed Masaud	Electronics	1
	Samuel Dass	Economics	1
	Samreen Fatima	Calculus	1
	Farzana	Probability	2
	Adnan	Math I	1

Table No. 4

Teacher timing and schedule TT_1

Time Slot	Mod	Tue	Wed	Thu	Fri	Sat
1: 8-9	1	1	1	1	1	0
2: 9-10	1	1	1	1	1	0
3: 10-11	1	1	1	1	1	0
4: 11-12	1	1	1	1	1	0
5: 12-13	1	1	1	1	0	0
6: 13-14	1	1	1	1	0	0
7: 14-15	1	1	1	1	0	0

Table No. 5

Mathematical Presentation of Table no. 5.

Matrix TTi x j x k
i = Number of Teachers, j = Time Slat, k = Day
if i = 1,
TT1, 1, 1 = 0, it means "Farooq Ahmed is not available on the time slot 8-9 on Monday"
TT1, 1, 1 = 1, it means "Farooq Ahmed is available on the time slot 8-9 on Monday"
TT1, 1, 1 = 2, it means "Farooq Ahmed is busy on the time slot 8-9 on Monday"

Room number 1 vacant or not

Time Slot	Mod	Tue	Wed	Thu	Fri	Sat
1: 8-9	1	1	1	1	1	0
2: 9-10	1	1	1	1	1	0
3: 10-11	1	1	1	1	1	0
4: 11-12	1	1	1	1	1	0
5: 12-13	0	0	0	0	0	0
6: 13-14	0	0	0	0	0	0
7: 14-15	0	0	0	0	0	0

Table No. 6

Mathematical Presentation of Table no. 5.

Matrix R u x v x w
u = room number, v = time slot, w = day
R 1, 1, 1 = 1, it means "Room number is free at time 1 on Monday"
R 1, 1, 1 = 0, it means "Room number is not free at time 1 on Monday"

Final timetable for a Selected Class

Time Slot	Mod	Tue	Wed	Thu	Fri	Sat
1: 8-9	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)
2: 9-10	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)
3: 10-11	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)
4: 11-12	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)
5: 12-13	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)
6: 13-14	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)
7: 14-15	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)	a(s, c, r, t, d)

Table No. 7

A(S,C,R,T,D)
A = Teacher, S = Subject, C = Class
T = Time, D = Day, R = Room

The expression A(S,C,T,D,R) state that the Teacher A teaches Subject S in class C at Time T in Room R.

Result:

The purpose of algorithm is to provide the best decision to the management. The algorithm was implemented in a computer program at SiSTech (institute) to design timetable of the eight (8) running batches BCIT, BBA, MBA in three (3)

different days. The data was fed into the database in the form of total number of classes, available classrooms, subjects, credit hours, shifts' time, available teachers and their expertise in subject. The computer programs automatically assigned subjects to respective teachers according to their expertise and resolve the scheduling problems in very short time and produced satisfactory timetable.

About Author:

Muhammad Rafiullah Arain: Department of Mathematics & Basic Sciences, NED University of Engineering & Technology, Karachi – Pakistan.
Contact No. Cell: +92 300 7090082, Office: +92 21 9243261-8 Ext 2209.
E-mail: rafiullaharain@yahoo.com

[i] The idea of this algorithm has been presented as a Poster in World Conference on "21 Century Mathematics 2005" arranged by SMS, GCU Lahore.

AN ANALYTICAL STUDY OF NATURAL CONVECTION IN THE INNER BOUNDARY-LAYER SUBJECT TO OSCILLATING TEMPERATURE

R. Roslan[a], I. Hashim[b], K. Ghazali[c]

[a,c] Universiti Malaysia Sabah, Kota Kinabalu, Malaysia
[b] Universiti Kebangsaan Malaysia, Bangi, Malaysia

Abstract. An analytical study of an oscillatory natural convection about an infinite horizontal circular cylinder in an unbounded region of a Newtonian fluid was considered for Prandtl number to be unity. The temperature of the cylinder is oscillating with frequency ω about a mean temperature T_∞, the temperature of the ambient fluid. By using the method of matched asymptotic expansion, the flow field was divided into two regions: an inner region adjacent to the cylinder and an outer region far from the cylinder. However, only the inner region was considered in this study. The separation technique between steady and unsteady in the second-order terms of temperature and stream function that we proposed were found to be directly leading to the desired result, thus simplifying the process of the solution in the inner region.
Key-words: matched asymptotic expansion

1 Introduction

The analysis of heat transfer through a boundary-layer over a body of arbitrary shape and surface temperature constitutes an important problem. However, there is a very limited literature for the natural convection around a horizontal circular cylinder when the temperature of the cylinder is oscillating harmonically. [2] was the first to investigate analytically the problem of fluid and heat flows caused by an oscillating temperature on the surface of the circular cylinder for various Prandtl number P, while [1, 5] for the case $P \neq 1$ and [4] for the case $P = 1$.
It is postulated that, the method of separation of the steady and the unsteady components of flow field presented by [1] from the governing equations can be used without encountering many difficulties. Another approach is to disregard separations on the governing equations as in the work by [2, 4]. Apart from separating the temperature T and the stream function ψ from the governing equations, another expansion approach is by separating the particular terms, i.e. the second-order term in the expansions of T and ψ. The steady and unsteady parts of the expansion, although without T, was proposed by [3] in the problem of oscillating cylinder. For the case $P = 1$, [2, 1] found that the steady temperature and stream function from the second-order term were generated at the outer edge of the inner region. The boundary conditions for the steady temperature and velocity, however, are satisfied only on the surface, not at the outer edge of the inner layer. However, for the case $P = 1$, [4] found that only the condition of the steady temperature agreed previous studies, not for the steady velocity. Recently, [5] made

some corrections to the conclusion of [1] for the insignificant of the second-order approximation in the outer region.

2 Governing equations and boundary conditions

Consider an unsteady natural convection flow past around a circular cylinder of radius a in a stationary fluid at a uniform and constant temperature $T_{\infty}^{'}$. The infinite cylinder is fixed with its axis horizontal, and so the problem is considered two-dimensional. Assume the temperature of the cylinder $T_{c}^{'}$ oscillates harmonically with a frequency ω, such that $T_{c}^{'} = T_{\infty}^{'}(1 + b\cos\omega t')$, where b is the non-dimensional amplitude in the temperature oscillations and t' is the time. The appropriate velocity scale in this problem is, $U_c = (g\beta b T_{\infty}^{'})/\omega$, where g is the magnitude of the acceleration due to gravity and β is the coefficient of the thermal expansion of the fluid. It is further assumed that the Boussinesq approximation is applicable. Now we take cylindrical polar coordinates (r',θ) in which coordinate r' is defined as the distance measured outwards from the origin of the cylinder with the axis of the cylinder at $r' = 0$ and θ is defined to be anticlockwise angle made by the outward normal with the downward vertical from the origin of the cylinder with $\theta = 0$ in the direction of gravity. The non-dimensional governing equations for $P = 1$ and the boundary conditions for all time for $0 \leq \theta < 2\pi$, can be written as follows, see [1, 4, 5],

$$\frac{\partial(\nabla^2\psi)}{\partial t} - \frac{\epsilon}{r}\frac{\partial(\nabla^2\psi,\psi)}{\partial(r,\theta)} = \epsilon^2\gamma\nabla^4\psi + \frac{1}{r}\frac{\partial(r\cos\theta,T)}{\partial(r,\theta)}, \tag{1}$$

$$\frac{\partial T}{\partial t} - \frac{\epsilon}{r}\frac{\partial(T,\psi)}{\partial(r,\theta)} = \epsilon^2\gamma\nabla^2 T, \tag{2}$$

$$\begin{aligned} &\psi = 0, \ \ \tfrac{\partial\psi}{\partial r} = 0, \ \ T = \cos t, && \text{at } r = 1, \\ &\psi \to \text{constant}, \ \ \tfrac{\partial\psi}{\partial r} \to 0, \ \ T \to 0, && \text{as } r \to \infty, \end{aligned} \tag{3}$$

where $\nabla^2 = \left(\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial\theta^2}\right)$. In the above equations the non-dimensional variables are $t = \omega t'$ the time, $r = r'/a$ the radial distance, $T = (T_{c}^{'} - T_{\infty}^{'})/bT_{\infty}^{'}$ the temperature, ψ the stream function which is related to the velocity components $(u,v)=(u'/U_c,v'/U_c)$ in the (r,θ) coordinates system by usual manner, namely $u = -\frac{1}{r}\frac{\partial\psi}{\partial\theta}$, $v = \frac{\partial\psi}{\partial r}$, while $\gamma = \nu\omega/U_c^2$ and $\epsilon = \frac{U_c}{a\omega}$, where ν is the coefficient of kinematic viscosity of the fluid. Here, the Prandtl number, $P = \frac{\nu}{K} = 1$ and we can write γ in terms of Reynolds number and Strouhal number, in which $\gamma = \frac{1}{\epsilon R_c} = \frac{S}{R_c}$ where $R_c = U_c a/\nu$ and $\epsilon = 1/S$.

3 The Inner Boundary-layer Equations

Since the temperature is the first that govern the flow, the asymptotic expansion for the stream function ψ and the temperature T in the inner region are assumed

to be of the form,

$$\begin{aligned} T^{(i)}(t,\eta,\theta) &= T_0^{(i)}(t,\eta,\theta) + \epsilon[T_1^{(i)s}(\eta,\theta) + T_1^{(i)u}(t,\eta,\theta)] \\ &+ \epsilon^2 T_2^{(i)}(t,\eta,\theta) + h.o.t, \end{aligned} \tag{4}$$

$$\begin{aligned} \psi^{(i)}(t,\eta,\theta) &= \epsilon\psi_0^{(i)}(t,\eta,\theta) + \epsilon^2[\psi_1^{(i)s}(\eta,\theta) + \psi_1^{(i)u}(t,\eta,\theta)] \\ &+ \epsilon^3\psi_2^{(i)}(t,\eta,\theta) + h.o.t, \end{aligned} \tag{5}$$

where superscripts s and u denote the steady and unsteady parts, respectively and the radial non-dimensional coordinate is stretched as follows:

$$\eta = \frac{r-1}{\epsilon\sqrt{2\gamma}}. \tag{6}$$

Substitution of expansions (5) and (4) into equations (1) and (2) results in the following set of equations in the inner region,

$$\frac{\partial^2 T_0^{(i)}}{\partial\eta^2} = 2\frac{\partial T_0^{(i)}}{\partial t}, \tag{7}$$

$$\frac{\partial^4\psi_0^{(i)}}{\partial\eta^4} - 2\frac{\partial^3\psi_0^{(i)}}{\partial t\partial\eta^2} = -2\sqrt{2\gamma}\sin\theta\frac{\partial T_0^{(i)}}{\partial\eta}, \tag{8}$$

$$\frac{\partial^2 T_1^{(i)u}}{\partial\eta^2} - 2P\frac{\partial T_1^{(i)u}}{\partial t} = \sqrt{\frac{2}{\gamma}}\left[\frac{\partial(\psi_0^{(i)}, T_0^{(i)})}{\partial(\eta,\theta)}\right]^u - \sqrt{2\gamma}\frac{\partial T_0^{(i)}}{\partial\eta}, \tag{9}$$

$$\frac{\partial^2 T_1^{(i)s}}{\partial\eta^2} = \sqrt{\frac{2}{\gamma}}\left[\frac{\partial(\psi_0^{(i)}, T_0^{(i)})}{\partial(\eta,\theta)}\right]^s, \tag{10}$$

$$\begin{aligned} \frac{\partial^4\psi_1^{(i)u}}{\partial\eta^4} &- 2\frac{\partial^3\psi_1^{(i)u}}{\partial t\partial\eta^2} = \sqrt{\frac{2}{\gamma}}\left[\frac{\partial(\psi_0^{(i)}, \partial^2\psi_0^{(i)}/\partial\eta^2)}{\partial(\eta,\theta)}\right]^u \\ &+ 2\sqrt{2\gamma}\left(\frac{\partial^2\psi_0^{(i)}}{\partial t\partial\eta} - \frac{\partial^3\psi_0^{(i)}}{\partial\eta^3} - \sin\theta\frac{\partial T_1^{(i)u}}{\partial\eta}\right) \\ &+ 4\gamma\cos\theta\frac{\partial T_0^{(i)}}{\partial\theta}, \end{aligned} \tag{11}$$

$$\begin{aligned} \frac{\partial^4\psi_1^{(i)s}}{\partial\eta^4} &= \sqrt{\frac{2}{\gamma}}\left[\frac{\partial(\psi_0^{(i)}, \partial^2\psi_0^{(i)}/\partial\eta^2)}{\partial(\eta,\theta)}\right]^s \\ &- 2\sqrt{2\gamma}\sin\theta\frac{\partial T_1^{(i)s}}{\partial\eta}, \end{aligned} \tag{12}$$

where $[\,]^s$ and $[\,]^u$ denote respectively the steady and unsteady parts of the product. Further, this set of equations needs to be solved subject to boundary conditions (3). We observe that equations (7) and (8) produce an unsteady temperature and stream function, equations (9) and (11) produce only an unsteady temperature and stream function, while equations (10) and (12) produce only a steady temperature and stream function and all the solutions depend on parameter γ.

4 Solution in the Inner Layer

Hence, on solving equations (7) and (8) subject to boundary conditions (3) we have the unsteady solutions for the first-order temperature and stream function in the inner region as follows,

$$T_0^{(i)} = e^{-\eta}\cos(t-\eta), \tag{13}$$

$$\begin{aligned}\psi_0^{(i)} &= \frac{\sin\theta\cos t}{2}\sqrt{\frac{\gamma}{2}}\left\{e^{-\eta}\left(\cos\eta+[1+2\eta]\sin\eta\right)\right\}\\ &+ \frac{\sin\theta\sin t}{2}\sqrt{\frac{\gamma}{2}}\left\{e^{-\eta}\left(\sin\eta-[1+2\eta]\cos\eta\right)\right\}\\ &+ \frac{\sin\theta}{2}\sqrt{\frac{\gamma}{2}}\left(\sin t-\cos t\right).\end{aligned} \tag{14}$$

Since our objective was to investigate the steady flow from the oscillating temperature problem, the solution for equations (9) and (11) will be omitted. For the steady second-order temperature and stream function in the inner region we should rewrite, the terms $\cos^2 t$ and $\sin^2 t$ in the forms $\frac{1+\cos 2t}{2}$ and $\frac{1-\cos 2t}{2}$, respectively. Further, all the terms $\cos t, \cos 2t, \sin t$ and $\sin 2t$ will be omitted, leaving only the steady terms. Thus, from equations (10), (13), (14) and the trigonometric rules, the solution for the steady second-order inner temperature $T_1^{(i)s}$ is given by

$$\begin{aligned}T_1^{(i)s} &= \cos\theta\left\{\frac{1}{4}e^{-\eta}\sin\eta+\frac{1}{8}e^{-2\eta}\eta+\frac{1}{4}e^{-2\eta}-\frac{1}{4}\right\}\\ &+ \eta B_1\cos\theta.\end{aligned} \tag{15}$$

As found by [2, 1, 4] the inner region solution $T_1^{(i)s}$ does not tend to zero as $\eta\to\infty$, even if B_1 is assumed to be zero. Hence the boundary condition at infinity cannot be satisfied. From equation (15) we have,

$$\lim_{\eta\to\infty} T_1^{(i)s} = -\frac{1}{4}\cos\theta,$$

From equations (12), (14) and (15) the solution for the steady second-order stream

function in the inner region is given by

$$\psi_1^{(i)s} = -\frac{\sin 2\theta}{384}\sqrt{2\gamma}\left\{e^{-2\eta}\left(33+42\eta+12\eta^2\right)+24\eta e^{-\eta}\cos\eta \right. \\ \left. +C_1\eta^4+C_2\eta^3+C_3\eta^2-33\right\}.$$

Hence, motivated by [1] for $P \neq 1$, if it is assumed that $C_1 = C_2 = C_3 = 0$ when $\eta \to \infty$ then we note that the steady velocity now tends to zero at the outer edge of the inner layer, that is

$$\lim_{\eta\to\infty}\frac{\partial\psi_1^{(i)s}}{\partial\eta} = 0.$$

However, this result contradicts the results of the previous research on oscillatory cylinder and oscillatory natural convection about a circular cylinder as presented by [6] and [1, 2, 4], respectively. Further, this could be proved by applying L'Hopital rule on the steady second-order tangential velocity from the solution by [1, 4], i.e. for the case $P \neq 1$.

5 Conclusions

Motivated by the previous method proposed by [2, 1], we found that the first-order solution of temperature and stream function in the inner region are the unsteady term. While, the second-order temperature and stream function contain the steady and unsteady terms. However, only the steady term from these two flow properties plays the important role for the flow in the outer region. Thus, compared to the conventional method employed by [2, 4], which is no separation technique was applied, the method we proposed proved to be much simpler. Further, compared to the separation technique between steady and unsteady about the temperature and stream function in the governing equations as employed by [1], we found that the separation technique between steady and unsteady second-order terms of temperature and stream function that we proposed is direct to the desire result, that is the steady second-order term in the inner region and neglecting the unsteady second-order term for temperature and stream function in the inner region.
We mentioned that the steady term are much more important compared to the unsteady term for the temperature and stream function. As found by numerous studies in the oscillating flows we found that these terms do not satisfy the boundary conditions when far from the cylinder. The steady temperature and the steady tangential velocity are persisted outside the thin layer and remains to be non-zero at the outer edge of the inner layer. However, these results contradict to that the the problem of natural convection in a Newtonian fluid when $P = 1$. We found that, only the steady temperature does not satisfy the boundary conditions at the outer edge of the inner region. While, this situation does not occur for the steady tangential velocity. This is due to $P = 1$ as the inflection point for the problem of a Newtonian fluid for the case of $P \neq 1$.

References

[1] Chatterjee, A.K. and Debnath, L., (1979), Double Boundary Layers in Oscillatory Convective Flow, *Il Nuovo Cimento*, **52**, 29–44.

[2] Merkin, J.H., (1967), Oscillatory Free Convection From an Infinite Horizontal Cylinder, *J. Fluid Mech.*, **30**, 561–575.

[3] Riley, N. Oscillaory Viscous Flows. Review and Extension. (1967), *J. Inst. Maths. Applics.*, **3**, 419–434.

[4] Roslan, R., Zainodin, H.J., Jusoh, A.W. and Hashim, S.R., (2003), Steady Boundary-Layer in Oscillatory Convective Flow, *Proc. of the 1st Int. Scientific Conf. on Modern Problems of Mathematical physics and informational technologies*, Tashkent, Uzbekistan, 194–201.

[5] Roslan, R., Zainodin, H.J. and Jusoh, A.W., On "Oscillatory Convective Flow *Il Nouvo Cimento* 52(1979)29 − −44", *Technical Report Universiti Malaysia Sabah* LT/SST-UMS/2004/009.

[6] Schlichting, H., Berechnung ebner Periodischer Grenzschichtströmungen, *Physikalische Zeitschrift*, **33**, 1932, 327–335.

R. Roslan: School of Science and Technology, Univerisiti Malaysia Sabah, 88502 Kota Kinabalu, Sabah, Malaysia.
E-mail: rozaini@ums.edu.my

I. Hashim: School of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia.
E-mail: ishak_h@pkrisc.cc.ukm.my

K. Ghazali: School of Science and Technology, Univerisiti Malaysia Sabah, 88502 Kota Kinabalu, Sabah, Malaysia.
E-mail: khadizah@ums.edu.my

ON LINEAR STABILITY ANALYSIS OF BÉNARD-MARANGONI CONVECTION IN A HORIZONTAL FLUID LAYER

M. N. Mohammed-Pauzi, I. Hashim

Universiti Kebangsaan Malaysia, Malaysia

Abstract. Linear stability theory is applied to the problem of the onset of steady Bénard-Marangoni convection in a horizontal layer of fluid heated from below subject to a uniform vertical temperature gradient. The fluid layer is bounded from below by a stress-free boundary kept at a fixed temperature and above by a deformable free surface. The non-dimensional Rayleigh number and Marangoni number are assumed to be linearly dependent. Stability diagrams are numerically computed and the critical Rayleigh number and the critical wavenumber are also obtained. In particular, we present a stability diagram showing regions for long-wavelength instability modes and short-wavelength instability modes.
Key-words: Convection, capillarity, heat and mass transfer

1 Introduction

Thermal convection driven by either buoyancy (Bénard) or thermocapillary (Marangoni) effects have been the subject of a great deal of theoretical and experimental investigation since the pioneering theoretical works of Rayleigh [8] and Pearson [5] respectively. In many real physical situations convection is driven by a combination of both buoyancy and thermocapillary forces, and so Nield [4] extended Pearson's [5] theory to include buoyancy effects. Nield [4] studied the onset of steady Bénard-Marangoni convection in a horizontal layer of fluid without free-surface deformation which is heated from below and subject to a vertical temperature gradient. Nield [4] found that the instability mechanisms reinforce each other. After an extensive numerical search Takashima [10] concluded that oscillatory Bénard-Marangoni convection is not possible in a fluid layer heated from below with a non-deformable free surface. Davis and Homsy [2] extended Nield's work [4] on steady convection to include a deformable free surface and found that weak surface deformation can stabilise buoyancy-dominated convection and destabilise thermocapillary-dominated convection.

In all the early work the effects of buoyancy and thermocapillary, represented by the non-dimensional Rayleigh number R and Marangoni number M respectively, were taken to be independent. However, in a typical physical experiment the control parameter is the temperature difference across the layer which appears linearly in both R and M, and so recent work has focused on the case in which M and R are linearly dependent. Benguria and Depassier [1] investigated the onset of coupled Bénard-Marangoni convection in a planar layer heated from below numerically and found that overstable convection is possible when thermocapillary effects are

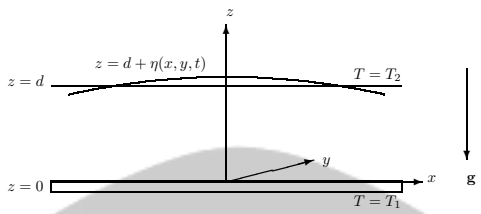


Figure 1: Sketch of problem geometry.

sufficiently strong and the free surface is sufficiently deformable. Independently Pérez-García and Carneiro [6] also considered the same problem numerically and found situations in which competition between a steady and an overstable and between two overstable modes is possible. The existence of a long-wavelength instability in thermocapillary-driven convection was predicted theoretically by Scriven and Sternling [9] and was verified experimentally by VanHook *et al.* [11].

In this paper we use linear stability theory to investigate the onset of steady Bénard-Marangoni convection in a horizontal planar layer of fluid heated from below with stress-free lower boundary. We extend the previous numerical results of Benguria and Depassier [1] for steady Bénard-Marangoni convection and, in particular, present a stability diagram showing regions for long-wavelength instability modes and short-wavelength instability modes.

2 Problem formulation

We wish to examine the stability of a horizontal layer of quiescent fluid of thickness d which is unbounded in the horizontal x- and y-directions. The fluid layer is bounded below by a stress-free planar boundary maintained at a constant temperature T_1 and above by a free surface initially at temperature T_2 and subject to a uniform vertical temperature gradient (see Figure 1). The fluid is Newtonian and incompressible with density

$$\rho = \rho_0 \left[1 - \alpha(T - T_2)\right],$$

where the constant ρ_0 is the value ρ at $T = T_2$ and $\alpha > 0$ is the coefficient of thermal volume expansion. The free surface is in contact with a passive gas at constant pressure and constant temperature T_∞ and has surface tension given by the simple linear law

$$\tau = \tau_0 - \gamma\,(T - T_2)\,,$$

where the constant τ_0 is the value of τ in the undisturbed state and the constant γ is positive for normal fluids.

In the reference state, the fluid is at rest with respect to the rotating axes and heat propagates only by conduction. When motion sets in, the velocity $\mathbf{v} = (u, v, w)$,

pressure p and temperature T fields obey the usual balance equations of mass, momentum and energy,

$$\nabla \cdot \mathbf{v} = 0, \tag{1}$$

$$\rho_0 \left(\frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla \right) \mathbf{v} = -\nabla p + \mu \nabla^2 \mathbf{v} - \rho \mathbf{g} \mathbf{e}_z, \tag{2}$$

$$\left(\frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla \right) T = \kappa \nabla^2 T, \tag{3}$$

where $\mathbf{g} = (0, 0, -g)$ is the gravitational field, $\mathbf{e}_z = (0, 0, 1)$ is a unit vector in the z-direction, μ is the viscosity, κ is the thermal diffusivity and $\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$ is the Laplacian operator.

At the free surface we have the usual kinematic condition,

$$w = \eta_t + u\eta_x + v\eta_y, \tag{4}$$

condition of continuity of the normal stress,

$$\begin{aligned} p - p_a = {} & \frac{2\mu}{N}[u_x\eta_x^2 + (u_y + v_x)\eta_x\eta_y - (u_z + w_x)\eta_x \\ & + v_y\eta_y^2 - (v_z + w_y)\eta_y + w_z] - \frac{\tau}{N\sqrt{N}}[\eta_{xx} + \eta_{yy}], \end{aligned} \tag{5}$$

and condition of continuity of the tangential stresses,

$$\begin{aligned} & \frac{\mu}{N}[-2u_x\eta_x(1 + \eta_y^2) + (u_y + v_x)\{-\eta_y(1 + \eta_y^2) + \eta_y\eta_x^2\} + 2v_y\eta_x\eta_y^2 \\ & \quad + (u_z + w_x)(1 + \eta_y^2 - \eta_x^2) - 2\eta_x\eta_y(v_z + w_y) + 2w_z\eta_x] \\ & \quad = [\tau_x - \eta_x\eta_y\tau_y + \eta_x\tau_z], \end{aligned} \tag{6}$$

$$\begin{aligned} & \frac{\mu}{\sqrt{N}}[-\eta_x(v_x + u_y) - 2\eta_y v_y + (v_z + w_y)(1 - \eta_y^2) \\ & \quad - (w_x + u_z)\eta_x\eta_y + 2w_z\eta_y] = [\tau_y + \eta_y\tau_z], \end{aligned} \tag{7}$$

where $\mathbf{n} = (-\eta_x, -\eta_y, 1)/N$ is the outward unit normal to the free surface, $N = (1 + \eta_x^2 + \eta_y^2)^{1/2}$, and the subscripts x, y, z and t denote differentiation with respect to the variables x, y, z and t respectively. The temperature obeys Newton's law of cooling,

$$-k\frac{\partial T}{\partial \mathbf{n}} = h(T - T_\infty), \tag{8}$$

where k and h are the thermal conductivity of the fluid and the heat transfer coefficient between the free surface and the air, respectively. The lower planar boundary is assumed to be isothermal and stress-free.

We shall investigate the linear stability of a basic state in which the fluid is at rest, $\mathbf{v} = \mathbf{0}$, the free surface is flat, $\eta = 0$, the temperature gradient across the layer is

uniform, $\bar{T} = T_1 - \frac{\Delta T}{d} z$, and the pressure is hydrostatic,

$$\bar{p} \quad = \quad p_a - \rho_0 g \left[1 + \frac{\alpha \Delta T}{2d} z \right] z,$$

where $\Delta T = T_1 - T_2$.

We non-dimensionalise the governing equations and boundary conditions (1)–(8) using d, κ/d^2, κ/d, $\rho_0 \nu \kappa / d^2$ and ΔT as appropriate scales for distance, time, velocity, pressure and temperature respectively. The non-dimensional groups appearing in the problem are the Rayleigh number, $R = g\alpha \Delta T d^3 / \nu\kappa$, the Marangoni number, $M = \gamma \Delta T d / \rho_0 \kappa \nu$, the Prandtl number, $P_r = \nu/\kappa$, the capillary number, $C_r = \rho_0 \nu \kappa / \tau_0 d$, the Galileo number $G = g d^3 / \nu^2$ and the Biot number, $B_i = hd/k$, where, in addition to the parameters defined above, ν denotes the viscosity of the fluid.

We investigate the linear stability of the basic state in the classical manner by seeking perturbed solutions for any quantity $\Phi(x, y, z, t)$ in terms of normal modes in the form

$$\Phi(x, y, z, t) = \Phi_0(x, y, z) + \phi(z) \mathrm{e}^{i(a_x x + a_y y)} \mathrm{e}^{st}, \tag{9}$$

where Φ_0 is the value of Φ in the basic state, ϕ is the amplitude of the perturbation, and $a = (a_x^2 + a_y^2)^{1/2}$ is the total horizontal wave number of the disturbance. The temporal exponent s will, in general, be complex with a real part representing the growth rate of the instability and an imaginary part representing its frequency.

Substituting (9) into the governing equations and neglecting second-order and higher terms in the perturbed quantities, we obtain the corresponding linearised equations involving only $W(z)$, the z-dependent part of the z-component of the perturbation to the velocity, and $\Theta(z)$, the perturbation to the temperature, respectively,

$$(D^2 - a^2) \left(D^2 - a^2 - \frac{s}{P_r} \right) W - a^2 R \Theta \quad = \quad 0, \tag{10}$$

$$(D^2 - a^2 - s)\Theta + W \quad = \quad 0, \tag{11}$$

subject to

$$sf - W \quad = \quad 0, \tag{12}$$

$$\left(D^2 - 3a^2 - \frac{s}{P_r} \right) DW - a^2 \left(P_r G + \frac{a^2}{C_r} \right) f \quad = \quad 0, \tag{13}$$

$$(D^2 + a^2) W + a^2 M (\Theta - f) \quad = \quad 0, \tag{14}$$

$$D\Theta + B_i (\Theta - f) \quad = \quad 0, \tag{15}$$

evaluated on the undisturbed position of the upper free surface $z = 1$, and

$$W \quad = \quad 0, \tag{16}$$

$$D^2 W \quad = \quad 0, \tag{17}$$

$$\Theta \quad = \quad 0, \tag{18}$$

evaluated on the stress-free lower planar boundary $z = 0$, where the operator $D = \mathrm{d}/\mathrm{d}z$ denotes differentiation with respect to the vertical coordinate z and $a^2 = a_x^2 + a_y^2$ is the total wave number in the horizontal x-y plane.

Note that the variables Θ and f can be calculated directly from equation (10) and boundary condition (13) respectively. Eliminating Θ between equations (10) and (11) then gives a single linear sixth-order ordinary differential equation for W,

$$\left[(D^2 - a^2)(D^2 - a^2 - s)\left(D^2 - a^2 - \frac{s}{P_r}\right) + a^2 R\right] W = 0. \tag{19}$$

The Rayleigh number R and the Marangoni number M are related by $M = \Gamma R$ where $\Gamma = \gamma/\rho_0 g \alpha d^2$. We shall investigate the case in which Γ is constant and so R and M are linearly dependent. For large values of Γ, convection is mainly thermocapillary-driven. While small Γ corresponds to buoyancy-dominated convection.

3 Solution of Linearised Problem

Equations (10) and (11) together with the boundary conditions (12)–(18) constitute a linear eigenvalue problem for the unknown temporal exponent s. The general solution of equation (19) is

$$W(z) = \sum_{i=1}^{6} A_i \mathrm{e}^{\xi_i z},$$

where $\xi_1, \ldots, \xi_6$ are the six distinct roots of the sixth-order algebraic equation

$$(\xi^2 - a^2)(\xi^2 - a^2 - s)\left(\xi^2 - a^2 - \frac{s}{P_r}\right) + a^2 R = 0$$

and A_i $(i = 1, \ldots, 6)$ are arbitrary constants. Imposing boundary conditions (12), (14) and (15)–(18), where expressions for Θ and f are obtained from equations (10) and (13) respectively, yields a linear system of the form $\mathrm{P}\mathbf{A} = 0$, where $\mathbf{A} = [A_1, \ldots, A_6]^T$. In general, the 6×6 coefficient matrix P, whose entries depend on a, R, s, G, C_r, Γ, P_r and B_i, is complex and may be rather complicated, and so, in general, has to be calculated using a computer numerically. We use a FORTRAN 77 program employing the NAG routine F03ADF and running on a PC to evaluate the determinant of P using an LU factorization with partial pivoting. A modification of the Powell [7] hybrid algorithm, which is a combination of Newton's method and the method of steepest descent, implemented using NAG routine C05NBF is then used to find the eigenvalues of P by solving the two nonlinear equations obtained from the real and imaginary parts of the determinant of P.

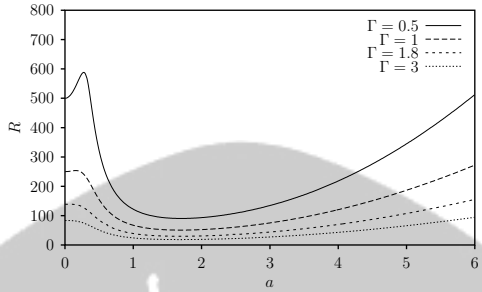


Figure 2: Marginal curves for the onset of steady convection in the case $C_r = 10^{-3}$, $G = 250$, $P_r = 1$ and $B_i = 0$ for several values of Γ.

4 Results and Discussions

The marginal stability curves in the (a, R) plane on which $s_r = 0$ separate regions of unstable modes with $s_r > 0$ from those of stable modes with $s_r < 0$. The critical Rayleigh number for the onset of steady ($s = 0$) convection, denoted by R_c, is the global minimum of R over $a \geq 0$. We denote the corresponding critical wavenumber by a_c.

Benguria and Depassier [1] solved the problem in the special case when the free surface is strongly deformable, i.e. $C_r = \infty$. However, this case is not physically realistic. While in practice the value of C_r may be very small (for a 1 cm layer of water open to air at 20^o C we have $C_r \sim 10^{-7}$) it will inevitably be non-zero. All numerical calculations reported in this paper are done for the case $C_r = 10^{-3}$, $P_r = 1$ and $B_i = 0$. The case when $B_i = 0$ is still representative since the value of B_i for a thin layer with which we are concerned is at most 0.1 for most fluids and such a small value of B_i does not affect appreciably the results for $B_i = 0$.

In Fig. 2 we plot some typical marginal curves for the onset of steady convection in the case $C_r = 10^{-3}$, $G = 250$, $P_r = 1$ and $B_i = 0$ for several values of Γ. Clearly the effect of increasing Γ is to shift the curves downwards. Numerically calculated values of R_c and a_c are plotted as functions of Γ in Fig. 3 in the case $C_r = 10^{-3}$, $P_r = 1$, $B_i = 0$ and $G = 250$. Evidently R_c and a_c are motonically decreasing functions of Γ, suggesting that the fluid layer is destabilised as Γ increases.

Figure 4 shows examples of situations in which two different modes of instabilities co-exist and convection setting in as long-wavelength modes for the case $C_r = 10^{-3}$, $P_r = 1$, $B_i = 0$ and $\Gamma = 1.8$. In this case, competition between the two modes (of different cell sizes) occurs when $G \approx 53.2$. If $G < 53.2$, then convection sets in at $a_c = 0$. Whereas convection sets in at $a_c = O(1)$ when $G > 53.2$. The Galileo number G seems to have very minute effect on this critical wavenumber $a_c = O(1)$ as depicted in Fig. 4. Numerically calculated values of G_c at which two different modes co-exist are plotted as an increasing function of Γ in Fig. 5 for the case $C_r = 10^{-3}$, $P_r = 1$ and $B_i = 0$. The region below the curve corresponds

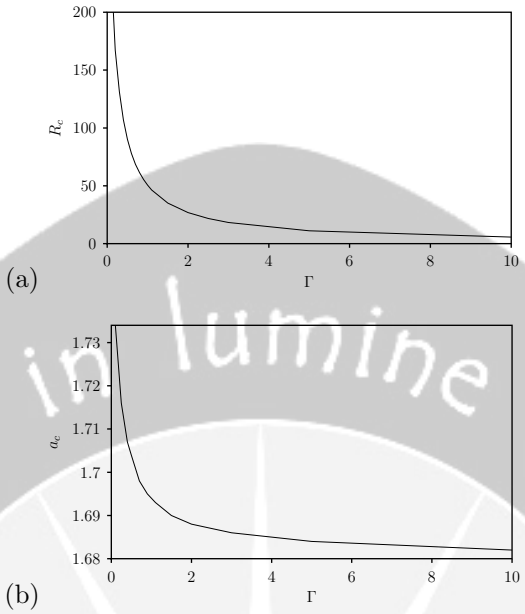


Figure 3: Computed critical values (a) R_c and (b) a_c as functions of Γ for the case $C_r = 10^{-3}$, $P_r = 1$, $B_i = 0$ and $G = 250$.

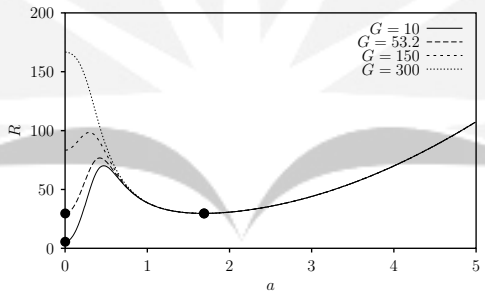


Figure 4: Marginal curves for the onset of steady convection in the case $C_r = 10^{-3}$, $P_r = 1$, $B_i = 0$ and $\Gamma = 1.8$ for several values of G. The dots (•) mark the global minima.

to the case R_c occurring at $a_c = 0$, i.e. long-wavelength modes; while the region above the curve corresponds to the case R_c occurring at $a_c = O(1)$, i.e. short-wavelength modes. Determining which modes are the dominant ones at the onset of convection requires a non-linear analysis which we do not undertake in this paper. The alternative possibility of the instability setting in via an oscillatory mode similar to that found by Hashim [3] will constitute the subject of our subsequent investigation.

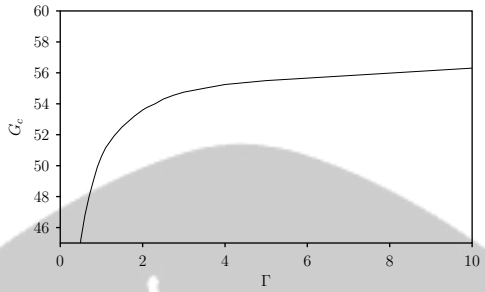


Figure 5: Computed critical values G_c plotted as a function of Γ at which two different modes coexist in the case $C_r = 10^{-3}$, $P_r = 1$ and $B_i = 0$.

5 Conclusions

In this paper we used classical linear stability theory to investigate the onset of Bénard-Marangoni convection in a horizontal planar layer of fluid heated from below in the case when R and M are linearly dependent. We obtained the marginal stability curves for the onset of steady convection which extended the numerical results of earlier authors. Determining the behaviour of the steady marginal curves for long-wavelength disturbances is particularly important because these can be the most unstable modes. We in particular presented the results of numerical calculations which show examples of situations in which two different modes of instabilities co-exist.

Acknowledgments

I. Hashim gratefully acknowledges Universiti Kebangsaan Malaysia for financial support through UKM grant no. ST-024-2003.

References

[1] Benguria, R.D. and M.C. Depassier (1989), On the linear stability theory of Bénard-Marangoni convection, *Phys. Fluids A*, **1**, 1123 – 1127.

[2] Davis, S.H. and G.M. Homsy (1980), Energy stability theory for free surface problems: Buoyancy-thermocapillary layers, *J. Fluid Mech.*, **98**, 527 – 553.

[3] Hashim, I. (2002), On competition between modes at the onset of Bénard-Marangoni convection in a layer of fluid, *Austral. & New Zealand Indust. Appl. Math. J.*, **43**, 387 – 395.

[4] Nield, D.A. (1964), Surface tension and buoyancy effects in cellular convection, *J. Fluid Mech.*, **19**, 341 – 352.

[5] Pearson, J.R.A. (1958), On convection cells induced by surface tension, *J. Fluid Mech.*, **4**, 489 – 500.

[6] Pérez-García, C. and G. Carneiro (1991), Linear stability analysis of Bénard-Marangoni convection in fluids with a deformable free surface, *Phys. Fluids A*, **3**, 292 – 298.

[7] Powell, M.J.D. (1970), A hybrid method for nonlinear equations, in P. Rabinowitz, editor, *Numerical Methods for Nonlinear Algebraic Equations*, Editor: P. Rabinowitz, Gordon and Breach, London, 87 – 114.

[8] Lord Rayleigh (1916), On convection currents in a horizontal layer of fluid, when the higher temperature is on the under side, *Phil. Mag.*, **32**, 529 – 546.

[9] Scriven, L.E. and C.V. Sternling (1964), On cellular convection driven by surface-tension gradients: Effects of mean surface tension and surface viscosity, *J. Fluid Mech.*, **19**, 321 – 340.

[10] Takashima, M. (1970), Nature of the neutral state in convective instability induced by surface tension and buoyancy, *J. Phys. Soc. Japan*, **28**, 810.

[11] VanHook, S.J., M.F. Schatz, J.B. Swift, W.D. McCormick, and H.L. Swinney (1997), Long-wavelength surface-tension-driven Bénard convection: experiment and theory, *J. Fluid Mech.*, **345**, 45 – 78.

M. N. Mohammed-Pauzi: School of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi Selangor, Malaysia.

I. Hashim: School of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi Selangor, Malaysia.
Fax: +603 8925 4519
E-mail: ishak_h@ukm.my

LAGRANGIAN DYNAMICS OF THE NAVIER-STOKES EQUATION

A. Sulaiman[a] and L.T. Handoko[b]

[a] Geomathematics Group, Geostech Laboratory BPPT, Kompleks Puspiptek Serpong, Tangerang 15310, Indonesia

[b] Group for Theoretical and Computational Physics, Research Center for Physics LIPI, Kompleks Puspiptek Serpong, Tangerang 15310, Indonesia

Abstract. The equation of motion governs fluid flow is well known as the Navier-Stokes equation. Most researches on fluid dynamics are mostly dedicated to obtain the solutions of this equation with particular boundary conditions and approximations. We propose an alternative approach to deal with fluid dynamics without solving the equation using the lagrangian. We attempt to develop a gauge invariant lagrangian and reconstruct the Navier-Stokes equation through the Euler-Lagrange equation. The lagrangian consists of gauge boson field $\mathcal{A}_\mu$ with appropriate content describing the fluid dynamics, *i.e.* $\mathcal{A}_\mu = (\Phi, -\vec{v})$. An example to apply it to the interaction of fluid in a solitonic medium is also given.
Key-words: bosonic Lagrangian, gauge field, Navier-Stokes

1 Introduction

The fluid dynamics still remains as an unsolved problem. Mathematically, a fluid flow is described by the Navier-Stokes (NS) equation [1]:

$$\frac{\partial \vec{v}}{\partial t} + (\vec{v}\cdot\vec{\nabla})\vec{v} = -\frac{1}{\rho}\vec{\nabla}P - \mu\vec{\nabla}^2\vec{v}\,, \tag{1}$$

where $\vec{v}$ is fluid velocity, P is pressure, ρ is density and μ is the coefficient of viscosity.

In principle, the study of fluid dynamics is focused on solving the Navier-Stokes equation with particular boundary conditions and / or some approximations depend on the phenomenon under consideration. Mathematically it has been known as the boundary value problem. The most difficult problem in fluid dynamics is turbulence phenomenon. In the turbulence regime, the solution for the Navier-Stoke equation has a lot of Fourier modes, such that the solution is untrackable numerically or analytically. It is predicted that the strong turbulence has 10^{10} numerical operation [2]. This motivates us to look for another approach rather than the conventional ones. This paper treats the fluid dynamics differently than the conventional point of view as seen in some fluid dynamics textbooks. In this approach, the fluid is described as a field of fluid buch. We use the gauge field theory to construct a lagrangian describing fluid dynamics by borrowing the gauge principle. The Navier-Stoke equation can be obtained from this Lagrangian as its equation of motion through the Euler-Lagrange principles.

2 Maxwell-like equation for ideal fluid

The abelian gauge theory $U(1)$ is an electromagnetic theory that reproduces the Maxwell equation. To build a lagrangian that is similar with the abelian gauge theory, we should 'derive' the Maxwell-like equation from the Navier-Stokes equation [3]. The result can be used as a clue to construct a lagrangianan for fluid that satisfies gauge principle. Considering the Navier-Stokes equation Eq. (1) for an ideal and incompressible fluid,

$$\rho\left(\frac{\partial\vec{v}}{\partial t}+(\vec{v}.\vec{\nabla})\vec{v}\right) = -\vec{\nabla}P\,, \tag{2}$$

$$\vec{\nabla}\cdot\vec{v} = 0\,. \tag{3}$$

Using the identity $\vec{v}\times(\vec{\nabla}\times\vec{v})=\vec{\nabla}(\frac{1}{2}\vec{v}^2)-(\vec{v}\cdot\vec{\nabla})\vec{v}$, it can be rewritten as,

$$\frac{\partial\vec{v}}{\partial t}+\vec{\nabla}\left(\frac{1}{2}\vec{v}^2\right)-\vec{v}\times(\vec{\nabla}\times\vec{v})=-\frac{1}{\rho}\vec{\nabla}P\,, \tag{4}$$

and then,

$$\frac{\partial\vec{v}}{\partial t}=\vec{v}\times(\vec{\nabla}\times\vec{v})-\vec{\nabla}\left(\frac{1}{2}\vec{v}^2+\frac{P}{\rho}\right)\,. \tag{5}$$

Putting the scalar potenstial $\Phi=\frac{1}{2}\vec{v}^2+\frac{P}{\rho}$, the vorticity $\vec{\omega}=\vec{\nabla}\times\vec{v}$ and the Lamb's vector $\vec{l}=\vec{\omega}\times\vec{v}$, the equation becomes,

$$\begin{aligned}\frac{\partial\vec{v}}{\partial t} &= -\vec{\omega}\times\vec{v}-\vec{\nabla}\Phi\\ &= -\vec{l}-\vec{\nabla}\Phi\,.\end{aligned} \tag{6}$$

Imposing curl operation in Eq. (6) we obtain the vorticity equation as follow,

$$\frac{\partial\vec{\omega}}{\partial t}=-\vec{\nabla}\times(\vec{\omega}\times\vec{v})\,. \tag{7}$$

In order to get the Maxwell-like equation for an ideal fluid, let us take divergence operation for Eq. (6), that is

$$\begin{aligned}\frac{\partial}{\partial t}(\vec{\nabla}\cdot\vec{v}) &= -\vec{\nabla}\cdot\vec{l}-\vec{\nabla}^2\Phi\\ \vec{\nabla}\cdot\vec{l}=-\vec{\nabla}^2\Phi &= \tilde{\rho}\,.\end{aligned} \tag{8}$$

Here we have used the incompressible condition, while by definition the divergence of vorticity is always zero, *i.e.* $\vec{\nabla}\cdot\vec{\omega}=0$. Imposing again curl operation, we have,

$$\begin{aligned}\frac{\partial}{\partial t}(\vec{\nabla}\times\vec{v}) &= -\vec{\nabla}\times\vec{l}-\vec{\nabla}\times(\vec{\nabla}\Phi)\,,\\ \frac{\partial\vec{\omega}}{\partial t} &= -\vec{\nabla}\times\vec{l}\,,\\ \vec{\nabla}\times\vec{l} &= -\frac{\partial\vec{\omega}}{\partial t}\,,\end{aligned} \tag{9}$$

using the identity $\vec{\nabla} \times (\vec{\nabla} \cdot \phi) = 0$.

Now, let us consider the definition of the Lamb's vector $\vec{l} = \vec{\omega} \times \vec{v}$. Taking the derivative $\partial/\partial t$ in the definition we obtain,

$$\frac{\partial \vec{l}}{\partial t} = \frac{\partial \vec{\omega}}{\partial t} \times \vec{v} + \vec{\omega} \times \frac{\partial \vec{v}}{\partial t} \,. \tag{10}$$

Substituting Eq. (6) and (7), we get,

$$\vec{\nabla} \times \vec{\omega} = \alpha \vec{j} + \alpha \frac{\partial \vec{l}}{\partial t} \,, \tag{11}$$

where,

$$\alpha = \frac{1}{\vec{v}^2} \,, \tag{12}$$

$$\vec{j} = -\vec{v}\vec{\nabla}^2\Phi + \left[\vec{\nabla} \times (\vec{v} \cdot \vec{\omega})\right] \vec{v} + \vec{\omega} \times \vec{\nabla}(\Phi + \vec{v}^2) + 2\left[(\vec{\nabla} \times \vec{v}) \cdot \vec{\nabla}\right] \vec{v} \,. \tag{13}$$

These results induce a series of equations,

$$\vec{\nabla} \cdot \vec{l} = \tilde{\rho} \,, \tag{14}$$

$$\vec{\nabla} \times \vec{l} = -\frac{\partial \vec{\omega}}{\partial t} \,, \tag{15}$$

$$\vec{\nabla} \cdot \vec{\omega} = 0 \,, \tag{16}$$

$$\vec{\nabla} \times \vec{\omega} = \alpha \vec{j} + \alpha \frac{\partial \vec{l}}{\partial t} \,, \tag{17}$$

that is clearly the Maxwell-like equation for fluids. If the fluid velocity is time independent, then $\vec{l} = -\vec{\nabla}\Phi$. This is the "electrostatic" condition. We use these results to develop gauge field theory approach for fluid dynamics in the next section.

3 Bosonic Lagrangian for Fluid

The correspondences of the electromagnetism and the ideal fluid can be written as follow,

$$\begin{aligned} \overrightarrow{B} &\leftrightarrow \overrightarrow{\omega} \,, \\ \overrightarrow{E} &\leftrightarrow \overrightarrow{l} \,, \\ \overrightarrow{A} &\leftrightarrow \overrightarrow{v} \,, \\ \phi &\leftrightarrow \Phi \,, \end{aligned} \tag{18}$$

where $\overrightarrow{B}$ is the magnetic field, $\overrightarrow{E}$ is the electric field, $\overrightarrow{A}$ is the electromagnetics vector, ϕ is a scalar function, $\overrightarrow{\omega}$ is the fluid vorticity, $\overrightarrow{l}$ is the Lamb's vector, $\vec{v}$ is fluid velocity and Φ is the scalar potential. The same as the electromagnetics field, we have a four vector $A_\mu = (\phi, \vec{A})$ which can be interpreted as the four vector

for fluid dynamics, $\mathcal{A}_\mu = (\Phi, -\vec{v})$. In the electromagnetics field, the scalar and vector potentials, ϕ and $\vec{A}$, are auxiliary fields. On the other hand, in the fluid dynamics the scalar potential $\Phi = \frac{1}{2}\vec{v}^2 + V$ describes the kinetic energy of fluid, while the vector potential $\vec{v}$ is fluid velocity. Similar to the electromagnetics field, the lagrangian density has the form of [4, 5],

$$\mathcal{L}_{NS} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + g\mathcal{J}_\mu\mathcal{A}^\mu \ , \tag{19}$$

where,

$$\mathcal{F}_{\mu\nu} \equiv \partial_\mu\mathcal{A}_\nu - \partial_\nu\mathcal{A}_\mu \ . \tag{20}$$

This Lagrangian obeys the gauge principles, *i.e.* it is invariant under a particular local gauge transformation,

$$\mathcal{A}_\mu \rightarrow \mathcal{A}'_\mu \equiv e^{-i\theta}\mathcal{A}_\mu \ , \tag{21}$$

where $\theta = \theta(x)$ is an arbitrary real constant. It is easy to show that the lagrangian density in Eq. (19) is invariant under this transformation.

The equation of motion governed by this lagrangian can be derived using the Euler-lagrange equation in term of $\mathcal{A}_\mu$,

$$\partial^\nu \frac{\partial\mathcal{L}_{NS}}{\partial(\partial^\nu\mathcal{A}^\mu)} - \frac{\partial\mathcal{L}_{NS}}{\partial\mathcal{A}^\mu} = 0 \ . \tag{22}$$

After a straightforward calculation, we obtain,

$$\partial^\nu(\partial_\mu\mathcal{A}_\nu - \partial^\nu\partial_\nu\mathcal{A}_\mu) - g\mathcal{J}_\mu = 0 \ . \tag{23}$$

Now integrating it over x^ν and considering only the non-trivial relation as $\nu \neq \mu$ gives,

$$\partial_0 A_i - \partial_i A_0 = -g \oint \mathrm{d}x_0 J_i = g \oint \mathrm{d}x_i J_0 \ . \tag{24}$$

Since $A_i = -\vec{v}$, $A_o = \Phi$, $\partial_o = \partial/\partial t$ and $\partial_i = \vec{\nabla}$. we have,

$$-\frac{\partial\vec{v}}{\partial t} - \vec{\nabla}\Phi = -g\vec{\tilde{J}} \ , \tag{25}$$

where $\tilde{J}_i \equiv \oint \mathrm{d}x_0 J_i = -\oint \mathrm{d}x_i J_0$. Concerning the scalar potential given by $\Phi = \frac{1}{2}\vec{v}^2 + V$, we obtain,

$$-\frac{\partial\vec{v}}{\partial t} - \frac{1}{2}\vec{\nabla}\left|\vec{v}\right|^2 - \vec{\nabla}V = -g\vec{\tilde{J}} \ . \tag{26}$$

Borrowing the identity $\frac{1}{2}\vec{\nabla}\left|\vec{v}\right|^2 = (\vec{v}\cdot\vec{\nabla})\vec{v} + \vec{v}\times(\vec{\nabla}\times\vec{v})$, we get,

$$\frac{\partial\vec{v}}{\partial t} + (\vec{v}\cdot\vec{\nabla})\vec{v} = -\vec{\nabla}V - \vec{v}\times\vec{\omega} - g\vec{\tilde{J}} \ , \tag{27}$$

where $\vec{\omega} \equiv \vec{\nabla}\times\vec{v}$ is the vorticity. This result reproduces the general NS equation with arbitrary conservative forces ($\vec{\nabla}V$). The potential V can be associated with

some known forces, for example, P/ρ, $(Gm)/r$ and $\eta(\vec{\nabla}\cdot\vec{v})$. Here, $P, \rho, G, \nu+\eta$ denote pressure, density, gravitational constant and viscosity as well.

From Eq. (23) we can write explicitly the expression for the 4-vector current as,

$$J_0 = \rho = \frac{1}{g}\vec{\nabla}\cdot\left(\partial_0\vec{A} - \vec{\nabla}\phi\right) , \tag{28}$$

$$\vec{J} = -\frac{1}{g}\vec{\nabla}\times\left(\vec{\nabla}\times\vec{A}\right) - \partial^0\left(\partial_0\vec{A} - \vec{\nabla}\phi\right) , \tag{29}$$

Taking the time derivative operation to Eq. (28) and divergence operation to Eq. (29) we get,

$$\frac{\partial\rho}{\partial t} = \frac{1}{g}\left[\partial_t^2(\vec{\nabla}\cdot\vec{A}) - \vec{\nabla}^2(\partial_t\phi)\right] , \tag{30}$$

$$\vec{\nabla}\cdot\vec{J} = -\frac{1}{g}\vec{\nabla}\cdot\left[\vec{\nabla}\times(\vec{\nabla}\times\vec{A})\right] - \frac{1}{g}\left[\partial_t^2(\vec{\nabla}\cdot\vec{A}) - \vec{\nabla}^2(\partial_t\phi)\right] , \tag{31}$$

Using vector identity $\vec{\nabla}\cdot\vec{\nabla}\times\vec{a} = 0$,

$$\frac{\partial\rho}{\partial t} + \vec{\nabla}\cdot\vec{J} = 0 , \tag{32}$$

that is the continuity equation, or in the four-vector formalism it can be written as $\partial_\mu\mathcal{J}^\mu = 0$.

4 Interaction between Soliton with Fluid

In this section we describe an idea to apply the theory described in the preceeding section. We give an example on applying the theory to provide a consistent way for the interaction between soliton and fluid system. Soliton is a pulse-like nonlinear wave which forms a collision with similar pulse having unchanged shape and speed [10]. The wave equations that exhibit soliton are the KdV equation, the Nonlinear Schrodinger equation, the Sine-Gordon equation, Nonlinear Klein-Gordon equation, the Born-Infeld equation, the Burger equation and the Boussiness equation. Considering the Nonlinear Klein-Gordon as follow:

$$\frac{\partial^2\phi}{\partial t^2} - \frac{\partial^2\phi}{\partial x^2} - m^2\phi + \frac{\lambda}{3!}\phi^3 = 0 . \tag{33}$$

The equation is a continuum version of that describes a propagation of molecular vibration (vibron) in $\alpha-$helical protein [9]. The vibration excitation in the $\alpha-$helix protein propagates from one group to the next because of the dipole-dipole interaction between the group. The wave is called the Davidov soliton [9]. Davydov has shown that in $\alpha-$helical protein soliton can be formed by coupling the propagation of amide$-I$ vibrations with longitudinal phonons along spines and that such entities are responsible for mechanism of energy transfer in biological system

[9]. If $\alpha-$helical protein immersed in Bio-fluid, then the phenomenon can be described by the interaction of soliton with fluid system. In standard technique in fluid dynamics, the problem will be done by solving of the Navier-Stokes equation and nonlinear Klein-Gordon simultaneously.

In our current approach the problem is treated as follow. First, let us rewrite Eq. (33) into four vector formalism,

$$\partial_\mu \partial^\mu \phi - m^2 \phi + \frac{\lambda}{3}\phi^3 = 0 \, . \tag{34}$$

Using the Euler-Lagrange equation, the lagrangian density is,

$$\mathcal{L} = \frac{1}{2}(\partial_\mu \phi)(\partial^\mu \phi) + \frac{m^2}{2!}\phi^2 - \frac{\lambda}{4!}\phi^4 \, . \tag{35}$$

In order to couple this lagrangian with the Navier-Stoke lagrangian in Eq. (**??**), it is sufficient to replace the covariant derivative in Eq. (35) [5],

$$\mathcal{D}_\mu \phi = (\partial_\mu + ig\mathcal{A}_\mu)\phi \, . \tag{36}$$

The covariant derivative is invariant under local gauge transformation[7]. Then the interaction between soliton and fluid system obeys the lagrangian,

$$\mathcal{L} = -\frac{1}{4}\mathcal{F}_{\mu\nu}\mathcal{F}^{\mu\nu} + \frac{1}{2}(\mathcal{D}_\mu \phi)(\mathcal{D}^\mu \phi) + \frac{m^2}{2}\phi^2 - \frac{\lambda}{4!}\phi^4 \, . \tag{37}$$

One interesting case is when we consider a static condition, *i.e.* $\partial_t f = 0$ with f is an arbitrary functions. Substituting $\mathcal{A}_\mu = (\Phi, -\vec{v})$ into Eq. (37) then the Lagrange density becomes,

$$\mathcal{L} = -\frac{1}{2}(\nabla \times \vec{v})^2 + \frac{1}{2}|(\nabla - ig\vec{v})\phi|^2 + \frac{m^2}{2!}\phi^2 - \frac{\lambda}{4!}\phi^4 \, . \tag{38}$$

The lagrangian is nothing else similar with the Ginzburg-Landau free energy lagrangian that is widely used in superconductor theory [8]. We have seen that the phenomenon of $\alpha-$helical protein immersed in fluid similar with quantum electrodynamics for boson particle, while for static case it is similar with the Ginzburg-Landau model for superconductor.

In order to perform an explicit calculation, suppose we have one-dimensional velocity in x direction $\vec{v} = (u(x), 0, 0)$ and $\phi = \phi(x)$. Then the lagrangian in Eq.(38) reads,

$$\mathcal{L} = \frac{1}{2}\phi_x^2 - \frac{1}{2}g^2u^2\phi^2 + \frac{m^2}{2!}\phi^2 - \frac{\lambda}{4!}\phi^4 \, . \tag{39}$$

Substituting it into Euler-lagrangian equation we arrive at,

$$\frac{d^2\phi}{dx^2} - \gamma(x)\phi + \frac{\lambda}{3!}\phi^3 = 0 \, , \tag{40}$$

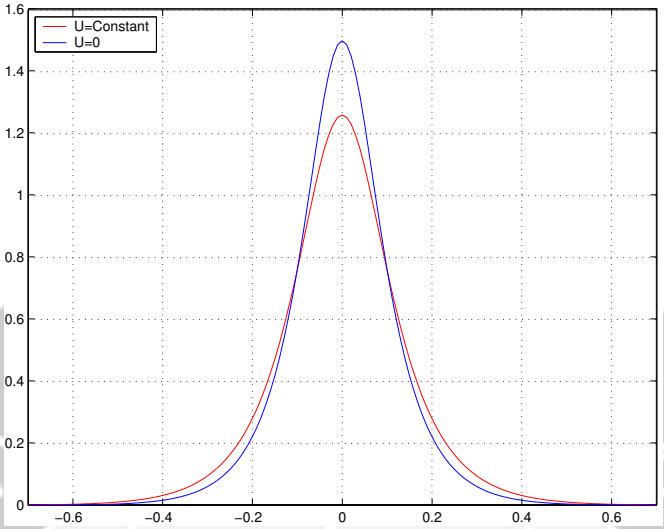


Figure 1: Single soliton solution of the nonlinear Klein-Gordon equation.

where $\gamma(x) = m^2 - g^2 u(x)^2$. The equation is called the variable coefficient of nonlinear Klein-Gordon equation.

Further, we can consider a special case when the fluid velocity is constant, $u(x) = U$, to obtain

$$\frac{d^2\phi}{dx^2} - \gamma\phi + \alpha\phi^3 = 0 \,, \tag{41}$$

with $\gamma = m^2 - g^2U^2$ and $\alpha = \lambda/3!$. To solve the equation, we can use a mathematical trick as follows. First multiply it by $d\phi/dx$,

$$\frac{d\phi}{dx}\frac{d^2\phi}{dx^2} - \gamma\phi\frac{d\phi}{dx} + \alpha\phi^3\frac{d\phi}{dx} = 0 \,, \tag{42}$$

then integrating out over x and putting the integration constant as zero due to integrable condition $\lim_{x\to\pm\infty}\phi = 0$. Finally we obtain,

$$\left(\frac{d\phi}{dx}\right)^2 - \gamma\phi^2 + \frac{\alpha}{2}\phi^4 = 0 \,, \tag{43}$$

and it can be rewritten further as,

$$\int \frac{d\phi}{\phi(\delta^2 - \phi^2)^{\frac{1}{2}}} = \int \sqrt{\frac{\alpha}{2}}dx \,, \tag{44}$$

where $\delta^2 = 2\gamma/\alpha$. Integration of the left hand side and solving the equation for ϕ

provide the result,

$$\begin{aligned}\phi &= \frac{2\delta e^{-\sqrt{\frac{\alpha}{2}}\delta x}}{1+e^{-2\sqrt{\frac{\alpha}{2}}\delta x}} = \frac{2\delta}{e^{\sqrt{\frac{\alpha}{2}}\delta x}+e^{-\sqrt{\frac{\alpha}{2}}\delta x}} \\ &= \frac{\delta}{\cosh(\sqrt{\frac{\alpha}{2}}\delta x)}0 = \delta\mathrm{sech}(\sqrt{\frac{\alpha}{2}}\delta x)\ . \end{aligned} \tag{45}$$

Thus, the solution for a homogeneous nonlinear Klein - Gordon equation is,

$$\phi(x) = A\mathrm{sech}(\Lambda x)\ , \tag{46}$$

where $A = (12\gamma)/\lambda$ and $\Lambda = (12\sqrt{3}\gamma)/\lambda^{3/2}$. This result is depicted in Fig. 1. The figure shows that the soliton propagation will be damped by fluid. This theory also can be applied in turbulence phenomenon [3].

5 Conclusion

We have shown an analogy between electromagnetics field and fluid dynamics using the Maxwell-like equation for an ideal fluid. The results provide a clue that we might be able to build a gauge invariant lagrangian density, the so-called Navier-Stokes lagrangian in term of scalar and vector potentials $\mathcal{A}_\mu$. Then the Navier-Stokes equation is obtained as its equation of motion through the Euler-lagrange principle. The application of the theory is wide, for instance the interaction between Davydov soliton with fluid system that can be described by the lagrangian density which is similar to quantum electrodynamics for boson particle. In the static condition, the lagrangian density is similar with the Ginzburg-Landau lagrangian. If the fluid flow is parallel with soliton propagation we also obtain the variable coefficient Nonlinear Klein-Gordon equation. Single soliton solution has been obtained in term of a second hyperbolic function. The result showed that the present of fluid flow will give a damping in solitary wave propagation.

Acknowledgment

The authors thank Terry Mart, Anto Sulaksono and all of the theoretical group members (Ketut Saputra, Ardy Mustafa, Handhika, Fahd, Jani, Ayung) for so many valuable discussion. This research is funded by DIP P3-TISDA BPPT and Riset Kompetitif LIPI (fiscal year 2005).

References

[1] P. Kundu (1996), *Fluids Mechanics*, Addison-Wesley, New York.

[2] T. Mulin (1995), *The Nature of Chaos*, Clarendon Press, Oxford.

[3] A. Sulaiman (2005), *Contruction of The Navier-Stokes Equation using Gauge Field Theory Approach*, Master Theses at Department of Physics, University of Indonesia.

[4] A. Sulaiman and L.T. Handoko (2005), Gauge field theory approach to construct the Navier-Stokes equation, *Acta Phys. Pol.*, **A**, in press.

[5] A. Sulaiman and L.T. Handoko (2005), Relativistic Fluid Dynamics through Gauge Invariant Lagrangian, *Fluid Res. Dynamics*, under submission.

[6] Huang.K (1992), *Quarks, Leptons and Gauge Fields*, Worlds Sceintific, Singapore.

[7] Muta.T (2000), *Foundation of Quantum Chromodynamics*, Worlds Sceintific, Singapore.

[8] Binney. J.J et.al. (1995), *The Theory of Critical Phenomena*, Clarendon press, Oxford.

[9] Takeno, S (1987), Vibron Soliton and Coherent Polarization, *Collected paper Dedicated to prof K Tomita*, Editor:Takeno.S et al , Kyoto University Press. Kyoto.

[10] A. Scott, et al (1973), Soliton: A New Concepts in Applied Science,*Proceeding of the IEEE*, **61**, 1443-1464.

Albert Sulaiman: Geomathematics Group, Geostech Laboratory, Badan Pengkajian dan Penerapan Teknologi (BPPT). Kompleks Puspiptek Serpong, Tangerang, Indonesia.
E-mail: lyman@tisda.org

L.T. Handoko: Group for Theoretical and Computational Physics, Research Center for Physics, Indonesian Institute of Sciences, Kompleks Puspiptek Serpong, Tangerang 15310, Indonesia.
E-mail: handoko@lipi.fisika.net

An Analytical Study of Hydromagnetic Natural Convection Subject to Radiation

Pallath Chandran and Nirmal C. Sacheti

Dept. of Math. & Statistics, College of Science, Sultan Qaboos University, Sultanate of Oman

Abstract: Magnetohydrodynamic flows are known to play important roles in various industrial and technological fields. One of the key aspects of such flows stems from the likely response of the boundary layer to externally applied forces due to gravity and magnetic field. In addition to these forces, the inclusion of radiation effects becomes essential in applications such as high temperature processing and space technology. In this paper, we have thus considered natural convection in an electrically conducting radiating fluid. The flow, in the presence of an externally applied magnetic field, is caused by the impulsive motion of an infinite vertical flat plate bounding the fluid. Under the Boussinesq approximation, the governing momentum and energy equations for natural convection, subject to isothermal or constant heat flux conditions at the boundary, have been solved analytically. The solutions have been obtained corresponding to the cases of magnetic field being fixed either relative to the fluid or to the boundary. In all situations considered in this work, the expressions for the temperature and velocity of the fluid have been obtained explicitly. There arises a number of nondimensional parameters, describing the physical processes considered, whose effects on the developing temperature and velocity profiles have been analysed in detail. Furthermore, the influence of the parameters on the shear stress at the moving vertical plate has also been discussed.

Keyword: Mathematical physics, natural convection

An Analytic Solution of the Ordinary Differential Equation of Green's Function for Transient Wave-Body Interaction Problems

Aries Sulisetyono

Department of Naval Architecture and Shipbuilding Engineering, Institute Technology of Sepuluh Nopember (ITS), Kampus ITS Sukolilo Surabaya, Indonesia

Abstract: It is well known that the principal difficulties in the prediction of floating body motion in the time domain are mainly due to, the evaluation of the memory part of Green's function and the convolution integral of the boundary integral equation. Based on a differential approach of the Green's function by Clement [Journal of Engineering Mathematics,33,1998], this paper introduces an analytical approach to evaluate the memory part of time domain Green's function as well as the convolution integral involved in the boundary integral equation. This solution can speed up the computational process and reduce the numerical error. Results of this approach are compared to those obtained by Runge-Kutta and Tabulation methods.

Keyword: power series expansion, free surface hydrodynamics, time domain Green's function, floating body motion

Heuristical Point Of View In the Wave Theory Of Hydrodynamics

Gunawan Nugroho

Department of Engineering Physics, Faculty of Industrial Engineering, Sepuluh Nopember Institute of Technology, Indonesia

Abstract: Wave behaviour in hydrodynamics is reviewed in this study. Reconstruction of Navier-Stokes equation in equivalent form to Maxwell equation of electrodynamics and the wave equation is generated. The result shows that wave solution hydrodynamics has property as electrodynamics in case of free flow. Statistic mechanical method is applied to theoretically prove that wave characteristics inherently existed in continuum system of hydrodynamics. In accordance with the method, it is theoretically shown that black body radiation is able to yield kinetic energy distribution when fluid is opposed to radiation filed.

THE DESIGN OF BLOCK-CIPHER-RESISTANT TO CRYPTANALYSIS

Yusuf Kurniawan[1], Adang Suwandi A[2], M Sukrisno Mardiyanto[2], Iping Supriana S[2], Sarwono Sutikno[2]

[1] Universitas Pasundan, Bandung, Indonesia
[2] ITB, Bandung, Indonesia

Abstract. Differential and linear cryptanalysis are two of the most important attacks to test the strength of block cipher. Several researches to increase the security of block cipher are based on those cryptanalysis. The variant of the attack grows quickly like higher order and truncated attack as the variant of differential attack, and multiple approximation as variant of linear attack. Unfortunately, the resistance to differential and linear attack don't mean that the cipher is resistant to all other attacks. For example, since AES is provable to resistance to these attacks, researchers try to attack AES using algebraic attacks. The attack tries to approximate S-Box with many equations and then simplify the equations and then solve the equations to get keys. This paper will try to describe how to strengthen cipher without reducing the implementation speed at various softwares and hardwares .
Key-words: block cipher, differential, linear, cryptanalysis, S-Box

1 Introduction

Since Data Encryption Standard (DES) is used in the seventies, academic world began to examine the strength of this block cipher. Many of cryptographers tried to break DES and investigated its property. DES is designed to be fast in hardware, so its performance in software is rather disappointed. DES uses many bits permutation, so its performance is better in hardware than software. DES use Feistel structure, so the encryption structure is same as the decryption structure, except the order of subkey used. With Feistel structure, the security level of encryption is same as decryption [1].

For two decades, no one can break DES algorithm practically in published article with time which is shorter than brute force attack. However, the 56-bit key size of DES is too short to face exhaustive key search. With dedicated hardware costing about 1 million US$, a DES key can be recovered in about an hour[2]. On Tuesday[3], January 19, 1999, Distributed net break DES in 22 hours and 15 minutes with nearly 100,000 PCs on the internet.[4] proposed an FPGA implementation completing the attack DES in 12-15 hours, using hardware roughly worth $3500. We can use triple DES to increase key size of DES, but this mode will slow process of encryption and decryption three times. Therefore, in 1997 till 2001, the process to replacement DES was conducted. The result was Rijndael algorithm

became the winner in Advanced Encryption Standard (AES) contest. Rijndael uses Substitution Permutation Network (SPN) rather than Feistel. Process of encryption is not same as decryption, so the security level of them are also different. SPN has high degree in parallelism rather than Feistel.

The organization of the remainder of the paper is as follows : Section 2 discuss DES-like ciphers and their cryptanalysis. Section 3 describes AES-like ciphers and their attacks. Section 4 describes the design of block cipher, and section 5 talks about conclusion.

2 DES-like ciphers

For a full specification of DES, we refer reader to [6]. Let $(G, +)$ be a finite Abelian group, G' a subgroup on G, so $F_i : G \rightarrow G'$ mappings and $E_i : G' \rightarrow G$. We define an r-round DES-like cipher over G as follows [7]:

Given a plaintext $p = (p_L, p_R) \in G'$ x G' and key $k = (k_1, k_2, ..k_r) \in G^r$ Ciphertext $c = (c_L, c_R)$ is computed in r iterative rounds. If input to first round $= x_L(0) = p_L$ and $x_R(0) = p_R$ and i = 1,2,...r, then

$x_i = (x_L(i), x_R(i))$ where
$x_L(i) = x_R(i-1)$ and
$x_R(i) = F_i(E_i(x_R(i-1)) \oplus k_i) \oplus x_L(i-1)$
So $c_L = x_R(r)$ and $c_R = x_L(r)$

2.1 Differential Cryptanalysis

Differential cryptanalysis is introduced by Biham and Shamir [9]. The characteristic of t rounds $\chi = \chi(\psi(0), ...\psi(t))$ of a DES-like cipher has a sequence of XOR input of each round $\psi(i) = (\psi_L(i), \psi_R(i)) \in G'$ x G' where $0 \leq i \leq t$ and $\psi_L(i) = \psi_R(i-1)$, i=1,2,...t

Given $x, x^* \in G'$ x G' and $k = (k_1, ..., k_r) \in G^r, r \geq t, \chi$ holds for x and k if $x \oplus x^* = \psi(0)$ and $x(i) \oplus x^*(i) = \psi(i)$

In the each round function there are some S-boxes. For any given $\Delta w_i, \Delta w_{i+1} \in GF(2^m)$, where w_i is the m-bit input of S-Box and w_{i+1} is m-bit output, the differential probabilities of $s_j - box : GF(2^m) \rightarrow GF(2^m)$ are defined as [8] :

$$P_{dif}[s_j(w_i) \oplus s_j(w_i \oplus \Delta w_i) = \Delta w_{i+1}] = \frac{\#\{w_i \in GF(2^m) | s_j(w_i) \oplus s_j(w_i \oplus \Delta w_i) = \Delta w_{i+1}\}}{2^m}$$

The probability of difference of input will give difference of output is the number of that occurrence which is possible, divided by the number of all possible input values. We can use characteristics [9]

$$0 \longleftarrow 0$$

$$0 \longleftarrow \psi = 19600000_{hex}$$

to break DES with probability of every two rounds are $\frac{1}{234}$. And characteristics probability of round 3 till 13 is $(\frac{1}{234})^6$. We have to use trick in first round so that we only use this probability to break 16 rounds of full DES.

To prove that security level of DES-like ciphers can against differential cryptanalysis, at least, it must be ensured that there is no differential with a probability high enough to enable succesfull attack [5]. Let G_1 and G_2 be finite Abelian groups. A mapping $F : G_1 \rightarrow G_2$ is called differentially δ-uniform if for all $\phi \in G_1$, $\phi \neq 0$, and $\varphi \in$G$_2$

$$|\{x \in G_1 | F(x \oplus \phi) \oplus F(x) = \varphi\}| \leq \delta$$

We have to be sure that δ of each S-Box used in cipher as small as possible, so we can believe that a block cipher is secure against differential attack. For example, consider function $F(x) = x^{2^k+1}$ and $F(x) = x^{-1}$

Proposition 1 . Let $F(x) = x^{2^k+1}$ in $GF(2^n)$ and let $s = gcd(k, n)$. Then F is differentially 2^s-uniform.

Proof : Given $\phi, \varphi \in GF(2^n)$,$\phi \neq 0$ then

$$(x \oplus \phi)^{2^k+1} \oplus x^{2^k+1} = \varphi \tag{1}$$

has either zero or at least two solutions. Let x_1 and x_2 be two different solutions, then

$$\begin{gathered}
(\mathrm{x}_1 \oplus \phi)^{2^k+1} \oplus x_1^{2^k+1} = (x_2 \oplus \phi)^{2^k+1} \oplus x_2^{2^k+1} \\
(\mathrm{x}_1 \oplus \phi)^{2^k} (x_1 \oplus \phi) \oplus x_1^{2^k} x_1 = (x_2 \oplus \phi)^{2^k} (x_2 \oplus \phi) \oplus x_2^{2^k} x_2 \\
x_1^{2^k} x_1 \oplus x_1^{2^k} \phi \oplus x_1 \phi^{2^k} \oplus \phi^{2^k+1} \oplus x_1^{2^k} x_1 = x_2^{2^k} x_2 \oplus x_2^{2^k} \phi \oplus x_2 \phi^{2^k} \oplus \phi^{2^k+1} \oplus x_2^{2^k} x_2 \\
(x_1 \oplus x_2)^{2^k} \phi \oplus (x_1 \oplus x_2) \phi^{2^k} = 0 \\
(x_1 \oplus x_2)^{2^k-1} = \phi^{2^k-1}
\end{gathered}$$

Then $$x_1 \oplus x_2 = \phi(G \backslash \{0\})$$

Where G is subfield of $(GF2^n)$ of order 2^s. Therefore given one solution x_0 of (1) the set of all solutions is $x_0 \oplus \phi G$ of cardinality 2^s. This completes proof.

Proposition 2. $F(x) = x^{-1}$,if $x \neq 0$ and $F(x) = 0$ if $x = 0$, is differentially 4-uniform in $GF(2^n)$

Proof: At first we consider in (F,+). Given $\phi, \varphi \in F$ and $\phi \neq 0$, then

$$(x + \phi)^{-1} + x^{-1} = \varphi \tag{2}$$

If $x \neq 0$ and $x \neq \phi$ then with multiplication (2) with $x(x + \phi)$ at two sides will result

$$x + (x + \phi) = x\varphi(x + \phi)$$
$$\varphi x^2 + \phi\varphi x + \phi = 0 \quad (3)$$

which has at most two solutions. If both $x = 0$ and $x = \phi$ are solution to (2) (we define that $\frac{1}{0} = 0$), so $\phi = \frac{1}{\varphi}$. If this is a case, so (3) becomes

$$x^2 + \phi x + \phi^2 = 0 \quad (4)$$

that give two more solutions to (2)

Now we consider special case in $F = GF(2^n)$ to solve (4). We rearrange (4) becomes

$$(x^2)^2 = (\phi x \oplus \phi^2)^2$$
$$x^4 = \phi^2 x^2 \oplus \phi^4$$
$$x^4 \oplus \phi^2(\phi x \oplus \phi^2) = \phi^4$$
$$x(x^3 \oplus \phi^3) = 0$$

which has only two solutions: $x = 0$ or ϕ if $gcd(3, 2^n - 1) = 1$ or if n is odd. If n is even, then 3 divides $2^n - 1$. If we set d = $\frac{1}{3}(2^n - 1)$, then there are two more solutions, $x = \alpha^{1+d}$ or $x = \alpha^{1+2d}$. If we use the function that has δ low enough for S-boxes, then we can be sure that our cipher can against differential attack. For example, Advanced Encryption Standard (AES) [11] uses inversion function in $GF(2^8)$, as well as Camellia that became one of the winner in NESSIE.

DES used S-Boxes that is hidden from public in its design. Several researchers try to reveal the design of DES. Although DES apparently don't use the known functions have low differentially uniform mappings, however DES has strength enough to against differential attack. But in academic world, attack that needs complexity lower than exhaustive search is considered as a successful attack, although this attack may not be practical. So, the differential attack of DES that needs about 2^{47} chosen plaintext is considered as success.

2.2 Linear Cryptanalysis

Linear attack is proposed by Matsui [10]. For any given $\Gamma w_i, \Gamma w_{i+1} \in GF(2^m)$, where w_i is the m-bit input of S-Box and w_{i+1} is m-bit output, Γw_i is input mask value, and Γw_{i+1} is output mask value, the linear probabilities of $s_j - box$: $GF(2^m) \rightarrow GF(2^m)$ are defined as :

$$P_{lin}[w_i.\Gamma w_i = s_j(w_i).\Gamma w_{i+1}] = \frac{\#\{w_i \in GF(2^m) | w_i.\Gamma w_i = s_j(w_i).\Gamma w_{i+1}\} - 2^{m-1}}{2^m}$$

The linear attack leads to a new design criterion in cryptographic design : nonlinearity. Nonlinearity measures the a function (Hamming) distance to linear (affine)

functions. It measures how well a function under consideration may be linearly approximated. Linear attack searches the best linear approximation, called effective linear expression of an cipher. At the start, we have to find good approximations to the nonlinear component of cipher (these are usually S-Boxes), and then we extend these approximations to the round function and then to the other rounds of cipher.

Meier [12] defined nonlinearity of function f as $N_f = \min_{i=0,1,2,...2^{n+1}-1} d(f, \vartheta_i)$ where $\vartheta_0, \vartheta_1, ...\vartheta_2^{n+1} - 1$ denote all affine functions, so that the first half consists of linear functions ordered according to the relation $\vartheta_i = f_{\alpha_i}$ for all i = 0,1,2,...,$2^n - 1$ and second half consists of the complements of the function in the first half. Or $\vartheta_i = f_{\alpha_i}$ for all $i = 2^n, 2^n + 1, ...2^{n+1} - 1$

Lemma 1[13]. Let f, g be functions with truth table of the real-valued function ζ_f, ζ_g, respectively. Then $d(f, g) = 2^{n-1} - \frac{1}{2} \langle \zeta_f, \zeta_g \rangle$

Proof. Let V_n be the set of all n-tuples of elements of the field $GF(2)$, d is hamming distance, w is hamming weight and $\langle f, g \rangle$ is inner product of two functions. We have

$$\begin{aligned} \langle \zeta_f, \zeta_g \rangle &= \sum_{x \in V_n} (-1)^{(f+g)(x)} \\ &= 2^n - 2w(f + g) \end{aligned}$$

since w(f+g)=d(f,g), the result follows.

Lemma 2 [14]. For any function f, its nonlinearity N_f satisfies the relation $N_f \leq 2^{n-1} - 2^{\frac{n}{2}-1}$

Proof. Let $H_n = \begin{bmatrix} l_0 \\ l_1 \\ \vdots \\ l_{2^n-1} \end{bmatrix}$ denote the Sylvester-Hadamard matrix of order 2^n where l_i denotes the i-th row of H_n for $i = 0, 1, ...2^n - 1$. Since H_n is a symmetric matrix, we has

$$\zeta_f.H_n = (\langle \zeta_f, l_0 \rangle, \langle \zeta_f, l_1 \rangle, ... \langle \zeta_f, l_{2^n-1} \rangle) \tag{5}$$

and

$$(\zeta_f.H_n)(\zeta_f.H_n)^T = H_n.H_n^T \zeta_f.\zeta_f^T = 2^n \zeta_f.\zeta_f^T = 2^{2n} \tag{6}$$

Computing the left hand side of (6) using (5) we get :

$$(\zeta_f.H_n)(\zeta_f.H_n)^T = \sum_{j=0}^{2^n-1} \langle \zeta_f, l_j \rangle^2$$

By combining these result, we obtain that

$$\sum_{j=0}^{2^n-1} \langle \zeta_f, l_j \rangle^2 = 2^{2n} \tag{7}$$

From (7), there exists a j_k satisfying $0 \leq j_k \leq 2^n - 1$ such that $\langle \zeta_f, l_{j_k} \rangle^2 \geq 2^n$. So, $\langle \zeta_f, l_{j_k} \rangle \geq 2^{\frac{n}{2}}$ or $\langle \zeta_f, l_{j_k} \rangle \leq -2^{\frac{n}{2}}$

- If $\langle \zeta_f, l_{j_k} \rangle \geq 2^{\frac{n}{2}}$ is true then by lemma 1, the distance between affine functions ϑ_{j_k} and a function f is $d(f, \vartheta_{j_k}) \leqslant 2^{n-1} - 2^{\frac{n}{2}-1}$
- If $\langle \zeta_f, l_{j_k} \rangle \leq -2^{\frac{n}{2}}$ holds, then $\langle \zeta_f, -l_{j_k} \rangle = \langle \zeta_f, l_{j_k+2^n} \rangle \geqslant 2^{\frac{n}{2}}$ where $l_{j_k} + 2^n = -l_{j_k}$. And by lemma 1, $d(f, \vartheta_{j_k+2^n}) = d(f, \bar{\vartheta}_{j_k}) \leq 2^{n-1} - 2^{\frac{n}{2}-1}$

With j_k, both $\langle \zeta_f, l_{j_k} \rangle$ and $\langle \zeta_f, -l_{j_k} \rangle$ are the largest among all j's for j=0,1,...,$2^{n+1}-$ 1. This makes that either $d(f, \vartheta_{j_k})$ or $d(f, \bar{\vartheta}_{j_k})$ is the smallest among all affine functions $\vartheta_0, \vartheta_1, ...\vartheta_{2^{n+1}-1}$. So $N_f \leq 2^{n-1} - 2^{\frac{n}{2}-1}$

Matsui [15] describes an 2R linear attack on 16-round DES, using a 14-round linear relation with probability $\frac{1}{2} - 1,19\text{x}2^{-21}$, require about 2^{43} known plaintext and ten active S-boxes, covering rounds 2 till 15 with success probability about 97,7%:

$$R_2[7,18,24] \oplus L_{15}[7,18,24,29] \oplus R_{15}[15] = K_2[22] \oplus K_3[44] \oplus K_4[22] \oplus K_6[22] \oplus K_7[44]$$

$$\oplus K_8[22] \oplus K_{10}[22] \oplus K_{11}[44] \oplus K_{12}[22] \oplus K_{14}[22] \quad (8)$$

DES designers had known about differential attack however they didn't know about linear attack, so linear attack can cryptanalyse DES more successful.

3 AES-Like Ciphers

Several proposals of strategies and methods to make block ciphers immune to cryptanalysis, have been suggested by many researchers. One of the most famous methods is wide trail strategy that is proposed by Daemen in his Doctoral thesis [16]. This strategy is used in Rijndael Algorithm which is chosen as the winner of Advanced Encryption Standard (AES). Before Rijndael, some algorithms such as Shark, Square and BKSQ also used this strategy. After the success of Rijndael as the winner of AES, several ciphers are design with this strategy, like Anubis and Khazad. Crypton and Rijndael are derived from Square. Camellia (was selected as a recommended cryptographic primitive by the EU NESSIE) and Aria use the similar S-Box with Rijndael, that is inversion function in $GF(2^8)$.

In the Wide Trail Strategy, the round transformation of a block cipher consists of different invertible transformation, each with its own functionality. The first transformation is nonlinear layer which is usually performed by S-Boxes. The second is linear layer that ensures that after a few rounds, all the output bits depend on all the input bits. This strategy proposed that we have to make transformations are independent for each other, so the different components can be specified quite independently from one another. The third layer is round key addition.

The AES-like Ciphers are designed to immune against conventional cryptanalysis like differential, linear, truncated differential, square, related key, interpolation

attack, and so forth. So, until now, it isn't known how to attack them with conventional attacks. These ciphers use simple algebraic expressions that have some advantages and disadvantages. The advantages are that the cipher has component with good properties that can against conventional attack, and the cipher can be implemented easily and efficiently at various platforms. The disadvantage is the cipher can be represented by simple algebraic expression. This fact motivates a new attack that is called algebraic attack.

Due to the design criteria of Rijndael, it can be expressed with elegant equations in several ways. Ferguson et al [18] derive a closed formula for Rijndael that can bee seen as a generalization of continued fractions. Any byte of the intermediate result after 5 rounds can be expressed as follows:

$$x = K + \sum \frac{C_1}{K^* + \sum \frac{C_2}{K^* + \sum \frac{C_3}{K^* + \sum \frac{C_4}{K^* + \sum \frac{C_5}{K^* + p_*^*}}}}} \tag{9}$$

Here every K is a byte depending on several bytes of the expanded key, each C_i is a known constant and each * is a known exponent. A fully expanded version of (9) has 2^{55} terms. The first equation would express the intermediate variables after 5 rounds as function of the plaintext bytes. The second equation would cover rounds 6-10 (For AES-128) by expressing the same intermediate variables as a function of the ciphertext bytes. Combining both equations would result in an equation with 2^{26} unknowns. It is unknown what a practical algorithm to solve this type of equations would look like.

Courtois and Pieprzyk [20] showed that Rijndael can be written as an overdefined system of multivariate quadratic equations (MQ). For 128-bit Rijndael, the problem of recovering the secret key from one single plaintext can be written as a system of 8000 quadratic equations with 1600 binary unknowns. Thus the security of Rijndael requires that there are no efficient algorithms for solving such systems. If these equations can be solved, it is a major revolution in cryptanalysis, that are all classical attack require a number of plaintexts/ciphertexts that is very big.

Biryukov and Canniere[19] compare systems of multivariate polynomials, which completely define several block ciphers. Because it isn't known the most efficient way of solving such systems of equation, they give criterion based on some intuitive assumptions to construct systems of equations :

1. Minimize the total number of free terms (free monomials). In order to achieve this, try to:
   - Minimize the degree of the equations. This reduces the total number of possible monomials
   - Minimize the difference between the total number of terms and the total number of equations.

2. Minimize the size of individual equations

Their results are shown at table 1. For Rijndael with 128 bits key, there are 3296 variables, 6296 equations, and 16096 terms that have to be solved to get the keys. Until now it isn't known how to solve this such equation.

Table 1: A Comparison of the complexities of the systems of equation in GF(2)

	Khazad	Misty1	Kasumi	Camellia-128	Rijndael-128	Serpent-128
Variables	**6464**	**3856**	**4264**	**3584**	**3296**	**16640**
Linear eq.	1664	2008	2264	1920	1696	8320
Nonlinear eq	6000	1848	2000	4304	4600	9360
Equations	**7664**	**3856**	**4264**	**6224**	**6296**	**17680**
Linear terms	6464	3856	4264	3584	3296	16640
Quadratic terms	10800	6832	7952	11520	12800	13000
Terms	**17264**	**10688**	**12216**	**15104**	**16096**	**29640**

4 The Design of Block Cipher

In this section we will describe some components of block cipher that can against cryptanalysis without lower performance in at various platform. Many of block ciphers are considered immune to classical attack, but many of them have no good performance enough when they are implemented at various platforms. For example, IDEA, RC2, AES finalists and Blowfish. IDEA uses three incompatible types of arithmetic operations and Blowfish use S-Box dependent-key. After a long time, no attack can be done to IDEA and Blowfish in open literature. But IDEA and Blowfish aren't suitable to be implemented in 8-bit smartcard, because IDEA uses 16-bit word for arithmetic operations and Blowfish needs huge memories to be implemented.

4.1 Linear Component

Linear component is used to give avalanche effect of cipher. Daemen [16] defines *branch number* to quantify the effect of an invertible linear mapping θ. If $w_h(a)$ is Hamming weight of a, and this component can be bits or elements from $GF(2^n)$ then

$$B(\theta) = \min_{a \neq 0}(w_h(a) + w_h(\theta(a)))$$

B gives a measure for the worst case diffusion. Note that $w_h(a) \leq$ n, for every choice of θ; if $w_h(a) = 1$, this implies that $B \leq$ n + 1. We call an invertible linear mapping for which $B =$ n + 1 optimal. We can use MDS (Maximal Distance Separable) codes to achieve this. A linear code C of length n, dimension k, and with minimum distance d between the codewords is denoted as an (n,k,d)-code. Codes with d = n - k + 1 are called MDS codes.

Proposition 3[17]. Let C be a (2n,n,n+1)-code over the Galois field $GF(2^m)$. Let G_m be the generator matrix of C in echelon form:

$$G_m = [I_{nXn} A_{nXn}]$$

Then C defines an optimal invertible linear mapping γ:

$$\gamma : GF(2^m)^n \rightarrow GF(2^m)^n : X \rightarrow Y = A.X$$

Proof: First we show that $B = n + 1$. The definition of B gives :

$$\begin{aligned} B(\gamma) &= \min_{X \neq 0}(w_h(X) + w_h(\gamma(X))) \\ &= \min_{X \neq 0}(w_h(X, \gamma(X))) \end{aligned}$$

The minimum of the Hamming weights of the non-zero codewords is by definition equal to $d = n + 1$. Then, to prove that γ is invertible, we consider that γ isn't invertible. So we have two vectors j, k such that $j \neq k$ and $\gamma(j) = \gamma(k)$. Since γ is linear, then $\gamma(j - k) = \gamma(j) - \gamma(k) = 0$. And then

$$B = \min_{l \neq 0}(w_h(l) + w_h(\gamma(l))) \leqslant w_h(j - k) + w_h(\gamma(j - k)) \leqslant n + 0 < n + 1$$

This is in contradiction with $B = n + 1$. So, γ is invertible. The branch number of a diffusion layer is very important. The low diffusion of the DES is used by Matsui[15] to construct a linear approximation of the cipher with high correlation.

4.2 Nonlinear Component

To get confusion, we can use inversion mapping in $GF(2^n)$ like S-Box which is used in SHARK cipher. To against algebraic attack, we can add linear transformation with different fields to this inversion function, such as Rijndael(S-Box in Rijndael is composed of $GF(2)$ and $GF(2^8)$). But, there is disadvantage with this design. The S-Box in Rijndael isn't involutional, so we must make two different algorithms to encrypt and decrypt. To overcome this problem we can use involutional S-Box in SPN cipher (Feistel don't need invertible F function), like Khazad and Anubis which have been submitted to NESSIE project [21]. Motivations behind such extensive use of involutional components is twofold: efficient implementation and equal security of encryption and decryption. But, involutional S-Box often cause implementation inefficiently and resistance to classical attack become suboptimal.

By considering that there isn't solution for algebraic attack for AES-like Ciphers, I prefer using inversion mapping in $GF(2^n)$ to compose nonlinear component. We can keep on using inversion function and linear transformation with different n, which is located before and after inversion function like Camellia cipher, to make implementation efficiently and resistance against classical attack (it is hoped that this way can make cipher immune against algebraic attack), but we must use new structure like in Aria Cipher (the future Korean national standard block-cipher) to make involutional cipher [22]

5 Conclusion

In this section we will conclude what we have discussed in section before :

- Linear and nonlinear component in block cipher have very important role to against all attaks. Linear component have a function to achieve diffusion, whereas nonlinear component have a function to give confusion.
- Rijndael is designed to against conventional attacks and to be implemented at various software and hardware efficiently, flexible and easily. Therefore, Rijndael only use simple components, however, these components cause Rijndael easily to be represented in algebraic expression in several ways. Furthermore, This fact make researches for constructing equations to define Rijndael and solving the equations to get the keys. Even though there are many researches in algebraic attack, at the time of writing this paper, no shortcut attacks on Rijndael (AES) have been reported openly.
- It is a easy to make complicated Block Cipher to obtain secure cipher (this type of cipher will run slowly at various platform and will be implemented inefficiently), but it is more difficult to assess its security, and more difficult to make cipher that can against every known attack and can be implemented efficiently at various software and hardware.

References

[1] Paulo S.L.M Barreto & Vincent Rijmen (2000), *The Anubis block Cipher*, Primitive submitted to NESSIE .

[2] B. Preenel,V. Rijmen & A. Bosselaers (1998), *Recent Developments in the Design of Conventional Cryptographic Algorithms*, LNCS 1528, pp 105-130, Springer-Verlag.

[3] http://www.eff.org/Privacy/Crypto/Crypto_misc/DESCracker/ last access : July 26, 2005

[4] Francois Koeune et al(2002), *A FPGA Implementation of the Linear Cryptanalysis*, FPL , Volume 2438 of Lecture Notes in Computer Science, pages 845-852, Springer-Verlag.

[5] K Nyberg (1992)& L.R. Knudsen, Provable Security Against Differential Cryptanalysis, *Proceedings of Crypto '92*

[6] Man Young Rhee(1994),*Cryptography and Secure Communications*,Mc Graw-Hill

[7] Kaisa Nyberg(1994), Differentially uniform mapping for Cryptography, *Advances in Cryptology, Proc. Eurocrypt'93,LNCS765* T. Helleseth, Ed.,Springer-Verlag, pp 439-444.

[8] K. Aoki et al(2000), *Camellia: A 128 Bit Block Cipher Suitable for Multiple Platforms*. NTT and Mitsubishi Electric Corporation

[9] E. Biham and A. Shamir(1993), *Differential Cryptanalysis of the Data Encryption Standard* Springer-Verlag, Berlin, Heidelberg, New York

[10] M Matsui(1994), Linear Cryptanalysis Method for DES Cipher, T. Helleseth, Editor , *Advances in Cryptology, -EUROCRYPT93*, Volume 765 of Lecture Notes in Computer Science, pp. 386-397. Springer-Verlag, Berlin, Heidelberg, New York. (A preliminary version written in Japanese was presented at SCIS93-3C).

[11] J. Daemen, V. Rijmen (2002). *The Design of Rijndael : AES The Advanced Encryption Standard*, Springer-Verlag.

[12] W. Meier and O. Staffelbach. (1990), *Nonlinearity criteria for cryptographic functions*, LNCS Springer-Verlag.

[13] J. Seberry, X. M. Zhang, and Y. Zheng(1995).Nonlinearity and propagation characteristics of balanced boolean functions. *Information and Computation* 119(1) 1-13.

[14] J. Seberry, X. M. Zhang(1993).Highly nonlinear 0-1 balanced boolean functions satisfying strict avalanche criterion. *Advances in Cryptology - AUSCRYPT92*, Lecture Notes in Computer Science. Springer-Verlag, Berlin, Heidelberg, New York, 718 145-155.

[15] M. Matsui(1994). The First Experimental Cryptanalysis of the DES. In Y.G. Desmedt, editor, *Advanced in Cryptology, Crypto'94,* LNCS 839, pages 1-11. Springer-Verlag.

[16] J. Daemen (1995), Cipher and Hash Function Design, Strategies based on linear and differential cryptanalysis, *Doctoral Dissertation*. K.U. Leuven

[17] V. Rijmen, J. Daemen, B. Preneel, A. Bosselaers, and E. De Win(1996), The cipher SHARK, in *Fast Software Encryption: Third International Workshop*, D. Gollman, ed., Springer-Verlag, Berlin, pp. 99112.

[18] Niels Ferguson et al(2001). A simple algebraic representation of Rijndael. In Serge Vaudenay and Amr M. Youssef, editors, *Proceeding of Selected Areas in Cryptography- SAC'01*, number 2259 in LNSC, pages 103-111. Springer Verlag.

[19] Alex Biryukov and Christophe De Cannire(2003). Block Ciphers and Systems of Quadratic Equations. In Thomas Johansson,editors. *Fast Software Encryption: 10th International Workshop, FSE 2003*, Lund, Sweden.

[20] Nicolas T. C. and Josef Pieprzyk (2002). Cryptanalysis of Block ciphers with overdefined systems of equations. IACR eprint server. Available at *http://eprint.iacr.org/2000/044/*

[21] NESSIE project, New European Schemes for Signatures, Integrity and Encryption. Homepage-available at http://cryptonessie.org

[22] Kwon et al (2003). *New Block Cipher ARIA.* National Security Research Institute, Specification of ARIA.

Yusuf Kurniawan: Ph D student at Department of Informatics, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia.
Department of Informatics, Universitas Pasundan, Jl Setiabudi 193 Bandung, Indonesia
E-mail: if33998@students.if.itb.ac.id

Adang Suwandi Ahmad: Professor at Department of Electronic Engineering, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia.
E-mail: asaisrg@yahoo.com

M. Sukrisno Mardiyanto : Department of Informatics,Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia.
E-mail: sukrisno@informatika.org

Iping Supriana S : Department of Informatics, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia.
E-mail: iping@informatika.org

Sarwono Sutikno : Department of Electronic Engineering, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia.
E-mail: ssarwono@ieee.org

On the new results of expanding super edge-magic total graphs*

I W. Sudarsana[1], E. T. Baskoro[2], D. Ismaimuza[1], H. Assiyatun[2]

[1] Department of Mathematics, Tadulako University
Jalan Sukarno-Hatta Palu, Indonesia
isudarsana203@yahoo.com, dasaismaimuza@yahoo.co.uk
[2] Department of Mathematics, Institut Teknologi Bandung
Jalan Ganesha 10 Bandung, Indonesia
{ebaskoro, hilda}@dns.math.itb.ac.id

Abstract. We denote by (p,q)-*graph* G a graph with p vertices and q edges. *An edge-magic total labeling* on a (p,q)-graph G is a bijection $\lambda : V(G) \cup E(G) \to \{1,2,...,p+q\}$ with the property that, for each edge xy of G, $\lambda(x)+\lambda(xy)+\lambda(y) = k$, for a fixed positive integer k. Moreover, λ is *a super edge-magic total labeling* if it has the property that the vertex labels are the integers $1,2,\ldots,p$ the smallest possible labels. *A* (p,q)-*graph* G is called edge-magic total (super edge-magic total) if there exists an edge-magic (super edge-magic, respectively) total labeling of G. In this paper, we study the properties of super edge-magic total graphs. We give some further necessary conditions for such graphs. Based on this condition we provide some theorems how to construct new super edge-magic total graphs from the old. In particular, we construct the new graph from two super edge-magic total graphs.

Key words and phrases: *edge-magic total labeling, super edge-magic total labeling, graph, dual labeling, the magic constant*

1 Introduction

In this paper we consider finite undirected graphs without loops and multiple edges. $V(G)$ and $E(G)$ stand for the vertex set and edge set of graph G, respectively. We denote by $K_{1,n-1}$ a star on n vertices. The general references for graph-theoretic ideas can be seen in [8] and [14].

We denote by (p,q)-*graph* G a graph with p vertices and q edges. *An edge-magic total labeling* on a (p,q)-graph G is a bijection $\lambda : V(G) \cup E(G) \to \{1,2,...,p+q\}$ with the property that, for each edge xy of G, $\lambda(x) + \lambda(xy) + \lambda(y) = k$, for a fixed positive integer k. Moreover, λ is

* Supported by Hibah Pekerti DP3M-DIKTI Indonesia, Contract Number: 337/P4T/DPPPM/HPTP/IV/2004

a super edge-magic total labeling if it has the property that the vertex labels are the integers $1, 2, \ldots, p$, the smallest possible labels.

A (p, q)-graph G is called edge-magic total (super edge-magic total) if there exists an edge-magic (super edge-magic, respectively) total labeling of G. We shall follow [13] to call $\lambda(x) + \lambda(xy) + \lambda(y)$ the *edge sum* of xy, and k the *magic constant* of the graph G. Edge-magic total graphs were first discussed by Kotzig and Rosa [10] (under the name of graph with magic valuation). Super edge-magic graphs were introduced by Enomoto et al. [3].

A number of classification studies on edge-magic total (resp. super edge-magic total) graphs has been intensively investigated. In [10] and [7] it is proved that every cycle C_n and caterpillar are edge-magic total. Kotzig and Rosa [11] show that no complete graph K_n with $n > 6$ is edge-magic total and give some edge-magic total labeling for K_n, $3 \leq n \leq 6$, $n \neq 4$. Wallis et al. [13] showed that all paths P_n and all n-suns are edge-magic total. In [15] and [16] are exhibited the relationships between super edge-magic total labelings and other well studied classes of labelings (harmonius, cordial, graceful and antimagic).

Some conjectures remain open, namely that all trees are edge-magic total [10] (super edge-magic total [3]) and all wheels W_n are edge-magic total if $n \not\equiv 3 \pmod 4$ [2]. Enomoto et al. [3] have checked (by a computer) that their conjectures are true for all trees with less than or equal 16 vertices and wheels W_n for $n \leq 30$. Philips et al. [17] showed that a wheel W_n for $n \equiv 0$ or $1 \pmod 4$ is edge-magic total. Slamin et al. [18] proved that for $n \equiv 6 \pmod 8$, every wheel W_n has an edge-magic total labeling.

For disconnected graphs, in [10] there is proved that nP_2 is super edge-magic total if and only if n is odd. Kotzig [9] showed that if G is a trichromatic graph and G is edge-magic total then a disjoint union of n (n odd) identical copies of G is also edge-magic total. In particular, if $G = P_3$ then graph nP_3, for n odd, is edge-magic total. In [19] it is shown that nP_3 is super edge-magic total when n is odd. Yegnanarayanan [19] also conjectured that for all n, nP_3 has an edge-magic total labeling. Baskoro and Ngurah [20] proved that nP_3 is super edge-magic total for n even, $n \geq 4$.

Figueroa-Centeno et al. [4] showed that $P_3 \cup nP_2$ is super edge-magic total for every $n \geq 1$; $P_2 \cup P_n$ is super edge-magic total for every $n \geq 3$; $mK_{1,n}$ is super edge-magic total for m odd and for every $n \geq 1$; and graph mP_n is super edge-magic total for any m and for any odd n. The super edge-magic total characterization of the $nP_3 \cup kP_2$ and $K_{1,m} \cup K_{1,n}$ can be found in [21]. Recently, E.T. Baskoro et al. [1] and I W. Sudarsana et

al. [2] gave some theorem how to construct new super edge-magic total graphs from old ones. More comprehensive information on edge-magic total and super edge-magic total graphs is given in [6].

In this paper, we study super edge-magic total labelings. We derive more results how to contruct a new super edge-magic total labeling from some old ones.

2 The Results

In this section, we give an expansion technique on super edge-magic total graphs. The double star $S_{n,m}$ is a graph formed from two stars $K_{1,n-1}$ and $K_{1,m-1}$ by joining two centers by an edge. Let $S_{n,m}$ has the vertex-set $\{u_1, u_2, ..., u_n, v_1, v_2, ..., v_m\}$ with the edge-set $\{u_1u_i : i = 2, ..., n\} \cup \{v_1v_j : j = 2, ..., m\} \cup \{u_1v_1\}$.

2.1 Further Necessary Conditions

Let us start with a necessary and sufficient conditions for a graph of being super edge-magic total in the following lemmas.

Lemma 1. *[4] A (p, q)-graph G is super edge-magic total if and only if there exists a bijective function $f : V(G) \to \{1, 2, ..., p\}$ such that the set $S = \{f(u) + f(v) : uv \in E(G)\}$ consists of q consecutive integers. In such a case, f extends to a super edge-magic total labeling of G with the magic constant $k = p + q + s$, where $s = min(S)$ and $S = \{k - (p + 1), k - (p + 2), \cdots, k - (p + q)\}$.*

Furthermore, in order to know what possible values of the magic constant for graph G to be super edge-magic total, E.T. Baskoro et al. [1] proved the following lemma.

Lemma 2. *[1] Let a (p, q)-graph G be super edge-magic total. Then, the magic constant k of G satisfies $p + q + 3 \leq k \leq 3p$.*

Corollary 1 *If k is the magic constant of a tree with p vertices then $2p + 2 \leq k \leq 3p$. Furthermore, the magic constant of a (p, q)-graph G with c components ranges between $2p - c + 3$ to $3p$.*

Next, we introduce the property of edge-magic total labelings.

Lemma 3. *Let a (p, q)-graph G be an edge-magic total. Let λ be an edge-magic total labeling of G with the magic constant k. Let $n \geq 0$. Then, the labeling λ_1 defined:*

$$\lambda_1(x) = \lambda(x) + n, \ \forall x \in V(G), \text{ and}$$
$$\lambda_1(xy) = \lambda(xy) + n, \ \forall xy \in E(G)$$

has the magic constant $k_1 = k + 3n$.

Proof. Let $uv \in E(G)$. Then,
$k_1 = \lambda_1(u) + \lambda_1(uv) + \lambda_1(v)$

$= \lambda(u) + n + \lambda(uv) + n + \lambda(v) + n$
$= \lambda(u) + \lambda(uv) + \lambda(v) + 3n$
$= k + 3n$

□

The labeling λ_1 is called *translation labeling of* λ on G with distance n. Moreover, λ_1 is a *translation super edge-magic total labeling* if it is obtained from the labeling λ is a super edge-magic total.

Lemma 4. *Let* $n \geq 0$. *A* (p, q)*-graph* G *is super edge-magic total if and only if there exists a bijective function* $f_n : V(G) \to \{n+1, n+2, ..., n+p\}$ *such that the set* $S = \{f_n(u)+f_n(v) : uv \in E(G)\}$ *consists of* q *consecutive integers. In such a case,* f_n *extends to a translation super edge-magic total labeling of* G *with the magic constant* $k_n = p+q+n+s$, *where* $s = min(S)$ *and* $S = \{k + 2n - (p + 1), k + 2n - (p + 2), \cdots, k + 2n - (p + q)\}$.

The proof, we can proceed analogously as in the proof lemma 1 (see [4]).

2.2 Further Results of Expanding Super Edge Magic

Theorem 1 *Let* $p \geq 2$. *Let a (p,q)-graph* G *be a super edge-magic total with the magic constant* k *and* $k \geq 2p + 2$. *Let* m *be any positive integer and* $m \leq 3p + 1 - k$. *Then a new graph, formed from* G *by adding exactly* m *distinct vertices* $\{x_1, x_2, ..., x_m\}$ *adjacent to* m *couples vertices* $\{\{z_1, z_2\}_1, \{z_2, z_3\}_2, ..., \{z_m, z_{m+1}\}_m\}$ *of* G *labeled by* $\{\{k - 2p - 1, k - 2p\}, \{k-2p, k-2p+1\}, ..., \{k-2p-2+m, k-2p-1+m\}\}$ *respectively, is super edge-magic total with the magic constant* $k_1 = k + 3m$.

Proof. By Lemma 1 there exists a vertex labeling λ on G such that $S = \{k - (p + q), k - (p + q - 1), ..., k - (p + 1)\}$. Let $\{x_1, x_2, ..., x_m\}$ be the new vertex. Let G_1 be the new graph. In G_1, define a vertex labeling in the following way.

$$\lambda_1(u) = \lambda(u) \text{ for } u \in V(G),$$
$$\lambda_1(x_i) = p + i \text{ for } i = 1, 2, ..., m.$$

Let $S_1 = \{\lambda_1(u) + \lambda_1(v) : \forall uv \in E(G_1)\}$. Clearly,
$S_1 = S \cup \{\cup_{i=1}^{m}\{x_i z_i, x_i z_{i+1}\}\}$, where $x_i z_i$ are the new edges. So, we have $S_1 = \{k-(p+q), ..., k-(p+1)\} \cup \{\cup_{i=1}^{m}\{k-(p+1)+2i-1, k-(p+1)+2i\}\}$. This implies that the new graph is super edge-magic total with the magic constant $k_1 = k-(p+q)+p+m+q+2m = k+3m$ (by Lemma 1). The theorem holds only if the highest label of the vertex adjacent to x_m is less then or equal to p, namely $k-2p-1+m \leq p$. So, $m \leq 3p+1-k$. □

If the magic constant k of a (p,q)-graph G is exactly $2p+2$, then m can be equal to $p-1$ (by Theorem 1). In this case, we add exactly $p-1$ distinct vertices adjacent to $p-1$ couples vertices of G. The resulting graph has the magic constant $k+3p-3$. Thus, the following corollary holds.

Corollary 2 *Let a (p,q)-graph G be a super edge-magic total with the magic constant $k = 2p+2$. Then a new graph, formed from G by adding exactly $p-1$ distinct vertices $\{x_1, x_2, ..., x_{p-1}\}$ adjacent to $p-1$ couples vertices $\{\{z_1, z_2\}_1, \{z_2, z_3\}_2, ..., \{z_{p-1}, z_p\}_{p-1}\}$ of G labeled by $\{\{k-2p-1, k-2p\}, \{k-2p, k-2p+1\}, ..., \{k-p-3, k-p-2\}\}$ respectively, is super edge-magic total with the magic constant $k_1 = k+3p-3$.* □

Theorem 2 *Let (p_1, q_1)-graph G_1 and (p_2, q_2)-graph G_2 be two graph are super edge-magic total with the magic constant $k1$ and k_2. Let $k_1 \geq 2p_1+2$. The new graph G formed from G_1 and G_2 in the following way.*

1. Joining the vertex x_1 of G_2 which has the smallest label to m distinct vertices $z_1, z_2, ..., z_m$ of G_1 labeled by $k_1-2p_1-1, k_1-2p_1, ..., k_1-2p_1-2+m$ respectively.
2. Joining the vertex z_m of G_1 to all vertices of G_2 except x_1.
If $m = 3p_1+2-k_1$ and $k_2 = 2p_2+q_2+1$ then the graph G is super edge-magic total with the magic constant $k = 3p_1+2p_2+q_2+1$.

Proof. By Lemma 1 there exists a vertex labeling λ_1 on G_1 such that $S_1 = \{k_1-(p_1+q_1), k_1-(p_1+q_1-1), \cdots, k_1-(p_1+1)\}$. By Lemma 4 there exists a vertex translation labeling of λ_2 on G_2 with distance p_1 such that $S_2 = \{k_2+2p_1-(p_2+q_2), k_2+2p_1-(p_2+q_2-1), \cdots, k_2+2p_1-(p_2+1)\}$. Define a vertex labeling $\lambda : V(G) \to \{1, 2, ..., p_1, p_1+1, ..., p_1+p_2\}$ in the following way.

$$\lambda(z) = \lambda_1(z) \text{ for } z \in V(G_1),$$
$$\lambda(x) = \lambda_2(x) + p_1 \text{ for } x \in V(G_2).$$

Let $S = \{\lambda(z)+\lambda(x) : \forall zx \in E(G)\}$. Clearly, $S = S_1 \cup \{\lambda(x_1)+\lambda(z_i) : 1 \le i \le m\} \cup \{\lambda(z_m) + \lambda(x_i) : 2 \le i \le p_2\} \cup S_2$, where x_1z_i, z_mx_i are the new edges of G. Clearly, $S = \{k_1-(p_1+q_1), \cdots, k_1-(p_1+1)\} \cup \{k_1-(p_1+1)+1, \cdots, k_1-(p_1+1)+m\} \cup \{k_1-(p_1+1)+m+1, \cdots, k_1-(p_1+1)+m+p_2-1\} \cup \{k_2+2p_1-(p_2+q_2), k_2+2p_1-(p_2+q_2-1), \cdots, k_2+2p_1-(p_2+1)\}$. Since $m = 3p_1 + 2 - k_1$ and $k_2 = 2p_2 + q_2 + 1$, then it can be easily verified that $S = \{k_1 - (p_1 + q_1), \cdots, k_1 - (p_1 + 1)\} \cup \{k_1 - (p_1 + 1) + 1, \cdots, 2p_1+1)\} \cup \{2p_1+2, \cdots, 2p_1+p_2\} \cup \{2p_1+p_2+1, \cdots, 2p_1+p_2+q_2\}$. By Lemma 1, G is super edge-magic total with the magic constant $k = k_1 + 2p_2 + q_2 + m - 1 = 3p_1 + 2p_2 + q_2 + 1$. □

If (p_2, q_2)-graph G_2 is a star then the magic constant k_2 can be equal to $3p_2$. Therefore, the magic constant of the new graph is equal to $3(p_1+p_2)$ (by theorem 2).

Theorem 3 *Let (p_1, q_1)-graph G_1 and (p_2, q_2)-graph G_2 be two graph are super edge-magic total with the magic constant $k1$ and k_2. Let $k_1 \ge 2p_1+2$. If $k_2 = 2p_2 + q_2 + k_1 - 3p_1$ then the new graph, formed from G_1 and G_2 by joining all vertices of G_2 to a vertex u_0 of G_1 labeled by $k_1 - 2p_1 - 1$, is super edge-magic total with the magic constant $k = 2p_2 + q_2 + k_1$.*

Proof. By Lemma 1 there exists a vertex labeling λ_1 on G_1 such that $S_1 = \{k_1 - (p_1 + q_1), k_1 - (p_1 + q_1 - 1), \cdots, k_1 - (p_1 + 1)\}$. By Lemma 4 there exists a vertex translation labeling of λ_2 on G_2 with distance p_1 such that $S_2 = \{k_2 + 2p_1 - (p_2 + q_2), k_2 + 2p_1 - (p_2 + q_2 - 1), \cdots, k_2 + 2p_1 - (p_2 + 1)\}$. Let G be the new graph formed from G_1 and G_2 by joining all vertices of G_2 to a vertex u_0 of G_1. Define a vertex labeling $\lambda : V(G) \to \{1, 2, ..., p_1, p_1 + 1, ..., p_1 + p_2\}$ in the following way.

$$\lambda(u) = \lambda_1(u) \text{ for } u \in V(G_1),$$
$$\lambda(v) = \lambda_2(v) + p_1 \text{ for } v \in V(G_2).$$

Let $S = \{\lambda(u)+\lambda(v) : \forall uv \in E(G)\}$. Clearly, $S = S_1 \cup \{\lambda(u_0)+\lambda(v_i) : 1 \le i \le p_2\} \cup S_2$, where u_0v_i are the edges connecting the graph G_1 with G_2. Clearly, $S = \{k_1 - (p_1 + q_1), \cdots, k_1 - (p_1 + 1)\} \cup \{k_1 - (p_1 + 1) + 1, \cdots, k_1-(p_1+1)+p_2\} \cup \{k_2+2p_1-(p_2+q_2), k_2+2p_1-(p_2+q_2-1), \cdots, k_2+2p_1-(p_2+1)\}$. Since $k_2 = 2p_2+q_2+k_1-3p_1$, then it can be easily verified that $S = \{k_1-(p_1+q_1), \cdots, k_1-(p_1+1)\} \cup \{k_1-(p_1+1)+1, \cdots, k_1-(p_1+1)+p_2\} \cup \{k_1-(p_1+1)+p_2+1, \cdots, k_1-(p_1+1)+p_2+q_2\}$. By Lemma 1, G is super edge-magic total with the magic constant $k = 2p_2+q_2+k_1$. □

Corollary 3 *Let a (p_1, q_1)-graph G_1 be a super edge-magic total with the magic constant $k1$ and $k_1 \ge 2p_1 + 2$. Let T_{p_2} be a tree on p_2 vertices. If*

$k_2 = 3p_2 + k_1 - (3p_1 + 1)$ then the new graph, formed from G_1 and tree T_{p_2} by joining all vertices of T_{p_2} to a vertex u_0 of G_1 labeled by $k_1 - 2p_1 - 1$, is super edge-magic total with the magic constant $k = k_1 + 3p_2 - 1$.

Recently, I W. Sudarsana et. al. [2] have showed that the new graph is super edge-magic total with the magic constant $k_1 + 3p_2 - 1$ for the graph (p_2, p_2)-graph G_2 are trees, especially, path P_n and star $K_{1,n-1}$.

References

1. E. T. Baskoro, I W. Sudarsana and Y. M. Cholily, How to construct new super edge-magic graphs from some old ones, to appear in *MIHMI* (2005).
2. I W. Sudarsana, E. T. Baskoro, D. Ismaimusa and H. Assiyatun, Creating new super edge-magic total labelings from old ones, submitted to *Journal of Combinatorial Mathematics and Combinatorial Computing (JCMCC)*.
3. H. Enomoto, A.S. Llado, T. Nakamigawa and G. Ringel, Super edge-magic graphs, *SUT J. Math.* 34 (1998), 105-109.
4. R. M. Figueroa-Centeno, R. Ichishima and F.A. Muntaner-Batle, On super edge-magic graphs, *Ars Combin.*, 64 (2002) 81-95.
5. Y. Fukuchi, A recursive theorem for super edge-magic labelings of trees, *SUT J. Math.* 36 (2000), 279-285.
6. J. A. Gallian, A dynamic survey of graph labelings, *Electronic Journal Combinatorics*, #DS6, 2003.
7. R. D. Godbold and P. J. Slater, All cycles are edge-magic, *Bull. ICA* 22 (1998).
8. N. Hartsfield and G. Ringel, *Pearls in Graph Theory*, Academic Press, San Diego (1994).
9. A. Kotzig, On Magic valuations of trichromatic graphs, *report of the CRM* (1971), CRM-148.
10. A. Kotzig and A. Rosa, Magic valuations of finite graphs, *Canad. Math. Bull.* 13 (1970), 451-461.
11. A. Kotzig, and A. Rosa, Magic valuations of complete graphs, *Centre de Recherches Mathematiques, Universite de Montreal* (1972), CRM-175.
12. G. Ringel and A. S. Llado, Another tree conjecture, *Bull. Inst. Combin. Appl.* 18 (1996), 83-85.
13. W. D. Wallis, E.T. Baskoro, M. Miller and Slamin, Edge-magic total labelings, *Australasian Journal of Combinatorics* 22 (2000), 177-190.
14. W. D. Wallis, Magic Graphs, Birkhauser, Boston - Basel- Berlin, (2001).
15. R.M. Figueroa-Centeno, R. Ichishima and F.A. Muntaner-Batle, *The place of super edge-magic labelings among other classes of labelings*, Discrete Math. 231 (2001), 153-168.
16. M. Baca, Y. Lin, M. Miller and R. Simanjuntak, *New constructions of magic and antimagic graph labelings*, Utilitas Math. 60 (2001), 229-239.
17. N.C.K. Philips, R.S. Rees and W.D. Wallis, *Edge-magic total labelings of wheels*, Bull. ICA 31 (2001), 21-30.
18. Slamin, M. Baca, Y. Lin, M. Miller and R. Simanjuntak, *Edge-magic total labelings of wheels, fans and friendship graphs,* Bull. ICA 35 (2002), 89-98.
19. V. Yegnanarayanan, *On magic graphs*, Utilitas Math. 59 (2001), 181-204.

20. E.T. Baskoro and A.A.G. Ngurah, *On super edge-magic total labeling of* nP_3, Bull. ICA 37 (2003), 82-87.
21. J. Ivanco and I. Luckanicova, *On edge-magic disconnected graphs*, SUT J. Math. 38 (2002), 175-184.

A Class of Iterated Function Systems That Produces *m*-ary Gray Codes

L. Haryanto and A. J. van Zanten

Department of Applied Mathematics, TUDelft, The Netherlands

Abstract: For some appropriate values $y = (1 - x)/2$ with $x \in [0,1)$, the discrete version of a quite simple *iterated function systems* (IFS), which consists of two functions $f_0, f_1 : [0,1) \mapsto [0,1)$ with $f_0(y) = \frac{y}{2}$ and $f_1(y) = \frac{y}{2} + \frac{1}{2}$, can be used to construct the standard binary Gray code by applying only a finite number of steps of the limit process that produces the corresponding attractor, usually a fractal, of the IFS. We generalize this procedure for a class of IFS's consisting of *m* functions $f_0, f_1, \ldots, f_{m-1} : [0,1) \mapsto [0,1)$ with $f_j(y) = \frac{y}{m} + \frac{j}{m}$, where $m \geq 2$ and $j = 0, 1, 2, \ldots, m-1$. The resulting algorithm can be used to construct both the *m*-ary (standard) Gray codes **G** and the sequence of integers $x = a_0 + a_1 m + \ldots + a_{n-1} m^{n-1}$ corresponding to the *m*-ary words $x = a_0 a_1 \ldots a_{n-1}$ of **G** simultaneously, i.e. it solves the index problem of the constructed *m*-ary Gray code as well.

Keywords: *m*-ary Standard Gray codes, iterated function systems (IFS).

TRANSFORMED BINARY PULSE EXCITATION : A NOVEL SOLUTION FOR EXHAUSTIVE SEARCH IN SPEECH CODING

Riko Arlando Saragih

Universitas Kristen Maranatha, Bandung, Indonesia

Abstract. Code-Excited Linear Predictive (CELP) has been proven to be the promising coder for producing high quality speech at bit rate 4.8 kbps by vector quantizing the excitation sequence using a large stochastic codebook. Even though, a CELP coder must run a huge computational to find the optimal excitation vector (search exhaustively) because it uses a large stochastic codebook (e.g. 1024 as in 4.8 kbps coding). The reduction of the computational load for the CELP coder has been done by using efficient codebook structure, but it still searches the optimal excitation vector exhaustively. Using Transformed Binary Pulse Excitation (TBPE) at the CELP coder significantly reduces the computational load because the optimal excitation vector can be generated by using a smaller codebook with an appropriate transformation matrix.
Key-words: CELP, excitation vector, transformation matrix, search, binary pulse

1 Introduction

In the last two decades, there is an acceleration in speech coding evolution. Speech coders with near-toll quality are available right now for 4.8 kbps until 8 kbps. These low bit rate voice coders are required for future applications such as digital mobile radio telephony, mobile satellite links, and the emerging ISDN service. Voice with good quality for these low bit rate voice coders has become possible with the introduction of a new generation of speech coding techniques known as *analysis-by-synthesis predictive coding*.

The CELP coder has proven to be the most promising candidate for producing high quality speech at bit rates as low as 4.8 kbps, where bit reduction is achieved by vector quantizing the excitation using a large stochastic codebook. The CELP principle is based on the observation that the residual signal, after short-term and long-term prediction, is a noise-like signal and it is assumed, therefore, that the residual can be modeled by a zero-mean Gaussian with slowly-varying power spectrum. A large excitation codebook (usually 1024 entries) is used and the optimum innovation sequence is determined by exhaustively searching the codebook for the address (and the corresponding gain) which minimizes the mean-squared weighted error criterion.

The need to exhaustively search the CELP excitation codebook has resultedin a computationally demanding algorithm which, until recently, hindered the real time implementation of CELP systems. The CELP complexity has been reduced by using efficient codebook structures such as ternary codebooks, overlapping sparse codebooks, algebraic codebooks, and vector sum excitation. Despite the reduction in the complexity offered by the above mentioned efficient codebook structures, the exhaustive search of the excitation codebook has still to be performed.

This paper will describe a new approach for representing the stochastic excitation sequence, called *Transformed Binary Pulse* Excitation *(*TBPE) [1], which significantly reduces the excessive complexity associated with CELP coders. This paper will describe the coder structure and the excitation definition, before deriving an efficient excitation determination procedure which eliminates the need for the exhaustive search of an excitation codebook.

2 CELP Coder Algorithm Description

Like all vector quantization techniques, CELP coding is a frame-oriented technique that breaks a sampled input signal into blocks of samples (i.e., vectors) that are processed as one unit. CELP coding is based on analysis-by-synthesis search procedures, perceptually weighted vector quantization (VQ), and linear prediction (LP). A 10th order LP filter is used to model the speech signal's short-term spectrum, or formant structure. Long-term signal periodicity, or pitch, is modeled by an adaptive code book VQ. The residual from the short-term LP and pitch VQ is vector quantized using a fixed stochastic code book. The optimal scaled excitation vectors from the adaptive and stochastic code books are selected by minimizing a time varying, perceptually weighted distortion measure that improves subjective speech quality by exploiting masking properties of human hearing.

The CELP coder's computational requirements are dominated by the two code book searches. The computational complexity and speech quality of the coder depend upon the search sizes of the code books. Any subset of either code book can be searched to fit processor constraints, at the expense of speech quality.

Fed-Std-1016 [2] uses an 8 kHz sample rate and a 30 ms frame size with four 7.5 ms sub frames. CELP analysis consists of three basic functions: 1) short-term linear prediction, 2) long-term adaptive code book search, and 3) innovation stochastic code book search. CELP synthesis consists of the corresponding three synthesis functions performed in reverse order with the optional addition of a fourth function, called a post filter, to enhance the output speech. The transmitted CELP parameters are the stochastic code book index and gain, the adaptive code book index and gain, and 10 line spectral parameters (LSP). The following description of our CELP coder represents only one of many possible implementations that would comply with Fed-Std-1016.

3 CELP Receiver

The CELP receiver is shown in Figure 1. After achieving frame synchronization, the receiver decodes the CELP parameters, including forward error correction decoding, as specified in Fed-Std-1016. Adaptive smoothing of and stability constraints upon the received CELP parameters are recommended to derive parameters suitable for driving the synthesizer. The receiver synthesizes speech by a parallel gain-shape code excitation of a linear prediction filter. The excitation is formed using a fixed stochastic code book and an adaptive code book. The stochastic code book contains sparse, overlapping, ternary valued, pseudo randomly generated code words. Both code books are overlapped and can be represented as linear arrays, where each 60 sample code word is extracted as a contiguous block of samples. In

the stochastic code book, the code words overlap by a shift of –2 (each code word contains all but two samples of the previous code word and two new samples). The adaptive code book has a shift of one sample or less between its code words. The code words with shifts of less than one sample are interpolated and correspond to non integer pitch delays. The linear prediction filter's excitation is formed by adding a stochastic code book vector, given by index i_s and scaled by g_s, to an adaptive code book vector, given by index i_a and scaled by g_a. The adaptive code book is then updated by this excitation for use in the following sub frame. Thus, the adaptive code book contains a history of past excitation signals, and the delay indexes the code word containing the best block of excitation from the past for use in predicting the present. The number of samples delayed in time is called the pitch delay; which corresponds to an adaptive code book index. For delays less than the sub frame length, a full vector of previous excitation does not exist, so the short vector is replicated to the full vector length to form a code word. Finally, an adaptive post filter may be added to enhance the synthetic output speech.

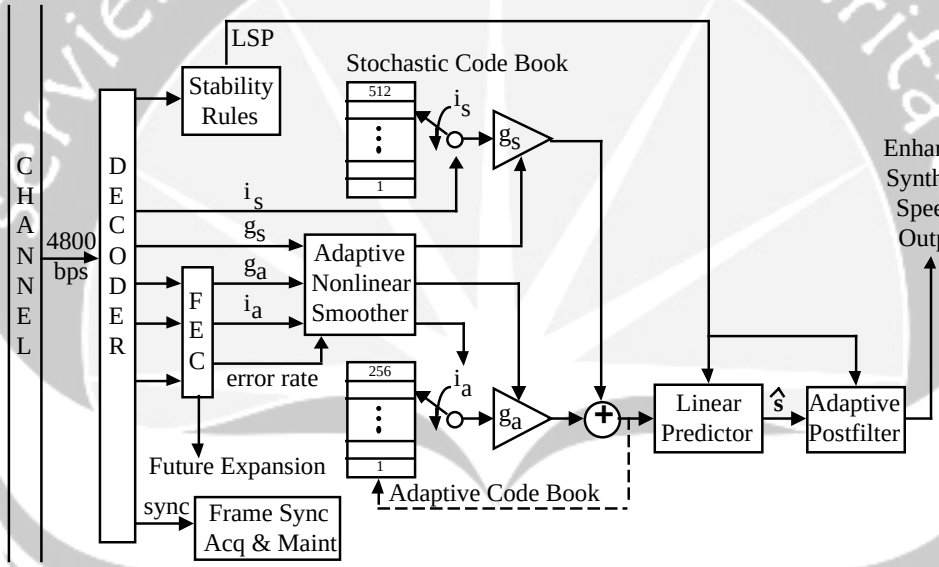


Figure 1. CELP Receiver

4 CELP Transmitter

The CELP transmitter, shown in Figure 2, contains a replica of the receiver's synthesizer (minus the post filter) that, in the absence of channel errors, generates speech identical to the receiver's. This approximation, $\hat{\mathbf{s}}$, is subtracted from the input speech and the difference is perceptually weighted. This perceptually weighted error is then used to drive an analysis-by-synthesis (closed-loop) error minimization gain-shape VQ search procedure. The search procedure finds the adaptive and stochastic code book indices and gains that minimize the perceptually weighted error. The linear prediction filter can be determined by conventional open-loop short-term LP analysis techniques on the input speech. The CELP

parameters, an alternating sync bit, and a future expansion bit are then encoded as specified by Fed-Std-1016 [2] for transmission.

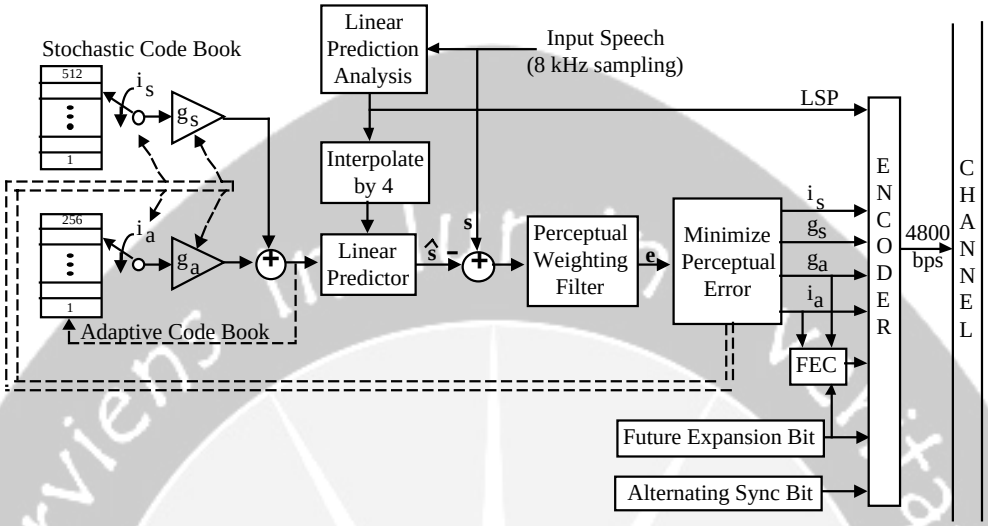


Figure 2. CELP Transmitter

5 Transformed Binary Pulse Excitation Coder

The Transformed Binary Pulse Excited LPC coder structure is shown in the block diagram of Figure 3. The synthesized speech $\hat{\mathbf{s}}(n)$ is found by filtering the excitation signal v(n) through the pitch synthesis filter 1/P(z) and the LPC synthesis filter 1/A(z). The error between the original speech s(n) and the synthesized speech $\hat{\mathbf{s}}(n)$ is then weighted by the perceptual weighting filter W(z) and the excitation signal is determined by minimizing the mean square of the weighted error $e_w(n)$. The LPC synthesis filter is an all-pole time-varying filter of the form :

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}}$$

where $\{a_i\}$ are the LPC parameters (short-term predictor coefficients) and p is the predictor order. The filter parameters are determined outside the analysis-by-synthesis loop by minimizing the mean squared short-term prediction residual, and they are updated every 20-30 ms. Short-term prediction removes the correlation between successive speech samples, while pitch prediction removes the long-term correlation in the speech (corresponding to the pitch periodicity). Assuming one-tap long-term prediction (LTP), the pitch synthesis filter is given by :

$$F(z) = \frac{1}{P(z)} = \frac{1}{1 - G.z^{-\alpha}}$$

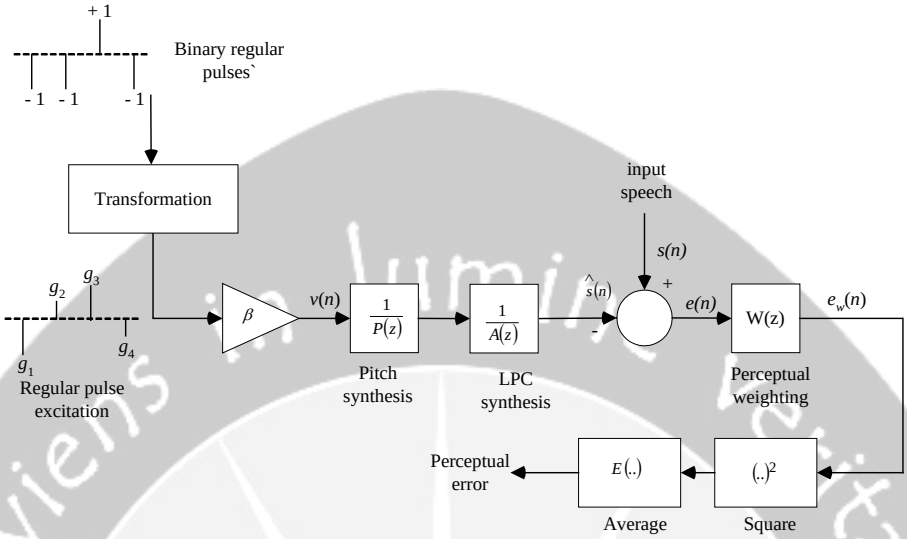


Figure 3. Transformed Binary Pulse Excitation Coder

where G is the LTP gain and a is the LTP delay. Although the pitch synthesis filter parameters can be determined outside the closed optimization loop, a significant speech quality improvement is obtained when the long-term prediction parameters are determined inside the analysis-by-synthesis loop. The selection of the error weighting filter is perceptually motivated. Its role is to shape the spectrum of the quantization noise such that it is concentrated in the frequency regions where the speech has high energy (the formant regions), thereby masking the noise by the speeds signal. The error weighting filter is given by :

$$W(z) = \frac{A(z)}{A\left(z/\gamma\right)} = \frac{1 - \sum_{i=1}^{p} a_i z^{-i}}{1 - \sum_{i=1}^{p} a_i \gamma^i z^{-i}}$$

where γ is a fraction between 0 and 1 which determines the degree by which the error spectrum is de-emphasized in the format regions. A commonly used value of γ is 0.8.

6 Excitation Definition

In the TBPE approach, the excitation signal consists of a number of pseudo stochastic pulses with predefined pulse positions. In an excitation frame of length *N*, suppose that there are *M* nonzero pulses separated by *D* - 1 zeros, where *M* = *N* DIV *D*, and DIV denotes integer division. The excitation vector is given by :

$$v(n) = \beta \sum_{i=1}^{M} g_i \, \delta(n - m_i)$$

where δ(n) is the Kronecker delta, g_i are the pulse amplitudes, m_i are the pulse positions, and β is a scalar gain similar to that which appears in the CELP. As in the RPE approach, there *are D* sets of pulse positions given by :

$$m_i^{(k)} = k + (i - 1)D$$

where *D* is the pulse spacing, and *k* is the position of the first pulse. In RPE coders, the optimum pulse amplitudes and first pulse position are determined by minimizing the mean-squared weighted error between the original and synthesized speech, and this requires solving a set of *M x M* equations *D* times. Further, the pulse amplitudes in RPE are each quantized with 3 bits after scaling by the maximum pulse, or the rm9 value of the pulses, which is quantized with 5 or 6 bits. The large number of bits needed to quantize the pulse amplitudes in RPE makes it difficult to achieve high quality speech below 9.8 kbps. In TBPE approach, the pulses are pseudo-stochastic random variables, similar to the CELP concept, and they are quantized only with one bit per pulse, in addition to the scaling gain β.

The pulse amplitudes g_i, i= l,..., M, are not obtained from a large stochastic codebook as in the CELP approach. Instead, they are determined by the transformation of a binary vector. That is :

$$\mathbf{g} = \mathbf{A}\,\mathbf{b}$$

where b is an M x 1 binary vector with elements -1 or 1, **A** is an *M x M* transformation matrix, and g is the excitation vector containing the pulse amplitudes. The vector b could be one of 2^M possible binary patterns, which means 2^M different excitation vectors can be obtained using the transformation in Equation (6). Thus, this transformation is equivalent to *a* 2^M sized codebook with the need to only store an *M x M* matrix. The equivalent of smaller codebook sizes can be obtained by setting some of the binary pulses to fixed values, or by omitting some of the columns of the matrix A. If the hypothetic codebook size is to be reduced by a factor m, either m pulses in the binary vector are made fixed (say -1), or *m* columns are omitted from the matrix **A** resulting into an M x Q transformation matrix and Q x 1 binary vector, where *Q = M - m.* On the other hand, the equivalent of larger codebooks is obtained by utilizing several transformation matrices. Using *m* different transformation matrices is equivalent to a book of size 2^{M+m}.

For the special case where the transformation matrix is equal to the identity matrix **I**, the excitation pulses are binary with values -1 or 1. In this case the excitation vectors can be viewed as 2^M points regularly distributed over the surface of a sphere in N-dimensional space. When the matrix **A** is orthogonal (i.e. $\mathbf{A}^T\,\mathbf{A} = \mathbf{I}$), the transformation results into a vector containing Gaussian random variables. Generating binary pulses at random and examining the distribution of the pulses g_i resulting from the orthogonal transformation reveals that the variables g_i follow a Gaussian distribution with zero mean and unit variance. Applying an orthogonal

transformation to the binary vectors rotates the vectors without changing their distribution in the N-dimensional space. Both identity and orthogonal transformations exhibited similar performances.

7 Coder Performance

The TBPE coder was evaluated at different bit rates in the range from 4.8 to 8 kbps. Good communications quality speech was obtained at 4.8 kbps and near-toll quality speech was obtained at 8 kbps.

Codebook Population	SEGSNR (dB)
Gaussian	14.03
Sparse	14.06
Ternary	13.81
Overlapping sparse	14.09
Binary regular pulses	13.71
Transformed pulses	13.85

Table 1. SEGSNR SEGSNR for the TBPE and different CELP approaches.

Table 1 shows a comparison between the segmental SNRs of the TBPE and different CELP approaches at 7.8 kbps. The 20 ms speech frame is divided into 5 sub frames of 32 samples length. For the CELP approaches a 9-bit stochastic codebook was used with the gain quantized with 5 bits (4 bits for the magnitude and 1 for the sign), and with the index and gain jointly optimized [3]. In case of TBPE, 8 binary pulses are used (decimation factor of 4) with the first pulse position quantized with 2 bits and the gain with 4 bits (the BPE gain is always positive as the sign information is carried by the pulses themselves). It is clear from the SNR figures in Table 1 that the objective quality of TBPE is very close to CELP. In fact, subjective listening tests did not show any difference in speech quality in either case. When the excitation vectors were not transformed (binary regular pulses), there was slight degradation in speech quality compared to the transformed case.

8 Summary

The TBPE coder has several advantages over the CELP. The main advantage is the significant reduction in the computational complexity. As shown in previous sections, the search of an excitation codebook of size 2^M is reduced to searching a local book of size $M+1$. In case of untransformed vectors, it requires about $M^2 + M$ instructions and for the next M vectors in the local codebook, about M $(M+3)$ instructions are needed to update the term.

9 References

[1] Redwan A. Salami, "Binary Pulse Excitation : A Novel Approach to Low Complexity CELP Coding," *Advances in Speech Coding*, Kluwer Academic Publishers, Boston, 1991, p.145-156.

[2] Fenichel, R., *Federal Standard 1016, Telecommunications: Analog to Digital Conversion of Radio Voice by 4,800 bit/second Code Excited Linear Prediction (CELP)*, National Communications System, Office of Technology and Standards, Washington, DC 20305-2010, 14 February 1991.

[3] Campbell, J., T. Tremain and V. Welch, "The DoD 4.8 kbps Standard (Proposed Federal Standard 1016)," *Advances in Speech Coding*, Kluwer Academic Publishers, Boston, 1991, p.121-133.

RIKO ARLANDO SARAGIH : Department of Electrical Engineering, Universitas Kristen Maranatha, Bandung, Jl. Soeria Soemantri 65 Bandung 40164, Indonesia.
Phone/Fax: +62 +22 2012186 ext 257
E-mail: riko.as@eng.maranatha.edu

THE APPLICATION OF RECURSIVE LEAST SQUARES LINEAR REGRESSION ALGORITHM IN THE DEVELOPMENT OF ANALOG SCALES WEIGHT MEASUREMENT INTERFACE MODULE FOR COMMUNITY HEALTH CENTRES

Trie Maya Kadarina, Soegijardjo Soegijoko

ITB, Bandung, Indonesia

Abstract. This paper describes the application of recursive least squares (RLS) linear regression algorithm in the development of analog scales weight measurement interface module for Community Health Centers (CHCs). The main function of such module is to transfer the weight measurement data of pregnant mothers and children under five years from existing analog mother and child scales to a PC. This interface module consists of both hardware and software modules. The hardware module converts and transfers weight measurement data from analog scales to the PC. We use a shift potentiometer sensor for a weight sensor placed in mechanical part of modified analog scales. The sensor converts spring deflection from the analog scale to resistance value due to weight difference. Then the hardware module converts the resistance value into digital pulses and transfers them to the PC. The C-based software modules record, process, calibrate and present the data. In the PC software module, an identification method using RLS linear regression algorithm is applied to model the weight measurement system. The relationship between inputted reference weight values and the measured periods of the digital pulses yields linear regression constants. Weight values in kg will be obtained based on the linear regression equation from the constants, measured periods of the digital pulses, and the inputted linear regression order (from 1st to 4th order). A number of simulations and clinical experiments have been completed in our laboratory and three different CHCs. The results show that by using the 4th order RLS linear regression, non-linearity of the sensor can be optimized and the minimum error of measurement is obtained. By implementing the RLS linear regression algorithm, the interface module is capable of transferring the weight measurement data from various types of analog scales and shift potentiometer sensors.

Key-words: recursive least squares, linear regression, identification, weight measurement interface

1 Introduction

Community Health Centre (CHC) or *Puskesmas* has an important role in delivering health care to community in Indonesia. Lack of facilities and ineffective of coordination, administration, data recording and reporting systems are some major problems in CHC. In order to improve the quality of community health care in Indonesia, our biomedical engineering laboratory has continuously developed an internet-based community telemedicine system [4]. A CHC equipped with PC-based medical workstation(s) and telecommunication interface, will be able to send or

receive medical information quickly and precisely to/from referral hospitals, health offices or another Community Health Centers.

In general, Indonesia has relatively high percentage of both mother mortality rate (MMR) and children mortality rate (CMR), among ASEAN countries. By implementing the internet-based telemedicine system for CHC, it is expected that in the long run, both MMR and CMR could be reduced through improving the quality of mother and child health care. One of the important physiological parameters that have to be measured to monitor the condition of pregnant mother, fetus and to evaluate nutrient of children under five is body weight. In our laboratory, a low cost microcontroller-based analog scales weight measurement interface modules has been developed. The main function of this interface module is to transfer the weight measurement data of pregnant mothers and children under five years from existing analog mother and child scales to a PC. Block diagram of analog scales weight measurement interface module is shown in figure 1. The interface module consists of both hardware and software modules. The main function of hardware module is to convert and transfer the weight measurement data from analog scales (both mother and child scales). Shift potentiometer sensor is used as a weight sensor placed in mechanical part of modified analog mother and child scales. The sensor converts spring deflection from the analog scale to resistance value due to the weight difference. Then the hardware module converts the resistance value into digital pulses and transfers them to the PC. The C-based software modules record, process and store the data and will be integrated with mother and child database software (in PC) developed using web-based programming. In the PC software module, an identification method using recursive least squares (RLS) linear regression algorithm is applied to model the weight measurement system in accordance with the types and specifications of the analog scales and the shift potentiometer sensors. The weight value in kg will be resulted based on linear regression equation. Further discussion about the application and implementation of the RLS linear regression algorithm in the development of analog scales weight measurement interface module will be describe in this paper.

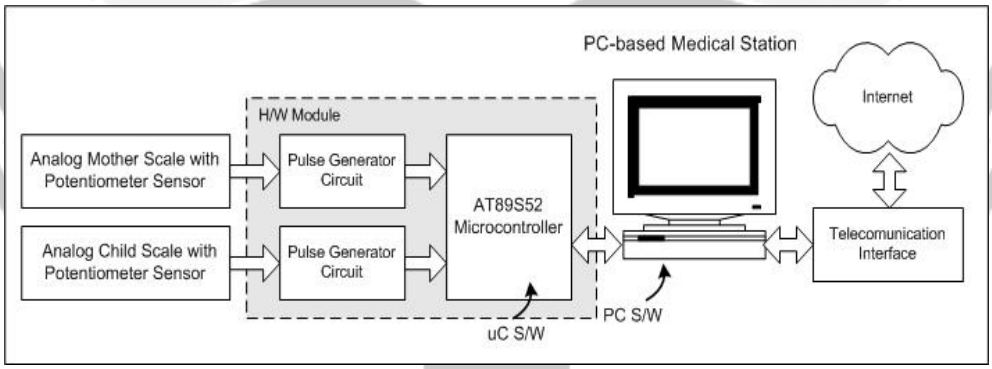


Figure 1 Block diagram of analog scales weight measurement interface modules

The weight measurement data can be recorded automatically, stored in PC, evaluated, further processed, or can be sent to other medical stations (different Community Health Centres, referral hospitals, and health offices). The required minimum PC specifications are as follows: Pentium III – 500MHz processor, 128 MB memory, 20 GB Hard disk, and joystick or serial interface port.

2 Methods

This part will briefly describe the identification method using RLS linear regression algorithm which is applied to model the weight measurement system. The basic principles of the system identification methods is to find the relationships, if any, that exist in a set of variables when at least one is random or unknown, being subject to random (unknown) fluctuations and possible measurement errors. In regression, typically one of the variables, often called the response or dependent variable, is particular interest and is denoted by y. The other variables ϕ_1, ϕ_2,.....ϕ_p, usually called explanatory, or independent variables, or regressors, are primarily used to predict or explain the behavior of y [1]. A linear regression analysis is based on the model:

$$y(t) = \phi^T(t)\theta + e(t) \tag{1}$$

where additive errors e are used and where $\phi^T(t)$ is an n-dimensional vector and θ is parameter estimate. In matrix notation we obtain the estimation model for linear regression is:

$$\mathrm{M}: \quad Y_N = \Phi_N \theta + e \tag{2}$$

By using least squares estimation, the sum of the squared errors between the model output and the observations can be minimized. The least squares solution (optimal parameter estimate) is:

$$\hat{\theta} = (\Phi_N^T \Phi_N)^{-1} \Phi_N^T Y_N \tag{3}$$

Least squares method is not suitable for modeling the weight measurement system. A lot of inputted data sets and a large matrix order are required for obtaining a good estimation results and it can be difficult to organize. Therefore we applied recursive least squares identification algorithm as follows:

$$\hat{\theta}_k = \hat{\theta}_{k-1} + P_k \phi_k \varepsilon_k$$

$$\varepsilon_k = y_k - \phi_k^T \hat{\phi}_{k-1} \tag{4}$$

$$P_k = P_{k-1} - \frac{P_{k-1}\phi_k \phi_k^T P_{k-1}}{1 + \phi_k^T P_{k-1} \phi_k}, \quad Po \text{ given}$$

where $\hat{\theta}_k$ is the parameter estimate, regressor ϕ_k, ε_k is the prediction error and the matrix P_k, which are all evaluated at time k = 1,2,3,... With this algorithm only few data need to be stored and the solution $\hat{\theta}_k$ can be updated.

3 Implementation

The analog scales weight interface module has the following technical specifications: using AT89S52 microcontroller: 8 bit, 8kB EEPROM and serial interface port, to be used with both mother and child scales, shift potentiometer type sensor is required (attached to a moving part of the analog scales), 0 – 120 kg measurement range (for mother) and 0 – 20 kg (for children), ~0.08 kg accuracy range (for mother) or ~0.01 kg (for children). Pulse generator circuits transfer and convert weight measurement data from both mother and child analog scales. These circuits convert resistance value into digital pulses. Then microcontroller circuit records, processes and sends the weight measurement data to PC via the serial port. There are two different types of software modules, namely: the microcontroller software and software for the PC. The microcontroller software module calculates the periods of the digital pulses by using microcontroller timer and sends the results to the PC.

The RLS linear regression algorithm is applied in PC software module using Turbo C software. The analog scales weight measurement interface module completed with PC software module which consists of 2 two main units, i.e.: RLS linear regression constants determination unit and weight measurement unit. The function of RLS linear regression constants determination unit is to identify the weight measurement system in accordance with the types and specifications of the analog scale and the shift potentiometer sensor. The RLS algorithm is applied in this software unit by implementing the equation (4). It is realized in form of a function in C-based programming which is developed from several sub-functions. The application file (.exe) size of the RLS linear regression constants determination unit that consists of the implemented RLS function is 53 kB. We have to run this application software as an initialization step for each type of analog scale and the shift potentiometer.

According to the equation (4) the linear regression constants ($\hat{\theta}_k$) will be resulted from relation between inputted reference weight values (y_k) and periods of digital pulses calculated and sent by the microcontroller to the PC (ϕ_k). Where k = 1, 2, …, to the length of inputted data set (y_k ,ϕ_k). The solution $\hat{\theta}_k$ can be updated as we add a new data set input. The number of the resulted linear regression constant is in proportion to the inputted linear regression order. The resulted constants are stored in a text file, namely constants.txt.

After the constants.txt file is yielded, the weight measurement process can be done. The weight value will be obtained by using the weight measurement software unit based on the linear regression model (equation 1). This unit is also completed with calibration function, to calibrate the sensor at 0 kg. The application file (.exe) size of the weight measurement unit is 39 kB. The weight value will be resulted based on the following model (equation (6)).

$$\text{Weight} = (\phi_p - \Delta_p)^{n}\theta_1 + (\phi_p - \Delta_p)^{n-1}\theta_2 + \ldots \theta_{n+1} \quad \ldots\ldots\ldots\ldots\ldots\ldots (6)$$

where θ is linear regression constant, n is linear regression order , ϕ_p is period of digital pulse value (ticks), and Δ_p is periods of digital pulses shift resulted from calibration process (ticks).

4 Simulation and Experimental Results

A number of simulations, laboratory experiments as well as clinical experiments have been completed in our laboratory and in three different Community Health Centres (CHCs). A series of simulation tests have been conducted by using MATLAB 6.1 software to find the best RLS order. The results show that the best RLS order for both analog mother and child scales is the 4th order (figure 2 and 3). By using the 4th order RLS linear regression, non-linearity of the sensor could be optimized.

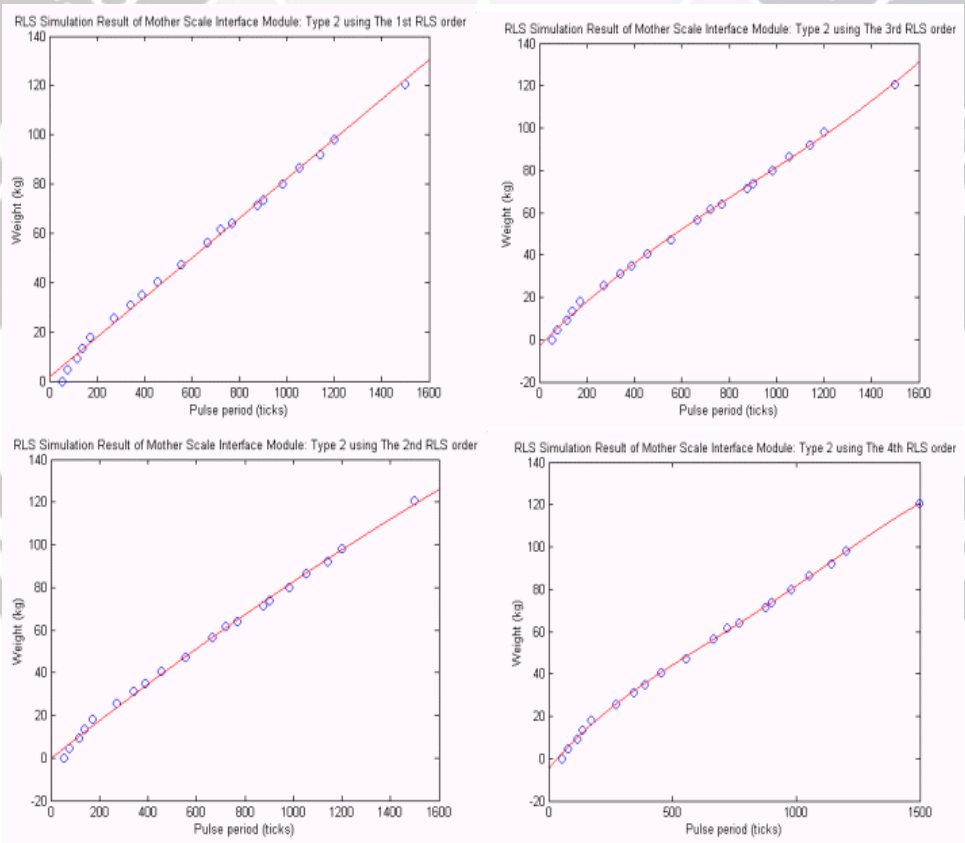


Figure 2 RLS simulation results using the 1st – 4th RLS order analog mother scale weight measurement interface module

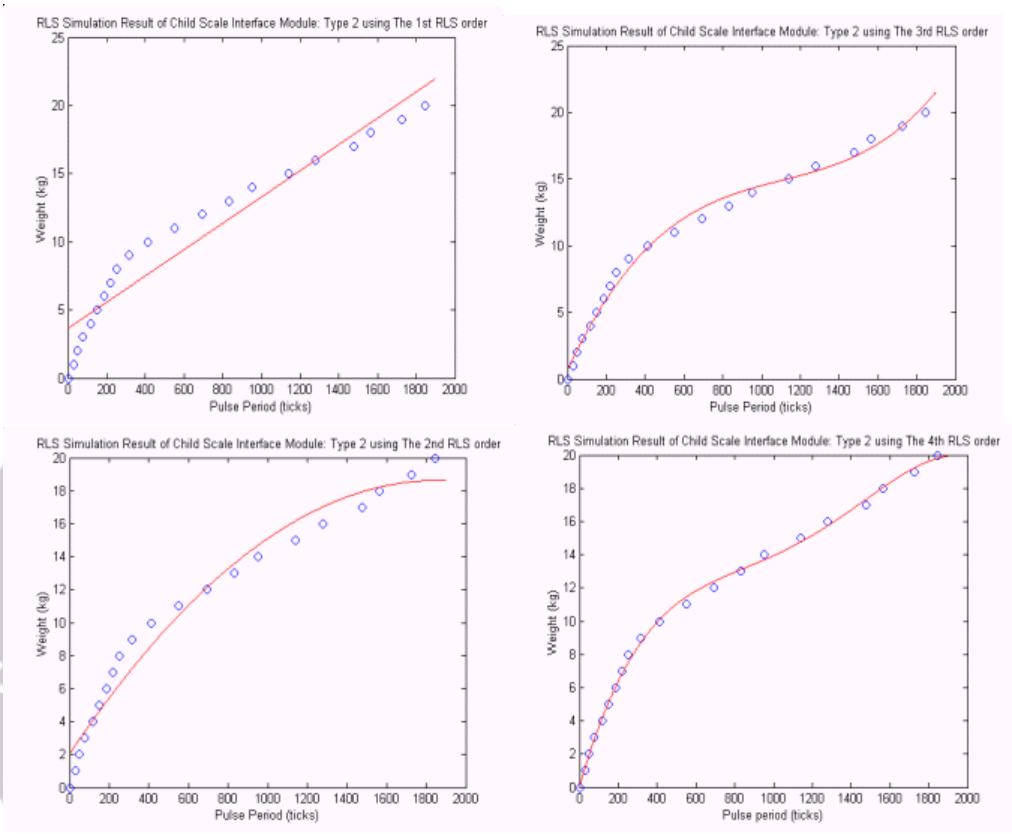


Figure 3 RLS simulation results using the 1st – 4th RLS order analog child scale weight measurement interface module

The measurement tests have been done in our laboratory. First, we run the RLS linear regression constants determination software unit to obtain the linear regression constants by inputting the reference (real) weight values (kg) and RLS order. Then the constants will be resulted and stored in a text file. Afterward we do the measurement process by weighing a number of loads using the weight measurement software unit for both of analog mother and child scales.

The weight value will be presented and can be stored in a text file, namely weight.txt. From the results we calculated the weight measurement error. The experimental results from several samples of weight measurement data are listed in Table 1 for analog mother scale and Table 2 for analog child scale.

The measurement results show that by using the 4th order RLS linear regression, the minimum error of measurement is obtained. It is relevant to the MATLAB simulation results.

Table 1 Weight Measurement Test Results of Analog Mother Scale Interface Module

Reference weight value (kg)	Weight value on the scale (kg)	RLS Order			
		1st order RLS		2nd order RLS	
		Weight (kg)	Error	Weight (kg)	Error
0	0	4.694	-	2.544	-
4.75	4.75	6.138	0.292211	4.084	0.140210526
12.5	12.5	10.633	0.14936	10.145	0.1884
17.1	17.1	14.886	0.129474	12.905	0.245321637
26.3	26.3	25.591	0.026958	25.321	0.037224335
39.9	39.9	37.182	0.06812	38.891	0.025288221
56	56	53.058	0.052536	56.308	0.0055
63.6	55.3	63.284	0.004969	63.755	0.002437
66.7	66.5	66.345	0.002331	66.47	0.000451
75	75	74.575	0.005667	76.119	0.01492
119.5	119.5	122.537	0.025414	116.851	0.22167
Reference weight (kg)	Weight value on the scale (kg)	RLS Order			
		3rd order RLS		4th order RLS	
		Weight (kg)	Error	Weight (kg)	Error
0	0	1.582	-	0.591	-
4.75	4.75	4.537	0.044842105	4.393	0.075157895
12.5	12.5	9.564	0.23488	12.132	0.02944
17.1	17.1	13.603	0.204502924	17.286	0.010877193
26.3	26.3	24.788	0.057490494	27.743	0.05486692
39.9	39.9	40.52	0.015538847	40.729	0.020776942
56	56	56.828	0.014785714	56.091	0.001625
63.6	55.3	64.584	0.015471698	62.252	0.021194969
66.7	66.5	66.864	0.002458771	66.578	0.001829085
75	75	75.337	0.004493333	75.093	0.00124
119.5	119.5	119.946	0.003732218	119.986	0.004066946

Clinical experiments have also been done in three different Community Health Centres (*Dago, Salam and Moch. Ramdhan* CHCs) by measuring body weight of pregnant mothers and children under five years. The weight measurement data of the patients (pregnant mothers and children) are successfully recorded automatically in the PC in form of text (.txt) file.

Table 2 Weight Measurement Test Results of Analog Child Scale Interface Module

Reference weight value (kg)	Weight value on the scale (kg)	RLS Order			
		1st order RLS		2nd order RLS	
		Weight (kg)	Error	Weight (kg)	Error
0	0	2.659	-	1.161	-
2.5	2.5	3.336	0.3344	2.397	0.0412
3.8	3.8	3.886	0.022631579	3.422	0.09947368
4.75	4.75	4.244	0.106526316	4.086	0.13978947
6	6	4.658	0.223666667	4.849	0.19183333
7.8	7.8	4.658	0.316666667	5.816	0.25435897
11	11	8.713	0.207909091	12.125	0.10227273
13.4	13.4	12.587	0.060671642	14.405	0.075
17.1	17.1	17.553	0.026491228	17.944	0.04935673
18	18	18.986	0.054777778	18.281	0.01561111
20	20	22.107	0.10535	18.353	0.08235
Reference weight (kg)	Weight value on the scale (kg)	RLS Order			
		3rd order RLS		4th order RLS	
		Weight (kg)	Error	Weight (kg)	Error
0	0	0.275	-	-0.018	-
2.5	2.5	2.411	0.0356	2.459	0.0164
3.8	3.8	3.768	0.0084211	3.984	0.0484211
4.75	4.75	4.684	0.0138947	4.867	0.0246316
6	6	5.701	0.0498333	6.021	0.0035
7.8	7.8	7.05	0.0961538	7.464	0.0430769
11	11	11.625	0.0568182	11.286	0.026
13.4	13.4	14.149	0.0558955	13.48	0.0059701
17.1	17.1	16.546	0.0323977	17.155	0.0032164
18	18	17.682	0.0176667	18.148	0.0082222
20	20	21.033	0.05165	19.762	0.0119

5 Conclusions

Based on the simulation, laboratory and clinical experimental results, we can derive the following conclusions:

- An identification method using recursive least squares (RLS) linear regression algorithm is applied in PC software of analog weight measurement interface module to model the weight measurement system in accordance with the types and specifications of the analog scales and the shift potentiometer sensors. We do the identification process by finding the relationship between inputted reference weight values and periods of digital

pulses measured and sent by microcontroller to the PC. The weight values in kg are successfully resulted based on the modified linear regression model (equation 6).

- By implementing the algorithm, the interface module is capable of transferring the weight measurement data from various types of analog scales and shift potentiometer sensors.
- The MATLAB simulation, measurement and clinical experimental results show that by using the 4th order RLS linear regression, non-linearity of the sensor can be optimized and the minimum error of measurement is obtained.

Acknowledgment

The authors would like to thank to head and staff of *Dago, Salam*, and *Moch. Ramdhan* Community Health Centres for supporting the clinical experiments.

References

[1] Rolf Johansson (1993), *System Modeling and Identification*, Prentice Hall International Editions.

[2] Norman Draper, and Harry Smith (1992), *Applied Regression Analysis*, Second Edition, John Wiley and Sons.

[3] Willis J. Tompkins, and John G. Webster (1988), *Interfacing Sensors to IBM PC*, Prentice Hall International Editions.

[4] Yoke Saadia Irawan & Soegijardjo Soegijoko (2002), Development of Internet – based Community Telemedicine System to Improve the Quality of Maternal Health –Care, *Proceedings of The First APT Telemedicine Workshop MCMT* on Mobile Technology for Medical Care and Triage, Jakarta, Indonesia.

[5] Trie Maya Kadarina & Soegijardjo Soegijoko (2005), Development of Weight Measurement Interface Modules for Telemedicine PC-based Medical Workstation, *Proceedings of The Third APT Telemedicine Workshop 2005 on Broadband Access to Health Aspects of Mental Analysis and Pathos Sharing via Broadband Network*, Kuala Lumpur, Malaysia.

TRIE MAYA KADARINA: Biomedical Engineering Program, Department of Electrical Engineering, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia. Phone/Fax: +62 +22 253 4117
E-mail: maya_mww@yahoo.com

SOEGIJARDJO SOEGIJOKO: Biomedical Engineering Program, Department of Electrical Engineering, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia. Phone/Fax: +62 +22 253 4117
E-mail: soegi@ieee.org

NASH EQUILIBRIUM OF TWO PLAYER LINEAR QUADRATIC DYNAMIC GAME DISCRETE SYSTEM

Salmah[a], Ari Suparwanto[a], Solikhatun[a]

[a] Universitas Gadjah Mada, Yogyakarta, Indonesia

Abstract. In this paper open loop Nash equilibrium of the linear quadratic dynamic game discrete system is considered. With Hamilton method optimal Nash solution is derived. Generalized difference Riccati equation is studied for finite time problem. Then the relationship between the existence of solution of the pair of generalized Riccati differential equation and optimal Nash equilibrium solution is studied.

Key-words: linear quadratic, dynamic game, discrete, Nash equilibrium

1 Introduction

In this paper we study open-loop discrete dynamic game. Strategy in open-loop game here is based on the assumptions that the parties act non cooperatively and that they have information on the model only its present state and the model structure.

For the continous game, the problem were studied in [1] and [2]. For discrete case, using Hamilton method the existence and uniqueness of optimal Nash equilibrium are studied in [1]. In paper [2], the relationship between the existence of generalized differential Riccati equation solution and optimal Nash solution of dynamic game in finite planning horizon for continous game are studied. Assimptotic behaviour of difference Riccati equation for discrete game with infinite planning horizon were studied in [3].

In this paper discrete game with finite planning horizon is studied. The relationship between the existences of generalized differential Riccati equation solution and optimal Nash solution is considered.

Two player linear quadratic discrete dynamic game with the players giving control to the system

$$x(k+1) = Ax(k) + b_1u_1(k) + b_2u_2(k), \quad x(0) = x_0, \tag{1.1}$$

With $x(k) \in \Re^n, u_i(k) \in \Re^{r_i}, 1 \le i \le 2, 0 \le k \le N-1.$

The two players minimizing objective function in the Nash sense in the form

$$J_1 = \frac{1}{2}x^T(N)K_{1N}x(N) + \frac{1}{2}\sum_{k=0}^{N-1}\left[x^T(k)Q_1x(k) + u_1^T(k)R_{11}u_1(k) + u_2^T(k)R_{12}u_2(k)\right], \tag{1.2}$$

$$J_2 = \frac{1}{2}x^T(N)K_{2N}x(N) + \frac{1}{2}\sum_{k=0}^{N-1}\left[x^T(k)Q_2x(k) + u_1^T(k)R_{21}u_1(k) + u_2^T(k)R_{22}u_2(k)\right], \tag{1.3}$$

With all matrices are symmetric, further more Q_1 , Q_2 and K_{1N} ,K_{2Ni} semi positive definite and R_{11} ,R_{12} ,R_{21} ,R_{22} positive definite.

It has been shown in [1] that the necessary conditions for an open loop Nash equilibrium for the game satisfy (1.1), (1.2) and (1.3) are

$$u_1(k) = -R_{11}^{-1}B_1^T\psi_1(k+1), \qquad u_2(k) = -R_{22}^{-1}B_2^T\psi_2(k+1),$$

with $\psi_i(k), i = 1,2$ satisfy

$$\begin{aligned}\psi_i(k) &= Q_i x(k) + A^T\psi_i(k+1), \\ \psi_i(N) &= K_{iN}x(N), \quad 1 \le i \le 2, \quad 0 \le k \le N-1.\end{aligned} \tag{1.4}$$

The generelized differential Riccati equation will be derived, following [3]. Suppose that

$$\psi_i(k) = K_i(k)x(k), \quad 1 \le i \le 2, \quad 0 \le k \le N-1. \tag{1.5}$$

Then (1) can be rewritten as

$$Ax(k) = \left[I + S_1K_1(k+1) + S_2K_2(k+1)\right]x(k+1), \tag{1.6}$$

With $x(0) = x_0$ and $S_i = B_iR_{ii}^{-1}B_i^T$, $1 \le i \le 2, \quad 0 \le k \le N-1$. If $\left[I + S_1K_1(k+1) + S_2K_2(k+1)\right]$ invertible, from (1.4), (1.5), and (1.6) we can derive the generelized differential Riccati equations

$$K_1(k) = Q_1 + A^TK_1(k+1)\left[I + S_1K_1(k+1) + S_2K_2(k+1)\right]^{-1}A, \quad K_1(N) = K_{1N}, \tag{1.7}$$

$$K_2(k) = Q_2 + A^TK_2(k+1)\left[I + S_1K_1(k+1) + S_2K_2(k+1)\right]^{-1}A, \quad K_2(N) = K_{2N}. \tag{1.8}$$

As has been stated in [3] if A invertible from (1.4), (1.6) we have

$$\begin{pmatrix}\tilde{x}\\ \tilde{\psi}_1\\ \tilde{\psi}_2\end{pmatrix}(m+1) = M_{N_a}\begin{pmatrix}\tilde{x}\\ \tilde{\psi}_1\\ \tilde{\psi}_2\end{pmatrix}(m), \tag{1.9}$$

With $m = N-k$, $\tilde{x}(m) = x(N+1-m)$, $\tilde{\psi}_i(m) = \psi_i(N+1-m), \quad 1 \le i \le 2$ and

$$M_{N_a} = \begin{pmatrix} A^{-1} & A^{-1}S_1 & A^{-1}S_2 \\ Q_1A^{-1} & A^T + Q_1A^{-1}S_1 & Q_1A^{-1}S_2 \\ Q_2A^{-1} & Q_2A^{-1}S_1 & A^T + Q_2A^{-1}S_2 \end{pmatrix}.$$

If the initial value is given the sequence $\begin{pmatrix} \tilde{x} \\ \tilde{\psi}_1 \\ \tilde{\psi}_2 \end{pmatrix}(m)$, $m \geq 1$, is uniquely determined. Note that we also have relationship

$$\tilde{\psi}_i(m) = \psi_i(N+1-m) = K_{iN}x(N+1-m) = K_{iN}\tilde{x}(m),$$

or

$$\tilde{\psi}_i(m) = K_{iN}\tilde{x}(m).$$

2 Relationship of solvability of Riccati difference equation and Nash solution for fixed point in time

The theorem below is discussed about the relationship between the existence of the solution of couple Riccati difference equation (1.7), (1.8) and Nash equilibrium solution for two player discrete game for fixed point in time case.

Denoting $(\tilde{x}^T(m), \tilde{\psi}_1^T(m), \tilde{\psi}_2^T(m))^T$ by $\tilde{y}(m)$, we can rewrite (1.9) in the form

$$\tilde{y}(m+1) = M_{Na}\tilde{y}(m) \tag{2.1}$$

Take $P = \begin{pmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, and $Q = \begin{pmatrix} 0 & 0 & 0 \\ -K_{1N} & I & 0 \\ -K_{2N} & 0 & I \end{pmatrix}$,

Then

$$P\tilde{y}(0) = (\tilde{x}_0^T, 0, 0)^T \text{ and } Q\tilde{y}(N) = (0, 0, 0)^T,$$

So we have $P\tilde{y}(0) + Q\tilde{y}(N) = \begin{pmatrix} \tilde{x}_0^T & 0 & 0 \end{pmatrix}^T$. From discrete system theory we have $\tilde{y}(N) = M_{N_a}^N \tilde{y}(0)$, therefore

$$P\tilde{y}(0) + Q\left(M_{N_a}\right)^N \tilde{y}(0) = \begin{pmatrix} \tilde{x}_0 \\ 0 \\ 0 \end{pmatrix},$$

or, if M_{Na} is invertible,

$$\left(P\left(M_{N_a}\right)^{-N} + Q\right)\left(M_{N_a}\right)^N \tilde{y}(0) = \begin{pmatrix} \tilde{x}_0 \\ 0 \\ 0 \end{pmatrix}. \tag{2.2}$$

From (2.1) and (2.2) we have the following proposition:

Proposition 1: *The two-player linear –quadratic difference game has a unique open loop Nash equilibrium for every initial state iff (2.1) and (2.2) is uniquely solvable for every* $\tilde{x}_0$, *provided the inverses in (2.2) exist.*

Let

$$\left(M_{N_a}\right)^{-N} = \begin{pmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{pmatrix}, \tag{2.3}$$

and

$$\left(M_{N_a}\right)^{N} \tilde{y}(0) = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix}, \tag{2.4}$$

By defining the following notation :

$$H(N) = W_{11}(N) + W_{12}(N)K_{1N} + W_{13}(N)K_{2N},$$

we can derive the following theorem.

Theorem 1: *The two-player linear –quadratic difference game has a unique open loop Nash equilibrium for every initial state iff* $H(N)$ *is invertible.*

Proof. From (2.2) and (2.3), we have

$$H(N)z_1 = \tilde{x}_0,$$
$$z_2 = K_{1N} z_1,$$
$$z_3 = K_{2N} z_1.$$

If $H(N)$ is invertible, we have immediately from the equations above that (2.1) with (2.2) is uniquely solvable for every $\tilde{x}_0$. Conversely, if (2.1) with (2.2) has a unique solution for every $\tilde{x}_0$, then by taking $\tilde{x}_0 = (1,\ldots,0)^T, \ldots, \tilde{x}_0 = (0,\ldots,1)^T$, we have $H(N)$ is invertible.

3 Relationship of solvability of Riccati difference equation and Nash solution for fixed time interval

Theorem 1 only discussed about local existence of Nash equilibrium for fixed point in time. Theorem 2 below will state that the existence of an equilibrium strategy for every point in time during some fixed time interval $[0, N]$ is equivalent to the existence of a solution to the couple Riccati difference equation (1.7), (1.8) on this interval.

Theorem 2: *The three statements below are equivalent.*

(1). For all $m \in [0, k_1]$ *there exist a unique open-loop Nash equilibrium for the two player linear quadratic differential game.*

(2). Matrix $H(m)$ *is invertible for all* $m \in [0, k_1]$.

(3). The set of Riccati differential equation (1.7), (1.8) has a solution on $[0, k_1]$.

Proof. From equation (2.1) and from Theorem 1, *H(N)* is invertible therefore we have

$$\tilde{y}_0 = \left(M_{N_a}\right)^{-N} \begin{pmatrix} I \\ K_{1N} \\ K_{2N} \end{pmatrix} H(N)^{-1} \tilde{x}_0 .$$

From discrete system theory we have $\tilde{y}(m) = M_{N_a}^{m} \tilde{y}_0$, then

$$\tilde{x}(m) = \begin{pmatrix} I & 0 & 0 \end{pmatrix} M_{N_a}^{m-N} \begin{pmatrix} I \\ K_{1N} \\ K_{2N} \end{pmatrix} H(N)^{-1} \tilde{x}_0 , \quad (3.1)$$

$$\tilde{\psi}_1(m) = \begin{pmatrix} 0 & I & 0 \end{pmatrix} M_{N_a}^{m-N} \begin{pmatrix} I \\ K_{1N} \\ K_{2N} \end{pmatrix} H(N)^{-1} \tilde{x}_0 , \quad (3.2)$$

$$\tilde{\psi}_2(m) = \begin{pmatrix} 0 & 0 & I \end{pmatrix} M_{N_a}^{m-N} \begin{pmatrix} I \\ K_{1N} \\ K_{2N} \end{pmatrix} H(N)^{-1} \tilde{x}_0 . \quad (3.3)$$

From the notation introduce above we have

$$\begin{pmatrix} I & 0 & 0 \end{pmatrix} M_{N_a}^{(m-N)} \begin{pmatrix} I \\ K_{1N} \\ K_{2N} \end{pmatrix} = H(N-m) .$$

Therefore from (3.1) we get $\tilde{x}(m) = \begin{pmatrix} I & 0 & 0 \end{pmatrix} H(N-m) H(N)^{-1} \tilde{x}_0$. From Theorem 1, $H(N-m)$ invertible, therefore $H^{-1}(N)\tilde{x}_0 = H^{-1}(N-m)\tilde{x}(m)$.

Substitute to (3.1), (3.2) we can expressed

$$\tilde{\psi}_1(m) = G_1(N-m) H^{-1}(N-m) \tilde{x}(m) ,$$

$$\tilde{\psi}_2(m) = G_2(N-m) H^{-1}(N-m) \tilde{x}(m) ,$$

For some G_1, G_2. Denote

$$G_1(N-m) H^{-1}(N-m) = K_1(m) ,$$

$$G_2(N-m)H^{-1}(N-m) = K_2(m),$$

We can write

$$\tilde{\psi}_1(m) = K_1(m)\tilde{x}(m), \quad (3.4)$$

$$\tilde{\psi}_2(m) = K_2(m)\tilde{x}(m). \quad (3.5)$$

Then from (1.4) we have

$$\tilde{\psi}_i(m+1) = Q_i\tilde{x}(m+1) + A^T\tilde{\psi}_i(m). \quad (3.6)$$

From (1.6) we have

$$A\tilde{x}(m) = \left[I + S_1K_1(m) + S_2K_2(m)\right]\tilde{x}(m+1) \quad (3.7)$$

Substitute (3.4), (3.5) and (3.7) to (3.6) we get

$$\tilde{\psi}_i(m+1) = Q_i\tilde{x}(m+1) + A^TK_i(m)\tilde{x}(m),$$

or

$$K_i(m+1)\tilde{x}(m+1) = Q_i\tilde{x}(m+1) + A^TK_i(m)\left[I + S_1K_1(m) + S_2K_2(m)\right]^{-1}A\tilde{x}(m+1),$$

i=1,2. Then we get Riccati equation (1.7), (1.8).

Acknowledgment

The writer would like to thank to PHK A3 research grant project in Jurusan Matematika Universitas Gadjah Mada who support this research.

References

[1] Basar, Tamer and Olsder, Geert Jan (1995), *Dynamic noncooperative Game Theory*, Second edition, Academic Press, London, San Diego, New York, Boston, Sydney, Tokyo and Toronto, pp.317-345,.

[2] Engwerda, Jacob J, (1998), On the open-loop Nash Equilibrium in the LQ-Games, *Journal on Economic Theory*.

[3] Freiling G, Jank G, Abou-Kandil H, Discrete Time Riccati Equations in Open-Loop Nash and Stackelberg Games, *http://www.uni-duisburg.de/FB11/FGS/F7/Publications/ab57/ab57.html* Universitat Duisburg, D-47048 Duisburg, Germany.

Salmah: Ph D student at Department of Mathematics, Universitas Gadjah Mada, Sekip Utara, Bulaksumur, Yogyakarta 55281, Indonesia.
Department of Mathematics, Universitas Gadjah Mada, Sekip Utara, Bulaksumur, Yogyakarta 55281, Indonesia. Phone/Fax +62 +22 522443
E-mail: syalmah@yahoo.com

Ari Suparwanto: Department of Mathematics, Universitas Gadjah Mada, Sekip Utara, Bulaksumur, Yogyakarta 55281, Indonesia. Phone/Fax +62 +22 522443
E-mail: ari@math-ugm.web.id

Solikhatun: Department of Mathematics, Universitas Gadjah Mada, Sekip Utara, Bulaksumur, Yogyakarta 55281, Indonesia. Phone/Fax +62 +22 522443

serviens in lumine veritatis

A GENERALIZED INTEGRATION FORMULA FOR DISCRETE - TIME SIMULATION BASED ON PIECEWISE POLYNOMIAL SIGNAL APPROXIMATION

Bambang Sridadi[a]

[a] Indonesian Aerospace (IAe), Bandung, Indonesia

Abstract. Signal and system in the real world can be considered to be continuous-time one defined on the time-axis. On the other hand, in practice, a discrete-time model simulating a continuous-time system has been successfully applied to many applications, e.g. real-time discrete-time flight simulation, by using digital computer. Numerical integration is important stage in the simulation of dynamical system. In this paper, a fluency tunable integration technique is derived. It is revealed that the fluency integration method includes and generalizes the conventional one, i.e. Euler's and Trapezoidal-like integration methods. The tunable parameters m (order of fluency approximation) and h (sampling interval) can be choosen adaptivelly according with smoothness property (continuously differentiable) of signal $s(t)$ we deal with. Thus, this concept provides a better generalized family of integration formula of the relationship between the discrete-time and continuous-time signal.
Key-words: discrete-time simulation, numerical method, signal and system, approximation theory, B-spline function.

1 Introduction

Signal and system in the real world can be considered to be continuous-time one defined on the time-axis. On the other hand, in practice, a discrete-time model simulating a continuous-time system has been successfully applied to many applications, e.g. real-time discrete-time flight simulation, by using digital computer. Consequently, there must be a fundamental theory to relate mathematically the discrete-time signal and system to the original continuous-time one.

On the other hand, structures, communication systems, control systems, design of aircraft, and the design of chemical plants are a few of the areas where the simulation of system with very different type of signal and concepts of numerical integration have been produced. Numerical integration is important stage in the simulation of system. Figure 1 shows the example of numerical integration role in the discrete-time system simulation process. Structural dynamicists have developed special numerical integration formulas for integrating their stiff differential equations. Controls analysts have produced such formulas based solely on frequency-domain considerations. And special single-step real-time numerical integration formulas have been developed by simulation scientists.

However, one of the problems in constructing this kind of integration theory, which relates a continuous-time system with a discrete-time one, is what kind of signal

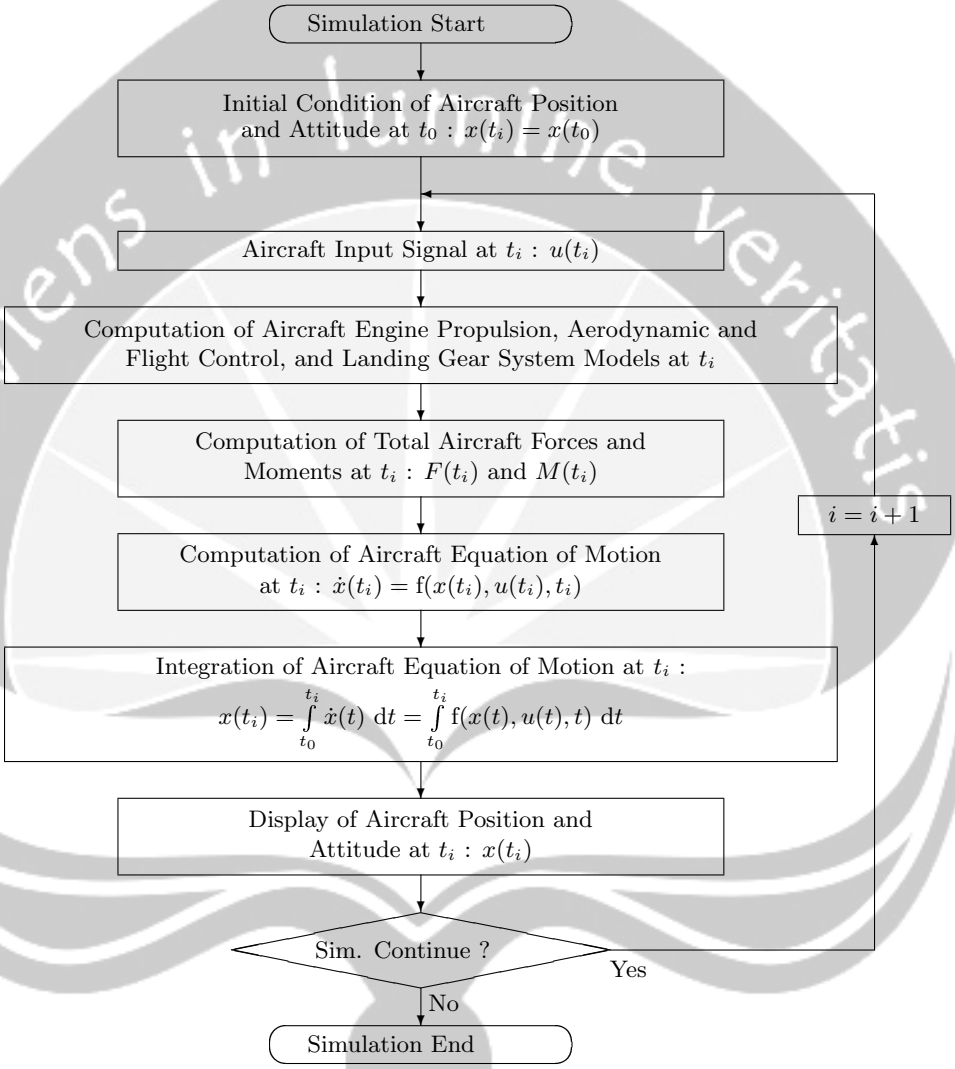


Figure 1: Numerical integration role in the discrete - time system simulation process.

space is to be assumed. This integration formulation has a problem that the signal space to which the continuous-time signal belongs is not mathematically defined. We first have to define the signal space. Then we have a mathematically strict relationship between the continuous-time signal and the discrete-time signal. The relationship gives us a discrete-time integration model approximating a continuous-time one with a clear meaning.

We will be basically concerned with two types of integration : the definite and indefinite integrals. The definite integral for the continuous-time differential equation [1] [2]

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = \mathrm{f}(x(t), u(t), t) = s(t) \tag{1}$$

is given by

$$x(b) = x(a) + \int_a^b s(t)\ \mathrm{d}t \tag{2}$$

while the indefinite integral is defined by

$$x(t) = x(a) + \int_a^t s(\tau)\ \mathrm{d}\tau \tag{3}$$

Some take the position that the definite integral computes a single number which is the area under the curve of a bounded signal and that the indefinite integral performs an antiderivative operation on the integrand, thus generating a sequence of numbers that are values of the antiderivative signal.

In the classical development of the numerical integration, numerical integration formulas from the sampled-data viewpoint are derived by : 1. Synthesizing a discrete-time approximation to continuous-time integration, 2. Writing the difference equation that descibes the discrete-time system. The difference equation is the numerical integration formula. For example, the difference equations that describe discrete approximation of continuous integration are Euler's integration formula, which describes the process of sampling the continuous integration of a zero-order reconstructed integrand

$$x_k = x_{k-1} + h\dot{x}_{k-1} \tag{4}$$

and trapezoidal integration formula

$$x_k = x_{k-1} + \frac{h}{2}(\dot{x}_k + \dot{x}_{k-1}) \tag{5}$$

In this paper, we shall derive a generalized integration method by assuming that the signal $s(t)$ is the m order fluency signal. The organization of this paper is as follows. In Section 2, the signal space composed of fluency signals is prepared. In Section 3, the integration of a continuous-time signal by assuming a fluency signal is discussed. Finally, Section 4 gives some examples of a integration formulation for $m = 1$, 2 and 3.

2 Notation and Mathematical Preliminaries

This section prepares some notation and definition of signal spaces composed of fluency signals.

The fluency signal [4] of order m is an (m-2)-times continuously differentiable piecewise polynomial of degree m-1, and it is identical with a staircase or polygonal signal when its order is 1 or 2, respectivelly. A fluency signal approaches a band-limited signal when its order approaches infinity as its smoothness increases in proportion to its order.

The totality of continuous-time signals is considered to be in the Hilbert space (and called *finite-energy signal space*)

$$L_2(R) := \{f| \int_{-\infty}^{\infty} |f(t)|^2 \mathrm{d}t < +\infty\} \tag{6}$$

with the inner product

$$(f,g)_{L_2} := \int_{-\infty}^{\infty} f(t)\ \overline{g(t)} \mathrm{d}t \tag{7}$$

where R denotes the field of real numbers.

Then we consider a signal space composed of fluency signals as a subspace of $L_2(R)$. We can define the signal space composed of fluency signals of order m (for $m = 1, 2, ...$) as follows

$$^{m}S := [{}^{m}_{[s]}\psi_k(t)]_{k=-\infty}^{\infty} \tag{8}$$

where we call ^{m}S the *fluency signal space* of order m. The basis $\{{}^{m}_{[s]}\psi_k(t)\}_{k=-\infty}^{\infty}$ is a fluency sampling basis, which is defined as

$${}^{m}_{[s]}\psi_k(t) := \sum_{\ell=-\infty}^{\infty} {}^{m}\beta(\ell - k)\ {}^{m}_{[b]}\psi_\ell(t) \quad \text{for} \quad k = 0, \pm1, \pm2, ... \tag{9}$$

Here $\{{}^{m}_{[b]}\psi_\ell(t)\}_{\ell=-\infty}^{\infty}$ is a B-spline basis of order m defined by [4]

$${}^{m}_{[b]}\psi_\ell(t) := \int_{-\infty}^{\infty} (\frac{\sin(\pi f h)}{\pi f h})^m\ \exp^{j2\pi(t-(\ell+\frac{m}{2})h)f}\,\mathrm{d}f \quad \text{for } \ell = 0, \pm1, \pm2, ... \tag{10}$$

and $\{{}^{m}\beta(k)\}_{k=-\infty}^{\infty}$ is defined by

$$\begin{aligned} {}^{m}\beta(k) &:= h\int_{-\frac{1}{2h}}^{\frac{1}{2h}} {}^{m}_{f}B(f)\ \exp^{j2\pi fkh}\,\mathrm{d}f \\ {}^{m}_{f}B(f) &:= \frac{h}{\sum_{p=-\infty}^{\infty}[\frac{\sin(\pi(fh-p))}{\pi(fh-p)}]^m} \end{aligned} \tag{11}$$

The B-spline basis can be represented in the form of a piecewise polynomial of degree m-1 as follows

$$ {}^{m}_{[b]}\psi_{\ell}(t) = \frac{m}{h^{m-1}} \sum_{p=0}^{m} \frac{(-1)^{p}(t-(\ell+p)h)_{+}^{m-1}}{p!(m-p)!} \tag{12} $$

which is $(m-2)$-times continuously differentiable over the t-axis, where

$$ (t-a)_{+}^{m-1} = (t-a)^{m-1}, \quad (t>a)0, \qquad (t \le a) $$

A B-spline signal ${}^{m}_{[b]}\psi_{\ell}(t)$ satisfies the following elementary properties:

(a) a time-limited (locally supported) property

$$ {}^{m}_{[b]}\psi_{\ell}(t) = 0, \quad t \notin (\ell h, (\ell+m)h) \tag{13} $$

(b) a shifting property

$$ {}^{m}_{[b]}\psi_{\ell}(t+kh) = {}^{m}_{[b]}\psi_{\ell-k}(t) \tag{14} $$

(c) symmetry property

$$ {}^{m}_{[b]}\psi_{\ell}(-t) = {}^{m}_{[b]}\psi_{-m-\ell}(t) \quad \text{or} \quad {}^{m}_{[b]}\psi_{\ell}(h-t) = {}^{m}_{[b]}\psi_{-m+1-\ell}(t) \tag{15} $$

The fluency sampling basis satisfies

$$ s(t) = \sum_{k=-\infty}^{\infty} s_k \; {}^{m}_{[s]}\psi_{k}(t) \tag{16} $$

for any signal $s(t) \in {}^{m}S$, where $s_k = s(t_k)(t_k = kh + (mh/2))$, is a sampling value. We call a signal $s(t)$, formed with a fluency sampling basis with the appropriate order m, the *fluency signal*. The waveform of fluency signal $s(t)$ and the sampling basis are shown in [3].

However, the fluency sampling basis is not given explicitly. The B-spline basis shall be used as a medium to investigate it, which is given by

$$ s(t) = \sum_{\ell=-\infty}^{\infty} w_{\ell} \; {}^{m}_{[b]}\psi_{\ell}(t) \tag{17} $$

where ${}^{m}_{[b]}\psi_{\ell}(t)$ means the B-spline basis and w_{ℓ} is the B-spline coefficient.

Let ${}^{m}_{[s]}\varphi$ denote the mapping which transforms a continuous-time signal $s(t)$ into its corresponding discrete-time one s_k, and ${}^{m}_{[b]}\varphi$ denotes the mapping that transforms $s(t)$ into w_{ℓ}. Then ${}^{m}_{[s]}\varphi$ is a linear bijection on ${}^{m}S$ onto l_2, and ${}^{m}_{[b]}\varphi$ is a linear bijection on ${}^{m}S$ onto l_2. In order to represent the relation between ${}^{m}_{[s]}\varphi$ and ${}^{m}_{[b]}\varphi$, the coordinate transform operator from ${}^{m}_{[s]}\varphi$ to ${}^{m}_{[b]}\varphi$ shall be defined as follows

$$ {}^{m}B := {}^{m}_{[b]}\varphi \; {}^{m}_{[s]}\varphi^{-1} \tag{18} $$

Mutual relations between ${}^{m}_{[s]}\varphi$, ${}^{m}_{[b]}\varphi$ and ${}^{m}B$ are discussed in [5].

3 Integration of a Continuous - Time Signal by Assuming Fluency Signals

In this section, the integration process by assuming $s(t)$ the m order fluency signal to obtain a generalized formulation is discused.

Consider the given s_k sample value of continuous - time differential equation

$$s(t) \;=\; \frac{\mathrm{d}x(t)}{\mathrm{d}t} \;=\; \mathrm{f}(x(t), u(t), t) \tag{19}$$

The problem is to find x_k, the value of $x(t)$ at $t = kh$

$$x_k = \int_{-\infty}^{kh} s(t)\ \mathrm{d}t \tag{20}$$

In the present method, the signal $s(t)$ is fluency signal interpolating the vector of the discrete-time sequence s_k by using the m-order fluency sampling basis ${}^{m}_{[s]}\psi_k(t)$, i.e.

$$s(t) = \sum_{k=-\infty}^{\infty} s_k\ {}^{m}_{[s]}\psi_k(t) \quad \text{where } s(.) \in {}^{m}S, \quad \text{for } m = 1, 2, 3, ... \tag{21}$$

For exact computation, because of the time-limitted (locally supported) property of the B-spline signal, ${}^{m}_{[s]}\psi_k(t)$ can be replaced by the B-spline basis ${}^{m}_{[b]}\psi_\ell(t)$ with the algorithm that has been proposed by [5]. We can then denote $s(t)$ in the form

$$s(t) = \sum_{\ell=-\infty}^{\infty} w_\ell\ {}^{m}_{[b]}\psi_\ell(t) \tag{22}$$

where w_ℓ is the B-spline coefficient sequence.

The following is detail of fluency integration process. We can rewrite (20) as follows.

$$\begin{aligned} x_k &= \int_{-\infty}^{kh} s(t)\ \mathrm{d}t \\ &= \int_{-\infty}^{(k-1)h} s(t)\mathrm{d}t \;+\; \int_{(k-1)h}^{kh} s(t)\ \mathrm{d}t \\ &= x_{k-1} + \int_{(k-1)h}^{kh} s(t)\ \mathrm{d}t \end{aligned} \tag{23}$$

where $s(t)$ is fluency function of order m, can be represented in the form of B-spline bases of order m defined by (22) and then reordering, we obtain the second

term of Equation (23)

$$\begin{aligned}\int_{(k-1)h}^{kh} s(t)\ \mathrm{d}t &= \int_{(k-1)h}^{kh} \sum_{\ell=-\infty}^{\infty} w_\ell\ {}^{m}_{[b]}\psi_\ell(t)\ \mathrm{d}t \\ &= \sum_{\ell=-\infty}^{\infty} w_\ell \int_{(k-1)}^{kh} {}^{m}_{[b]}\psi_\ell(t)\ \mathrm{d}t \end{aligned} \tag{24}$$

Using time - limited (locally supported) property of B - spline signal, then we can rewrite

$$\int_{(k-1)h}^{kh} s(t)\ \mathrm{d}t = \sum_{\ell=\frac{a-(m-1)h}{h}}^{\frac{b-h}{h}} w_\ell \int_{(k-1)h}^{kh} {}^{m}_{[b]}\psi_\ell(t)\ \mathrm{d}t \tag{25}$$

Integration part of Equation (25) can be computed using Equation (12) which is representation of B-spline bases of order m in the form of piecewise polynomials of order m as follows

$$\begin{aligned}{}^{m}_{[b]}I_\ell &= \int_{(k-1)h}^{kh} {}^{m}_{[b]}\psi_\ell(t)\ \mathrm{d}t \\ &= \int_{(k-1)h}^{kh} \frac{m}{h^{m-1}} \sum_{p=0}^{m} \frac{(-1)^p (t-(\ell+p)h)_+^{m-1}}{p!(m-p)!}\ \mathrm{d}t \\ &= \frac{m}{h^{m-1}} \sum_{p=0}^{m} \frac{(-1)^p}{p!(m-p)!} \int_{(k-1)h}^{kh} (t-(\ell+p)h)_+^{m-1}\ \mathrm{d}t \\ &= \frac{m}{h^{m-1}} \sum_{p=0}^{m} \frac{(-1)^p}{p!(m-p)!} \int_{(k-1)h}^{kh} (t-\alpha)_+^{m-1}\ \mathrm{d}t \quad \text{for } \alpha=(\ell+p)h \\ &= \frac{m}{h^{m-1}} \sum_{p=0}^{m} \frac{(-1)^p}{p!(m-p)!} \frac{(t-\alpha)_+^{m}}{m}\Big|_{(k-1)h}^{kh} \end{aligned} \tag{26}$$

Hence, we obtain the integration formula, as follows

$${}^{m}_{h}Int(t): \begin{cases} x_k = x_{k-1} + \sum_{\ell=k-m}^{k-1} w_\ell\ {}^{m}_{[b]}I_\ell \\ {}^{m}_{[b]}I_\ell = \frac{m}{h^{m-1}} \sum_{p=0}^{m} \frac{(-1)^p}{p!(m-p)!} \int_{(k-1)h}^{kh} (t-\alpha)_+^{m-1}\ \mathrm{d}t \\ \text{for } \alpha=(\ell+p)h \end{cases} \tag{27}$$

That is ${}^{m}_{h}Int(t)$ a fluency integration technique with tunable parameters m (order of fluency approximation) and h (sampling interval) can be choosen according with

smoothness property (continuously differentiable) of signal $s(t)$ we deal with. The algorithm of fluency integration techniques which can be used for the discrete-time system simulation is decribed in the following Figure 2.

Each fluency integrator ${}_h^m Int(t)$ may have many values of h and m. Aircraft flight-training simulators are an example of simulations where different nominal flight conditions can require different values of sampling rate h and order of fluency approximation m for each flight condition (e.g., flaps up or down, landing gear up or down, and high and low Mach number). The h and m for each condition are selected by repeating the fluency tuning procedure. These h and m are stored and used in the simulation when the appropriate condition has been reached (e.g., when a flap handle has been put in the down position, a gear handle is placed in the up position, or Mach > 0.5 etc.). A slow-varying input signal (e.g., cruise condition of aircraft) will be simulated at slow sampling rate (high h) or low order of fluency approximation (low m), while a input signal with a high-frequency content (e.g., maneuvering, landing or tracking condition of aircraft) should be simulated at a high sampling rate (low h) or high order of fluency approximation (high m).

The fluency tunable integration is obtained by selecting the appropriate order m within a given tolerance of the approximation error incurred in assuming the signal, according to the characteristics of the signal of the continuous-time system we are dealing with. Thus, this concept provides a better generalized family of integration formula of the relationship between the discrete-time and continuous-time signal.

4 Examples for Order of Fluency Approximation $m = 1$, 2 and 3

An example of empirical m selection for the integration formulation will be discused here. In the case of order of fluency approximation $m = 1$, i.e. $s(t)$ is a staircase signal, then from (27) the integration result is as follows

$$\begin{aligned} x_k &= \int_{-\infty}^{kh} s(t)\ \mathrm{d}t \\ &= \int_{-\infty}^{(k-1)h} s(t)\mathrm{d}t\ + \int_{(k-1)h}^{kh} s(t)\ \mathrm{d}t \\ &= x_{k-1} + \delta_x \\ &= x_{k-1} + \sum_{\ell=k-m}^{k-1} w_\ell\ {}_{[b]}^m I_\ell \end{aligned}$$

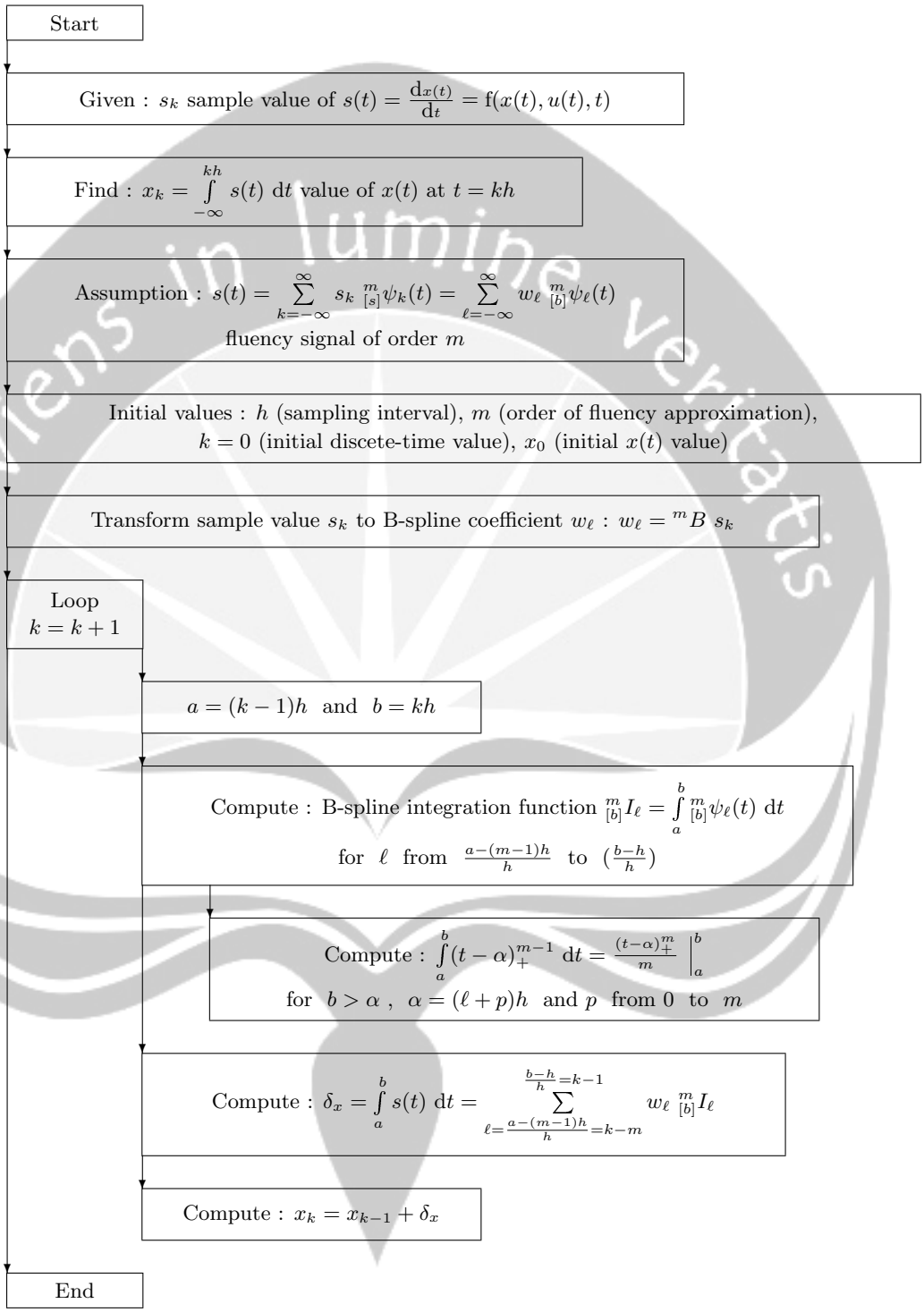


Figure 2: Algorithm of fluency integration techniques for the discrete - time system simulation.

$$
\begin{aligned}
&= x_{k-1} + \sum_{\ell=k-m}^{k-1} w_\ell \, \frac{m}{h^{m-1}} \sum_{p=0}^{m} \frac{(-1)^p}{p!(m-p)!} \int_{(k-1)h}^{kh} (t-\alpha)_+^{m-1} \, \mathrm{d}t \quad \text{for } \alpha = (\ell+p)h \\
&= x_{k-1} + w_{k-1} \sum_{p=0}^{1} \frac{(-1)^p}{p!(1-p)!} \int_{(k-1)h}^{kh} (t-(k-1+p)h)_+^0 \, \mathrm{d}t \\
&= x_{k-1} + w_{k-1} \left(\frac{(-1)^0}{0!1!} \int_{(k-1)h}^{kh} (t-(k-1)h)_+^0 \, \mathrm{d}t \; + \; \frac{(-1)^1}{1!0!} \int_{(k-1)h}^{kh} (t-kh)_+^0 \, \mathrm{d}t \right) \\
&= x_{k-1} + w_{k-1} \int_{(k-1)h}^{kh} 1 \, \mathrm{d}t \\
&= x_{k-1} + w_{k-1} \; (kh-(k-1)h) \\
&= x_{k-1} + hw_{k-1}
\end{aligned} \tag{28}
$$

This formulation is identical with formula of Euler's integration method or first order Runge-Kutta method with w_{k-1} equal to sample value of s_{k-1}.

If signal $s(t)$ is a polygonal signal, then from (27) with order of fluency approximation $m = 2$ the integration result becomes

$$
\begin{aligned}
x_k &= \int_{-\infty}^{kh} s(t) \, \mathrm{d}t \\
&= \int_{-\infty}^{(k-1)h} s(t)\mathrm{d}t \; + \int_{(k-1)h}^{kh} s(t) \, \mathrm{d}t \\
&= x_{k-1} + \delta_x \\
&= x_{k-1} + \sum_{\ell=k-m}^{k-1} w_\ell \; {}_{[b]}^{m}I_\ell \\
&= x_{k-1} + \sum_{\ell=k-m}^{k-1} w_\ell \, \frac{m}{h^{m-1}} \sum_{p=0}^{m} \frac{(-1)^p}{p!(m-p)!} \int_{(k-1)h}^{kh} (t-\alpha)_+^{m-1} \, \mathrm{d}t \quad \text{for } \alpha = (\ell+p)h \\
&= x_{k-1} + \left(w_{k-2} \, \frac{2}{h} \sum_{p=0}^{2} \frac{(-1)^p}{p!(2-p)!} \int_{(k-1)h}^{kh} (t-(k-2+p)h)_+^1 \, \mathrm{d}t \right.
\end{aligned}
$$

$$
\begin{aligned}
&+ w_{k-1}\,\frac{2}{h}\sum_{p=0}^{2}\frac{(-1)^p}{p!(2-p)!}\int_{(k-1)h}^{kh}(t-(k-1+p)h)_+^1\,\mathrm{d}t\Bigg)\\
&= x_{k-1}+w_{k-2}\,\frac{2}{h}\,\frac{h^2}{4}+w_{k-1}\,\frac{2}{h}\,\frac{h^2}{4}\\
&= x_{k-1}+\frac{h}{2}(w_{k-2}+w_{k-1}) \qquad (29)
\end{aligned}
$$

This formula is like trapezoidal integration method or closed Newton-Cotes integration method with w_{k-1} equal to sample value of s_{k-1} and w_k equal to sample value of s_k, respectivelly. In this example, the formulation is more flexible because a polygonal signal generally better approximates the continuous-time signal.

In the case of $s(t)$ is a quadratic spline signal with order of fluency approximation $m = 3$, then the integration result is as follows

$$
\begin{aligned}
x_k &= \int_{-\infty}^{kh} s(t)\,\mathrm{d}t\\
&= \int_{-\infty}^{(k-1)h} s(t)\mathrm{d}t\; + \int_{(k-1)h}^{kh} s(t)\,\mathrm{d}t\\
&= x_{k-1}+\delta_x\\
&= x_{k-1}+\sum_{\ell=k-m}^{k-1} w_\ell\;{}^{m}_{[b]}I_\ell\\
&= x_{k-1}+\sum_{\ell=k-m}^{k-1} w_\ell\,\frac{m}{h^{m-1}}\sum_{p=0}^{m}\frac{(-1)^p}{p!(m-p)!}\int_{(k-1)h}^{kh}(t-\alpha)_+^{m-1}\,\mathrm{d}t \text{ for } \alpha=(\ell+p)h\\
&= x_{k-1}+\sum_{\ell=k-3}^{k-1} w_\ell\,\frac{3}{h^{3-1}}\sum_{p=0}^{3}\frac{(-1)^p}{p!(3-p)!}\int_{(k-1)h}^{kh}(t-\alpha)_+^{3-1}\,\mathrm{d}t \quad \text{for } \alpha=(\ell+p)h\\
&= x_{k-1}+\Bigg(w_{k-3}\frac{3}{h^2}\sum_{p=0}^{3}\frac{(-1)^p}{p!(3-p)!}\int_{(k-1)h}^{kh}(t-(k-3+p)h)_+^2\,\mathrm{d}t\\
&\quad + w_{k-2}\frac{3}{h^2}\sum_{p=0}^{3}\frac{(-1)^p}{p!(3-p)!}\int_{(k-1)h}^{kh}(t-(k-2+p)h)_+^2\,\mathrm{d}t\\
&\quad + w_{k-1}\frac{3}{h^2}\sum_{p=0}^{3}\frac{(-1)^p}{p!(3-p)!}\int_{(k-1)h}^{kh}(t-(k-1+p)h)_+^2\,\mathrm{d}t\Bigg)
\end{aligned}
$$

Table 1: A series of integration formulation for the discrete - time system simulation.

	Continuous-time signals	Continuous-time integrations
	$s(t) \in L_2(R)$	Given : $s(t) = \frac{\mathrm{d}x(t)}{\mathrm{d}t} = \mathrm{f}(x(t), u(t), t)$ Find : $x(kh) = \int_{-\infty}^{kh} s(t)\ \mathrm{d}t$
Order	Assumptions	Discrete-time integrations
1	$s(t) \in {}^1S$	$x_k = x_{k-1} + hw_{k-1}$
2	$s(t) \in {}^2S$	$x_k = x_{k-1} + \frac{h}{2}(w_{k-2} + w_{k-1})$
3	$s(t) \in {}^3S$	$x_k = x_{k-1} + \frac{h}{6}(w_{k-3} + 4w_{k-2} + w_{k-1})$
...	...	...
m	$s(t) \in {}^mS$	${}^m_h Int(t) : x_k = x_{k-1} + \sum_{\ell=k-m}^{k-1} w_\ell\ {}^m_{[b]}I_\ell$, where ${}^m_{[b]}I_\ell = \frac{m}{h^{m-1}} \sum_{p=0}^{m} \frac{(-1)^p}{p!(m-p)!} \int_{(k-1)h}^{kh} (t-\alpha)_+^{m-1}\ \mathrm{d}t$ for $\alpha = (\ell + p)h$

$$
\begin{aligned}
&= x_{k-1} + w_{k-3}\frac{3}{h^2}\frac{h^3}{18} + w_{k-2}\frac{3}{h^2}\frac{4h^3}{18} + w_{k-1}\frac{3}{h^2}\frac{h^3}{18} \\
&= x_{k-1} + \frac{h}{6}(w_{k-3} + 4w_{k-2} + w_{k-1}) \qquad (30)
\end{aligned}
$$

This example presents the advantage of the present series of integration formulation. This is formula of third order fluency integration method with w_{k-2}, w_{k-1} and w_k are B-spline coefficients computed from sample value of $s(t)$. The continuous-time signal is naturally modelled as a smooth function. The Table 1 shows the effectiveness of this fluency integration method. It is evident that the series of formulation includes and generalizes the conventional one, i.e. Euler's and Trapezoidal-like integration methods.

5 Conclusions

In this paper, a fluency tunable integration technique is derived. From the formulation, it is revealed that the fluency integration method includes and generalizes the conventional one, i.e. Euler's and Trapezoidal-like integration methods. The tunable parameters m (order of fluency approximation) and h (sampling interval) can be choosen adaptivelly according with smoothness property (continuously differentiable) of signal $s(t)$ we deal with. Thus, this concept provides a better generalized family of integration formula of the relationship between the discrete-time and continuous-time signal.

Because the discrete signal (w_k) is the B - spline coefficient computed from sampled value signal (s_k), such computation time leads to some problems in the real - time integration application. These problems are left for further investigation in the near future. Also, further interesting research is to apply this method to the real - time discrete - time system field, for example real - time digital flight simulation.

References

[1] H.P. William, A.T. Saul, T.V. William and P.F. Brian, "Numerical Recipes in C : the Art of Scientific Computing", *Cambridge Univ. Press*, pp. 129 - 164, 1992.

[2] Jon M. Smith, "Mathematical Modeling and Digital Simulation for Engineers and Scientists", *John Wiley & Sons*, New York, 1977.

[3] K. Toraichi, Bambang Sridadi and H. Inaba, "A Series of discrete-time models of a continuous-time system based on fluency signal approximation", *International Journal of Systems Science*, vol. 26, no. 4, pp. 871 - 881, 1995.

[4] M. Kamada, K. Toraichi and R. Mori, "Periodic spline orthonormal bases", *Journal of Approximation Theory*, vol. 55, no. 1, pp. 27 - 34, 1988.

[5] M. Kamada, K. Toraichi and R. Mori, "Spline function approach to digital signal processing", *International Journal of Systems Science*, vol. 19, no. 12, pp. 2473 - 2490, 1988.

Bambang Sridadi : Department of Simulation Technology, Indonesian Aerospace (IAe), Jl. Pajajaran 154, Bandung 40174, Indonesia.
Phone/Fax: +62 +22 667 0374, +62 +22 605 5408
E-mail: bsridadi@indonesian-aerospace.com

SUBHARMONIC RESONANCE OF A NONLINEAR SECOND ORDER EQUATION

Hartono

State University of Yogyakarta, Yogyakarta, Indonesia

Abstract. Rain-wind induced vibrations of stay cables of cable-stayed bridges can be modeled as a simple oscillator. The rain drops that hit the cables generate a rivulet on the surface of the cable hence changes the cross section of the cable. During the motion of the cable the position of rivulet may vary periodically. By using a quasi-steady approach to model the aerodynamic forces one arrives at a nonlinear second order equation.
Subharmonic resonance evaluated at this equation and the saddle node bifurcation occurred when the amplitude of the movement of the rivulet was varied.
Key-words: subharmonic resonance, oscillator

1 Introduction

As known that the vibration of stay cables of cable-stayed bridges can induced by rain-wind. This phenomena was studied experimentally by Hikami and Shiraishi [3] in a wind tunnel. Additional experimental work has been done by Matsumoto a.o. [4]. As has been observed on scale models in wind-tunnels the raindrops that hit the inclined stay cable generate one or more rivulets on the surface of the cable. The presence of flowing water on the cable changes the cross section of the cable as experienced by the wind field. Accordingly the pressure distribution on the cable with respect to the direction of the (uniform) wind flow may became asymmetric, resulting in a lift force perpendicular to direction of the wind velocity. The case with water rivulets can also be characterized by the presence of the ridge of water with the difference that this water ridge is not fixed to the surface of the cable. As long as the water ridge is present, it may be blown or shaken off. It is implying that one may assume that the position of the ridge varies in time. This is a conclusion in the paper by Ruscheweyh [6], where it is remarked that the rhythmic movement of the water rivulets, cable and aerodynamic force seem to be the key point for understanding the phenomenon.
The main resonance in the second order linear equation occur when the frequency of external force is equal to the natural frequency of the system. This implying that the solution become unbounded. In case the equation is nonlinear for instance the Duffing equation the main resonance lead to the jump phenomenon (see [5],pp.7-9). Another characteristic of nonlinear system is the secondary resonance or the subharmonic resonance i.e. the frequency of the external force is the multiplication of the natural frequency of the system, for example the frequency of the external force is twice of the natural frequency of the system as will be discussed in this paper.

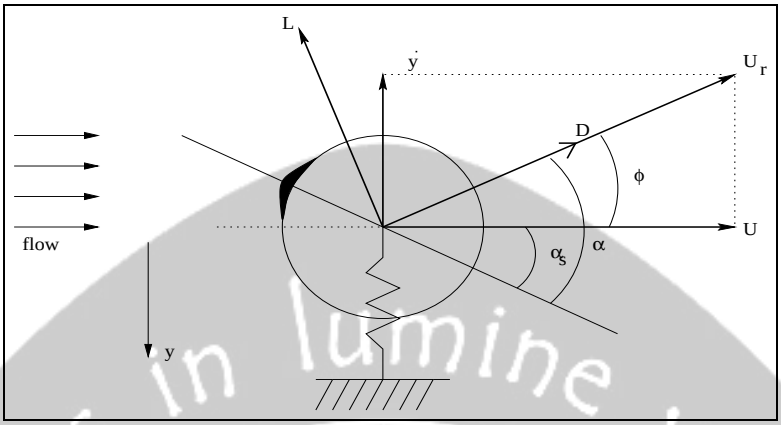


Figure 1: Cross-section of the cylinder-spring system, fluid flow with respect to the cylinder and wind forces on the cylinder

2 THE MODEL EQUATION FOR RAIN-WIND INDUCED VIBRATIONS OF A PROTOTYPE OSCILLATOR

The modeling principles we use are closely related to the quasi-steady approach as give in [1]. We consider a rigid cylinder with uniform cross-section supported by springs in a uniform rain-wind flow directed perpendicular to the axis of the cylinder. The oscillator is constructed in such a way that only vertical (one degree of freedom) oscillations are possible. The basic cross-section of the cylinder is circular, however on the surface of the cylinder there is a ridge able to carry out small amplitude oscillations. To model the rain-wind forces on the cylinder a quasi-steady approach is used; the type of oscillations which can be studied on the respective assumptions are known as galloping. A more detailed description of the quasi-steady approach can be found in [7]. The basic assumption of the quasi-steady approach is that at each moment in the dynamic situation the rain-wind force can be taken equal to the steady force exerted on the cylinder in static state. In the dynamic situation one should take into account that the flow-induced forces are based on the instantaneous flow velocity which is equal to the vector sum of flow velocity and the time varying vertical flow velocity induced by the (vertical) motion of the cylinder.

The steady rain-wind forces can be measured in a wind-tunnel and are expressed in the form of non-dimensional aerodynamic coefficients which depend on the angle of attack α. This angle, an essential variable for the description of the dynamics of the oscillator, is defined as the angle between the resultant flow velocity and an axis of reference fixed to the cylinder; measured positive in clockwise direction. The system we will study in more detail is sketched in fig 1.

The horizontal wind velocity is U and as the cylinder is supposed to move in the positive y direction, there is a virtual vertical wind velocity $-\dot{y}$. The drag

force D is indicated in the direction of the resultant wind-velocity U_r, whereas the lift force L is perpendicular to D in anti clockwise direction. The ridge on the cylinder bold indicated in fig 1. is able to carry out small amplitude oscillations. The aerodynamic force F_y in vertical direction can easily be derived from fig 1. :

$$F_y = -D\sin\phi - L\cos\phi \tag{1}$$

where ϕ is the angle between U_r and U, positive in clockwise direction, with $|\phi| \leq \pi/2$.
The drag and lift force are given by the empirical relations:

$$\begin{aligned} D &= \frac{1}{2}\rho\ d\ l\ U_r^2\ C_D(\alpha) \\ L &= \frac{1}{2}\rho\ d\ l\ U_r^2\ C_L(\alpha) \end{aligned} \tag{2}$$

where ρ is the density of air, d the diameter of the cylinder, l the length of the cylinder, $C_D(\alpha)$ and $C_L(\alpha)$ are the drag and lift coefficient curves respectively, determined by measurements in a wind-tunnel.

From fig 1. it follows that :

$$\begin{aligned} \sin\phi &= \dot{y}/U_r \\ \cos\phi &= U/U_r \\ \alpha &= \alpha_s + \arctan(\dot{y}/U) \end{aligned} \tag{3}$$

The equation of motion of the oscillator readily becomes :

$$m\ddot{y} + c_y\dot{y} + k_y y = F_y, \tag{4}$$

where m is the mass of the cylinder, $c_y > 0$ the structural damping coefficient of the oscillator, $k_y > 0$ the spring constant.
By using (2) and (3) we obtain for F_y:

$$F_y = -\frac{1}{2}\rho\ d\ l\sqrt{U^2 + \dot{y}^2}\ (C_D(\alpha)\dot{y} + C_L(\alpha)U) \tag{5}$$

Setting $\omega_y^2 = k_y/m$, $\tau = \omega_y t$ and $z = \omega_y y/U$ equation (4) becomes:

$$\begin{aligned} \ddot{z} + 2\beta\dot{z} + z &= -K\sqrt{1+\dot{z}^2}\ (C_D(\alpha)\dot{z} + C_L(\alpha)) \\ \alpha &= \alpha_s + \arctan(\dot{z}) \end{aligned} \tag{6}$$

where $2\beta = c_y/m\omega_y$ and $K = \rho\ d\ lU/2m\omega_y$ are non-dimensional parameters, and $\dot{z}$ now stands for differentiation with respect to τ.
We study the case where the drag and lift coefficient curve can be approximated by a constant and a cubic polynomial respectively:

$$\begin{aligned} C_D(\alpha) &= C_{D_o} \\ C_L(\alpha) &= C_{L_1}(\alpha - \alpha_o) + C_{L_3}(\alpha - \alpha_o)^3, \end{aligned} \tag{7}$$

where C_{D_o}, C_{L_1} and C_{L_3} are real parameters with $C_{D_o} > 0$ and for the interesting cases $C_{L_1} < 0$ and $C_{L_3} > 0$. By using $\alpha = \alpha_s + \arctan \dot{z}$ we obtain for $C_L(\alpha)$:

$$C_L(\alpha) = C_{L_1}(\alpha_s - \alpha_o + \arctan \dot{z}) + C_{L_3}(\alpha_s - \alpha_o + \arctan \dot{z})^3 \tag{8}$$

The cases that $\alpha_s = \alpha_o$ and $\alpha_s \neq \alpha_o$ where α_s and α_o are (time independent) parameters have been studied in [1]. Here we study the case that the position of the (water) ridge varies with time:

$$\alpha_s - \alpha_o = f(t) = f(\tau/\omega_y) \tag{9}$$

Substitution of (8) and (9) in (6) and expanding the right hand side with respect to $\dot{z}$ in the neighborhood of $\dot{z} = 0$ yields:

$$\begin{aligned} \ddot{z} + z &= -K[C_{L_1} f(t) + C_{L_3} f^3(t) + \\ &\quad (C_{D_o} + C_{L_1} + 2\beta/K + 3C_{L_3} f^2(t))\ \dot{z} + \\ &\quad (\frac{1}{2} C_{L_1} f(t) + \frac{1}{2} C_{L_3} f^3(t) + 3C_{L_3} f(t))\ \dot{z}^2 + \\ &\quad (\frac{1}{6} C_{L_1} + C_{L_3} + \frac{1}{2} C_{D_o} + \frac{1}{2} C_{L_3} f^2(t))\ \dot{z}^3] + 0(\dot{z}^4) \end{aligned} \tag{10}$$

This derivation can also be found on [2]. If the order $\dot{z}^4$ and more greater are neglected yield a cubical nonlinear second order equation.

2.1 The subharmonic resonance of a second order nonlinear equation

One may assume that the time varying position of the ridge has a similar character as the motion of the cable i.e. if the cable oscillates harmonically then one may expect that the water ridge moves accordingly. Furthermore, one can also assume that the frequency oscillation of the water ridge is twice the frequency of the motion of the cable. So in this case one can put $f(t) = A\cos\omega t = A\cos(\frac{\omega}{\omega_y}\tau) = A\cos\Omega\tau$ where $\Omega = \frac{\omega}{\omega_y}$ with

$$\begin{aligned} f^2(t) &= \frac{1}{2} A^2 (1 + \cos 2\Omega\tau), \\ f^3(t) &= \frac{3}{4} A^3 (\cos\Omega\tau + \frac{1}{3}\cos 3\Omega\tau) \end{aligned}$$

and $\Omega = 2 + \epsilon\eta$ where $|\epsilon| \ll 1$. By setting $(2 + \epsilon\eta)\tau = \theta$ (10) becomes:

$$\begin{aligned} \ddot{z} + \frac{1}{4} z &= -K[A_2 \cos\theta + A_3 \cos 3\theta + \\ &\quad \frac{1}{2}(A_o + A_1 \cos 2\theta)\ \dot{z} - (\frac{1}{4}\epsilon\eta/K) z + \\ &\quad (A_4 \cos\theta + \frac{1}{2} A_3 \cos 3\theta\ \dot{z}^2 + \\ &\quad 2(A_5 + \frac{1}{6} A_1 \cos 2\theta)\ \dot{z}^3] \end{aligned} \tag{11}$$

where

$$\begin{aligned} A_o &= C_{D_o} + C_{L_1} + 2\beta/K + \frac{3}{2}C_{L_3}A^2, \\ A_1 &= \frac{3}{2}C_{L_3}A^2, \\ A_2 &= C_{L_1}A + \frac{3}{4}C_{L_3}A^3, \\ A_3 &= \frac{1}{4}C_{L_3}A^3, \\ A_4 &= \frac{1}{2}C_{L_1}A + 3C_{L_3}A + \frac{3}{8}C_{L_3}A^3, \\ A_5 &= \frac{1}{6}C_{L_1} + C_{L_3} + \frac{1}{2}C_{D_o} + \frac{1}{4}C_{L_3}A^2, \end{aligned}$$

and a dot now stands for differentiation with respect to θ. This situation is called the subharmonic resonance of the second order cubical nonlinear equation.
The affect of A (the amplitude of the movement of the water ridge) to the system can be observed by setting A varies and where the other parameters are given the following numerical values: $C_{D_o} = 0.5$, $C_{L_1} = -6.0$, $\beta/K = 2.0$, and $C_{L_3} = 2.0$.
By application of transformation :

$$\begin{aligned} z &= y_1 \cos(\frac{1}{2}\theta) + 2y_2 \sin(\frac{1}{2}\theta), \\ \dot{z} &= -\frac{1}{2}y_1 \sin(\frac{1}{2}\theta) + y_2 \cos(\frac{1}{2}\theta), \end{aligned} \tag{12}$$

after first order averaging one obtain:

$$\begin{aligned} \dot{\bar{y}}_1 &= \frac{K}{8}\bar{y}_1[(F - G\bar{y}_1^2 - 4H\bar{y}_2^2] - \frac{1}{2}\epsilon\eta\bar{y}_2, \\ \dot{\bar{y}}_2 &= \frac{K}{8}\bar{y}_2[(F - H\bar{y}_1^2 - 4G\bar{y}_2^2] + \frac{1}{8}\epsilon\eta\bar{y}_1, \end{aligned} \tag{13}$$

where

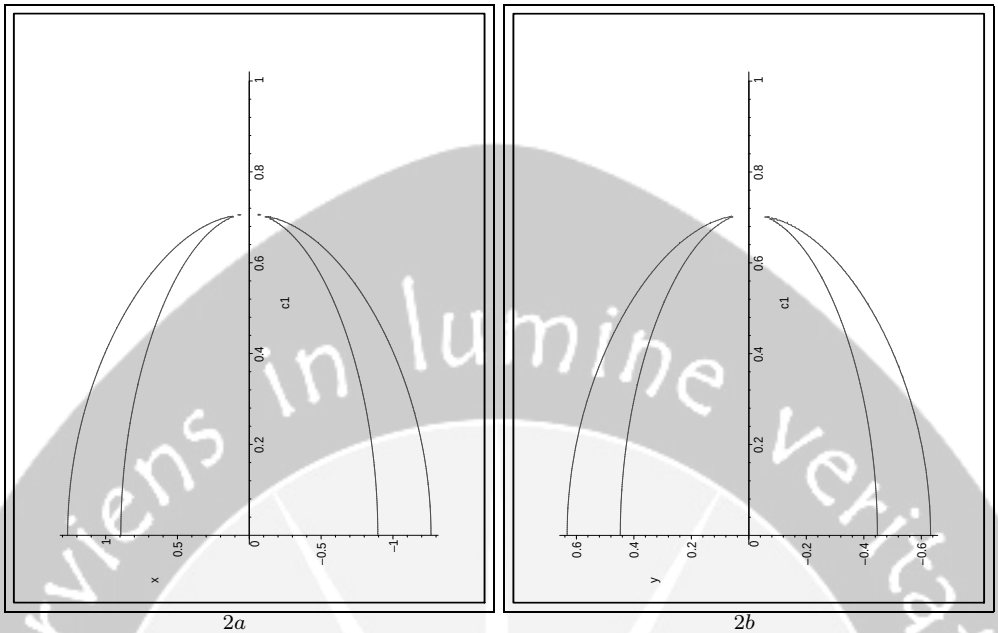
$$F = 3 - 6A^2, \qquad G = \frac{15}{8} + \frac{7}{8}A^2 \quad \text{and} \quad H = \frac{15}{8} + \frac{3}{8}A^2.$$

In case $\eta = 0$ and $0 < A < \frac{1}{2}\sqrt{2}$, the system (13) have 9 critical points that are

$$(0,0), \quad (\pm\sqrt{\frac{F}{G}}, 0), \quad (0, \pm\frac{1}{2}\sqrt{\frac{F}{G}}), \quad (\pm\sqrt{\frac{F}{G+H}}, \pm\frac{1}{2}\sqrt{\frac{F}{G+H}}).$$

Further, the Jacobian of the system (13) is :

$$\frac{K}{8}\begin{pmatrix} F - 3G\bar{y}_1^2 - 4H\bar{y}_2^2 & -8H\bar{y}_1\bar{y}_2 \\ -2H\bar{y}_1\bar{y}_2 & F - H\bar{y}_1^2 - 12G\bar{y}_2^2 \end{pmatrix}.$$



2a 2b

Figure 2: 2a. Relation curve between A and $\bar{y}_1$. Vertical axis is A and horizontal axis is $\bar{y}_1$. 2b. Relation curve between A and $\bar{y}_2$.

The linearization around these critical points yields that the origin is a unstable node because the Jacobian matrix evaluated in the origin have two real positive eigenvalues. The next four are saddle, because the Jacobian evaluated at these points have one real positive and one real negative eigenvalue. Then the rest are stable node, because the Jacobian evaluated at these points have two real negative eigenvalues.
In case $\eta = 0$ and $A \geq \frac{1}{2}\sqrt{2}$, the system (13) only have one stable node critical point that is the origin. The node stability of the origin due to the right hand side of the system (13) are definitely negative.
Finally, one can conclude that in case $\eta = 0$ the system (13) bifurcate at $A = \frac{1}{2}\sqrt{2}$ and the bifurcation is saddle node bifurcation because the saddle point and the node points are collapse. The diagram bifurcation is depicted in Figure 2.a and 2.b. These figure are plotted by using the software Maple and the calculation base on the Gröbner bases algorithm as shown in the following illustration.
Suppose that

$$\begin{aligned} p_1 &= \frac{K}{8}\bar{y}_1(F - G\bar{y}_1^2 - 4H\bar{y}_2^2), \\ p_2 &= \frac{K}{8}\bar{y}_2(F - H\bar{y}_1^2 - 4G\bar{y}_2^2) \end{aligned} \tag{14}$$

and I is an ideal generated by $\{p_1, p_2\}$. By using the Gröbner bases algorithm in

software Maple, one found another bases of I i.e. $\{q_1, q_2, q_3, q_4\}$ where

$$\begin{aligned} q_1 &= 225\bar{y}_1^5 + 180\bar{y}_1^5A^2 + 35\bar{y}_1^5A^4 + 876\bar{y}_1^3A^2 - 540\bar{y}_1^3 + \\ &\quad 408\bar{y}_1^3A^4 + 1152\bar{y}_1A^4 - 1152\bar{y}_1A^2 + 228\bar{y}_1, \\ q_2 &= 5\bar{y}_2\bar{y}_1^3A^2 + 24\bar{y}_1\bar{y}_2A^2 + 15\bar{y}_2\bar{y}_1^3 - 12\bar{y}_1\bar{y}_2, \\ q_3 &= 48\bar{y}_1A^2 - 24\bar{y}_1 + 15\bar{y}_1^3 + 7\bar{y}_1^3A^2 + 60\bar{y}_1\bar{y}_2^2 + 12\bar{y}_1\bar{y}_2^2A^2, \\ q_4 &= 48\bar{y}_2A^2 - 24\bar{y}_2 + 60\bar{y}_2^3 + 28\bar{y}_2^3A^2 + 15\bar{y}_1^2\bar{y}_2 + 3\bar{y}_1^2\bar{y}_2A^2. \end{aligned} \tag{15}$$

Furthermore, the system $p_1 = 0, p_2 = 0$ is equivalent to the system $q_1 = 0, q_2 = 0, q_3 = 0, q_4 = 0$. The polynomial q_1 is a polynomial in A and $\bar{y}_1$, in other words q_1 only depends on A and $\bar{y}_1$. The relation curve between A and $\bar{y}_1$ can be found by using the command implicit-plot at software Maple, that is the implicit-plot of $q_1 = 0$ and the result is depicted in Figure 2.a. Similarly one can found the Figure 2.b. If one make a horizontal line in Figure 2.a as well as in Figure 2.b, then the line will cross the curve at five points for value of A between 0 and $\frac{1}{2}\sqrt{2}$ and the line will cross the curve at one point for the value of A is greater than or equal to $\frac{1}{2}\sqrt{2}$.

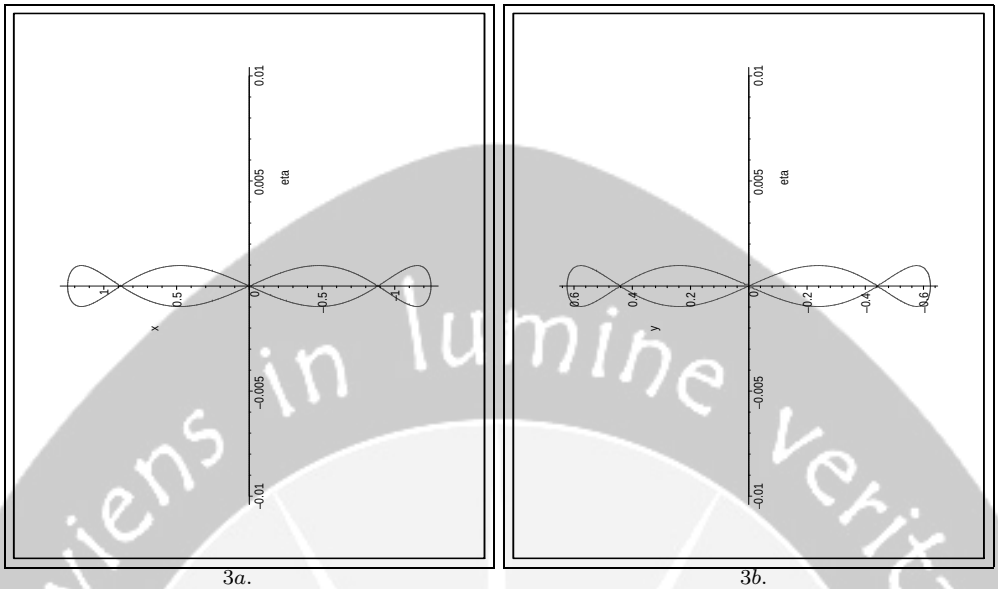
To analyzing the effect of the detuning one can setting A fixed and keep the η as a parameter. So the number of critical points only depends on parameter η. For instance $A = 0.1$ the relation between η and $\bar{y}_1$ depicted in fig. 3a. The number of critical points of equation (13) indicate by the number of intersection point between the curve in fig. 3a with the horizontal line. For instance $\eta = 0.0005$ there exist 9 critical points because the horizontal line $\eta = 0.0005$ crossing the curve at nine points, but for $\eta = 0.0015$ only one critical point i.e. the origin. This situation implying that there exist a bifurcation for value of η between 0.0005 and 0.0015.

3 Concluding and Remarks

The vibration of a stay cables of cable stayed bridges can be modeled as a simple oscillator and yielding a model equation of a nonlinear second order. The amplitude of movement of the water ridge on the cable make important role on this phenomenon. When the frequency oscillation of the water ridge is almost twice of the natural frequency of the cable (subharmonic resonance) and when the amplitude of the movement of the water ridge (A) is varied, then the saddle node bifurcation occur i.e. at value of A equal to $\frac{1}{2}\sqrt{2}$. Similarly, the saddle node bifurcation also occur when the detuning parameter (η) is varied and the amplitude of the movement of the water ridge keep fixed.

Acknowledgment

The author would like to thank to A.H.P. van der Burgh for interesting discussion on the derivation of the model equation.



3*a*. 3*b*.

Figure 3: 5a. Curve relation between $\bar{y}_1$ and η after using Gröbner basis algorithm. Horizontal axis is η and vertical axis is $\bar{y}_1$. 5b. Curve relation between $\bar{y}_2$ and η.

References

[1] Haaker,T.I. and Van der Burgh, A.H.P. (1994), "On the dynamics of aeroelastic oscillators with one degree of freedom, *SIAM J. Appl. Math.* , **54**, 1033 – 1047.

[2] Van der Burgh, A.H.P. and Hartono (2004), "Rain-wind induced vibrations of a simple oscillator", *Int. J. Nonlin. Mech.* **39**, 93 – 100.

[3] Hikami, Y. and Shiraishi, N. (1998), "Rain-wind induced vibrations of cables in cable stayed bridges", *J. Wind Engineer. and Industr. Aerodyn.* **29**, 409-418.

[4] Matsumoto, M., Shiraishi, N., Kitazawa, M., Knisely, C., Shirato, H., Kim, Y. and Tsujii, M. (1990), "Aerodynamic behavior of inclined circular cylinders-cable aerodynamics" *J. Wind Engineer. and Industr. Aerodyn.*, **33**, 63 – 72.

[5] Nayfeh, A.H. and Mook, D.T. (1985), *Nonlinear Oscillation* , John Wiley & Sons, New York.

[6] Ruscheweyh, H. (1999), "The mechanism of rain-wind induced vibration", Wind Engineering into the 21th century Vol. 2, pp. 1041-1047. *Proc. of the 10th Intern. Conf. on Wind Engineering*, Copenhagen 1999, Balkema, Rotterdam .

[7] Van der Burgh, A.H.P. (1999), *Nonlinear Dynamics of Structures Excited by Flows: Quasi-steady Modeling and Asymptotic Analysis in Fluid-Structure Interactions in Acoustics*, CISM Courses and Lectures No. 396, Springer Wien New York.

Analytical Study of Chaotic Solution of Autoparametric System with Parametric Excitation

S. Fatimah

Mathematics dept. of UPI, Bandung, Indonesia

Abstract: We analytically study the existence of chaotic dynamics on Autoparametric System with parametric excitation. The method of averaging is used to yield a set of autonomous equation of the approximation to the response of the system. We use a global perturbation method developed by Kovacic and Wiggins to analyze the parameter range for which a Shilnikov type Homoclinic orbit exists. This orbit gives rise to a well-described chaotic dynamics.

Keywords: Dynamical system

A Strongly Nonlinear Fractional Rayleigh Oscillator

S. B. Waluya

UNNES, Semarang, Indonesia

Abstract: In this paper a Strongly Nonlinear Fractional Rayleigh Oscillator will be studied. It will be shown that the recently developed perturbation method based on integrating factors can be used to approximate first integrals. Not only approximations of first integrals will be given, but it will also be shown how in a rather efficient way the existence and stability of time-periodic solutions can be obtained from these approximations. In particular the Strongly Nonlinear Fractional Rayleigh Oscillator equation

$$\ddot{X} + \alpha X + \mu X^2 = \varepsilon(1 - \dot{X}^2)\dot{X}^{1/3}$$

will be studied in detail.

Keywords: Strongly Nonlinear, Fractional Rayleigh Oscillator, Perturbation Method, First Integral

Analysis of a Class of a Nonlinear Mathieu Equation with an Application to Flow Induced Vibrations

H. Lumbantobing

Institute of Research, University of Cenderawasih, Jayapura-Papua, Indonesia

Abstract: This paper is concerned with an analysis of a nonlinear Mathieu equation with an application to flow induced vibrations. The nonlinear Mathieu equation can be interpreted as a model of equation of an oscillator with a parametric excitation and a nonlinear perturbation. Assuming the parametric excitation and the nonlinear perturbation are small then we can apply an averaging method to analyze the system's dynamic behaviour. Criteria for the stability of trivial solution, the existence and the stability of various nontrivial (periodic) solutions and their bifurcations are given. We apply some results obtained to a model describing aeroelastic oscillations of a structure with one degree of freedom.

Keywords: Mathieu equation, aeroelasticity, nonlinear oscillation, averaging method, bifurcation.

Free-Surface Flow Caused by a Source

L.H. Wiryanto

Department of Mathematics, Institut Teknologi Bandung, Indonesia

Abstract: A flow caused by a line source is considered in a channel of finite depth. The source is placed at the bottom of the channel, and it produces a free surface with a cusp pointing to the source. We assume that the fluid is inviscid and incompressible, and the flow is irrotational so that the flow can be expressed as a boundary value problem of potential function from Laplace equation. Numerical solutions of this problem are computed by an integral equation method, constructed by transforming the flow domain conformally and introducing a hodograph variable. The numerical procedure is then used to observe the relationship between the nondimensional parameter Froude number F, based on the downstream flow, and the distance y_a of the separation point of the cusp. When $F \to \infty$ we obtain the limiting solution with $y_a = 0.363$.

Keywords: Free-surface flow, line source, integral equation method, hodograph variable

ENVIRONMENTAL IMPACT MODELING OF SEAWATER DESALINATION PLANTS IN THE RED SEA

Anton Purnama[a], H.H. Al-Barwani[a], Ronald Smith[b]

[a] Sultan Qaboos University, Sultanate of Oman
[b] Loughborough University, United Kingdom

Abstract. With limited and depleting natural resources of groundwater, desalinating seawater can supplement some of the critically lacking amounts of water needed for sustainability in Saudi Arabia. If desalination plants were to operate along the coasts of the arid climate Red Sea, the additional loss of water and the continuous disposal of brine waste due to the plant's water production would then increase the salinity. A mathematical model is presented to calculate the impact of desalination plants on the salinity within a semi-enclosed sea of simple geometry. Due to the exponential sensitivity to the plant's location and its water production rate, the impact of desalination plants at the northern Red Sea is found to be more severe.
Key-words: Hypersaline, mathematical model, Red Sea, seawater desalination

1 Introduction

The Red Sea plays an essential role in providing Saudi Arabia with water produced by desalination of seawater to meet the kingdom's continuously growing demands for water as the consequences of its rapid industrial development and population growth [2,11]. As drought conditions worsen, almost all the existing underground water resources have been developed and are being exploited at an unsustainable rate. In fact, the excessive mining of the groundwater resources are being accompanied by increased urban industrial and agricultural pollution. To avert the real threat to resource sustainability, Saudi Arabia is steeping up efforts to boost long-term availability by building seawater desalination plants. Once a desalination plant is built, its daily water production capacity will subsequently be increased in line with the projected demands.

By 2006, the daily total production capacity of seawater desalination plants in Saudi Arabia will grow to 5.65 million m^3 of water. Some of these plants are located on the coast of the Arabian Gulf, and as shown in Fig.1, the majority of the plants are on the coast of the Red Sea [2]. The total water produced by the desalination plants in the Red Sea was estimated at 2.65 million m^3 of water per day [10]. The largest plants are Shuaiba with a capacity of 1.1 million m^3 of water per day and Jeddah with 0.4 million m^3 for supplying water to the two most populated cities of Jeddah and Makkah.

Seawater desalination processes also produce reject brine, highly concentrated salt water (up to factors of 2.5) to be disposed of into the Red Sea [1,10]. Assuming a 60% recovery rate of a desalination plant, the total volume of brine waste is continuously being discharged in excess of 1.76 million m^3 per day.

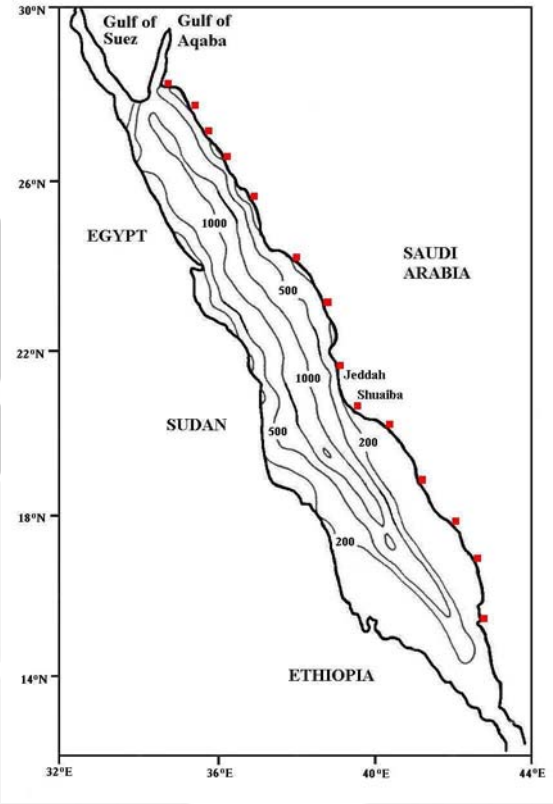


Fig. 1. Desalination plants of Saudi Arabia in the Red Sea.

The Red Sea is a part of the World Rift system separating the African continent from Arabia [8]. It is a long narrow basin, almost in a straight line, a distance of 2000 km with widths ranging from 145 km to 306 km, and eventually constricted to 26 km at the southern end in the Strait of Bab el Mandab (Fig. 1). The bottom topography is wedge shaped with large depths in the center of the basin. The deepest part of more than 2000 m has been recorded at the central Red Sea, but the average depth of the Red Sea is only 490 m. The area of the Red Sea is about 450000 km^2.

The exchange of water between the Red Sea and the Gulf of Aden occurs at the Strait of Bab el Mandab. There is no surface water run off because no rivers enter the Red Sea [5,11]. The rainfall over the Red Sea and its coasts is extremely small. Moisture exchange and solar heating across the air-sea interface is enhanced by the extremely arid nature of the bordering lands. The annual mean net evaporation (minus precipitation) is estimated at 2 m/yr [8], thus generating high salinity with the increase of surface salinity observed from north to south [3,5,8] and, in particular, surface salinity of more than 40 ppt observed at its northern end.

The semi-enclosed Red Sea is the most saline body of water in the world's oceans and is environmentally very fragile; therefore any further loss of water by

desalination plants and the returned discharge of brine waste would make the Red Sea become hypersaline. To assess the impact of seawater desalination on the salinity, a mathematical model is developed. Using a simple channel geometry representation of the semi-enclosed marginal sea, the model shows that seawater desalination at the northern Red Sea would deteriously change the salinity. Although the current production capacity of desalination plants are safely in the linear regime, special attention should be given in the long-term water planning as the impact depends exponentially on the plant's location and its volumetric rate of seawater extraction.

2 Solution of the model's equations

Due to lack of data and other physical properties, the scales of variability and many aspects of the currents in the Red Sea are still poorly understood [3,5,14]. The exchange of water through the Strait of Bab el Mandab [9,12] is not yet well measured. Thus, to develop a mathematical model, it is necessary to make many simplifying assumptions. We model the Red Sea as a semi-enclosed sea joining to the Gulf of Aden at $x = L$ with salinity s_L.

The equation of mass flux of water is a balance between the incoming tidally averaged current $U(x)$ through the cross-sectional area $A(x)$ with continuous depletion by evaporation at the rate μ and the seawater intake and brine waste discharges from a desalination plant located at $x = a$:

$$\frac{d}{dx}(AU) = -\mu B - rQ\delta(x-a),$$

where $B(x)$ is the channel width, rQ the rate of the plant's water production and δ is the Dirac delta function. The plant's recovery ratio is typically $r \le 0.6$. On integrating, we obtain

$$AU = \begin{cases} -\mu \int_0^x B(z)dz, & 0 \le x < a \\ -rQ - \mu \int_0^x B(z)dz, & a \le x < L \end{cases}. \qquad (1)$$

The surface salinity in the Red Sea increases roughly with distance from the Strait of Bab el Mandab [3,8], so a one-dimensional advection-diffusion approach can be adopted [6,7]:

$$\frac{d}{dx}(AUs) - \frac{d}{dx}\left(AD\frac{ds}{dx}\right) = Qs\,\delta(x-a), \qquad (2)$$

where $D(x)$ is the tidally averaged shear-dispersion coefficient. Note that if seawater of salinity s is removed at the volumetric rate Q, then the discharge rate of brine waste from the plant is $(1-r)Q$ with salt concentration $s/(1-r)$.

On integrating (2) and substituting for AU from (1), then matching the salinity to s_L at $x = L$, we finally obtain the logarithm of relative salinity

$$\ln\left(\frac{s}{s_L}\right) = \int_x^L \frac{dz}{AD}\left(\mu\int_0^z B(p)dp\right) + (1+r)Q\int_a^L \frac{dz}{AD}, \quad 0 \le x < a, \tag{3}$$

and

$$\ln\left(\frac{s}{s_L}\right) = \int_x^L \frac{dz}{AD}\left(\mu\int_0^z B(p)dp + (1+r)Q\right), \quad a \le x < L. \tag{4}$$

Thus, the salinity increase due to seawater desalination $\Delta s = s - s_*$ can be evaluated from

$$\frac{\Delta s}{s_*} = \begin{cases} \exp\left[(1+r)Q\int_a^L \frac{dz}{AD}\right] - 1, & 0 \le x < a \\ \exp\left[(1+r)Q\int_x^L \frac{dz}{AD}\right] - 1, & a \le x < L \end{cases}, \tag{5}$$

where s_* is the salinity for the case without seawater desalination. The increase is exponentially dependent on the seawater intake rate Q and the location of the plant $x = a$. Therefore, as illustrated in Fig. 3, it is more substantial the further the desalination plant is from the open sea $x = L$.

3 The impact on the salinity of the Red Sea waters

Along the Red Sea, the channel depth and width vary markedly. Thus, unless we can approximate the variations by simple functions, the solutions will have to be evaluated numerically. We model channel geometry of the semi-enclosed sea with $B(x) = B_L$ and $H(x) = 2H_L$ [13]. The channel depth profile is specified as the topographic surface

$$\frac{z}{2H_L} = -1 + \left(\frac{y}{B_L}\right)^2 .$$

As shown in Fig. 2, the channel is of constant width and of parabolic cross-sections.

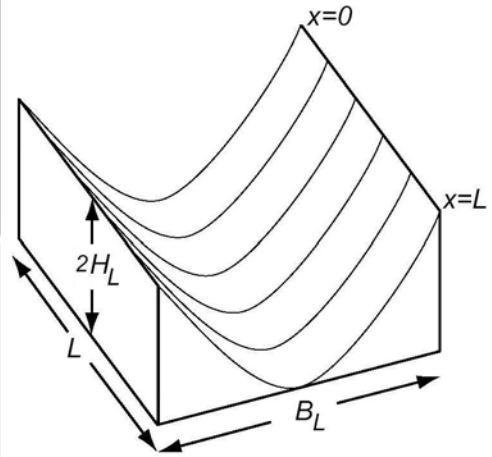


Fig. 2. Simple channel geometry of constant width with parabolic cross-sections.

To illustrate how a seawater desalination plant operated along its coast would change the salinity distribution within the semi-enclosed sea, we also need to model the longitudinal dispersion coefficient D. Assuming that the water exchange with the open sea at $x = L$ is the main source of water for the semi-enclosed sea, we have

$$A(x)U(x) = U_L \int_0^x B(z)dz ,$$

where U_L is the tidally averaged value of the rate of change of water depth. Next, as the vertical shear dispersion dominates [8,14], D is proportionally to HU [4], so that

$$D(x) = \frac{\alpha U_L}{B(x)} \int_0^x B(z)dz .$$

The model parameter αU_L can be estimated numerically from measurements of surface salinity in the Red Sea. Therefore, in the case without seawater

desalination, by taking the observed surface salinity s_* at $x = x_*$, the mathematical model is "scaled" to the Red Sea. By putting $Q = 0$ in (4), we have

$$\alpha U_L = \frac{\mu}{\ln(s_*/s_L)} \int_{x_*}^{L} \frac{dz}{H(z)}. \tag{6}$$

The typical values relevant to the Red Sea are $B_L = 225$ km, $H_L = 500$ m, $L = 2000$ km and the annual mean evaporation rate $\mu = 2$ m/yr. Using $s_* = 40$ ppt as the value of salinity at the central Red Sea $x_*/L = 0.5$ (approximating the location of the Shuaiba and Jeddah desalination plants) and the salinity at the Strait of Bab el Mandab $s_L = 37$ ppt [9,12], (6) gives $\mu L/\alpha U_L H_L = 4\ln(s_*/s_L) \approx 0.312$. The other parameters related to the seawater desalination plant are $r = 0.6$ and the Shuaiba and Jeddah plants' total annual water production rate $rQ_* \approx 0.5475$ km^3/yr. Hence, $q_* = (1+r)Q_*/\mu L B_L \approx 1.62 \times 10^{-3}$.

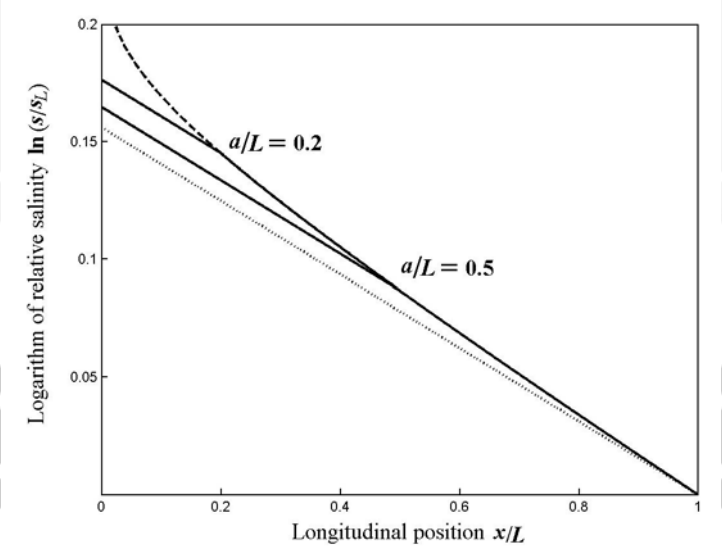


Fig. 3. Logarithm of relative salinity due to a desalination plant with $q = 50q_*$.

Thus, from (3) and (4), for a desalination plant with $q = (1+r)Q/\mu L B_L$ located at $x = a$, the logarithm of the relative salinity in the Red Sea is given by

$$\ln\left(\frac{s}{s_L}\right) \approx 0.156\left[1-\frac{x}{L}+\ q\ln\left(\frac{L}{a}\right)\right], \quad 0 \le \frac{x}{L} < \frac{a}{L},$$

and

$$\ln\left(\frac{s}{s_L}\right) \approx 0.156\left[1-\frac{x}{L}+q\ln\left(\frac{L}{x}\right)\right], \quad \frac{a}{L} \le \frac{x}{L} < 1.$$

Fig. 3 plots the logarithm of relative salinity with $q = 50q_*$ at two plant locations of $a/L = 0.2$ and 0.5. For comparison, the hypersaline condition without seawater desalination is shown by the dotted curve, and a plant at the head of the Red Sea $a = 0$ by the dashed curve. Note that from (5), the salinity increase is also due to both the plant's water production rate and its location.

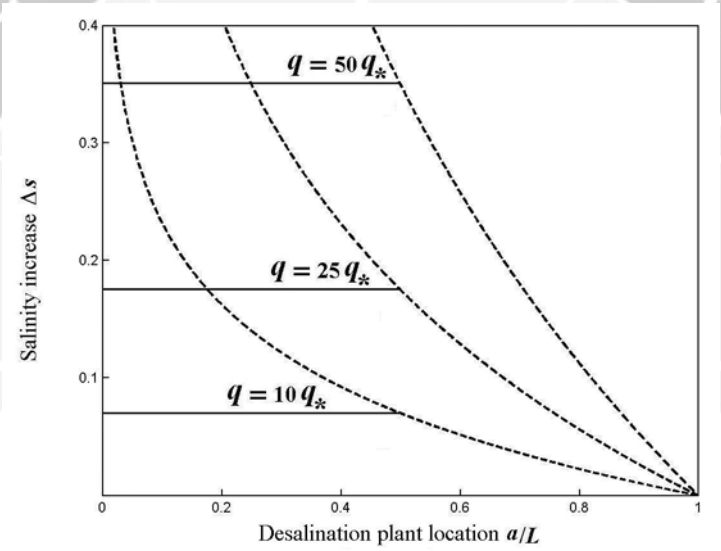


Fig. 4. The Red Sea salinity increase due to seawater desalination at $a/L = 0.5$.

The salinity increase (5) due to a desalination plant at $a/L = 0.5$ in the Red Sea with $s_* = 40$ ppt is shown in Fig. 4. The impact of a plant with the production rate $q = 10q_*$ is corresponding to the salinity increase of 0.07 ppt, and if the plant's capacity is increased further to $q = 50q_*$, the salinity increase is 0.35 ppt. The dashed curve represents the salinity increase due to a desalination plant located at a/L with the production rate q. Therefore, for the production rate of $q = 25q_*$ and by relocating the plant towards the head of the Red Sea to $a/L = 0.25$, the impact is the salinity increase of 0.35 ppt.

4 Conclusions

A desalination plant's reliability and service availability are essential to sustain and allow the continuing long-term socio-economic development in Saudi Arabia. By 2010, the domestic and industrial water demand is expected to double to more than 10 million m^3 per day [2,11]. Regrettably, as desalinated water is indispensably required at any cost, the potential impacts on the salinity of the Red Sea waters have so far been ignored. Lessons from the exploitation of the groundwater resources, which are not only consuming the natural resources but also contaminating them, should have been learned so that meeting the water demands by seawater desalination should not necessarily be at the expense of the Red Sea's fragile environment.

Due to its semi-enclosed nature and arid climate, in order to minimize the impact of the desalination plants on the hypersaline Red Sea, care should be taken to determine which of the desalination plants production rates to be increased. Higher seawater salinity also reduces the desalination plant's recovery ratio, and hence increases the cost of desalinated water.

Acknowledgment

This work was supported by the Internal Grant at Sultan Qaboos University under the research project code IG/SCI/DOMS/05/07.

References

[1] Al-Muataz, I.S. (1991), Environmental impact of seawater desalination plants, *Environmental Monitoring and Assessment*, **16**, 75-84.

[2] Akkad, A.A. (1990), Conservation in the Arabian Gulf countries, *Management and Operations, Journal of the American Water Works Association*, **May**, 40-50.

[3] Clifford, M., Horton, C., Schmitz, J. and Kantha, L.H. (1997), An oceanographic nowcast/forecast system for the Red Sea, *Journal of Geophysical Research*, **102**, 25101-25122.

[4] Elder, J. (1959), The dispersion of marked fluid in turbulent shear flow, *Journal of Fluid Mechanics*, **5**, 544-560.

[5] Johns, W.E., Jacobs, G.A., Kindle, J.C., Murray, S.P. and Carron, M. (2000), Arabian marginal seas and gulfs, *Technical Report 2000-01*, RSMAS, University of Miami.

[6] Largier, J.L., Hearn, C.J. and Chadwick, D.B. (1996), Density structures in low inflow estuaries, *Coastal and Estuaries Studies*, **53**, 227-241.

[7] Largier, J.L., Hollibaugh, J.T. and Smith, S.V. (1997), Seasonally hypersaline estuaries in Mediterranean-climate regions, *Estuarine, Coastal and Shelf Science*, **45**, 789-797.

[8] Morcos, S.A. (1970), Physical and chemical oceanography of the Red Sea, *Oceanography Marine Biology Annual Review*, **8**, 73-202.

[9] Murray, S.P. and Johns, W.E. (1997), Direct observations of seasonal exchange through the Bab al Mandab Strait, *Geophysical Research Letters*, **24**, 2557-2560.

[10] Purnama, A., Smith, R. and Al-Barwani, H.H. (2005), Assessing the impact of seawater desalination plants on the hypersaline Red Sea, *Arabian Journal for Science and Engineering*, submitted.

[11] Shahin, M. (1989), Review and assessment of water resources in the Arab region, *Water International*, **14**, 206-219.

[12] Smeed, D. (1997), Seasonal variation of the flow in the Strait of Bab al Mandab, *Oceanologica Acta*, **20**, 773-781.

[13] Smith, R. (1977), Long-term dispersion of contaminants in small estuaries, *Journal of Fluid Mechanics*, **82**, 129-146.

[14] Tragou, E. and Garrett, C. (1997), The shallow thermohaline circulation of the Red Sea, *Deep-Sea Research*, **44**, 1355-1376.

ANTON PURNAMA: Department of Mathematics and Statistics, College of Science, Sultan Qaboos University, PO Box 36, Al-Khod 123, Muscat, Sultanate of Oman.
E-mail: antonp@squ.edu.om

H.H. AL-BARWANI: Department of Mathematics and Statistics, College of Science, Sultan Qaboos University, PO Box 36, Al-Khod 123, Muscat, Sultanate of Oman.
E-mail: hamdi@squ.edu.om

RONALD SMITH: Department of Mathematical Sciences, Loughborough University, Leicestershire, LE11 3TU, England, United Kingdom.
E-mail: ron.smith@lboro.ac.uk

COMPUTATIONAL ANALYSIS OF SCAVENGING GAS FLOW IN A TWO-STROKE LINEAR ENGINE

Tulus[a] & A. K. Ariffin[b]

[a]Universitas Sumatera Utara, Medan Indonesia
[b]Universiti Kebangsaan Malaysia, Bangi, Malaysia

Abstract. This paper presents the simulation of gas flow during the scavenging processes in a linear combustion engine incorporating combustion chamber and kickback chamber. The computation is performed using a finite volume method incorporating the Navier-Stokes equation and analyzed transiently in the moving mesh for three-dimensional model. During the compression and expansion stroke, the mesh moves due to the piston displacement. The results of the analysis are the distribution of velocity and pressure in the inlet port, scavenging, intake ports, combustion chamber and exhaust port.
Key-words: free piston, Navier-Stokes equation, finite volume method, moving mesh

1 Introduction

Linear internal combustion engines may find application in the generation of electrical power using linear motion. The operation of this engine is distinct from that of a conventional slider-crank mechanism engine, insofar as the motion of the two horizontally opposed pistons is not externally constrained [1]. This technology is advantageous because it is mechanically simpler and allows for a great deal more freedom in defining a piston motion profile, enabling the use of novel combustion regimes [2].

In the two-stroke engine design, the inlet frequency of gases introduced and expelled during each cycle is a major factor. The inlet and exhaust gases are moving at speeds designed to cause pressure waves. To achieve this operating cycle, a fresh charge of the gas must be supplied to the engine cylinder at a high enough pressure to displace the burned gases from the previous cycle. The combined intake and exhaust process that clears the cylinder of burned gases and fills it with a fresh mixture is called scavenging. The waves assist the filling of the scavenging and the extraction of the waste gases. The exhaust system is tuned to assist in this filling and extraction process [3]. Poor in-cylinder mixing due to ineffective fuel delivery is believed to be problematic in the engine [4].

Computational Fluid Dynamic (CFD) of in-cylinder flows in internal combustion engines is demanding both in terms of physical modeling and geometrical mesh handling. From the modeling standpoint, the physical phenomena of interest span a wide spectrum. At the low end of requirements, compressible turbulent flow of a Newtonian fluid is simulated (cold-flow simulation) and the model may then be successively extended to include heat transfer, combustion, chemical kinetics, modeling of fuel sprays, which includes spray injection, automation, turbulence dispersion, drop breakup and collision and the interphase exchange of mass, momentum and energy [5].

From the geometrical point of view, the two-stroke free linear engine analysis is similarly complex. A two-stroke internal combustion engine is a 3-D geometry of complex shape and contains a moving piston. In order to accommodate the motion of engine components, computational mesh undergoes both geometrical (mesh motion) and topological changes. And in order to perform a successful flow simulation in an internal combustion engine, the range of models listed above needs to take into account the effects of mesh motion and topological changes.

CFD provides the methods for numerical simulation of fluid flows. In spite of the fact that CFD analysis is regularly used in some areas of engineering, it is still not a widely accepted design tool. The complexity of flow regimes in, for example, internal combustion engines, in such that an accurate and predictive simulation becomes very expensive in term of time and computer resources. In order to simulate the features of the flow well, complicated models and accurate solution are needed [6].

This study examines the velocities of gases in the scavenging processes that include inlet, intake ports, exhaust port, and outlet, and also the combustion chamber of the two-stroke linear engine transiently. Three-dimensional models of combustion chambers incorporated with scavenging are developed. Moving mesh are considered to compute the flow transiently in the chambers and ports. The calculations are performed for 3-D using Star-CD software.

2 Model

The linear engine considered in this study is depicted as in Figure 1, and the 3-D models of its combustion chamber incorporate with the scavenging ports, is depicted as in Figure 2. The modeling strategy used in this study is referred from [7]. The gas is assumed gasoline, initially at the room temperature and pressure.

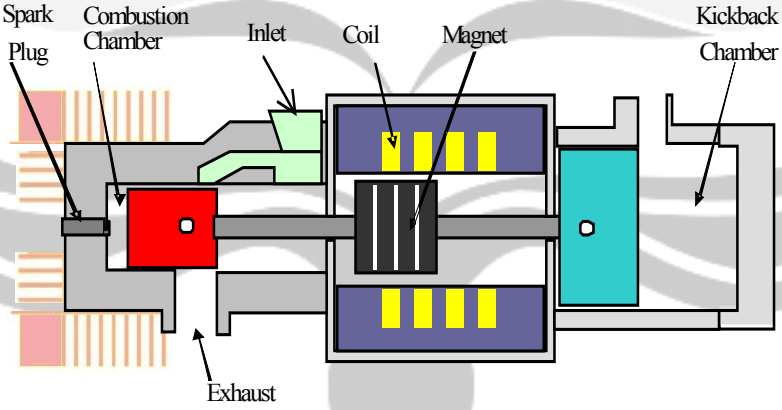


Figure 1 Linear generator engine model.

3 Flow field

The mass and momentum conservation equations (the Navier-Stokes equations) for general incompressible and compressible fluid flows are, in Cartesian tensor notation [8]:

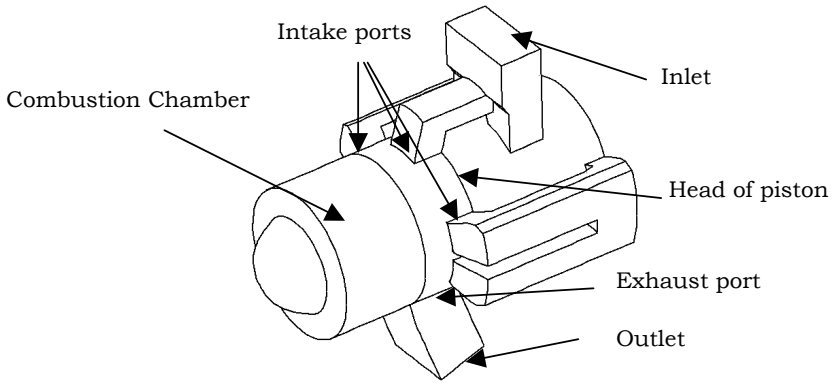


Figure 2 The 3-D model of combustion chamber, intake port and exhaust port.

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x_j}(\rho u_j) = s_m \tag{1}$$

$$\frac{\partial \rho u_i}{\partial t} + \frac{\partial}{\partial x_j}(\rho u_j u_i - \tau_{ij}) = -\frac{\partial p}{\partial x_i} + s_i \tag{2}$$

where t is time, x_i is Cartesian coordinate (i = 1, 2, 3), u_i is absolute fluid velocity component in direction x_i, p is piezometric pressure = $p_s - \rho_0 g_m x_m$ where p_s is static pressure, ρ_0 is reference density, the g_m are gravitational acceleration components and the x_m are coordinates relative to a datum where ρ_0 is defined, ρ is density, τ_{ij} are stress components, s_m is mass source, and s_i are momentum source components.

4 Finite volume method

The differential equations governing the conservation of mass, momentum and energy within the fluid are discretised by the finite volume method (FVM). Consider a general coordinate-free form of conservation equation [8]:

$$\frac{\partial}{\partial t}(\rho\phi) + \mathrm{div}(\rho \mathbf{u}\phi - \Gamma_\phi \mathrm{grad}\phi) = s_\phi \tag{3}$$

where **u** is the fluid velocity vector, ϕ stands for any of the dependent variables, and Γ_ϕ, s_ϕ are the associated diffusion and source coefficients, which can be deduced from the parent equations.

An exact form of equation (3) valid for an arbitrary time-varying volume V bounded by moving closed surface S can be written as

$$\frac{d}{dt}\int_V \rho\phi dV + \int_S (\rho \mathbf{u}_r \phi - \Gamma_\phi \mathrm{grad}\phi)\cdot d\mathbf{S} = \int_V s_\phi dV \tag{4}$$

where **S** is the surface vector and $\mathbf{u}_r$ is now the relative velocity between the fluid (**u**) and the surface $S(\mathbf{u}_c)$. If V and S are, respectively, taken to be volume V_p and discrete faces S_j (j = 1, N_f) of computational cell, equation (4) becomes

$$\underbrace{\frac{d}{dt}\int_{V_p}\rho\phi dV}_{T_1} + \underbrace{\sum_j \int_S (\rho \mathbf{u}_r \phi - \Gamma_\phi \mathrm{grad}\phi)\cdot d\mathbf{S}}_{T_2} = \underbrace{\int_V s_\phi dV}_{T_3} \tag{5}$$

The first term, T_1 of Equation (5) is discretised as

$$T_1 \approx \frac{(\rho\phi V)_P^n - (\rho\phi V)_P^o}{\delta t} \tag{6}$$

where the superscripts o and n refer to old and new time levels, respectively, separated by an interval δt. The second term of Equation (5) is split into the separate contributions C_j and D_j due to convection and diffusion, respectively, and each is expressed in terms of average values over cell faces, denoted by $(\,)_j$:

$$T_2 \approx \sum_j (\rho \mathbf{u}_r \phi \cdot \mathbf{S})_j - \sum_j (\Gamma_\phi \mathrm{grad}\phi \cdot d\mathbf{S})_j \tag{7}$$

$$\approx \sum_j (\rho u_r \cdot S)_j \, \phi_j - \sum_j \Gamma_{\phi,j} \;\; f_j^l(\phi_N - \phi_P) + \left\{\mathrm{grad}\phi \cdot \mathbf{S} - f_j^l \mathrm{grad}\phi \cdot \mathbf{d}_{PN}\right\}_j$$

where N and P are cell-centered nodes.

The third term of equation (5) can be written in the general quasi-linear form

$$T_3 \approx s_1 + s_2 \phi_P \tag{8}$$

5 Moving mesh

Mesh design in problems with a moving mesh and changing cell connectivity is dominated by need to keep the dynamic part of the grid simple so that they can be easily changed during the transient run. The formulation of mesh motion in such problems is divided into two conceptual steps. The first deals with connectivity changes, cell removal, addition, and reconnection. The second step is to specify the grid vertex positions as a function of time by supplying a set grid manipulation commands to be executed at each time step.

The mesh of the model is depicted as in Figure 2 and 3. The initial mesh contains all cells to be used in the analysis. Figure 2 shows that there are some cells of combustion chamber region and cells of scavenging chamber region. These cells are grouped into some layers and are deactivated in order to get the current mesh in the transient analysis. When cells are added, they are still deemed to be

connected to the neighbours they had at the time of their removal. In case any part of the solution domain become separated from the rest of the flow field during a transient run, the cell material type in that domain is to be changed.

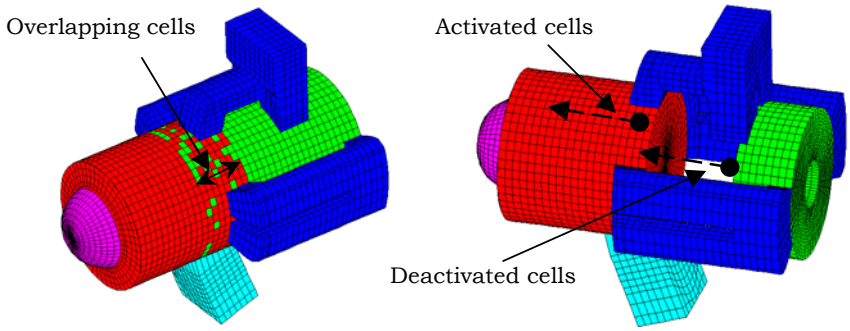


Figure 3 Moving mesh using activating and deactivating of cell layers

There are 20 cell layers in the cylinder region and 20 cell layers in the scavenging chamber region. The events are defined so that one layer is deactivated at every event, corresponding to the piston movement between top and bottom death center. Similarly, the events for reactivating the cell layers are developed, corresponding to the piston movement between bottom and top death center. Figure 3 shows the deactivated cells in the scavenging chamber.

6 Material properties and boundary conditions

The materials of the fluids are a mixture of air and gasoline. The mixture is filled in to the inlet ports. In this paper the effect of chemical reaction is not considered. The mixture is considered just as gas.

There are two types of boundary condition applied to this problem, attachment and pressure. The attachment boundaries belonging to the intake and exhaust ports are collected and temporarily assigned to a region, and the attachment boundaries belonging to the cylinder are assigned to another region. At the inlet and outlet regions with respect to intake and exhaust ports, respectively, the boundary conditions are pressures. The placement of the boundaries can be seen in the Figure 4. According to the problem, the inlet pressure of 500kPa and 100kPa are applied to the boundary condition at the inlet and outlet ports, respectively.

7 Results and discussion

Figures 5 (a) and (b) show the velocities distribution inside the engine at the piston position of 2.5mm from bottom dead center in the compression cycle. The gases from the inlet port flow to the scavenging chamber and to the intake port. Gases from the scavenging chamber flow also to the other intake ports and come into the combustion chamber with different directions.

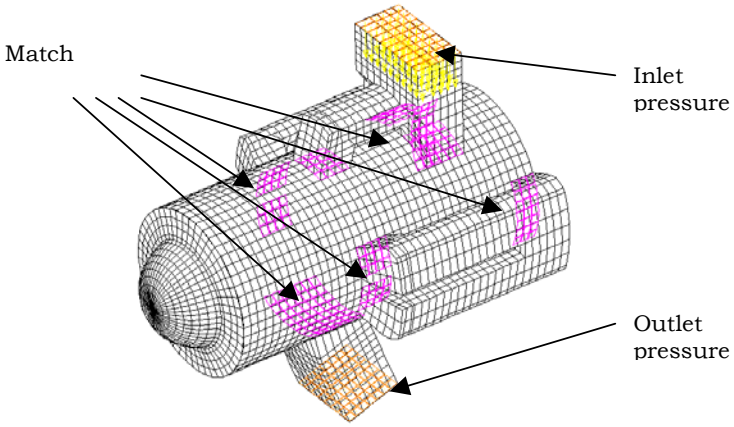


Figure 4 Boundary conditions of the problem

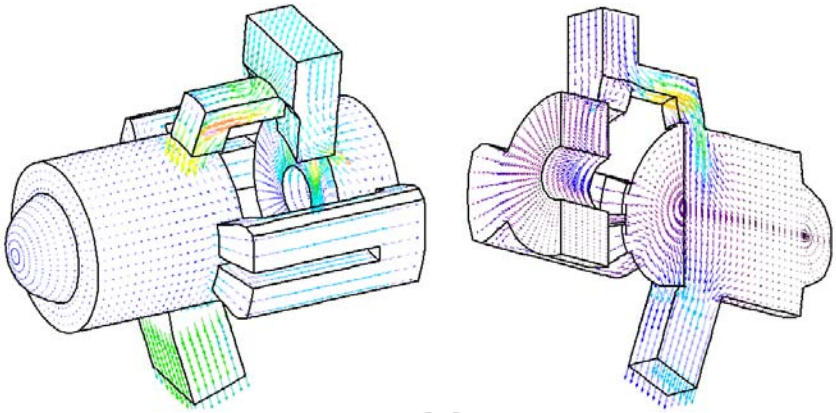
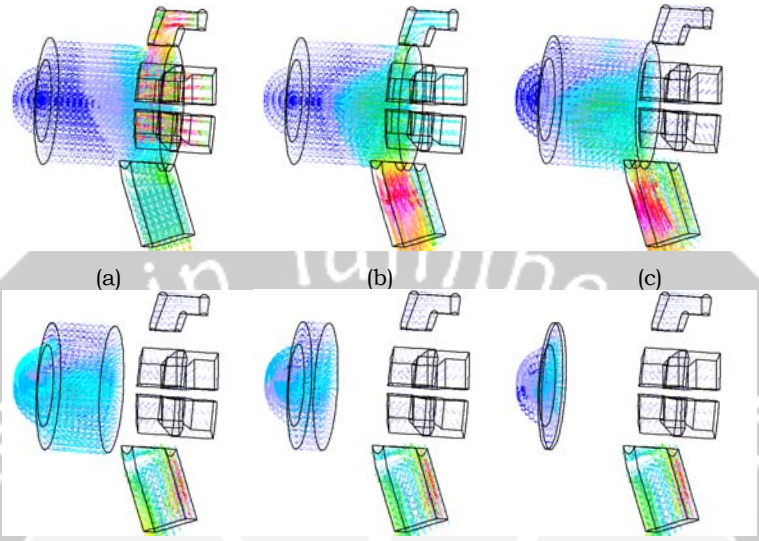


Figure 5 (a) Velocities distribution inside the engine, and
(b) Clip view the engine.

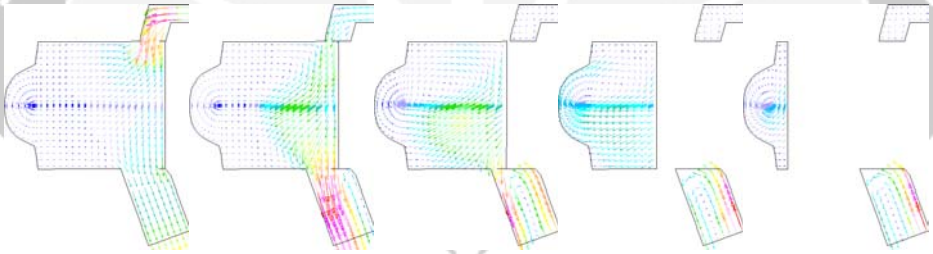
Figures 6(a) shows the velocities distribution at time 0.00033 seconds in the combustion chamber region, and the piston positions are after the start of compression. At this time, the faster flows occur in the intake ports. But after that, the Figures 6 (b) and (c) show that the higher velocities occur in the exhaust ports, until the outlet to the exhaust port is closed. The Figures 6 (b), (c) and (d) show that the gas flow in combustion chamber almost go toward the bottom side of the cylinders. These phenomena occur because of the flows from the intake ports. There are turbulences in the upper-middle regions of cylinders.



(a) (b) (c)

(d) (e) (f)

Figure 6 Flow distributions for flat piston crown



(a) (b) (c) (d) (e)

Figure 7 Flow distributions at the cross-sectional clip for flat piston crown

Figures 6 show how and where the turbulences occur. From those figures, it is known that the gases flow from piston side, go toward bottom side of the cylinders, and then make a rotation in the cup regions near cylinder heads, and back to the cylinders. It can be possible in the cylinder using flat piston crown that fresh gases come from the inlets move through exhaust port before burning.

Conclusions

It can be concluded that the gas flow problems in the scavenging of two stroke linear engine can be analyzed using Finite Volume Method. The fresh mixture gases from the intake ports flow in to the combustion chamber. The burnt gases flow from the combustion chamber to the exhaust port.

Acknowledgment

The authors would like to thank the Malaysian Ministry of Science, Technology and Innovation for sponsoring this work under the project IRPA 03-02-02-0056-PR0025/04-03, and the first author thank to the Center for Graduate Studies, Universiti Kebangsaan Malaysia, for sponsoring the author study under Fellowship Scheme.

References

[1] Nandkumar, S. (1998), *Two stroke linear engine*, M.Sc. Thesis Collection, West Virginia University.

[2] Prados, M. A. (2002), *Towards a Linear Engine*. M.Sc. Thesis, Stanford University.

[3] Pollard, B. (2005), *Tech Talk: Two-Stroke Engine Theory*, UK-Karting, http://www.karting.co.uk. (April 2005)

[4] Kim, G. H. & Kirkpatrick, A. (2004), Retrofitting a natural gas engine, *CD adapco Dynamic*, **23**, 21-22.

[5] Jasak H., Weller H. G. & Nordin N. (2004), In-Cylinder CFD Simulation using a C++ Object-Oriented Tollkit, SAE 2004-10-0110.

[6] Jasak, H. (1996), *Error analysis and estimation for the Finite Volume Method with Applications to fluid flows*, Ph.D. Thesis, Imperial College of Science, Technology and Medicine.

[7] CD adapco Group (2004), *Tutorials*, Star-CD Version 3.2, CD adapco Group.

[8] CD adapco Group (2004), *Methodology*, Star-CD Version 3.2, CD adapco Group.

Authors:
Drs. Tulus, M.Si.
Department of Mathematics
Universitas Sumatera Utara
Medan 20155, Indonesia
Currently study for PhD program at
Department of Mechanical and Materials Engineering
Universiti Kebangsaan Malaysia
43600 Bangi, Selangor DE
Email: tulus_jp@yahoo.com

Assoc. Prof. Dr. Ahmad Kamal Ariffin,
Department of Mechanical and Materials Engineering
Universiti Kebangsaan Malaysia
43600 Bangi, Selangor DE
Email: kamal@eng.ukm.my

HYDRODYNAMICS ON BOJONGSOANG FACULTATIVE POND USING MATHEMATICAL MODEL

Rositayanti Hadisoebroto[a], Suprihanto Notodarmojo[b]

[a] Universitas Trisakti, Jakarta, Indonesia
[b] ITB, Bandung, Indonesia

Abstract. Wastewater treatment performance in facultative pond could be analyzed by evaluating its hydrodynamics. Hydrodynamics study could be done by mathematical model which is built from 2 (two) government equations, momentum and continuity equations, integrated by using finite difference method of semi implicit Crank-Nicolson. The results of hydrodynamics simulation, which is built from mathematical model, show that flow rate influences pond hydrodynamics. The higher influent flow rate cause flow distribution at the whole part of the pond and the higher water velocity, so dead zone will be reduced. If hydrodynamics performance is better, the treatment performance of facultative pond will be better.
Key-words: hydrodynamics, mathematical model, facultative pond, flow rate

1 Introduction

Facultative pond as a part of stabilization pond system has been used to treat domestic wastewater from the small communities, as long as wide area is available. The system is chosen because it is simple to operate and maintain, and the cost is low too. Unfortunately, design criteria for facultative pond has not available yet. Its design and built are based on experience and expertise of the designer. Environmental and weather factors that affect the system are very complex, so biological process and physical phenomena inside it have been known yet. Therefore, there are many cases of inefficient operation, like dead zone or short-circuiting (Wood et.al.; 1995). Good mixing ensures a more uniform distribution of BOD, dissolved oxygen, bacteria and algae and hence a better degree of waste stabilization. (hamzeh ramadan)

One of the facultative ponds that operate is Bojongsoang WWTP, which is located in Bojongsoang and Bojongsari villages. To date, the performance of WWTP is not quite satisfactory. There are a number of possible things that may cause the less effectiveness of the WWTP. One of the possibilities is that the hydraulic characteristics of the pond is not properly met the criterion such as the existence of dead zone or eddy current. Therefore, a study on hydraulic characteristics of the pond is necessary.

Evaluation of hydraulic characteristic on the water body could be done directly by using tracer study (Dorego & Leduc, 1996), (Torres et.al., 1997), (Torres et.al., 1999), (Torres et.al., 2000) or by computational simulation using mathematical model |(Wood et.al., 1995), (Agunwamba, 1992), (Wood et.al., 1998). Since the pond is very wide, tracer study on it need many time and cost. So, the mathematical model is an alternative tool to evaluate the pond hydrodynamics. At

this research, mathematical model is built from governing equations consist of two hydrodynamics equations, continuity and momentum equations, which are solved by using finite difference numerical method of semi implicit (Crank-Nicolson).

Both of these equations are derived by mass conservation and momentum at control volume three dimension integrated to depth to obtain two dimension equation at x and y direction (horizontal to depth) (Pradiko, 2002).

1.1 Continuity Equation

Continuity law for unsteady water can be derived by conservation laws of mass in one space between two sections with a very small distance as a control volume. Continuity equation can be obtained by approximating mass enters and exits from a control volume. The existing phenomenon is presented in the following figure 1 (Pradiko, 2002).

Figure 1 describes space control volume along Δx to continuity equation of water, where (Pradiko, 2002):
U = average velocity flow in axis of the abscis direction in the middle of internodes (m/sec)
H = water depth (m)
ζ = water elevation (m)
ρ = mass density flow in the middle of internodes (kg/m^3)
q = input flow enter per set of width along Δx density mass assumed equal to water mass density (m^2/sec).

Hence input conservation laws of mass in control volume (Pradiko, 2002):

"Rate of water mass which flows into control volume – rate of water mass flows out of control volume = rate of increasing of volume in control volume space"

$$\left(\rho UH - \frac{\partial(\rho UH)}{\partial x}\frac{\Delta x}{2}\right) - \left(\rho UH + \frac{\partial(\rho UH)}{\partial x}\frac{\Delta x}{2}\right) + \rho q_x \partial x = \frac{\partial(\rho\zeta)}{\partial t}\Delta x \qquad [1]$$

$$\frac{\partial(\rho\zeta)}{\partial t} + \frac{\partial(\rho UH)}{\partial x} = \rho q_x \qquad [2]$$

Because water is assumed to be incompressible, ρ was constant, so that equation [2] can be written (Pradiko, 2002):

$$\frac{\partial\zeta}{\partial t} + \frac{\partial(UH)}{\partial x} = q_x \qquad [3]$$

If derived of equation is conducted for the *y* direction, hence the equation becomes (*Pradiko, 2002*):

$$\frac{\partial\zeta}{\partial t} + \frac{\partial(VH)}{\partial y} = q_y \qquad [4]$$

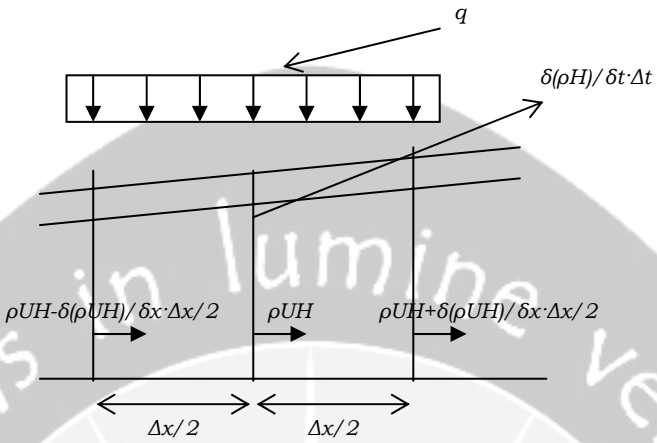


Figure 1. Control Volume of Continuity Calculation

If equation [3] and [4] combined, then they can be applied for both direction, *x* and *y* direction (*Pradiko, 2002*):

$$\frac{\partial \zeta}{\partial t} + \frac{\partial (UH)}{\partial x} + \frac{\partial (VH)}{\partial y} = q \qquad [5]$$

Where *q* represents input stream enter per set of width along Δ*x* and Δ*y*, or *q*=*Q*/*A*, so that equation [5] turns into (*Pradiko, 2002*):

$$\frac{\partial \zeta}{\partial t} + \frac{\partial (UH)}{\partial x} + \frac{\partial (VH)}{\partial y} = \frac{Q}{A} \qquad [6]$$

where
Q = water flow rate (m^3/sec)
A =section area (m^2)

1.2 Momentum Equation

Equation of water motion (momentum) is based on Newton second's law, which takes into account both acceleration (*a*) and force per mass unit (*F*/*m*) at one particular object. Newton second's law states the amount of working external force equals to the rate of changing of linear momentum, that is (*Pradiko, 2002*):

$$\sum F_X = m \cdot a_X = \frac{d(mU)}{dt} = m\frac{dU}{dt} + U\frac{dm}{dt} \qquad [7]$$

Where :
F_x = resultant force in x direction
a_x = acceleration (*dU*/*dt*)
m = object mass or fluids
U = velocity of *x* direction

Because force works in a steady volume and not having change of specific mass (density), hence the rate of mass to time is assumed zero or *dm/dt=0*. Equation [7] will become (*Pradiko, 2002*):

$$\sum F_X = m\frac{dU}{dt} \qquad [8]$$

Because velocity of *U* represents space and time function, *U=U(x,y,t)*, hence derivation equation of velocity of *U* completely is (*Pradiko, 2002*):

$$\frac{dU}{dt} = \frac{\partial U}{\partial t} + \frac{\partial U}{\partial x}\frac{dx}{dt} + \frac{\partial U}{\partial y}\frac{dy}{dt} \qquad [9]$$

If *dx/dt* is *U* and *dy/dt* is *V*, hence equation [9] becomes (*Pradiko, 2002*):

$$\frac{dU}{dt} = \frac{\partial U}{\partial t} + \overline{U}\frac{\partial U}{\partial x} + \overline{V}\frac{\partial U}{\partial y} \qquad [10]$$

Therefore, equation (8) will become (*Pradiko, 2002*):

$$\sum F_X = m\left[\frac{\partial U}{\partial t} + \overline{U}\frac{\partial U}{\partial x} + \overline{V}\frac{\partial U}{\partial y}\right] \qquad [11]$$

Newton second's law is used to determine the rate of external forces which have an effect to mass, which in this case is water mass in control volume. The forces per mass unit are :

a. Hydrostatic pressure force

Hydrostatic pressure force is the force that exists because of gravity force. In this case is water mass in control space, which pressing water below so that water volume weight will resist the water motion. Hydrostatic pressure force can be defined by equation :

$$F_h = -gH\frac{\partial \zeta}{\partial x} \qquad [12]$$

g represents the constant of gravity rate.

Negative sign indicates that the force is resisting water motion in control space (*Pradiko, 2002*).

b. Friction force to bottom

Friction force to bottom represents force that happens because of an object mass. In this case is water mass in control space, which presses bottom part of control space so this bottom part performs a reaction in the form of friction force. Friction force to this bottom can be defined in the following equation :

$$F_g = -\frac{rU}{H^2}\left(U^2 + V^2\right)^{1/2} \qquad [13]$$

r represents friction parameter to bottom.

Negative sign indicates that force has the character to resist water motion in control space (*Pradiko, 2002*).

c. Turbulent diffusion force

Turbulent diffusion force is happened because of the existence of molecular flow or turbulent. Molecule flow or molecular diffusion causes transfer of particles from a place with higher concentration to the place with lower concentration. Turbulent diffusion force can be defined with following equation:

$$F_d = A_H\left(\frac{\partial^2 U}{\partial x^2}+\frac{\partial^2 U}{\partial y^2}\right) \quad [14]$$

A_H represents coefficient of eddy turbulence (*Pradiko, 2002*).

d. Surface wind pressure force

Wind is one of the main source of energy to water dynamics. Energy transfer of wind to water surface will cause stream and wave. In a deep water, influence of wind only have an effect until a certain point of depth which is called the depth of Ekman. In general, velocity level of stream due to wind is around 2-3% from its wind velocity. Pressurized surface wind force can be defined with following equation :

$$F_a = \lambda W_x\left(W_x^2+W_y^2\right)^{1/2} \quad [15]$$

λ represents coefficient of pressurized surface wind (*Yustiani, 2000*).

By including forces per mass unit above hence equation [11] turns into (*Yustiani, 2000*):

$$F_d + F_a - F_h - F_g = \frac{\partial U}{\partial t}+\overline{U}\frac{\partial U}{\partial x}+\overline{V}\frac{\partial U}{\partial y} \quad [16]$$

Thereby momentum equation in x direction is (*Yustiani, 2000*):

$$\frac{\partial U}{\partial t}+U\frac{\partial U}{\partial x}+V\frac{\partial U}{\partial y}+gH\frac{\partial \zeta}{\partial x}+\frac{rU}{H^2}\left(U^2+V^2\right)^{1/2}=A_H\left[\frac{\partial^2 U}{\partial x^2}+\frac{\partial^2 U}{\partial y^2}\right]+\lambda W_x\left(W_x^2+W_y^2\right)^{1/2} \quad [17]$$

By applying the same forces in ordinate direction, momentum equation in y direction will be obtained by the following (*Yustiani, 2000*):

$$\frac{\partial V}{\partial t}+U\frac{\partial V}{\partial x}+V\frac{\partial V}{\partial y}+gH\frac{\partial \zeta}{\partial y}+\frac{rV}{H^2}\left(U^2+V^2\right)^{1/2}=A_H\left[\frac{\partial^2 V}{\partial x^2}+\frac{\partial^2 V}{\partial y^2}\right]+\lambda W_y\left(W_x^2+W_y^2\right)^{1/2} \quad [18]$$

where :

U = velocity in x direction integrated to depth (m/sec)
V = velocity in y direction integrated to depth (m/sec)
ζ = relative elevation to certain reference (m)
H = distance from ground until certain reference (m)
Q = debit (m^3/sec)
A = section area (m^2)
W_x = wind pressure component in x direction (m/sec)
W_y = wind pressure component in y direction (m/sec)
g = gravity constant (9,80 m/sec^2)

A_H = diffusion coefficient of turbulent horizontal (m^2/sec) = 0.001
r = parameter of bottom friction = 0,035
λ = surface wind friction coefficient = 0,00005

2 Research Metodhology

2.1 Governing Equations

The governing equations are the hydrodynamics equations consist of two equations; those are momentum equation (equation [17] and [18]) and continuity equation (equation [6]). Numerical method to solve the governing equations is semi-implicit Crank-Nicolson finite difference method. From continuity equation, the water level on next iteration time will be obtained, while from momentum equation, the water velocity in x and y direction on next iteration time will be obtained.

2.2 Model Discretization

The area of reviewed facultative pond is 74.000 m^2 with average depth of 2 m. In building the model, the pond is divided into grids with the area of 2,5x2,5 m^2, so that in its entirety there are 146x161 grids.

2.3 Model Application

The model can be applied to simulate the water velocity of facultative pond. Simulation is run in variation water flow rate (0,08, 0,25, and 0,7 m^3/sec), in condition of West wind velocity is constant, 10 m/sec. From simulation, facultative pond hydraulics characteristics can be obtained.

3 Results And Discussions

3.1 Solving of Governing Equation

Through solution of continuity equation (equation [6]) and momentum equations (equation [17] and [18]) by using finite difference method of Crank-Nicolson, we can obtain :

Continuity equation :

$$\underset{(1)}{\frac{\partial \zeta}{\partial t}} + \underset{(2)}{H\frac{\partial U}{\partial x}} + \underset{(3)}{H\frac{\partial V}{\partial y}} = \underset{(4)}{\frac{Q}{A}}$$

Numerical solution :

Term (1) : $\frac{\partial \zeta}{\partial t} = \frac{\zeta_{i,j}^{n+1} - \zeta_{i,j}^{n}}{\Delta t}$

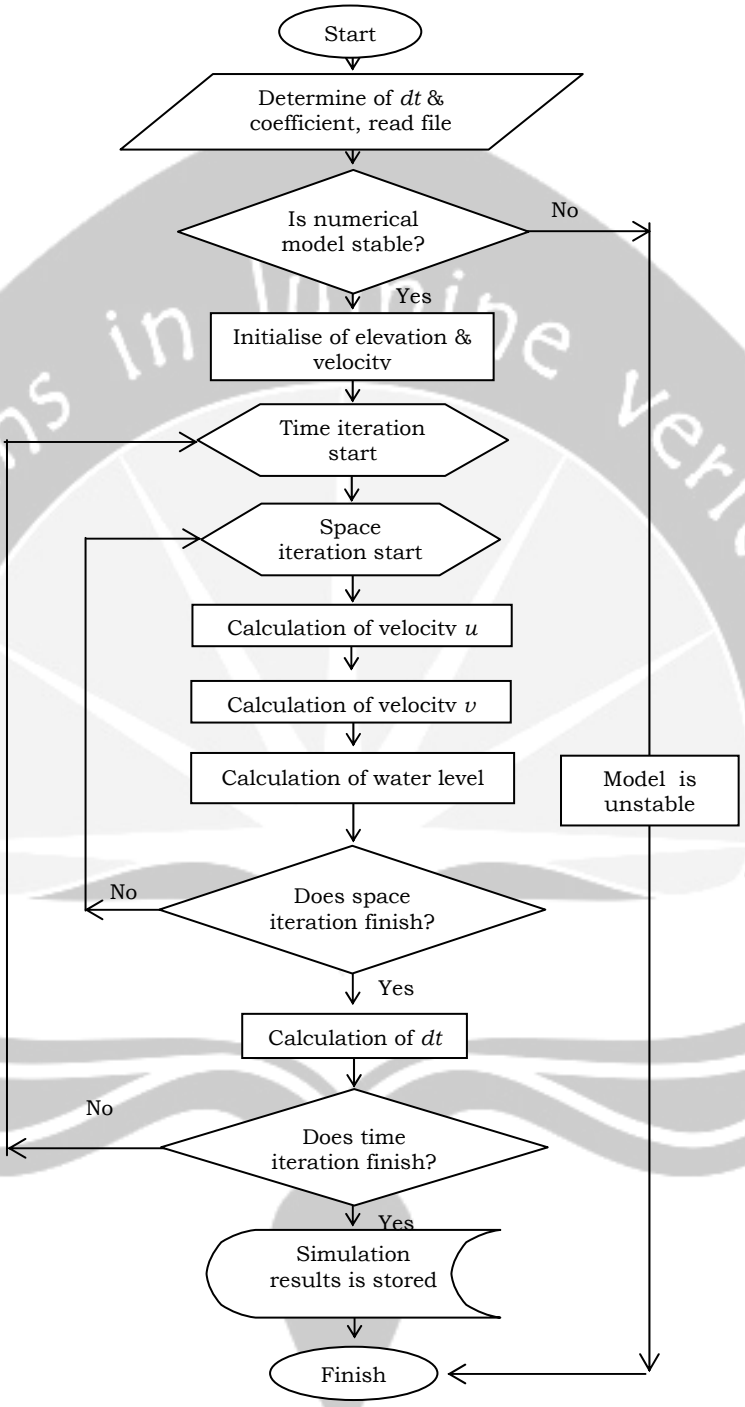


Figure 2. Programming Flowchart

Term (2) : $H\dfrac{\partial U}{\partial x} = H^{*}\dfrac{U^{n}_{i+1,j} - U^{n}_{i-1,j}}{2\Delta x} = S1$

Term (3) : $H\dfrac{\partial V}{\partial y} = H^{*}\dfrac{V^{n}_{i,j+1} - V^{n}_{i,j-1}}{2\Delta y} = S2$

where : $H^{*} = \dfrac{h_{i,j} + \zeta^{n}_{i,j} + h_{i+1,j} + \zeta^{n}_{i+1,j}}{2}$

Term (4) : $\dfrac{Q}{A} = \dfrac{Q^{n}_{i,j}}{\Delta x \times \Delta y} = S3$

If terms (1), (2), (3), and (4) are combined, the equation will become:

$$\zeta^{n+1}_{i,j} = \zeta^{n}_{i,j} - \Delta t(S1 + S2 - S3) \qquad [19]$$

The above equation works when there is flow enters (inlet). On the contrary, when there are flow exits (outlet), the equation will become:

$$\zeta^{n+1}_{i,j} = \zeta^{n}_{i,j} - \Delta t(S1 + S2 + S3) \qquad [20]$$

When there is no stream entering and exiting (Q=0), the equation becomes:

$$\zeta^{n+1}_{i,j} = \zeta^{n}_{i,j} - \Delta t(S1 + S2) \qquad [21]$$

Momentum equation of x direction

$$\frac{\partial U}{\partial t} + U\frac{\partial U}{\partial x} + V\frac{\partial U}{\partial y} + gH\frac{\partial \zeta}{\partial x} + \frac{rU}{H^{2}}\left(U^{2} + V^{2}\right)^{1/2} = \lambda W_{X}\left(W_{x}^{2} + W_{y}^{2}\right)^{1/2} + A_{H}\left[\frac{\partial^{2}U}{\partial x^{2}} + \frac{\partial^{2}U}{\partial y^{2}}\right]$$

(1) (2) (3) (4) (5) (6) (7)

Numerical solution :

Term (1) : $\dfrac{\partial U}{\partial t} = \dfrac{U^{n+1}_{i,j} - U^{n}_{i,j}}{\Delta t}$

Term (2) : $U\dfrac{\partial U}{\partial x} = U^{n}_{i,j}\left(\dfrac{U^{n}_{i+1,j} - U^{n}_{i-1,j}}{2\Delta x}\right) = A_{X}$

Term (3) : $V\dfrac{\partial U}{\partial y} = V^{*}\left(\dfrac{U^{n}_{i,j+1} - U^{n}_{i,j-1}}{2\Delta y}\right) = B_{X}$

where : $V^{*} = \dfrac{V^{n}_{i,j} + V^{n}_{i,j-1} + V^{n}_{i+1,j} + V^{n}_{i+1,j-1}}{4}$

$Sv_{X} = A_{X} + B_{X}$

Term (4) : $gH\dfrac{\partial \zeta}{\partial x} = gH^{*}_{x}\left(\dfrac{\zeta^{n}_{i+1,j} - \zeta^{n}_{i,j}}{\Delta x}\right) = Sp_{X}$

where : $H^{*}_{x} = \dfrac{h_{i,j} + \zeta^{n}_{i,j} + h_{i+1,j} + \zeta^{n}_{i+1,j}}{2}$

Term (5) : $\frac{rU}{H^2}\left(U^2+V^2\right)^{1/2} = \frac{rU_{i,j}^{n+1}}{\left(H_x^*\right)^2}\left[\left(U_{i,j}^{n}\right)^2+\left(V^*\right)^2\right]^{1/2} = Sk_X U_{i,j}^{n+1}$

where : $Sk_X = \frac{r}{\left(H_x^*\right)^2}\left[\left(U_{i,j}^{n}\right)^2+\left(V^*\right)^2\right]^{1/2}$

Term (6) : $\lambda W_X\left(W_x^2+W_y^2\right)^{1/2} = Sw_X$

Term (7) : $A_H\left(\frac{\partial^2 U}{\partial x^2}+\frac{\partial^2 U}{\partial y^2}\right) = Sf_X$

$$Sf_X = A_H\left[\frac{1}{2}\left(\frac{U_{i+1,j}^n-2U_{i,j}^n+U_{i-1,j}^n}{\Delta x^2}+\frac{U_{i+1,j+1}^n-2U_{i,j+1}^n+U_{i-1,j+1}^n}{\Delta x^2}\right)+\frac{1}{2}\left(\frac{U_{i,j+1}^n-2U_{i,j}^n+U_{i,j1}^n}{\Delta y^2}+\frac{U_{i+1,j+1}^n-2U_{i+1,j}^n+U_{i+1,j-1}^n}{\Delta y^2}\right)\right]$$

If terms (1), (2), (3), (4), (5), (6), and (7) are combined, the equation above will becomes:

$$\frac{U_{i,j}^{n+1}-U_{i,j}^{n}}{\Delta t}+Sv_X+Sp_X+Sk_X U_{i,j}^{n+1} = Sw_X+Sf_X \qquad [22]$$

$$U_{i,j}^{n+1}-U_{i,j}^{n}+Sv_X\Delta t+Sp_X\Delta t+Sk_X U_{i,j}^{n+1}\Delta t = Sw_X\Delta t+Sf_X\Delta t \qquad [23]$$

$$U_{i,j}^{n+1}\left(1+Sk_X\Delta t\right) = U_{i,j}^{n}-\Delta t\left(Sv_X+Sp_X-Sw_X-Sf_X\right) \qquad [24]$$

Therefore the numerical solution for momentum equation of *x* direction is:

$$U_{i,j}^{n+1} = \left[U_{i,j}^{n}-\Delta t\left(Sv_X+Sp_X-Sw_X-Sf_X\right)\right]R_X \qquad [25]$$

where : $R_X = \frac{1}{1+Sk_X\Delta t}$

Momentum equation of *y* direction

$$\underbrace{\frac{\partial V}{\partial t}}_{(1)}+\underbrace{U\frac{\partial V}{\partial x}}_{(2)}+\underbrace{V\frac{\partial V}{\partial y}}_{(3)}+\underbrace{gH\frac{\partial \zeta}{\partial y}}_{(4)}+\underbrace{\frac{rV}{H^2}\left(U^2+V^2\right)^{1/2}}_{(5)} = \underbrace{\lambda W_{Xy}\left(W_x^2+W_y^2\right)^{1/2}}_{(6)}+\underbrace{A_H\left[\frac{\partial^2 V}{\partial x^2}+\frac{\partial^2 V}{\partial y^2}\right]}_{(7)}$$

Numerical solution :

Term (1) : $\frac{\partial V}{\partial t} = \frac{V_{i,j}^{n+1}-V_{i,j}^{n}}{\Delta t}$

Term (2) : $U\frac{\partial V}{\partial x} = U^*\left(\frac{V_{i+1,j}^{n}-V_{i-1,j}^{n}}{2\Delta x}\right) = A_Y$

where : $U^* = \frac{U_{i,j}^{n}+U_{i-1,j}^{n}+U_{i,j+1}^{n}+U_{i-1,j+1}^{n}}{4}$

Term (3) : $V\frac{\partial V}{\partial y} = V_{i,j}^{n}\left(\frac{V_{i,j+1}^{n} - V_{i,j-1}^{n}}{2\Delta y}\right) = B_Y$

$Sv_Y = A_Y + B_Y$

Term (4) : $gH\frac{\partial \zeta}{\partial y} = gH_Y^{*}\left(\frac{\zeta_{i,j+1}^{n} - \zeta_{i,j}^{n}}{\Delta y}\right) = Sp_Y$

where : $H_Y^{*} = \frac{h_{i,j} + \zeta_{i,j}^{n} + h_{i,j+1} + \zeta_{i,j+1}^{n}}{2}$

Term (5) : $\frac{rU}{H^2}\left(U^2 + V^2\right)^{1/2} = \frac{rV_{i,j}^{n+1}}{\left(H_y^{*}\right)^2}\left[\left(U^{*}\right)^2 + \left(V_{i,j}^{n}\right)^2\right]^{1/2} = Sk_y V_{i,j}^{n+1}$

where : $Sk_y = \frac{r}{\left(H_y^{*}\right)^2}\left[\left(U^{*}\right)^2 + \left(V_{i,j}^{n}\right)^2\right]^{1/2}$

Term (6) : $\lambda W_y\left(W_x^2 + W_y^2\right)^{1/2} = Sw_y$

Term (7) : $A_H\left(\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2}\right) = Sf_Y$

$$Sf_Y = A_H\left[\frac{1}{2}\left(\frac{V_{i+1,j}^{n} - 2V_{i,j}^{n} + V_{i-1,j}^{n}}{\Delta x^2} + \frac{V_{i+1,j+1}^{n} - 2V_{i,j+1}^{n} + V_{i-1,j+1}^{n}}{\Delta x^2}\right) + \frac{1}{2}\left(\frac{V_{i,j+1}^{n} - 2V_{i,j}^{n} + V_{i,j1}^{n}}{\Delta y^2} + \frac{V_{i+1,j+1}^{n} - 2V_{i+1,j}^{n} + V_{i+1,j-1}^{n}}{\Delta y^2}\right)\right]$$

If terms (1), (2), (3), (4), (5), (6), and (7) are combined, the equation will become:

$$\frac{V_{i,j}^{n+1} - V_{i,j}^{n}}{\Delta t} + Sv_y + Sp_y + Sk_y V_{i,j}^{n+1} = Sw_y + Sf_y \qquad [26]$$

$$V_{i,j}^{n+1} - V_{i,j}^{n} + Sv_Y\Delta t + Sp_Y\Delta t + Sk_Y V_{i,j}^{n+1}\Delta t = Sw_Y\Delta t + Sf_Y\Delta t \qquad [27]$$

$$V_{i,j}^{n+1}\left(1 + Sk_Y\Delta t\right) = V_{i,j}^{n} - \Delta t\left(Sv_Y + Sp_Y - Sw_Y - Sf_Y\right) \qquad [28]$$

Therefore the numerical solution for momentum equation of *y* direction is:

$$V_{i,j}^{n+1} = \left[V_{i,j}^{n} - \Delta t\left(Sv_Y + Sp_Y - Sw_Y - Sf_Y\right)\right]R_Y \qquad [29]$$

where : $R_Y = \frac{1}{1 + Sk_Y\Delta t}$

3.2 Simulation Results

Figure 3, 4, and 5 show simulation results at variation influent flow rate. Figure 3 shows simulation at minimum flow rate (0,08 m^3/sec), while Figure 4 and 5 are at average (0,25 m^3/sec) and maximum (0,7 m^3/sec) flow rate, respectively. Arrow signs inside the pond indicate the water flow; the velocity height is indicated by arrow length, while the direction of arrow indicates the flow direction.

It can be seen in Figure 3 that the water velocity at the pond center is very small, since the arrow signs are small or even invisible. It means the dead zone exists at the pond center.

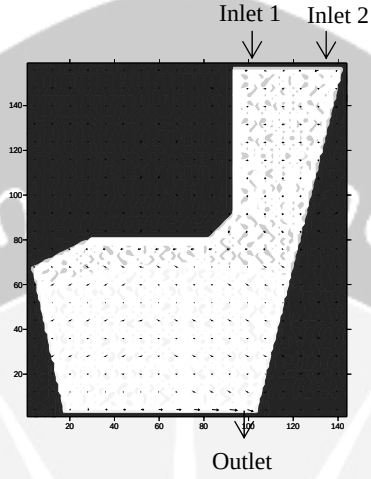


Figure 3. Flow pattern at flow rate of 0,08 m^3/sec

Figure 4 shows that the dead zone still exists at the pond center, but at the smaller area, since many arrows at the pond center become visible. It can be seen either that the arrow length increases, so the water velocity on the whole part of the pond are higher than at minimum flow rate. Therefore, the higher influent flow rate, the higher water velocity on the pond.

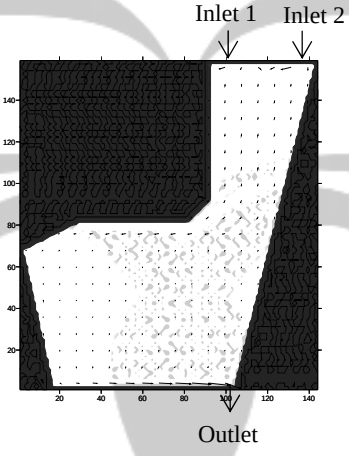


Figure 4. Flow pattern at flow rate of 0,25 m^3/sec

It can be seen in Figure 5 that water flows distributed evenly throughout the whole part of the pond, since there is no invisible arrow. Therefore, the dead zone becomes reduce, as the influent flow rate is higher. In short, the higher influent flow rate, the better hydraulic performance of the pond.

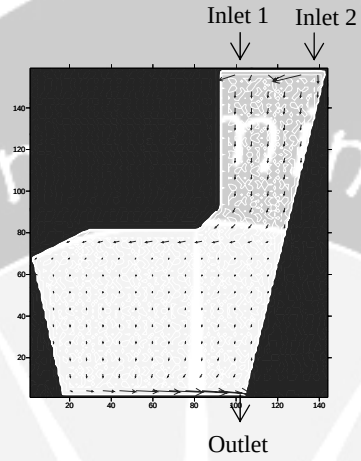


Figure 5. Flow pattern at flow rate of 0,7 m^3/sec

4 Conclusions

Two-dimensional hydrodynamics model is developed from governing equations that solved by numerical method. The governing equations conceived of 2 (two) hydrodynamics equations; continuity and momentum equations, can be solved by semi implicit Crank-Nicolson finite difference method. The hydrodynamics model can be applied to simulate water flow on the facultative pond.

Simulation results show that the dead zone exists at the pond center. When the influent flow rate is higher, water velocity become higher. At the maximum flow rate (0,7 m^3/sec), water flow is distributed evenly throughout the whole part of the pond. In other words, the dead zone is reduced since the influent flow rate is the highest. Therefore, the higher influent debit can improve hydraulics performance of the pond. Since then, the treatment performance is expected to be better.

References

[1] Agunwamba J.C. (1992), Field Pond Performance and Design Evaluation Using Physical Models. *Water Research,* **26(10)**, 1403-1407.
[2] Dorego, N.C., Leduc R. (1996), Characterization of Hydraulic Flow Patterns in Facultative Aerated Lagoons. *Water Science Technology*, **34(11)**, 99-106.
[3] Pradiko, Hary. (2002), *Model Dua Dimensi Gerak Arus di Perairan Pantai Muara Sungai Cisadane.* Tesis Magister Teknik Lingkungan ITB.
[4] Ramadan, Hamzah, http://ponce.sdsu.edu/ramadan/ stabilizationponds.htm

[5] Torres J.J., Soler A., Saez J., Leal L.M., Aguilar M.I. (1999), Study of Internal Hydrodynamics in Three Facultative Ponds of Two Municipal WSPS in Spain. *Water Research*, **33(5)**, 1133-1140.
[6] Torres J.J., Soler A., Saez J., Llorens M. (2000), Hydraulic Performance of a Deep Stabilisation Pond Fed at 3.5 m Depth. *Water Research*, **34(3)**, 1042-1049.
[7] Torres J.J., Soler A., Saez J., Ortuno J.F. (1997), Hydraulic Performance of a Deep Wastewater Stabilization Pond. *Water Research*, **31(4)**, 679-688.
[8] Wood M.G., Greenfield P.F., Howes T, Johns M.R., Keller J. (1995), Computational Fluid Dynamic Modelling of Wastewater Ponds to Improve Design. *Water Science Technology*, **31(12)**, 111-118.
[9] Wood M.G., Howes T., Keller J., Johns M.R. (1998), Two Dimensional Computational Fluid Dynamic Models for Waste Stabilisation Ponds. *Water Research*, **32(3)**, 958-963.
[10] Yustiani, Yonik M. (2000), *Model Dua Dimensi Penyebaran Ammonium, Nitrit, dan Nitrat di Perairan Pantai Semarang dengan Persamaan Kinetik Orde Satu Thomann.* Tesis Magister Teknik Lingkungan ITB.

ROSITAYANTI HADISOEBROTO: Department of Environmental Engineering, Universitas Trisakti, Jl. Kyai Tapa No. 1, Jakarta Barat 11440, Indonesia. Phone/Fax: +62 +21 560 2575
E-mail: rositayanti@yahoo.com

SUPRIHANTO NOTODARMOJO: Department of Environmental Engineering, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia.

An Estimation of Internal Solitary Waves in the Lombok Strait using Two-Layer Model

S. R. Pudjaprasetya

Department of Mathematics, Institut Teknologi Bandung, Indonesia

Abstract: We consider the Hamiltonian KdV-model for internal waves in the two-layer fluid, that was derived by Grimshaw & S.R. Pudjaprasetya in 1998. Different from the KdV internal waves that are common in literatures, this Hamiltonian KdV holds for two-layer fluid system, where the upper and bottom parts are bounded by rigid boundaries. This KdV equation has a Hamiltonian structure, and it is written explicitly in interface deviation variable, with coefficients that are depend explicitly on the depth and density of each layer. Here, we will apply this Hamiltonian KdV to oceanic internal waves. This is possible because, in studying internal wave, we often neglect the deviation on the sea surface. Mathematically speaking, this is the same with having rigid boundary for the upper part. In fact, this Hamiltonian KdV holds when the upper and bottom boundaries does not change in time.

When the upper and bottom boundaries are flat, the Hamiltonian KdV has a solitary wave solution, a wave that is travelling undisturbed in shape with constant velocity. The amplitude, wavelength and velocity of the internal solitary wave depend explicitly on the depth and density of each layer. Therefore, estimation of internal solitary wave in Lombok Strait using this Hamiltonian KdV is possible.

Fluid density stratification data in Lombok Strait is used here, and will be approximated by two-layer fluid by making use of a solution of the corresponding eigenvalue problem, that are known in the literature. Further, we consider the SAR image of Lombok Strait that contains bright and dark bands, as a signature of the presence of solitary internal waves. From the intensity plot of that SAR image, the wavelength of the solitary wave can be estimated. Hence, the amplitude and velocity dan be estimated as well.

Keywords: Hamiltonian KdV, two-layer approximation, internal solitary waves.

FREE SURFACE FLUID FLOWS INDUCED BY A SUBMERGED SINK IN A THREE-LAYER FLUID UNDER THE EFFECT OF SURFACE TENSION

Basuki Widodo

Mathematics Department, Faculty of Mathematics and Natural Sciences,
INSTITUT TEKNOLOGI SEPULUH NOPEMBER - SURABAYA

Abstract. Over the last 20 years the selective withdrawal of fluid from a reservoir has received a great deal of attention, see for example [1]. Another example of selective withdrawal of fluid is in the application of withdrawing, through a pump, pollution which occurs in rivers, channels and lakes. This matter attracts us to investigate it.

We therefore develop a mathematical model of the fluid flow which is induced by a submerged sink in a three-layer fluid. In our mathematical model, the sink of fluid is located on the vertical axis, the body force which is involved is surface tension. Further, a boundary integral technique has been implemented to solve that problem.

From the numerical results obtained in this investigation, when the effect of surface tension is included we conclude that it plays a very important role in the determination of the free surface fluid flow profile even at very small values of the Weber number.

Keywords: Submerged sink, boundary integral technique, surface tension, Weber number

1. Introduction

Over the last two decades the selective fluid withdrawal from a reservoir has received a great deal of attention, see for example [1]. One of the reasons for this wide range of applications is because of its use as a management technique for the supply of water with the desired water quality properties. Another example of the selective withdrawal of the fluid is in the removal through a pump of pollution that occurs in rivers, channels and lakes. In this situation, the pollution sources may come from an industrial cesspool, domestic waste, agriculture waste, etc. Further, the pollution may be in the form of a waste that endangers mankind, for example oil spillage on water that may be on a river, lake or ocean.

Further, some previous investigators, e.g. [2], [3] and [4], have considered the problem of the free surface fluid flow which is induced by a submerged line sink or source beneath a cusped free surface. [2] used a spectral method and a conformal mapping to obtain solutions for a submerged line source or sink in a fluid of infinite depth. This approach has subsequently proved very successful in a wide variety of free surface fluid flow problems in which surface waves are absent, see for example [5]. The

problem investigated by [2] has been further developed by other researchers who have been concerned with the presence of a bottom surface. In the three-layer configuration with the fluid flow steady, ideal and irrotational, the situation when there are cusp points on both free surfaces was considered by [6]. They found that solutions exist for values of the Froude numbers greater than unity.

In this paper we develop a mathematical model of fluid withdrawal through a submerged line sink from a three-layer fluid under the effects of both gravity and surface tension at infinite Froude numbers when fluid is withdrawn from the geometrical situation which consists of three homogeneous layers of fluid of different densities separated by two free surfaces. The fluids of lighter and heavier densities occupying the upper and lower regions, respectively. In our mathematical model we assume that the sink of fluid is located on the vertical axis where this axis is perpendicular to the horizontal x-axis.

2. Problem Formulation

The problem of fluid withdrawal through a submerged line sink from a three-layer fluid, incorporating the effects of surface tension at infinite Froude numbers, is considered. Each of three layers of fluid has a different constant density, i.e. ρ_1, ρ_2 and $\rho_3 \left(\rho_1 < \rho_2 < \rho_3\right)$. The light and heavy fluids occupy semi-infinite regions of space and are separated by an infinite strips of fluid of height H and density ρ_2 in the middle layer, see Figure 1(a). A source of volume flowrate 2Q is now taken place at the point S. We assume that on both free surfaces that cusp points occur vertically above and vertically below the point S, at the points A and C, respectively. The flow is assumed to be incompressible, irrotational, inviscid and steady to move under the effect of surface tension. In this fluid only the square of the fluid velocity is involved in the governing equation and therefore the solutions are insensitive to the direction in which the fluid flows along their free surfaces. Consequently, the direction of potential flow may be reversed and the results for a submerged line sink may be obtained from the solution for a submerged line source simply by changing the sign. We therefore consider the fluid flow due to a source.

Further, a coordinate system Z = X + iY is introduced with the X axis along the undisturbed lower free surface and The Y axis through the source S. In Figure 1(a), H_S, H_A and H_C are the heights of the source and cusp points A and C, respectively, above the undisturbed level of the lower free surface. Far from the source, the fluid flow in the middle layer of fluid is uniform and has speed U_∞ away from the source. The complex velocity potential is defined by $W = \Phi + i\Psi$, where Φ is the velocity potential and Ψ is the stream function. Without any loss of generality we choose Φ=0 at the point A, Ψ=0 on the lower free surface SCD′, and Ψ=UH on the upper free surface SAD, see Figure 1(a).

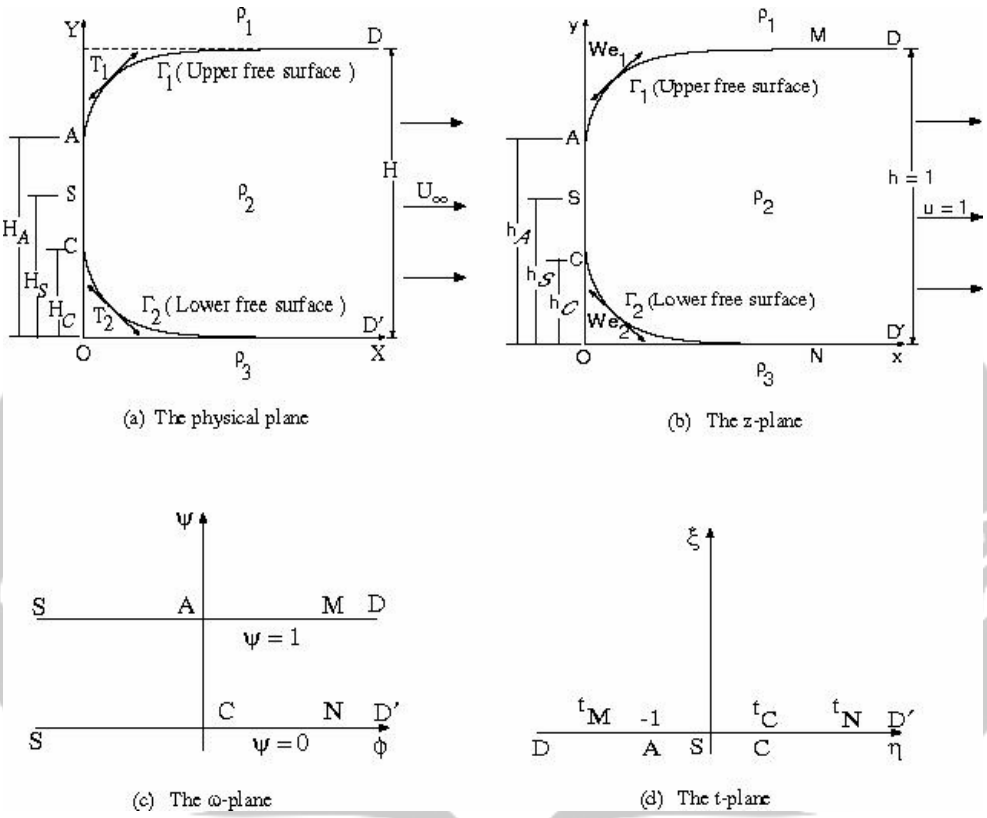


Figure 1*: (a) and (b) show a schematic diagram of the physical plane and the non-dimensional physical z-plane, whilst (c) and (d) show the complex velocity potential ω-plane and the transformed t-plane, respectively.*

We further implement a non-dimensional coordinate system $z = x + iy = (X + iY)/H$, and Figure 1(b) shows the flow in the physical z-plane, where $h_A = H_A/H$, $h_S = H_S/H$ and $h_C = H_C/H$. The complex velocity potential may be non-dimensional form as $\omega = \varphi + i\psi = (\Phi + i\Psi)/UH$, and the strip in Figure 1(c) represents the flow in the ω-plane where all speeds are now non-dimensional form with respect to U_∞.

We now apply Bernoulli's equation on the upper free surface AD to obtain the non-dimensional fluid speed, namely,

$$u_1 = \sqrt{1 + 2(1 - y_1)/Fr_1^2 - 2\,We_1(d\theta_1/ds_1)} \qquad (1)$$

where $u_1 = U_1 / U_\infty$, $y_1 = Y_1 / H$ and $s_1 = S_1 / H$. The Froude number and Weber number on the upper free surface far from the source are denoted by $Fr_1 = U_\infty / \sqrt{g_1 H}$ and $We_1 = T_1 / U_\infty{}^2 \rho_1 H$, respectively, where $g_1 = g(\rho_2 - \rho_1)/\rho_2$, T_1 is the surface tension coefficient of the upper free surface and s_1 is the non-dimensional of the upper free surface length. Whereas, on the lower free surface, CD′, Bernoulli's equation produces a non-dimensional fluid velocity in the form

$$u_2 = \sqrt{1 + 2(1 - y_2)/Fr_2{}^2 - 2We_2(d\theta_2 / ds_2)} \tag{2}$$

where $u_2 = U_2 / U_\infty$, $y_2 = Y_2 / H$ and $s_2 = S_2 / H$. The Froude number and Weber number on the lower free surface far from the source are denoted by $Fr_2 = U_\infty / \sqrt{g_2 H}$ and $We_2 = T_2 / U_\infty{}^2 \rho_2 H$, respectively, where $g_2 = g(\rho_3 - \rho_2)/\rho_2$, T_2 is the surface tension coefficient of the lower free surface and the non-dimensional lower free surface length is denoted by s_2.

We next consider the complex velocity which is related to the complex potential by $d\omega / dz$. We therefore have the identities namely,

$$d\omega / dz = u\, e^{-i\theta} \tag{3}$$

in which u is the non-dimensional fluid speed at some point in the flow field and θ represents the angle that the streamline makes with the positive x-axis at that point. To solve the problem of the free surface fluid flows induced by a submerged line source under the effect of surface tension, we apply the Riemann-Hilbert's technique, namely,

$$\Omega = i \ln\left(\frac{d\omega}{dz}\right) = \theta + i\tau \tag{4}$$

in which τ = ln u, and τ, dω/dz and Ω are analytical functions in the strip of the ω-plane, see figures 1(b) and 1(c). In the Riemann-Hilbert's technique, all calculations are based on a upper half plane of the t-plane and therefore the infinite strip in the complex ω-plane is transformed onto the upper half-plane of the auxiliary transformed t-plane by using the mapping function, see figure 1(d), namely

$$t = e^{\pi\omega} \tag{5}$$

Futher, we obtain a Riemann-Hilbert mixed boundary-value problem, see [7] and the boundary condition on the real η axis of the t-plane are as follow:

$$\Im m.\Omega(\eta) = \tau_1(\eta) = \frac{1}{2}\ln\left(1 + \frac{2(1 - y_1(\eta))}{Fr_1{}^2} - 2We_1 \frac{d\theta_1}{ds_1}\right), \qquad \eta < -1 \tag{6}$$

$$\Re e.\Omega(\eta) = \theta_1(\eta) = \frac{\pi}{2} \qquad -1 < \eta < 0, \tag{7}$$

$$\Re e.\Omega(\eta) = \theta_2(\eta) = -\frac{\pi}{2} \qquad 0 < \eta < t_C \tag{8}$$

$$\Im m.\Omega(\eta) = \tau_2(\eta) = \frac{1}{2}\ln\left(1 + \frac{2y_2(\eta)}{Fr_2{}^2} - 2We_1\frac{d\theta_2}{ds_2}\right) \qquad \eta > t_C \tag{9}$$

On the free surface at infinity, i.e. far away from the source, the non-dimensional speed u = 1 and the free surface become a horizontal line. Therefore,

$$\lim_{t\to\infty}\Omega(t) = \lim_{t\to\infty}\theta(t) + i\lim_{t\to\infty}\tau(t) = 0 \tag{10}$$

By referring to the general solution of the Riemann-Hilbert problem, see [7], we obtain the solution of Ω in the form

$$\begin{aligned}\Omega(t) = \frac{\sqrt{(t+1)(t-t_C)}}{\pi}\Bigg[&\int_{-\infty}^{-1}\frac{\tau_1(\eta)}{\sqrt{(\eta+1)(\eta-t_C)}(\eta-t)}d\eta \\ &+ \frac{\pi}{2}\int_{-1}^{0}\frac{1}{\sqrt{(\eta+1)(t_C-\eta)}(\eta-t)}d\eta - \frac{\pi}{2}\int_{0}^{t_C}\frac{1}{\sqrt{(\eta+1)(t_C-\eta)}(\eta-t)}d\eta \\ &- \int_{t_C}^{\infty}\frac{\tau_2(\eta)}{\sqrt{(\eta+1)(\eta-t_C)}(\eta-t)}d\eta\Bigg],\end{aligned} \tag{11}$$

Where X(t) = $\sqrt{(t+1)(t-t_C)}$ is a particular solution for the homogeneous solution of Ω(t) when its real part, $\Re e(\Omega)$, and the imaginary part, $\Im m(\Omega)$, are equal to zero on the real axis η. When t approaches the real axis η from the upper half plane, the value of X(t) on the real axis η is given by

$$X^+(\eta) = \sqrt{(\eta+1)(\eta-t_C)} \qquad \eta < -1, \tag{12}$$

$$X^+(\eta) = -i\sqrt{(\eta+1)(t_C-\eta)} \qquad -1 < \eta < t_C, \tag{13}$$

$$X^+(\eta) = -\sqrt{(\eta+1)(\eta-t_C)} \qquad \eta > t_C \tag{14}$$

The exact solution of the free surface fluid flow induced by a submerged source or sink in a three-layer problem can be obtain when the Froude numbers are infinite and the Weber numbers are zero. In fact, when Fr_1 and Fr_2 are infinite the speeds on both of the free surface are equal to unity and therefore τ = 0. Equation (11) then becomes

$$\Omega(t)=\frac{\sqrt{(t+1)(t-t_C)}}{\pi}\left[\frac{\pi}{2}\int_{-1}^{0}\frac{1}{\sqrt{(\eta+1)(t_C-\eta)}(\eta-t)}d\eta-\frac{\pi}{2}\int_{0}^{t_C}\frac{1}{\sqrt{(\eta+1)(t_C-\eta)}(\eta-t)}d\eta\right], \quad (15)$$

where t_C is an unknown parameter which has to be determined using the boundary condition at infinity.

Using the condition at infinity, equation (10) yields

$$\int_{-1}^{0}\frac{1}{\sqrt{(\eta+1)(t_C-\eta)}}d\eta-\int_{0}^{t_C}\frac{1}{\sqrt{(\eta+1)(t_C-\eta)}}d\eta=0 \quad (16)$$

or

$$\sin^{-1}\left(\frac{1-t_C}{1+t_C}\right)=0 \quad (17)$$

Hence,

$t_C = 1$, namely, $\phi_C = 0$ and $\phi_A = \phi_C$ (18)

Therefore, the flow field should be symmetrical about the horizontal plane which passes through the source and this implies $h_S = 0.5$. We now obtain the angle that the case surface make with the horizontal by taking the principal value of expression (15) on the intervals $-\infty < \eta < -1$ and $1 < \eta < \infty$, namely

$$\theta(\eta)=-\sin^{-1}\left(\frac{1}{\eta}\right) \qquad -\infty < \eta < -1 \text{ and } 1 < \eta < \infty. \quad (19)$$

Equation (3) may now be written in the form

$$dz=e^{i\theta}\frac{d\omega}{ud\eta}d\eta=\frac{1}{\pi}e^{i\theta}\frac{1}{u\eta}.d\eta \quad (20)$$

And this may be integrated, subject to the condition that on the upper free surface AD, $z \to \infty + I$, as $\eta \to -\infty$, to give

$$z=-\frac{\sqrt{(\eta^2-1)^3}}{\eta}-\eta\sqrt{\eta^2-1}-\ln\left(\sqrt{\eta^2-1}-\eta\right)+i\frac{1}{\pi}\left(\frac{1}{\pi}+\pi\right) \quad (21)$$

And on the lower free surface CD', $z \to \infty + I$, as $\eta \to \infty$, to give

$$z = -\frac{\sqrt{(\eta^2-1)^3}}{\eta} + \eta\sqrt{\eta^2-1} - \ln\left(\sqrt{\eta^2-1}-\eta\right) + i\frac{1}{\pi}\left(\frac{1}{\pi}+\pi\right) \tag{22}$$

The position of the cusp points A and C are now given by

$$h_A = \frac{\pi-1}{\pi} = 0.6817 \quad \text{and} \quad h_c = \frac{1}{\pi} = 0.3183 \tag{23}$$

We now consider the nonlinear solutions for $1 < Fr_1, Fr_2 < \infty$ and this case no analytical solution can be obtained. Therefore, we may obtain numerical solution by the interactive technique. First, we develop a numerical interactive technique to obtain the velocity potential at the point C and the profile of the free surface, and then we may determine the position of the cusp point A and C and the sink S. In the nonlinear integral equation (11), we let t be in the upper half-plane and then take the limit as t approaches the real axis η, together with the Cauchy principle value. Further, the real and imaginary parts of Ω(t) can be separated as follows. The real part, θ(t), is the angle the free surface makes with the horizontal x-axis which for the upper free surface AD which is expressed by

$$\begin{aligned}\theta_1(t) = {} & \frac{\sqrt{(t+1)(t-t_C)}}{\pi}\left[\int_{-\infty}^{-1}\frac{\tau_1(\eta)}{\sqrt{(\eta+1)(\eta-t_C)}(\eta-t)}d\eta - \int_{t_C}^{\infty}\frac{\tau_2(\eta)}{\sqrt{(\eta+1)(\eta-t_C)}}d\eta\right] \\ & + \frac{\sqrt{(t+1)(t-t_C)}}{\pi}\left[\frac{\pi}{2}\int_{-1}^{0}\frac{1}{\sqrt{(\eta+1)(\eta-t_C)}(\eta-t)}d\eta\right. \\ & \left. - \frac{\pi}{2}\int_{0}^{t_C}\frac{1}{\sqrt{(\eta+1)(t_C-\eta)}(\eta-t)}d\eta\right], -\infty < t < -1\end{aligned} \tag{24}$$

Equation (24) comprises of the *Cauchy Principle Value* and a finite integral. Therefore, the finite integral in equation (24) can be solved by using analytical solution, namely,

$$\begin{aligned}& -\frac{\sqrt{(t-1)(t-t_C)}}{\pi}\frac{\pi}{2}\int_{0}^{t_C}\frac{1}{\sqrt{(\eta+1)(t_C-\eta)}(\eta-t)}d\eta = \\ & -\frac{\sqrt{(t-1)(t-t_C)}}{2}\int_{0}^{t_C}\frac{1}{\sqrt{(\eta+1)(t_C-\eta)}(\eta-t)}d\eta\end{aligned} \quad , \quad -\infty < t < -1 \tag{25}$$

Further, the integral

$$\int \frac{d\eta}{\sqrt{(\eta+1)(t_C-\eta)}(\eta-t)} \tag{26}$$

Can be solve in the form of

$$\int \frac{dx^*}{x^*\sqrt{X^*}} \tag{27}$$

Where $X^* = a + bx^* + cx^*$ is a 2 degree polynomial. By using the substitution

$$X^* = \eta - t \tag{28}$$

We may obtain

$$X^* = -x^{*2} + x^* (-2t + t_C - 1) + (t+1)(t_C - t) \tag{29}$$

And hence

$$A = (t+1)(t_C - t), \quad b = -2t + t_C - 1 \quad \text{and } c = 1 \tag{30}$$

If $-\infty < t < -1$ and $1 < 0$ so that

$$\int \frac{dx^*}{x^*\sqrt{X^*}} = \frac{1}{\sqrt{-a}} \sin^{-1}\left(\frac{bx^* + 2a}{x^*\sqrt{b^2 - 4ac}}\right) \tag{31}$$

or

$$\int \frac{dx^*}{x^*\sqrt{X^*}} = \frac{1}{\sqrt{(t+1)(t-t_C)}} \tag{32}$$

$$\sin^{-1}\left[\frac{(-2t + t_C - 1)(\eta - t) + 2(t+1)(t_C - t)}{(\eta - t)\sqrt{(-2t + t_C - 1)^2 + 4(t+1)(t_{C-t})}}\right]$$

Therefore

$$\int_0^{T_c} \frac{d\eta}{\sqrt{(\eta+1)(t_C - \eta)}(\eta - t)} = \frac{1}{\sqrt{(t+1)(t-t_C)}} \tag{33}$$

$$\left[\sin^{-1}\left(\frac{(-2t + t_C - 1)(t_C - t) + (t+1)(t_C - t)}{(t_C - t)\sqrt{(-2t + t_C - 1)^2 + 4(t+1)(t_C - t)}}\right) - \sin^{-1}\left(\frac{(-2t + t_C - 1)(-1 - t) + 2(t+1)(t_C - t)}{(-1 - t)\sqrt{(-2t + t_C - 1)^2 + 4(t+1)(t_C - t)}}\right)\right]$$

and

$$\int_{-1}^{0}\frac{d\eta}{\sqrt{(\eta+1)(t_C-\eta)}(\eta-t)}=\frac{1}{\sqrt{(t+1)(t-t_C)}} \tag{34}$$

$$\left[\sin^{-1}\left(\frac{(-2t+t_C-1)(-t)+2(t+1)(t_C-t)}{(-t)\sqrt{(-2t+t_C-1)^2+4(t+1)(t_C-t)}}\right)-\right.$$
$$\left.\sin^{-1}\left(\frac{(-2t+t_C-1)(t-1)+2(t+1)(t_C-t)}{(-t)\sqrt{(-2t+t_C-1)^2+4(t+1)(t_C-t)}}\right)\right]$$

Further,

$$\frac{\sqrt{(t+1)(t-t_C)}}{2}\int_{-1}^{0}\frac{d\eta}{\sqrt{(\eta+1)(t_C-\eta)}(\eta-t)}-\frac{\sqrt{(t+1)(t-t_C)}}{2}$$
$$\int_{0}^{tc}\frac{d\eta}{\sqrt{(\eta+1)(t_C-\eta)}(\eta-t)}=\sin^{-1}\left(\frac{t(1-t_C)-2t_C}{t(t_C+1)}\right) \tag{35}$$

Therefore, the angle on the upper free surface AD, $\theta_1(t)$, can be expressed as

$$\theta_1(t)=\frac{\sqrt{(t+1)(t-t_C)}}{\pi}\left[\int_{-\infty}^{-1}\frac{\tau_1(\eta)}{\sqrt{(\eta+1)(\eta-t_C)}(\eta-t)}d\eta\right.$$
$$\left.-\int_{tc}^{\infty}\frac{\tau_2(\eta)}{\sqrt{(\eta+1)(\eta-t_C)(\eta-t)}}d\eta\right]+\arcsin\left(\frac{t(1-t_C)-2t_C}{t(t_C+1)}\right) \qquad -\infty < t < -1 \tag{36}$$

And for the lower free surface CD‘ is given by

$$\theta_1(t)=\frac{\sqrt{(t+1)(t-t_C)}}{\pi}\left[-\int_{-\infty}^{-1}\frac{\tau_1(\eta)}{\sqrt{(\eta+1)(\eta-t_C)}(\eta-t)}d\eta\right.$$
$$\left.+\int_{tc}^{\infty}\frac{\tau_2(\eta)}{\sqrt{(\gamma+1)(\eta-t_C)(\eta-t)}}d\eta\right]+\frac{\sqrt{(t+1)(t-t_C)}}{\pi}\left[\frac{\pi}{2}\int_{-1}^{0}\frac{1}{\sqrt{(\eta+1)(\eta-t_C)}(\eta-t)}d\eta\right.$$
$$\left.-\frac{\pi}{2}\int_{0}^{tc}\frac{1}{\sqrt{(\eta+1)(t_C-\eta)}(\eta-t)}d\eta-\right] \qquad t_C < t < \infty \tag{37}$$

Equation (37) also caries a Cauchy Principle Value integral and a finite integral and hence we may use an analytical solution to solve that finite integral by the following method:

$$\frac{\sqrt{(t+1)(t-t_C)}}{\pi}\frac{\pi}{2}\int_{-1}^{0}\frac{d\eta}{\sqrt{(\eta+1)(t_C-\eta)}(\eta-t)}-\frac{\sqrt{(t+1)(t-t_C)}}{2}\frac{\pi}{2}$$
$$\int_{0}^{tc}\frac{d\eta}{\sqrt{(\eta+1)(t_C-\eta)}(\eta-t)}=\frac{\sqrt{(t+1)(t-t_C)}}{2}\left[\int_{-1}^{0}\frac{d\eta}{\sqrt{(\eta+1)(t_C-\eta)}(\eta-t)}-\right.$$
$$\left.\int_{0}^{tc}\frac{d\eta}{\sqrt{(\eta+1)(t_C-\eta)}(\eta-t)}\right]$$

$t_C < t < \infty$ (38)

Further, by applying expressions (26), (27), (28), and (29), we obtain an expression similar with expression (30). If $t > t_C$ then $a > 0$ and hence we obtain expression similar with expression (31) and (32). Therefore, the angle on the lower free surface CD is given by

$$\theta_2(t)=\frac{\sqrt{(t+1)(t-t_C)}}{\pi}\left[-\int_{-\infty}^{-1}\frac{\tau_1(\eta)}{\sqrt{(\eta+1)(\eta-t_C)}(\eta-t)}d\eta\right.$$
$$\left.+\int_{tc}^{\infty}\frac{\tau_2(\eta)}{\sqrt{(\gamma+1)(\eta-t_C)(\eta-t)}}d\eta\right]+\arcsin\left(\frac{t(1-t_C)-2t_C}{t(t_C+1)}\right), \quad t_C < t < \infty \tag{39}$$

Whereas the imaginary part, τ (t), is the logarithm of the non-dimensional fluid speed on AC, which given by

$$\tau(t)=\ln u(t)=\frac{\sqrt{(t+1)(t_c-t)}}{\pi}\left[-\int_{-\infty}^{-1}\frac{\tau_1(\eta)}{\sqrt{(\eta+1)(\eta-t_C)}(\eta-t)}d\eta\right.$$
$$\left.+\int_{t_C}^{\infty}\frac{\tau_2(\eta)}{\sqrt{(\eta+1)(\eta-t_C)}(\eta-t)}d\eta\right]$$
$$+\frac{\sqrt{(t+1)(t_C-t)}}{\pi}\left[-\frac{\pi}{2}\int_{-1}^{0}\frac{d\eta}{\sqrt{(\eta+1)(t_C-\eta)}(\eta-t)}\right.$$
$$\left.+\frac{\pi}{2}\int_{0}^{t_C}\frac{d\eta}{\sqrt{(\eta+1)(t_C-\eta)}(\eta-t)}\right] \quad -1<t<t_C \tag{40}$$

For the non-dimension fluid speed on AS, expression (40) can be used but -1 < t < 0. It therefore can be expressed by

$$\tau(t) = \ln u(t) = \frac{\sqrt{(t+1)(t_c - t)}}{\pi}$$
$$\left[-\int_{-\infty}^{-1} \frac{\tau_1(\eta)}{\sqrt{(\eta+1)(\eta - t_c)(\eta - t)}} d\eta + \int_{t_c}^{\infty} \frac{\tau_2(\eta)}{\sqrt{(\eta+1)(\eta - t_c)(\eta - t)}} d\eta \right]$$
$$+ \frac{\sqrt{(t+1)(t_C - t)}}{\pi} \left[-\frac{\pi}{2} \int_{-1}^{0} \frac{d\eta}{\sqrt{(\eta+1)(t_C - \eta)}(\eta - t)} \right.$$
$$\left. + \frac{\pi}{2} \int_{0}^{t_C} \frac{d\eta}{\sqrt{(\eta+1)(t_C - \eta)}(\eta - t)} \right] \quad -1 < t < 0 \tag{41}$$

Further, in expression (40), every integral does not contain a singularity point and hence the Cauchy principal value integral is not included. Further, from expression (40) we may find the speed on AS where -1 < t < 0 and the speed on SC where 0 < t < tc. For obtaining the non-dimensional fluid speed on AS, we solve the integrals

$$-\frac{\sqrt{(t-1)(t_C - t)}}{\pi} \left[\frac{\pi}{2} \int_{-1}^{0} \frac{d\eta}{\sqrt{(\eta+1)(t_C - \eta)}(\eta - t)} \right] \quad -1 < t < 0 \tag{42}$$

And

$$\frac{\sqrt{(t+1)(t_C - t)}}{\pi} \left[\frac{\pi}{2} \int_{0}^{t_C} \frac{d\eta}{\sqrt{(\eta+1)(t_C - \eta)}(\eta - t)} \right] \quad -1 < t < 0 \tag{43}$$

For solving the integral (42), we use equation (28) to solve the integral in the forms of (26) and (27). We therefore obtain solutions in the same manner as we did for equation (29) and (30). Further, if -1< t < 0 then a > 0 so that

$$\int \frac{dx^*}{x^* \sqrt{X^*}} = -\frac{1}{\sqrt{a}} \ln \left(\frac{\sqrt{X^*} + \sqrt{x^*}}{x^*} + \frac{b}{2\sqrt{a}} \right) \tag{44}$$

Or

$$\int \frac{dx^*}{x^* \sqrt{X^*}} = -\frac{1}{\sqrt{(t-1)(t_C - t)}} \ln \left(\frac{-2t + t_C - 1}{2\sqrt{(t+1)(t_C - 1)}} + \right.$$

$$\left. \frac{\sqrt{(\eta+1)(t_C-\eta)}+\sqrt{(t-1)(t_C-t)}}{\eta-t} \right) \tag{45}$$

Therefore

$$-\frac{\sqrt{(t+1)(t_C-t)}}{2}\int_{-1}^{0}\frac{d\eta}{\sqrt{(\eta+1)(t_C-\eta)(\eta-t)}} = -\frac{1}{2}\ln\left(\frac{t(1+t_C)}{\left(\sqrt{t_C-t}+\sqrt{t_C(t+1)}\right)^2}\right) \tag{46}$$

Whereas, for the integral (43) we obtain

$$-\frac{\sqrt{(t+1)(t_C-t)}}{2}\int_{-1}^{0}\frac{d\eta}{\sqrt{(\eta+1)(t_C-\eta)}(\eta-t)} = -\frac{1}{2}\ln\left(\frac{t(1+t_C)}{\left(\sqrt{t_C-t}+\sqrt{t_C(t+1)}\right)^2}\right) \tag{47}$$

Further, if we sum equations (46) with (47), we obtain

$$-\ln\left(\frac{t(1+t_C)}{\left(\sqrt{t_C-t}+\sqrt{t_C(t+1)}\right)^2}\right) \tag{48}$$

Therefore, the non-dimensional fluid speed on AS can be expressed by

$$\tau(t) = \frac{\sqrt{(t+1)(t_C-t)}}{\pi}\left[-\int_{-\infty}^{-1}\frac{\tau(\eta)}{\sqrt{(\eta+1)(\mu-t_C)}(\eta-t)}d\eta + \int_{t_C}^{\infty}\frac{\tau(\eta)}{\sqrt{(\eta+1)(\eta-t_C)}(\eta-t)}d\eta\right] - \ln\left(\frac{t(1+t_C)}{\left(\sqrt{t_C-t}+\sqrt{t_C(t+1)}\right)^2}\right) \qquad -1 < t < 0. \tag{49}$$

Further, for non-dimensional fluid speed on CS can be expressed by

$$\tau(t) = \ln u(t) = \frac{\sqrt{(t+1)(t_c - t)}}{\pi}\left[-\int_{-\infty}^{-1} \frac{\tau_1(\eta)}{\sqrt{(\eta+1)(\eta - t_c)}(\eta - t)} d\eta \right.$$
$$\left. + \int_{t_c}^{\infty} \frac{\tau_2(\eta)}{\sqrt{(\eta+1)(\eta - t_c)}(\eta - t)} d(\eta)] \right]$$
$$+ \frac{\sqrt{(t+1)(t_c - t)}}{\pi}\left[-\frac{\pi}{2}\int_{-1}^{c} \frac{d(\eta)}{\sqrt{(\eta+1)(t_c - \eta)}(\eta - t)} \right. \quad (50)$$
$$\left. + \frac{\pi}{2}\int_{0}^{t_c} \frac{d(\eta)}{\sqrt{(\eta+1)(t_c - \eta)}(\eta - t)} \right] \quad -0 < t < t_c$$

Expression (50) looks similar to the expression (41) and hence we solve that expression using similar techniques. Therefore, expression (50) can be expressed in a similar way as for the expression (49) where $0 < t < t_c$ or it can be expressed by

$$\tau(t) = \ln u(t) = \frac{\sqrt{(t+1)(t_c - t)}}{\pi} - \left[\int_{-\infty}^{-1} \frac{\tau_1(\eta)}{\sqrt{(\eta+1)(\eta - t_c)}(\eta - t)} d\eta \right.$$
$$\left. + \int_{t_c}^{\infty} \frac{\tau(\eta)}{\sqrt{(\eta+1)(\eta - t_c)}(\eta - t)} d\eta \right] \quad (51)$$
$$- \ln\left(\frac{t(1+t_c)}{\left(\sqrt{t_c - t} + \sqrt{t_c(t+1)}\right)^2} \right) \quad -0 < t < t_c$$

Further, on the free surface profiles at infinity, i.e. Far away from the source, the non-dimensional fluid speed the horizontal line is zero. In this case, the boundary condition is given by

$$\lim_{t\to\infty} \Omega(t) = 0 \quad or \quad \lim_{t\to\infty} \tau(t) = 0 \quad \lim_{t\to\infty} \theta(t) = 0 \quad (52)$$

Further, on the upper free surface, i.e. Γ_1, at infinity the angle that the free surface makes with the horizontal is given by

$$\lim_{t\to-\infty}\theta_1(t)=\lim_{t\to-\infty}\frac{\sqrt{(t+1)(t-t_c)}}{\pi}\left[\int_{-\infty}^{-1}\frac{\tau_1(\eta)}{\sqrt{(\eta+1)(\eta-t_c)}(\eta-t)}d\eta+\int_{t_c}^{\infty}\frac{\tau_2(\eta)}{\sqrt{(\eta+1)(\eta-t_c)}(\eta-t)}d(\eta)\right]+\lim_{t\to-\infty}\frac{\sqrt{(t+1)(t_c-t)}}{\pi}\left[\frac{\pi}{2}\int_{-1}^{0}\frac{d(\eta)}{\sqrt{(\eta+1)(t_c-\eta)}(\eta-t)}-\frac{\pi}{2}\int_{0}^{t_c}\frac{d(\eta)}{\sqrt{(\eta+1)(t_c-\eta)}(\eta-t)}\right]=0 \tag{53}$$

This expression can be written in the form

$$\lim_{t\to-\infty}\frac{1}{\pi}\left[\int_{-\infty}^{-1}\frac{\tau_1(\eta)}{\sqrt{(\eta+1)(\eta-t_c)}}d\eta-\int_{t_c}^{\infty}\frac{\tau_2(\eta)}{\sqrt{(\eta+1)(\eta-t_c)}}d\eta\right]+\lim_{t\to-\infty}\frac{1}{2}\left[\int_{-1}^{0}\frac{d(\eta)}{\sqrt{(\eta+1)(t_c-\eta)}}-\int_{0}^{t_c}\frac{d(\eta)}{\sqrt{(\eta+1)(t_c-\eta)}}\right]=0 \tag{54}$$

Further, the second term on the left hand side of equation (54) can be solved using the following method:

We solve

$$\int\frac{d(\eta)}{\sqrt{(\eta+1)(t_c-\eta)}} \tag{55}$$

in the form of

$$\int\frac{dx^*}{\sqrt{X^*}} \tag{56}$$

Where $X^*=a+bx^*+cx^{*2}$. by using the substitution $x^*=\eta$ then

$$X^*=(x^*+1)(t_c-x^*)=-x^{*2}+x^*(t_c-1)+t_c. \tag{57}$$

and hence we may obtain

$$a = t_c\ , \quad b = t_c - 1\ , \quad and \quad c = -1 \tag{58}$$

And hence

$$\int \frac{dx^*}{\sqrt{X^*}} = \frac{1}{-\sqrt{c}} \sin^{-1}\left(\frac{-2cx^* - b}{\sqrt{b^2 - 4ac}}\right) \tag{59}$$

Further, equation (49) can be written in the form

$$\int \frac{d\eta}{\sqrt{(\eta+1)(t_c - \eta)}} = \sin^{-1}\left(\frac{-2\eta + 1 - t_c}{1 + t_c}\right) \tag{60}$$

Therefore we have

$$\frac{1}{2}\int_{-1}^{0} \frac{d\eta}{\sqrt{(\eta+1)(t_c - \eta)}} = \frac{1}{2}\left[\sin^{-1}\left(\frac{1-t_c}{1+t_c}\right) - \sin^{-1}(-1)\right] \tag{61}$$

And

$$\frac{1}{2}\int_{0}^{t_c} \frac{d\eta}{\sqrt{(\eta+1)(t_c - \eta)}} = \frac{1}{2}\left[\sin^{-1}(1) - \sin^{-1}\left(\frac{1-t_c}{1+t_c}\right)\right] \tag{62}$$

Further, if we subtract equation (61) from equation (62) then we obtain

$$\begin{aligned} &\frac{1}{2}\left[\sin^{-1}\left(\frac{1-t_c}{1+t_c}\right) - \sin^{-1}(-1)\right] - \\ &\frac{1}{2}\left[\sin^{-1}(1) - \sin^{-1}\left(\frac{1-t_c}{1+t_c}\right)\right] = \sin^{-1}\left(\frac{1-t_c}{1+t_c}\right) \end{aligned} \tag{63}$$

Therefore equation (54) can be expressed by

$$\begin{aligned} &\frac{1}{\pi}\left[\int_{-\infty}^{-1} \frac{\tau_1(\eta)}{\sqrt{(\eta+1)(\eta - t_c)}} d\eta - \int_{t_c}^{\infty} \frac{\tau_2(\eta)}{\sqrt{(\eta+1)(\eta - t_c)}} d\eta\right] \\ &+ arc\sin\left(\frac{1-t_c}{t_c+1}\right) = 0 \end{aligned} \tag{64}$$

Whereas Γ_2 can be expressed by

$$\lim_{t\to-\infty}\theta_2(t)=\lim_{t\to-\infty}\frac{\sqrt{(t+1)(t-t_c)}}{\pi}\left[-\int_{-\infty}^{-1}\frac{\tau_1(\eta)}{\sqrt{(\eta+1)(\eta-t_c)}(\eta-t)}d\eta+\int_{t_c}^{\infty}\frac{\tau_2(\eta)}{\sqrt{(\eta+1)(\eta-t_c)}(\eta-t)}d(\eta)\right]+\lim_{t\to-\infty}\frac{\sqrt{(t+1)(t_c-t)}}{\pi}\left[-\frac{\pi}{2}\int_{-1}^{0}\frac{d(\eta)}{\sqrt{(\eta+1)(t_c-\eta)}(\eta-t)}+\frac{\pi}{2}\int_{0}^{t_c}\frac{d(\eta)}{\sqrt{(\eta+1)(t_c-\eta)}(\eta-t)}\right]=0 \tag{65}$$

Further expression (65) can be written in the form

$$\lim_{t\to-\infty}\frac{1}{\pi}\left[-\int_{-\infty}^{-1}\frac{\tau_1(\eta)}{\sqrt{(\eta+1)(\eta-t_c)}}d\eta+\int_{t_c}^{\infty}\frac{\tau_2(\eta)}{\sqrt{(\eta+1)(\eta-t_c)}}d\eta\right]+\lim_{t\to-\infty}\frac{1}{2}\left[-\int_{-1}^{0}\frac{d(\eta)}{\sqrt{(\eta+1)(t_c-\eta)}}+\int_{0}^{t_c}\frac{d(\eta)}{\sqrt{(\eta+1)(t_c-\eta)}}\right]=0 \tag{66}$$

Further, expression (66) can be using the same technique as employed for expression (60), (61), (62) and (63). Therefore, expression (66) can be given by

$$\frac{1}{\pi}\left[-\int_{-\infty}^{-1}\frac{\tau_1(\eta)}{\sqrt{(\eta+1)(\eta-t_c)}}d\eta+\int_{t_c}^{\infty}\frac{\tau_2(\eta)}{\sqrt{(\eta+1)(\eta-t_c)}}d\eta\right]+\arcsin\left(\frac{t_c-1}{t_c+1}\right)=0 \tag{67}$$

When s, the distance measured along the free surface is taken as the independent variable, and on using equation (5), we obtain the expression

$$d\eta=\pi\eta(s)u(s)ds \tag{68}$$

Along both the vertical line AC and the free surfaces we have

$$\frac{d\phi(s)}{ds}=u(s) \tag{69}$$

And we also have the identities

$$\frac{dx(s)}{ds} = \cos\theta(s) \tag{70}$$

$$\frac{dx(s)}{ds} = \sin\theta(s) \tag{71}$$

Further, in the physical plane we may write expressions (36) and (39) as follows:

On the upper free surface:

$$\begin{aligned}\theta_1(s) = \sqrt{(t(s)+1)(t(s)-t_c}\Bigg[&\int_{\Gamma_2}\frac{\tau_2(l)\eta(l)u_2(l)}{\sqrt{(\eta(l)+1)(\eta(l)-t_c)}(\eta(l)-t_s)}dl \\ &-\int_{\Gamma_2}\frac{\tau_2(l)\eta(l)u_2(l)}{\sqrt{(\eta(l)+1)(\eta(l)-t_c)}(\eta(l)-t_s)}dl\Bigg] \\ &+\arcsin\left(\frac{t(s)(1-t_c)-2t_c}{t(s)(t_c+1)}\right) \qquad s\in\Gamma_2\end{aligned} \tag{72}$$

On the lower free surface:

$$\begin{aligned}\theta_2(s) = \sqrt{(t(s)+1)(t(s)-t_c}\Bigg[&-\int_{\Gamma_1}\frac{\tau_1(l)\eta(l)u_1(l)}{\sqrt{(\eta(l)+1)(\eta(l)-t_c)}(\eta(l)-t_s)}dl \\ &-\int_{\Gamma_2}\frac{\tau_2(l)\eta(l)u_2(l)}{\sqrt{(\eta(l)+1)(\eta(l)-t_c)}(\eta(l)-t_s)}dl \\ &+\arcsin\left(\frac{t(s)(1-t_c)-2t_c}{t(s)(t_c+1)}\right) \qquad s\in\Gamma_2\end{aligned} \tag{73}$$

In which $t_c = {}^{\pi\phi(c)}, \phi(c)$ is non-dimensional velocity potential at the point C, and on $,\Gamma_1,\quad \eta(l) = -e^{\pi\phi(s)}$ and $t(s) = -e^{\pi\phi(s)}$, whilst on $\Gamma_2,\quad \eta(l) = -e^{\pi\phi(s)}$ and $t(s) = -e^{\pi\phi(s)}$

The velocity potential on both the upper and lower free surface are given by

$$\phi_1(s) = \int_A^M u_1(l) \quad \text{on} \quad \Gamma_1 \tag{74}$$

and

$$\phi_2(s) = \phi_c + \int_C^N u_2(l)dl \qquad on \quad \Gamma_2 \tag{75}$$

respectively, and the coordinates of the upper surface are given by

$$y_1(s) = h_A + \int_A^M \sin\theta_1(l)dl, \tag{76}$$

$$x_1(s) = \int_A^M \cos\theta_1(l)dl, \tag{77}$$

Whilst the profile of the lower free surface is expressed by

$$y_2(s) = h_A + \int_C^N \sin\theta_2(l)dl, \tag{78}$$

$$x_2(s) = \int_C^N \cos\theta_2(l)dl, \tag{79}$$

In which h_A and h_c are given, subject to the condition that $y_1(s) \to 1$ as $s \to \infty$ and $y_2(s) \to 0$ as $s \to \infty$. In addition, the direction of all the linear integrations along the upper and lower free surface is along the direction of the fluid velocity. If the coordinate of the source may be expressed by

$$h(s) = h_C + \int_C^S e^{-\tau(l)} d\phi(l) \qquad l \quad on \quad CS, \tag{80}$$

or

$$h(s) = h_A + \int_A^S e^{-\tau(l)} d\phi(l) \qquad l \quad on \quad AS, \tag{81}$$

3. The Iterative Procedure

In each computation of the numerical scheme the value of $F_{r_1}, F_{r_2}, W_{e_1}$ and W_{e_2} were fixed, the velocity potential at the point C and the profile of free surfaces were obtain through the following iterative procedure:

(a) Assume the initial profiles of the free surfaces Γ_1 and Γ_2.

(b) Use expressions (1) and (2) to evaluate the speed on the upper free surface, i.e. $u_1^0(s)$, and on the lower free surface, i.e. $u_2^0(s)$, respectively.

(c) Assume the velocity potential at the point C, ϕ_C is its root. In our calculations the Newton_Raphson method was used to find the root of equation (64) in addition to obtain the velocity potentials on the upper, i.e. $\phi_1^0(s)$, upper and lower free surface, i.e. $\phi_2^0(s)$.

(d) Insert $u_1^0(s), u_2^0(s), \phi_1^0(s)$ and $\phi_2^0(s)$ into the right hand side of equation (72) to obtain the angle $\phi_1^0(s)$ on the upper free surface AM, and equation (73) to obtain the angle $\phi_2^0(s)$ on the lower free surface CN.

(e) From expressions (76), (77), (78) and (79) the new approximate free surface profile are determined.

These value of $u(s), \theta(s), \phi(s), x(s)$ and $y(s)$, are then used as new approximation and the iterative procedure from step (b) to step (e) is respected until the numerical procedure has converged when successive value at any point are smaller than 10^{-6}.

After obtaining the upper and lower free surface then the positions of the source can be obtain by using either expression (80) or (81).

4. Numerical Results and Discussion

The variation of the free surface profiles when surface tension applies, i.e. $We_1 = We_2 = We = 0$, 0.001, 0.002 and 0.003, and there is no surface tension, i.e. $We_1 = We_2 = We = 0$ and on the upper free surface when the velocity potential value, $\phi_C = 0$ are shown in Figure 2. The non-dimensional height of the cusp point A, h_A, above the horizontal x-axis gradually decreases as the Weber number increases. However, the non-dimensional height of the cusp point C, h_C, increases as the Weber number increases. The respective values of h_A are 0.6817, 0.6803, 0.6790 and 0.6778, whereas, those of h_C are 0.3183, 0.3197, 0.3210 and 0.3222, for We = 0, 0.001, 0.002 and 0.003, respectively. The non-dimensional height of the sink, h_S, is fixed although the Weber number increases, i.e. $h_S = 0.5$, from symmetry about y = 0.5, the bottom surface behaves in exactly the same manner as does the top surface. As expected, we observe that as the Weber number increases the free surface profiles become less curved due to the effects of the surface tension.

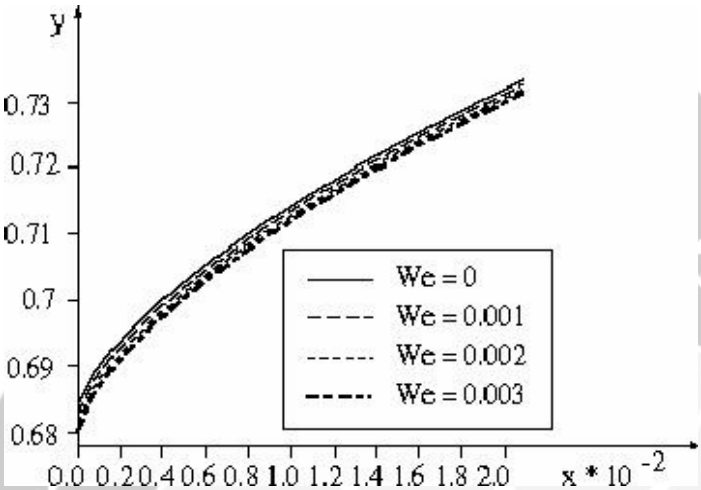


***Figure 2**: The variation of the free surface profile on the upper side when $We_1 = We_2 = We = 0$, 0.001, 0.002 and 0.003 when the Froude numbers are infinite, the velocity potential at the cusp point C is $\phi_C = 0$ and the free surface lengths are TL = 3.0.*

Figure 3 shows the variation that the free surface makes with the horizontal on the upper surface when there is surface tension, i.e. $We_1 = We_2 = We = 0$, 0.001, 0.002 and 0.003, and no surface tension, i.e. $We_1 = We_2 = We = 0$, the Froude numbers are infinite, the velocity potential at the cusp point C is $\phi_C = 0$ and the free surface lengths are TL = 3.0. From this figure it can be seen for $We \geq 0.003$ that the flow becomes unstable and the effects of capillary waves have been detected.

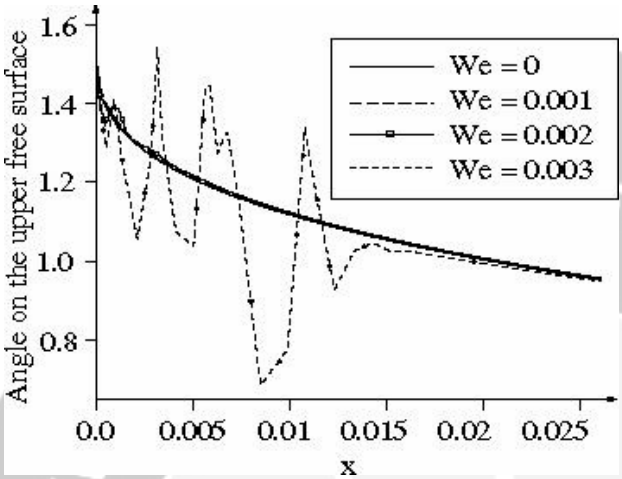


***Figure 3**: The variation that the free surface makes with the horizontal on the upper side when $We_1 = We_2 = We = 0$, 0.001, 0.002 and 0.003, the Froude numbers are infinite, the velocity potential at the cusp point C is $\phi_C = 0$ and the free surface lengths are TL = 3.0.*

5. Conclusion

A two-dimensional, steady, inviscid, incompressible and irrotational withdrawal of fluid through a submerged line sink from a three-layer fluid under the effect of surface tension at infinite Froude numbers problem has been considered using boundary integral technique. From the numerical results obtained in this investigation we conclude that when the Froude numbers are infinite we obtained that the surface tension plays a very important role in the determination of the free surface profiles, even at very small values of the Weber numbers.

Acknowledgement

We wish to acknowledge the financial support and the opportunity which is given to us from Mathematics Department - Sepuluh Nopember Institute of Technology, Surabaya-Indonesia through the Grant Project of PHK-A2 (The General Directorate of Higher Education – Indonesian Government).

References

[1] Imberger, J. and Patterson, J.C. (1989) Physical limnology, *Advanced in Applied Mechanics*, Vol. 27, pp. 302-475.

[2] Tuck, E.O and Vanden-Broeck, J.M. (1984) A cusp-like free surface flow due to a submerged source or sink, *Journal Australian Mathematics Soc. B.*, Vol. 25, pp. 443-450.

[3] King, A.C. and Bloor, M.I.G. (1988) A note on the free surface induced by a submerged source at infinite Froude number, *Journal Australian Mathematics Soc. B.*, Vol. 30, pp. 147-156.

[4] Hocking, G.C. (1985) Cusp-like free-surface flows due to a submerged source or sink in the presence of a flat or sloping bottom, *Journal Australian Mathematics Soc. B.*, Vol. 26, pp. 470-486.

[5] Yih, C.S. (1980) *Stratified Flows*, Academic Press, London.

[6] Wen, X. & Ingham, D. B. (1991) The free surface flow induced by a submerged source or sink from a three-layer fluid, Proceedings: *Computational Modeling of Free and Moving Boundary Problems, volume 1*, pp. 261-275.

[7] Muskhelishvilli, N. I. (1953) *Singular Integral Equations,* Edited by Radock, J. R. M. P. Noordhoff, Groningen, Holland.

BASUKI WIDODO: Mathematics Department, Faculty of Mathematics and Natural Sciences, INSTITUT TEKNOLOGI SEPULUH NOPEMBER – SURABAYA, Kampus ITS Keputih Sukolilo Surabaya (60111)- Indonesia. Telp.: 62-31-5943354/Fax.: 62-31-5996506

Email: b_widodo_2000@yahoo.com

A BEM FOR A CLASS OF ELLIPTIC BVPs OF FUNCTIONALLY GRADED MATERIALS

Mohammad Ivan Azis

Hasanuddin University, Makassar, Indonesia

Abstract. A Boundary Element Method (BEM) is derived for obtaining solutions to a class of elliptic boundary value problems (BVPs) of functionally graded, anisotropic media. Some particular examples are considered to illustrate the application of the BEM.
Key-words: Boundary Element Method, Elliptic BVP, Functionally Graded Materials

1 Introduction

Whereas the boundary element method (BEM) provides an effective numerical procedure for the solution of elliptic boundary value problems for homogeneous media the same is not generally true for inhomogeneous media. In recent years some progress has been made in using the method to solve problems for inhomogeneous isotropic materials (see for example Clements [5], Cheng [3, 4], Rangogni [9], Shaw [10], Gipson, Ortiz and Shaw [8] and Ang, Kusuma and Clements [1]). In the case of anisotropic inhomogeneous media there are few published studies. An elliptic equation which is relevant for a certain class of problems for anisotropic inhomogeneous media has been considered by Clements and Rogers [7]. They obtained a boundary integral equation for the case when the coefficients in the equation depend on one spatial variable only. Specifically the equation considered by Clements and Rogers [7] takes the form

$$\frac{\partial}{\partial x_i}\left[\lambda_{ij}(x_2)\,\frac{\partial\phi(x_1,x_2)}{\partial x_j}\right]=0,$$

where the coefficients λ_{ij} depend on x_2 only and the repeated summation convention (summing from 1 to 2) is employed.
This chapter is concerned with obtaining boundary integral equations for the solution of boundary value problems governed by equations of the form

$$\frac{\partial}{\partial x_i}\left[\lambda_{ij}(x_1,x_2)\,\frac{\partial\phi(x_1,x_2)}{\partial x_j}\right]=0. \tag{1}$$

Equations of this type govern the behavior of a wide class of boundary value problems of both isotropic and anisotropic inhomogeneous media. Antiplane strain in elastostatics and plane thermostatics for anisotropic inhomogeneous materials are two areas for which the governing equation is of the type (1).

Several techniques will be considered for obtaining boundary integral equations for the solution of (1). For each technique it is necessary to place some constraint on the class of coefficients λ_{ij} for which the solution obtained is valid. Some numerical examples are considered to illustrate the application of the boundary integral equations. The analysis of this chapter is purely formal; the main aim being to construct effective boundary element methods for classes of equations which fall within the type (1).

2 The boundary value problem

Referred to a Cartesian frame Ox_1x_2 a solution to (1) is sought which is valid in a region Ω in R^2 with boundary $\partial\Omega$ which consists of a finite number of piecewise smooth closed curves. On $\partial\Omega_1$ the dependent variable $\phi(\mathbf{x})$ ($\mathbf{x} = (x_1, x_2)$) is specified and on $\partial\Omega_2$

$$P(\mathbf{x}) = \lambda_{ij} \left(\partial\phi/\partial x_j\right) n_i \tag{2}$$

is specified where $\partial\Omega = \partial\Omega_1 \cup \partial\Omega_2$ and $\mathbf{n} = (n_1, n_2)$ denotes the outward pointing normal to $\partial\Omega$.

For all points in Ω the matrix of coefficients $[\lambda_{ij}]$ is a real symmetric positive definite matrix so that throughout Ω equation (1) is a second order elliptic partial differential equation. Further, the coefficients λ_{ij} are required to be twice differentiable functions of the two independent variables x_1 and x_2.

The method of solution will be to obtain boundary integral equations from which numerical values of the dependent variables ϕ and P may be obtained for all points in Ω. The analysis here is specially relevant to an anisotropic medium but it equally applies to isotropic media. For isotropy, the coefficients in (1) take the form $\lambda_{11} = \lambda_{22}$ and $\lambda_{12} = 0$ and use of these equations in the following analysis immediately yields the corresponding results for an isotropic material.

3 Reduction to a constant coefficient equation

The coefficients λ_{ij} are required to take the form

$$\lambda_{ij}(\mathbf{x}) = \lambda_{ij}^{(0)}\, g(\mathbf{x}) \tag{3}$$

where the $\lambda_{ij}^{(0)}$ are constants and g is a differentiable function of $\mathbf{x}$. Use of (3) in (1) yields

$$\lambda_{ij}^{(0)} \frac{\partial}{\partial x_i}\left(g \frac{\partial\phi}{\partial x_j}\right) = 0. \tag{4}$$

Let

$$\psi(\mathbf{x}) = g^{1/2}(\mathbf{x})\, \phi(\mathbf{x}) \tag{5}$$

so that (4) may be written in the form

$$\lambda_{ij}^{(0)} \frac{\partial}{\partial x_i} \left[g \, \frac{\partial \left(g^{-1/2} \psi \right)}{\partial x_j} \right] = 0. \tag{6}$$

That is

$$\lambda_{ij}^{(0)} \left[\left(\frac{1}{4} g^{-3/2} \frac{\partial g}{\partial x_i} \frac{\partial g}{\partial x_j} - \frac{1}{2} g^{-1/2} \frac{\partial^2 g}{\partial x_i \partial x_j} \right) \psi + g^{1/2} \frac{\partial^2 \psi}{\partial x_i \partial x_j} \right] = 0. \tag{7}$$

Use of the identity

$$\frac{\partial^2 g^{1/2}}{\partial x_i \partial x_j} = -\frac{1}{4} g^{-3/2} \frac{\partial g}{\partial x_i} \frac{\partial g}{\partial x_j} + \frac{1}{2} g^{-1/2} \frac{\partial^2 g}{\partial x_i \partial x_j}$$

permits (7) to be written in the form

$$g^{1/2} \lambda_{ij}^{(0)} \frac{\partial^2 \psi}{\partial x_i \partial x_j} - \psi \lambda_{ij}^{(0)} \frac{\partial^2 g^{1/2}}{\partial x_i \partial x_j} = 0. \tag{8}$$

It follows that if g is such that

$$\lambda_{ij}^{(0)} \frac{\partial^2 g^{1/2}}{\partial x_i \partial x_j} + k g^{1/2} = 0, \tag{9}$$

then the transformation (5) carries the variable coefficients equation (4) to the constant coefficients equation

$$\lambda_{ij}^{(0)} \frac{\partial^2 \psi}{\partial x_i \partial x_j} + k \psi = 0. \tag{10}$$

where k is a constant.

Also, substitution of (3) and (5) into (2) gives

$$P = -P^{[g]} \psi + P^{[\psi]} g^{1/2} \tag{11}$$

where

$$P^{[g]}(\mathbf{x}) = \lambda_{ij}^{(0)} \frac{\partial g^{1/2}}{\partial x_j} n_i, \tag{12}$$

$$P^{[\psi]}(\mathbf{x}) = \lambda_{ij}^{(0)} \frac{\partial \psi}{\partial x_j} n_i. \tag{13}$$

A boundary integral equation for the solution of (10) is given in Clements [6] in the form

$$\eta(\mathbf{x}_0) \, \psi(\mathbf{x}_0) = \int_{\partial \Omega} \left[\Gamma(\mathbf{x}, \mathbf{x}_0) \, \psi(\mathbf{x}) - \Phi(\mathbf{x}, \mathbf{x}_0) \, P^{[\psi]}(\mathbf{x}) \right] ds(\mathbf{x}) \tag{14}$$

where $\mathbf{x}_0 = (a, b)$, $\eta = 0$ if $(a, b) \notin \Omega \cup \partial\Omega$, $\eta = 1$ if $(a, b) \in \Omega$, $\eta = \frac{1}{2}$ if $(a, b) \in \partial\Omega$ and $\partial\Omega$ has a continuously turning tangent at (a, b).

The so called fundamental solution Φ in (14) is any solution of the equation

$$\lambda_{ij}^{(0)} \frac{\partial^2 \Phi}{\partial x_i \partial x_j} + k\Phi = \delta(\mathbf{x} - \mathbf{x}_0)$$

and the Γ is given by

$$\Gamma(\mathbf{x}, \mathbf{x}_0) = \lambda_{ij}^{(0)} \frac{\partial \Phi(\mathbf{x}, \mathbf{x}_0)}{\partial x_j} n_i,$$

where δ is the Dirac delta function. For two-dimensional problems Φ and Γ are given by

$$\begin{aligned}
\Phi(\mathbf{x}, \mathbf{x}_0) &= \begin{cases} \frac{K}{2\pi} \ln R, & \text{if } k = 0, \\ \frac{\imath K}{4} H_0^{(2)}(\omega R), & \text{if } k > 0, \\ \frac{-K}{2\pi} K_0(\omega R), & \text{if } k < 0, \end{cases} \\
\Gamma(\mathbf{x}, \mathbf{x}_0) &= \begin{cases} \frac{K}{2\pi} \frac{1}{R} \lambda_{ij}^{(0)} \frac{\partial R}{\partial x_j} n_i, & \text{if } k = 0, \\ \frac{-\imath K \omega}{4} H_1^{(2)}(\omega R) \lambda_{ij}^{(0)} \frac{\partial R}{\partial x_j} n_i, & \text{if } k > 0, \\ \frac{K\omega}{2\pi} K_1(\omega R) \lambda_{ij}^{(0)} \frac{\partial R}{\partial x_j} n_i, & \text{if } k < 0, \end{cases}
\end{aligned} \tag{15}$$

where

$$\begin{aligned}
K &= \ddot{\tau}/\zeta \\
\omega &= \sqrt{|k|/\zeta} \\
\zeta &= [\lambda_{11}^{(0)} + \lambda_{12}^{(0)}(\tau + \overline{\tau}) + \lambda_{22}^{(0)} \tau \overline{\tau}]/2 \\
R &= \sqrt{(\dot{x}_1 - \dot{a})^2 + (\dot{x}_2 - \dot{b})^2} \\
\dot{x}_1 &= x_1 + \dot{\tau} x_2 \\
\dot{a} &= a + \dot{\tau} b \\
\dot{x}_2 &= \ddot{\tau} x_2 \\
\dot{b} &= \ddot{\tau} b
\end{aligned}$$

where $\dot{\tau}$ and $\ddot{\tau}$ are respectively the real and the positive imaginary parts of the complex root τ of the quadratic

$$\lambda_{11}^{(0)} + 2\lambda_{12}^{(0)} \tau + \lambda_{22}^{(0)} \tau^2 = 0,$$

and $H_0^{(2)}$, $H_1^{(2)}$ denote the Hankel function of second kind and order zero and order one respectively. K_0, K_1 denote the modified Bessel function of order zero and order one respectively, $\imath$ represents the square root of minus one and the bar denotes the complex conjugate.

The derivatives $\partial R/\partial x_j$ needed for the calculation of the Γ in (15) are given by

$$\begin{aligned}\frac{\partial R}{\partial x_1} &= \frac{1}{R}(\dot{x}_1 - \dot{a}),\\ \frac{\partial R}{\partial x_2} &= \dot{\tau}\left[\frac{1}{R}(\dot{x}_1 - \dot{a})\right] + \ddot{\tau}\left[\frac{1}{R}(\dot{x}_2 - \dot{b})\right].\end{aligned}$$

Use of (5) and (11) in (14) yields

$$\eta(\mathbf{x}_0)\, g^{1/2}(\mathbf{x}_0)\, \phi(\mathbf{x}_0) = \int_{\partial\Omega} \Big\{ \Big[g^{1/2}(\mathbf{x})\, \Gamma(\mathbf{x}, \mathbf{x}_0) - P^{[g]}(\mathbf{x})\, \Phi(\mathbf{x}, \mathbf{x}_0) \Big]\, \phi(\mathbf{x}) - \Big[g^{-1/2}(\mathbf{x})\, \Phi(\mathbf{x}, \mathbf{x}_0) \Big]\, P(\mathbf{x}) \Big\}\, ds(\mathbf{x}). \tag{16}$$

This equation provides a boundary integral equation for determining ϕ and P at all points of Ω.

The analysis of the section requires that the coefficients λ_{ij} are of the form (3) with g satisfying (9). This condition on g allows for considerable choice in the coefficients. For example, when $k = 0$, g can assume a number of multiparameter forms with the parameters being employed to fit λ_{ij} to numerical data for the coefficients. Possible multiparameter forms include

$$g(\mathbf{x}) = (\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2)^2, \tag{17}$$

$$g(\mathbf{x}) = \left[\Re\{\alpha_0 + \alpha_1 z + \alpha_2 z^2 + \ldots + \alpha_n z^n\}\right]^2, \tag{18}$$

where the α_i are constants, $\Re$ denotes the real part of a complex number and $z = x_1 + \tau x_2$. More generally, the square of the real part of any analytical function of the complex variable z can serve as a possible form for g. For the case when $k \neq 0$ some possible multiparameter forms of g are

$$g(\mathbf{x}) = [A\cos(\alpha_m x_m)]^2 \quad \text{with} \quad \lambda_{ij}^{(0)} \alpha_i \alpha_j = k, \tag{19}$$

$$g(\mathbf{x}) = [A\exp(\alpha_m x_m)]^2 \quad \text{with} \quad \lambda_{ij}^{(0)} \alpha_i \alpha_j = -k, \tag{20}$$

where A, α_m are real constants.

4 A perturbation method

The boundary element procedure described in the previous section provides an effective numerical method for determining ϕ when g satisfies (9). In this section a procedure is obtained for the case when λ_{ij} is perturbed about $\lambda_{ij}^{(0)} g$.

The coefficient λ_{ij} is taken in the form

$$\lambda_{ij}(\mathbf{x}) = \lambda_{ij}^{(0)} g(\mathbf{x}) + \epsilon\, \lambda_{ij}^{(1)}(\mathbf{x}), \tag{21}$$

where ϵ is a small parameter and $\lambda_{ij}^{(1)}$ is a differentiable function of $\mathbf{x}$. Substitution of (21) into (1) and use of (5) gives

$$\lambda_{ij}^{(0)} \frac{\partial}{\partial x_i}\left[g\,\frac{\partial}{\partial x_j}\left(g^{-1/2}\psi\right)\right] = -\epsilon\,\frac{\partial}{\partial x_i}\left(\lambda_{ij}^{(1)}\frac{\partial \phi}{\partial x_j}\right). \tag{22}$$

Use of the analysis used to derive (10) from (6) for the left hand side and a simplification on the right hand side of (22) yields

$$\lambda_{ij}^{(0)}\frac{\partial^2\psi}{\partial x_i \partial x_j} + k\psi = -\epsilon g^{-1/2}\left(\lambda_{ij}^{(1)}\frac{\partial^2\phi}{\partial x_i\partial x_j} + \frac{\partial \lambda_{ij}^{(1)}}{\partial x_i}\frac{\partial\phi}{\partial x_j}\right) \tag{23}$$

where g satisfies (9).

A solution to equation (23) is sought in the form

$$\psi(\mathbf{x}) = \sum_{r=0}^{\infty} \epsilon^r\ \psi^{(r)}(\mathbf{x}). \tag{24}$$

From (5) and (24) we may write ϕ in a series form

$$\phi(\mathbf{x}) = \sum_{r=0}^{\infty} \epsilon^r\ \phi^{(r)}(\mathbf{x}), \tag{25}$$

where $\phi^{(r)}$ corresponds to $\psi^{(r)}$ according to

$$\psi^{(r)} = g^{1/2}\ \phi^{(r)} \qquad \text{for } r = 0, 1, \ldots.$$

Substitution of (24) into (23), equating coefficients of powers of ϵ, yields

$$\lambda_{ij}^{(0)}\frac{\partial^2\psi^{(r)}}{\partial x_i\partial x_j} + k\psi^{(r)} = h^{(r)} \qquad \text{for } r = 0, 1, \ldots, \tag{26}$$

where

$$\begin{aligned} h^{(0)}(\mathbf{x}) &= 0, \\ h^{(r)}(\mathbf{x}) &= -g^{-1/2}\left(\lambda_{ij}^{(1)}\frac{\partial^2\phi^{(r-1)}}{\partial x_i\partial x_j} + \frac{\partial\lambda_{ij}^{(1)}}{\partial x_i}\frac{\partial\phi^{(r-1)}}{\partial x_j}\right) \qquad \text{for } r = 1, 2, \ldots. \end{aligned} \tag{27}$$

The integral equation for (26) is

$$\begin{aligned} \eta(\mathbf{x}_0)\ \psi^{(r)}(\mathbf{x}_0) &= \int_{\partial\Omega}\left[\Gamma(\mathbf{x},\mathbf{x}_0)\ \psi^{(r)}(\mathbf{x}) - \Phi(\mathbf{x},\mathbf{x}_0)\ P^{[\psi^{(r)}]}(\mathbf{x})\right] ds(\mathbf{x}) \ + \\ &\quad \int_{\Omega} h^{(r)}(\mathbf{x})\ \Phi(\mathbf{x},\mathbf{x}_0)\ dS(\mathbf{x}) \qquad \text{for } r = 0, 1, \ldots, \end{aligned} \tag{28}$$

where

$$P^{[\psi^{(r)}]} = \lambda_{ij}^{(0)}(\partial\psi^{(r)}/\partial x_j)n_i.$$

We also have

$$P^{[\psi^{(r)}]} = g^{1/2} P^{(r)} + \phi^{(r)} P^{[g]} \qquad \text{for } r = 0, 1, \ldots,$$

where

$$P^{(r)}(\mathbf{x}) = \lambda_{ij}^{(0)} \, \frac{\partial \phi^{(r)}}{\partial x_j} \, n_i$$

and $P^{[g]}$ is given in (12).

Thus the integral equation (28) may be written in the form

$$\begin{aligned} \eta(\mathbf{x}_0)\, g^{1/2}(\mathbf{x}_0)\, \phi^{(r)}(\mathbf{x}_0) \;=\; & \int_{\partial\Omega} \Big\{ \Big[g^{1/2}(\mathbf{x})\, \Gamma(\mathbf{x}, \mathbf{x}_0) - P^{[g]}(\mathbf{x})\, \Phi(\mathbf{x}, \mathbf{x}_0) \Big] \phi^{(r)}(\mathbf{x}) - \\ & \Big[g^{1/2}(\mathbf{x})\, \Phi(\mathbf{x}, \mathbf{x}_0) \Big] P^{(r)}(\mathbf{x}) \Big\} \, ds(\mathbf{x}) \; + \\ & \int_{\Omega} h^{(r)}(\mathbf{x})\, \Phi(\mathbf{x}, \mathbf{x}_0) \, dS(\mathbf{x}). \end{aligned} \tag{29}$$

Now, the corresponding value of P may be written as

$$P = g P^{(0)} + \sum_{r=1}^{\infty} \epsilon^r \left[g P^{(r)} + G^{(r)} \right], \tag{30}$$

where

$$G^{(r)}(\mathbf{x}) = \lambda_{ij}^{(1)} \frac{\partial \phi^{(r-1)}}{\partial x_i} n_j. \tag{31}$$

To satisfy the boundary conditions in Section 2 it is required that

$$\begin{aligned} \phi^{(0)} &= \phi \qquad \text{on } \partial\Omega_1, \\ P^{(0)} &= g^{-1} P \qquad \text{on } \partial\Omega_2, \end{aligned}$$

where ϕ takes on its specified value on $\partial\Omega_1$ and P takes on its specified value on $\partial\Omega_2$. It then follows from (25) and (30) that for $r = 1, 2, \ldots$

$$\begin{aligned} \phi^{(r)} &= 0 \qquad \text{on } \partial\Omega_1, \\ P^{(r)} &= -g^{-1} G^{(r)} \qquad \text{on } \partial\Omega_2. \end{aligned}$$

The integral equation (29) may now be used to find the numerical values of the unknowns on the boundary $\partial\Omega$ and the numerical values of $\phi^{(r)}$ and derivatives in the domain Ω for $r = 0, 1, \ldots$. At each stage in using (29) to determine $\phi^{(r)}$, the $G^{(r)}$ occuring in the boundary condition $P^{(r)} = -g^{-1} G^{(r)}$ on $\partial\Omega_2$ may be obtained from (31) which is evaluated from the previous iteration. Equations (25) and (30) then provide the values of ϕ in the domain Ω and P on the boundary $\partial\Omega$.

5 Further reduction to constant coefficients

The analysis of this chapter has so far been concerned with coefficients which fall within the general class given by (21). The following analysis seeks to consider the case when the coefficients $\lambda_{ij}(\mathbf{x})$ are not all proportional to the same function of $\mathbf{x}$ and thus fall outside the general class given by (21).
Attention will be restricted to anisotropic materials for which the material has symmetry properties which lead to the coefficient λ_{12} being zero. In this case equation (1) becomes

$$\frac{\partial}{\partial x_1}\left[\lambda_{11}(x_1,x_2)\frac{\partial\phi}{\partial x_1}\right]+\frac{\partial}{\partial x_2}\left[\lambda_{22}(x_1,x_2)\frac{\partial\phi}{\partial x_2}\right]=0. \tag{32}$$

It will be further assumed that the coefficients λ_{11} and λ_{22} have variable separable forms so that they can be written in the form

$$\lambda_{11}(x_1,x_2) = X_1(x_1)Y_1(x_2), \tag{33}$$
$$\lambda_{22}(x_1,x_2) = X_2(x_1)Y_2(x_2). \tag{34}$$

Substitution of (33) and (34) into (32) yields

$$\frac{1}{X_2(x_1)}\frac{\partial}{\partial x_1}\left[X_1(x_1)\frac{\partial\phi}{\partial x_1}\right]+\frac{1}{Y_1(x_2)}\frac{\partial}{\partial x_2}\left[Y_2(x_2)\frac{\partial\phi}{\partial x_2}\right]=0.$$

The new independent variables

$$\xi_1=\int\left(\frac{X_2(x_1)}{X_1(x_1)}\right)^{1/2}dx_1,\quad \xi_2=\int\left(\frac{Y_1(x_2)}{Y_2(x_2)}\right)^{1/2}dx_2 \tag{35}$$

now provide

$$\frac{1}{M(\xi_1)}\frac{\partial}{\partial\xi_1}\left[M(\xi_1)\frac{\partial\phi}{\partial\xi_1}\right]+\frac{1}{N(\xi_2)}\frac{\partial}{\partial\xi_2}\left[N(\xi_2)\frac{\partial\phi}{\partial\xi_2}\right]=0, \tag{36}$$

where

$$M=(X_2X_1)^{1/2},\quad N=(Y_2Y_1)^{1/2}. \tag{37}$$

A new dependent variable ψ is now introduced according to

$$\phi=M^{-1/2}N^{-1/2}\psi. \tag{38}$$

Use of (38) in (36) yields

$$\nabla^2\psi-\left[\left(M^{-1/2}N^{-1/2}\right)\nabla^2\left(M^{1/2}N^{1/2}\right)\right]\psi=0, \tag{39}$$

where

$$\nabla^2\equiv\partial^2/\partial{\xi_1}^2+\partial^2/\partial{\xi_2}^2.$$

If

$$\nabla^2\left(M^{1/2}N^{1/2}\right)=0, \tag{40}$$

then from (39) ψ satisfies Laplace's equation

$$\nabla^2\psi = 0.$$

Since M is a function of ξ_1 only and N is a function of ξ_2 only (40) then M and N must adopt the forms

$$M(\xi_1) = (\alpha\xi_1 + \beta)^2, \quad N(\xi_2) = (\gamma\xi_2 + \delta)^2$$

where α, β, γ and δ are constants.
The boundary integral equations for ψ in the new frame $O\xi_1\xi_2$ may be obtained from (11)–(14) with $\lambda_{11}^{(0)} = \lambda_{22}^{(0)} = 1, \lambda_{12}^{(0)} = 0$. Specifically the equation for ψ is

$$\eta(\boldsymbol{\xi}_0)\,\psi(\boldsymbol{\xi}_0) = \int_{\partial\Omega_\xi} \Big[\Gamma(\boldsymbol{\xi},\boldsymbol{\xi}_0)\,\psi(\boldsymbol{\xi}) - \Phi(\boldsymbol{\xi},\boldsymbol{\xi}_0)\,P^{[\psi]}(\boldsymbol{\xi})\Big]\,ds(\boldsymbol{\xi})$$

where Φ and Γ are given by (15) with $k = 0$ and the region Ω_ξ with boundary $\partial\Omega_\xi$ denoting the domain and boundary under consideration referred to the $O\xi_1\xi_2$ frame.
The boundary integral equation for ϕ is given by

$$\begin{aligned}\eta(\boldsymbol{\xi}_0)\,g^{1/2}(\boldsymbol{\xi}_0)\,\phi(\boldsymbol{\xi}_0) \;=\; & \int_{\partial\Omega_\xi} \Big\{\Big[g^{1/2}(\boldsymbol{\xi})\,\Gamma(\boldsymbol{\xi},\boldsymbol{\xi}_0) - P^{[g]}(\boldsymbol{\xi})\,\Phi(\boldsymbol{\xi},\boldsymbol{\xi}_0)\Big]\,\phi(\boldsymbol{\xi}) - \\ & \Big[g^{-1/2}(\boldsymbol{\xi})\,\Phi(\boldsymbol{\xi},\boldsymbol{\xi}_0)\Big]\,P(\boldsymbol{\xi})\Big\}\,ds(\boldsymbol{\xi}) \qquad (41)\end{aligned}$$

where

$$g(\boldsymbol{\xi}) = M(\xi_1)N(\xi_2).$$

6 Numerical results

In this section some particular boundary value problems are solved numerically by employing the integral equations obtained in Sections 3, 4 and 5. In implementing this method to obtain numerical results standard boundary element procedure is employed (see for example Clements [6]).

Problem 2.1 Consider the boundary value problem governed by (1) for an inhomogeneous material of geometry as given in Figure 1 with coefficients

$$[\lambda'_{ij}] = \begin{bmatrix} (1+2\alpha'_2x'_2)^2 + 0.1\beta\sin(\pi x'_2) & 0 \\ 0 & 2(1+2\alpha'_2x'_2)^2 + 0.2\beta\sin(\pi x'_2) \end{bmatrix} \qquad (42)$$

where $\lambda'_{ij} = \lambda_{ij}/\widehat{\lambda}$, $\widehat{\lambda}$ is a reference coefficient, $x'_i = x_i/l$ $(i = 1, 2)$, l is a reference length, $\alpha'_2 = \alpha_2 l$, α_2 is given in (17) and β is a dimensionless constant. The boundary conditions are (see Figure 1)

$$\begin{array}{ll} \phi' = 0, & \text{on AB,} \\ P' = 0, & \text{on BC and AD,} \\ P' = 1, & \text{on CD} \end{array}$$

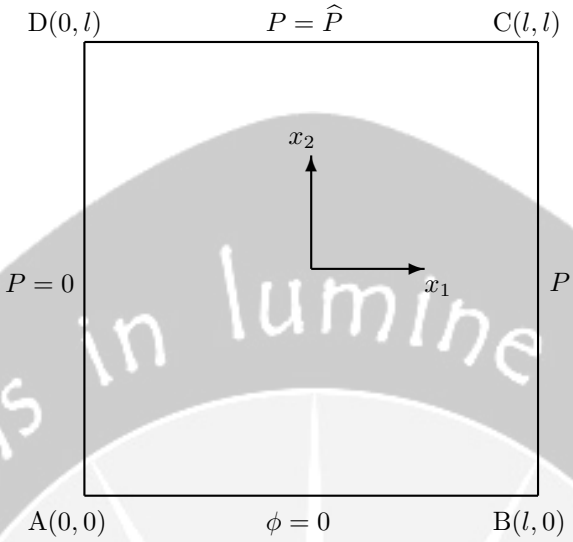


Figure 1: The geometry for Problem 2.1

where $\phi' = \phi/\widehat{\phi}$, $P' = P/\widehat{P}$, $\widehat{\phi}$ is a reference value of ϕ and $\widehat{P} = \widehat{\lambda}\,\widehat{\phi}/l$ is a reference value of P.

The coefficients (42) may be written in the form (21) with

$$\begin{aligned} \epsilon &= 0.1, \\ g &= (1 + 2\alpha_2' x_2')^2, \\ \left[\lambda_{ij}^{\prime(0)}\right] &= \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \\ \left[\lambda_{ij}^{\prime(1)}\right] &= \begin{bmatrix} \beta \sin(\pi x_2') & 0 \\ 0 & 2\beta \sin(\pi x_2') \end{bmatrix} \end{aligned} \tag{43}$$

where $\lambda_{ij}^{\prime(m)} = \lambda_{ij}^{(m)}/\widehat{\lambda}$ for $m = 0, 1$.

When $\beta = 0$ the exact solution to the problem is

$$\phi' = x_2'/[2(1 + 2\alpha_2' x_2')]$$

with $\sigma_i = \lambda_{ij}(\partial\phi/\partial x_j)$ given by

$$\sigma_1' = 0 \quad \text{and} \quad \sigma_2' = 1$$

where $\sigma_i' = \sigma_i/\widehat{P}$.

Table 1 and Table 2 show the numerical and analytical results for the solutions ϕ' and P' when $\alpha_2' = 1$, $\beta = 0, 1, 2, 3$. For $\beta = 0$ it is observed from Table 1

and Table 2 that the numerical solutions for ϕ' and P' converge to the analytical solutions as the number of segments increases. For $\beta \neq 0$ an analytical solution to the problem does not exist (DNE) and it is apparent from Table 1 and Table 2 that the numerical solutions also converge as the number of segments increases. Furthermore, it should be noted that the values of ϕ' in Table 1 decrease as the parameter β increases. This is to be expected as the value of P' is kept constant then an increase in the value of the coefficients λ'_{ij} resulted from an increase in the values of β will cause a decrease in the value of ϕ. However, the values of P' in Table 2 increase as the value of β increases. This is also to be expected as by the definition $P' = \lambda'_{ij}(\partial\phi'/\partial x'_i)n_j$, so that if the coefficients λ'_{ij} increase then the value of P' will also increases.

For the case $\beta \neq 0$ the numerical solutions for ϕ' and P' are taken in the series forms (25) and (30) respectively. The series solutions are approximated by taking the first two terms of each series, that is $\phi' \approx \phi'^{(0)} + \epsilon\phi'^{(1)}$ and $P' \approx gP'^{(0)} + \epsilon(gP'^{(1)} + G'^{(1)})$ where $G'^{(1)} = G^{(1)}/\widehat{P}$.

The domain integral in (29) is evaluated by dividing the domain into 15×15 equal square cells and the integrand is assumed to be constant, taking on its value at the mid point of each cell. The function $h^{(1)}$ in (27) for this particular problem is given by

$$h^{(1)}(\mathbf{x}) = -g^{-1/2}\left(-2g^{-1/2}\lambda_{ij}^{(1)}\frac{\partial g^{1/2}}{\partial x_i} + \frac{\partial \lambda_{ij}^{(1)}}{\partial x_i}\right)\frac{\partial \phi^{(0)}}{\partial x_j}$$

which is derived from the equation

$$\lambda_{ij}^{(0)}\frac{\partial^2 \psi^{(0)}}{\partial x_i \partial x_j} = 0.$$

Thus we only need the value of the derivative $\frac{\partial \phi^{(0)}}{\partial x_j}$ of the first iteration solution to find the value of the function $h^{(1)}(\mathbf{x})$.

Problem 2.2 Consider a problem governed by an equation of the type (32) for a medium of geometry as shown in Figure 2 with the coefficients of the forms (33) and (34). Specifically the problem is governed by

$$\frac{\partial}{\partial x_1}\left[\lambda_{11}(x_1, x_2)\frac{\partial \phi}{\partial x_1}\right] + \frac{\partial}{\partial x_2}\left[\lambda_{22}(x_1, x_2)\frac{\partial \phi}{\partial x_2}\right] = 0$$

with coefficients

$$\begin{aligned}\lambda'_{11}(x_1, x_2) &= (1 + 2x'_1)^{3/2}(2 + 3x'_2)^2 \\ \lambda'_{22}(x_1, x_2) &= (1 + 2x'_1)^{1/2}(2 + 3x'_2)^2.\end{aligned}$$

The boundary conditions are (see Figure 2)

$$\begin{aligned}\phi' &= 1/2 && \text{on AB,} \\ \phi' &= 1/(2 + 3x'_2) && \text{on BC,} \\ \phi' &= 1/5 && \text{on CD,} \\ \phi' &= 1/(2 + 3x'_2) && \text{on AD.}\end{aligned}$$

Table 1: Solution ϕ' to Problem 2.1, when $\alpha'_2 = 1$

Position (x'_1, x'_2)	ϕ' BEM 80 segments	ϕ' BEM 160 segments	ϕ' BEM 320 segments	ϕ' Analytical
		$\beta = 0$		
(1.0,0.1)	0.0390	0.0404	0.0411	0.0416
(1.0,0.3)	0.0909	0.0924	0.0931	0.0937
(1.0,0.5)	0.1219	0.1235	0.1243	0.1250
(1.0,0.7)	0.1425	0.1442	0.1450	0.1458
(1.0,0.9)	0.1573	0.1590	0.1599	0.1607
		$\beta = 1$		
(1.0,0.1)	0.0385	0.0399	0.0406	DNE
(1.0,0.3)	0.0890	0.0904	0.0912	DNE
(1.0,0.5)	0.1191	0.1207	0.1215	DNE
(1.0,0.7)	0.1393	0.1410	0.1418	DNE
(1.0,0.9)	0.1539	0.1557	0.1565	DNE
		$\beta = 2$		
(1.0,0.1)	0.0380	0.0396	0.0402	DNE
(1.0,0.3)	0.0871	0.0886	0.0893	DNE
(1.0,0.5)	0.1163	0.1179	0.1186	DNE
(1.0,0.7)	0.1361	0.1378	0.1386	DNE
(1.0,0.9)	0.1506	0.1523	0.1531	DNE
		$\beta = 3$		
(1.0,0.1)	0.0376	0.0391	0.0397	DNE
(1.0,0.3)	0.0852	0.0866	0.0873	DNE
(1.0,0.5)	0.1135	0.1151	0.1158	DNE
(1.0,0.7)	0.1329	0.1346	0.1353	DNE
(1.0,0.9)	0.1472	0.1490	0.1498	DNE

For this particular problem the coefficients λ_{11} and λ_{22} are of the type (33) and (34) with

$$X'_1(x_1) = (1 + 2x'_1)^{3/2}, \qquad Y'_1(x_2) = (2 + 3x'_2)^2,$$
$$X'_2(x_1) = (1 + 2x'_1)^{1/2}, \qquad Y'_2(x_2) = (2 + 3x'_2)^2,$$

where $X'_1 = X_1/\widehat{X}_1$, $X'_2 = X_2/\widehat{X}_2$, $Y'_1 = Y_1\widehat{X}_1/\widehat{\lambda}$, $Y'_2 = Y_2\widehat{X}_2/\widehat{\lambda}$, $\widehat{X}_1, \widehat{X}_2$ are reference values of X_1 and X_2 respectively. Thus from (35) and (37)

$$M' = 1 + 2x'_1 = {\xi'_1}^2, \qquad N' = (2 + 3x'_2)^2 = (2 + 3\xi'_2)^2,$$

where, $M' = M/\widehat{M}$, $\widehat{M} = (\widehat{X}_1\widehat{X}_2)^{1/2}$, $N' = N/\widehat{N}$, $\widehat{N} = \widehat{\lambda}/(\widehat{X}_1\widehat{X}_2)$, $\xi'_i = \xi_i/\widehat{\xi}$ (for $i = 1, 2$), $\widehat{\xi} = (\widehat{X}_1\widehat{X}_2)^{1/2}l$. The M' and N' satisfy equation (40). It follows from (38) that the transformation

$$\begin{aligned} \phi' &= M'^{-1/2}N'^{-1/2}\psi' \\ &= {\xi'_1}^{-1}(2 + 3\xi'_2)^{-1}\psi' \end{aligned}$$

Table 2: Solution P' to Problem 2.1 when $\alpha'_2 = 1$

Position (x'_1, x'_2)	P' BEM 80 segments	P' BEM 160 segments	P' BEM 320 segments	P' Analytical
		$\beta = 0$		
(0.1,0.0)	-0.9779	-0.9912	-0.9958	-1.0
(0.3,0.0)	-0.9829	-0.9918	-0.9960	-1.0
(0.5,0.0)	-0.9830	-0.9915	-0.9960	-1.0
(0.7,0.0)	-0.9829	-0.9918	-0.9960	-1.0
(0.9,0.0)	-0.9779	-0.9912	-0.9958	-1.0
		$\beta = 1$		
(0.1,0.0)	-0.9763	-0.9922	-0.9986	DNE
(0.3,0.0)	-0.9814	-0.9928	-0.9988	DNE
(0.5,0.0)	-0.9815	-0.9925	-0.9988	DNE
(0.7,0.0)	-0.9814	-0.9928	-0.9988	DNE
(0.9,0.0)	-0.9763	-0.9922	-0.9986	DNE
		$\beta = 2$		
(0.1,0.0)	-0.9749	-0.9932	-1.0014	DNE
(0.3,0.0)	-0.9800	-0.9938	-1.0016	DNE
(0.5,0.0)	-0.9801	-0.9936	-1.0016	DNE
(0.7,0.0)	-0.9800	-0.9938	-1.0016	DNE
(0.9,0.0)	-0.9749	-0.9932	-1.0014	DNE
		$\beta = 3$		
(0.1,0.0)	-0.9734	-0.9942	-1.0042	DNE
(0.3,0.0)	-0.9785	-0.9949	-1.0044	DNE
(0.5,0.0)	-0.9786	-0.9946	-1.0044	DNE
(0.7,0.0)	-0.9785	-0.9949	-1.0044	DNE
(0.9,0.0)	-0.9734	-0.9942	-1.0042	DNE

where $\psi' = \psi/(\widehat{\phi\lambda}^{1/2})$, transforms the original partial differential equation to

$$\nabla^2 \psi' = 0$$

where $\nabla^2 \equiv \partial^2/\partial {\xi'_1}^2 + \partial^2/\partial {\xi'_2}^2$.

Table 3 shows a comparison between BEM results obtained using equation (41) and analytical results for some interior points. The analytical solution to this problem is $\phi' = 1/(2 + 3x'_2)$.

Problem 2.3 Consider the known analytical solution to (1)

$$\phi' = \frac{\sin \beta' x'_2}{\cos(\alpha'_1 x'_1 + \alpha'_2 x'_2)} \tag{44}$$

where $\alpha'_i = \alpha_i l$ $(i = 1, 2)$, α_i is given in (19) and β' is a dimensionless constant, to a problem for an inhomogeneous material occupying the region of a unit square

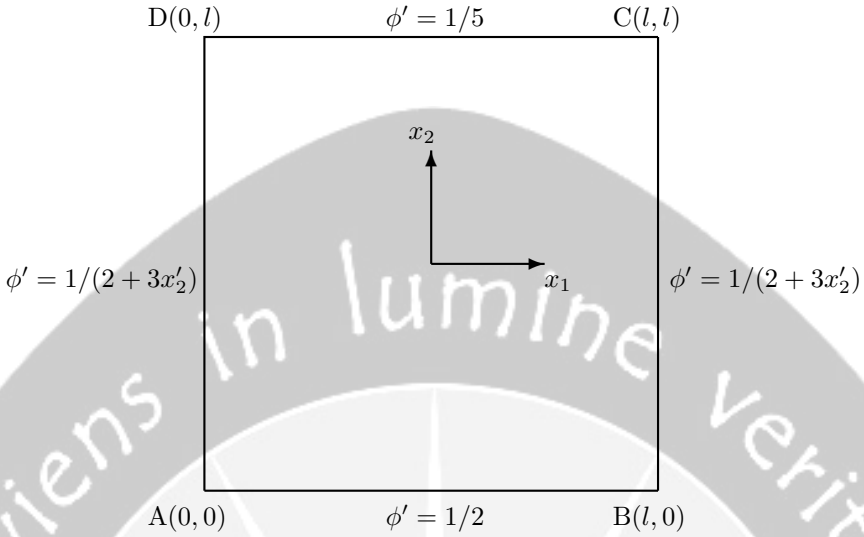


Figure 2: The boundary conditions for Problem 2.2

depicted in Figure 3. The coefficients of the material vary with position in the square according to

$$[\lambda'_{ij}] = \begin{bmatrix} \cos^2(\alpha'_1 x'_1 + \alpha'_2 x'_2) & 0.5\cos^2(\alpha'_1 x'_1 + \alpha'_2 x'_2) \\ 0.5\cos^2(\alpha'_1 x'_1 + \alpha'_2 x'_2) & 0.5\cos^2(\alpha'_1 x'_1 + \alpha'_2 x'_2) \end{bmatrix}. \tag{45}$$

The boundary conditions are (see Figure 3)

P', as may be obtained from (44), on AB, BC and CD,
ϕ', as given by (44), on AD.

The coefficients (45) may be written in the form (3) with

$$\begin{aligned} \left[\lambda_{ij}^{'(0)}\right] &= \begin{bmatrix} 1 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}, \\ g(\mathbf{x}) &= \cos^2(\alpha'_1 x'_1 + \alpha'_2 x'_2). \end{aligned}$$

Thus, $g(\mathbf{x})$ takes the form (19) with $A = 1$. The parameters k' ($k' = kl^2/\widehat{\lambda}$) and α'_1 are chosen to be 0.5 and the parameter α'_2 is required to satisfy the condition in (19) ($\lambda_{ij}^{'(0)}\alpha'_i\alpha'_j = k'$). Specifically, α'_2 satisfies

$$\alpha'_2 = \frac{1}{\lambda_{22}^{'(0)}}\left[-\lambda_{12}^{'(0)}\alpha'_1 + \sqrt{\lambda_{12}^{'(0)^2}\alpha_1'^2 - \lambda_{22}^{'(0)}\left(\lambda_{11}^{'(0)}\alpha_1'^2 - k'\right)}\right].$$

The parameter β' is defined as $\beta' = \sqrt{k'/\lambda_{22}^{'(0)}}$.

Table 3: Numerical and analytical results for Problem 2.2

Position (x_1', x_2')	ϕ' BEM 8 segments	ϕ' BEM 16 segments	ϕ' BEM 32 segments	ϕ' Analytical
(0.9,0.4)	0.285323	0.313721	0.312487	0.312500
(0.2,0.1)	0.431316	0.434713	0.434756	0.434783
(0.6,0.3)	0.344864	0.344839	0.344834	0.344828
(0.8,0.9)	0.207950	0.212652	0.212752	0.212766
(0.4,0.6)	0.263112	0.263144	0.263153	0.263158
(0.9,0.2)	0.396525	0.381003	0.384669	0.384615
(0.5,0.1)	0.433516	0.434958	0.434778	0.434783
(0.3,0.8)	0.226931	0.227248	0.227264	0.227273
(0.9,0.9)	0.185691	0.213604	0.212776	0.212766

Table 4 shows a comparison between the BEM and the analytical solutions. The BEM solutions converge to the analytical solutions as the number of segments increases.

Problem 2.4 Consider the known analytical solution to (1)

$$\phi' = \frac{\exp \beta' x_2'}{\exp(\alpha_1' x_1' + \alpha_2' x_2')} \tag{46}$$

where $\alpha_i' = \alpha_i l$ $(i = 1, 2)$, α_i is given in (20) and β' is a dimensionless constant, for a problem which is associated with an inhomogeneous material as shown in Figure 3 with coefficients

$$\left[\lambda_{ij}'\right] = \begin{bmatrix} \exp[2(\alpha_1' x_1' + \alpha_2' x_2')] & 0.5\exp[2(\alpha_1' x_1' + \alpha_2' x_2')] \\ 0.5\exp[2(\alpha_1' x_1' + \alpha_2' x_2')] & 0.5\exp[2(\alpha_1' x_1' + \alpha_2' x_2')] \end{bmatrix}. \tag{47}$$

The boundary conditions are (see Figure 3)

P', as may be obtained from (46), on AB, BC and CD,
ϕ', as given by (46), on AD.

The coefficients (47) may be written in the form (3) with

$$\begin{aligned} \left[\lambda_{ij}^{'(0)}\right] &= \begin{bmatrix} 1 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}, \\ g(\mathbf{x}) &= \exp[2(\alpha_1' x_1' + \alpha_2' x_2')], \end{aligned}$$

so that $g(\mathbf{x})$ takes the form (20) with $A = 1$. The parameters k' and α_1' are chosen to be -0.5 and the parameter α_2' is required to satisfy the condition in (20) $(\lambda_{ij}^{'(0)} \alpha_i' \alpha_j' = -k')$. Specifically, α_2' satisfies

$$\alpha_2' = \frac{1}{\lambda_{22}^{'(0)}} \left[-\lambda_{12}^{'(0)} \alpha_1' + \sqrt{\lambda_{12}^{'(0)^2} \alpha_1'^{\,2} - \lambda_{22}^{'(0)} \left(\lambda_{11}^{'(0)} \alpha_1'^{\,2} + k' \right)} \right].$$

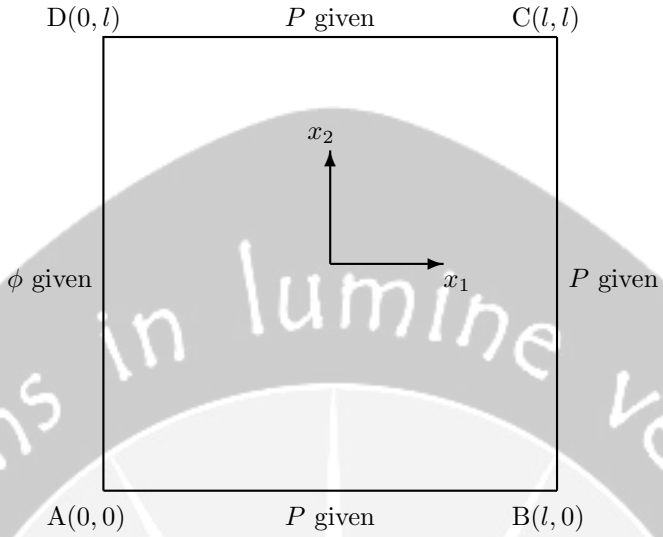


Figure 3: The boundary conditions for Problem 2.3 and Problem 2.4

The parameter β' is defined as $\beta' = \sqrt{-k'/\lambda_{22}^{'(0)}}$.

Again, Table 5 shows that the BEM solutions converge to the analytical solutions as the number of segments increases.

Problem 2.5 Consider the problem governed by (1) with coefficients

$$[\lambda_{ij}] = \begin{bmatrix} \exp(-\alpha x_2) & 0 \\ 0 & \exp(-\alpha x_2) \end{bmatrix}$$

where α is a constant, with boundary conditions (see Figure 4)

$$\begin{array}{l} x_1 = -(L+D), \quad 0 \le x_2 \le \infty, \quad f = 0, \\ -(L+D) \le x_1 \le -L, \quad x_2 = 0, \quad f = 0, \\ -L \le x_1 \le L, \quad x_2 = 0, \quad f = -f_1, \\ L \le x_1 \le L+D, \quad x_2 = 0, \quad f = 0, \\ x_1 = L+D, \quad 0 \le x_2 \le \infty, \quad f = 0, \\ -(L+D) \le x_1 \le L+D, \quad x_2 = \infty, \quad f = \alpha\phi, \end{array}$$

where L and D are constants,

$$\begin{array}{rcl} f & = & un_1 + vn_2 \\ u & = & -\partial\phi/\partial x_1 \\ v & = & \alpha\phi - \partial\phi/\partial x_2 \end{array}$$

and f_1 is a constant value of f.

Table 4: BEM and analytical solutions for Problem 2.3

Position	BEM			Analytical		
(x_1', x_2')	ϕ'	$\partial\phi'/\partial x_1'$	$\partial\phi'/\partial x_2'$	ϕ'	$\partial\phi'/\partial x_1'$	$\partial\phi'/\partial x_2'$
40 segments						
(0.3,0.5)	0.5064	0.0953	0.9886	0.5073	0.0877	0.9928
(0.5,0.5)	0.5290	0.1322	1.0475	0.5282	0.1221	1.0562
(0.7,0.5)	0.5596	0.1752	1.1274	0.5566	0.1642	1.1391
(0.9,0.5)	0.6031	0.3528	1.1864	0.5946	0.2182	1.2482
(0.5,0.1)	0.1107	0.0307	1.0296	0.1041	0.0153	1.0485
(0.5,0.3)	0.3190	0.0725	1.0488	0.3157	0.0594	1.0642
(0.5,0.7)	0.7362	0.2103	1.0199	0.7366	0.2042	1.0240
(0.5,0.9)	0.9350	0.3116	0.9500	0.9361	0.3063	0.9671
80 segments						
(0.3,0.5)	0.5068	0.0913	0.9912	0.5073	0.0877	0.9928
(0.5,0.5)	0.5285	0.1267	1.0526	0.5282	0.1221	1.0562
(0.7,0.5)	0.5580	0.1693	1.1340	0.5566	0.1642	1.1391
(0.9,0.5)	0.5968	0.2119	1.2446	0.5946	0.2182	1.2482
(0.5,0.1)	0.1072	0.0231	1.0389	0.1041	0.0153	1.0485
(0.5,0.3)	0.3171	0.0658	1.0571	0.3157	0.0594	1.0642
(0.5,0.7)	0.7364	0.2073	1.0223	0.7366	0.2042	1.0240
(0.5,0.9)	0.9357	0.3088	0.9659	0.9361	0.3063	0.9671
160 segments						
(0.3,0.5)	0.5071	0.0896	0.9921	0.5073	0.0877	0.9928
(0.5,0.5)	0.5284	0.1244	1.0545	0.5282	0.1221	1.0562
(0.7,0.5)	0.5573	0.1668	1.1365	0.5566	0.1642	1.1391
(0.9,0.5)	0.5959	0.2207	1.2448	0.5946	0.2182	1.2482
(0.5,0.1)	0.1057	0.0195	1.0433	0.1041	0.0153	1.0485
(0.5,0.3)	0.3165	0.0628	1.0606	0.3157	0.0594	1.0642
(0.5,0.7)	0.7366	0.2059	1.0232	0.7366	0.2042	1.0240
(0.5,0.9)	0.9359	0.3077	0.9663	0.9361	0.3063	0.9671

The governing equation (1) for this particular problem may be written as

$$\frac{\partial^2\phi}{\partial x_1^2} + \frac{\partial^2\phi}{\partial x_2^2} - \alpha\frac{\partial\phi}{\partial x_2} = 0. \tag{48}$$

Defining the dimensionless variables

$$\begin{aligned}
&X = \tfrac{\alpha x_1}{2}, \quad Z = \tfrac{\alpha x_2}{2}, \quad \theta = \tfrac{\pi\phi}{f_1 L},\\
&X_0 = \tfrac{\alpha L}{2}, \quad X_1 = \tfrac{\alpha D}{2},\\
&U = \tfrac{2\pi u}{f_1\alpha L} = -\tfrac{\partial\theta}{\partial X}, \quad V = \tfrac{2\pi v}{f_1\alpha L} = 2\theta - \tfrac{\partial\theta}{\partial Z},\\
&F = Un_1 + Vn_2 = \tfrac{\pi f}{f_1 X_0}
\end{aligned}$$

Table 5: BEM and analytical solutions for Problem 2.4

Position	BEM			Analytical		
(x_1', x_2')	ϕ'	$\partial\phi'/\partial x_1'$	$\partial\phi'/\partial x_2'$	ϕ'	$\partial\phi'/\partial x_1'$	$\partial\phi'/\partial x_2'$
40 segments						
(0.3,0.5)	0.9631	0.4751	-0.3487	0.9675	0.4838	-0.3541
(0.5,0.5)	1.0637	0.5305	-0.3921	1.0693	0.5346	-0.3914
(0.7,0.5)	1.1755	0.5886	-0.4356	1.1817	0.5909	-0.4325
(0.9,0.5)	1.3064	0.8583	-0.4897	1.3060	0.6530	-0.4780
(0.5,0.1)	1.2329	0.6192	-0.4397	1.2379	0.6189	-0.4531
(0.5,0.3)	1.1454	0.5717	-0.4240	1.1505	0.5752	-0.4211
(0.5,0.7)	0.9883	0.4909	-0.3646	0.9938	0.4969	-0.3638
(0.5,0.9)	0.9172	0.4580	-0.3572	0.9236	0.4618	-0.3381
80 segments						
(0.3,0.5)	0.9655	0.4807	-0.3550	0.9675	0.4838	-0.3541
(0.5,0.5)	1.0666	0.5321	-0.3914	1.0693	0.5346	-0.3914
(0.7,0.5)	1.1786	0.5889	-0.4330	1.1817	0.5909	-0.4325
(0.9,0.5)	1.3023	0.6306	-0.4774	1.3060	0.6530	-0.4780
(0.5,0.1)	1.2353	0.6183	-0.4536	1.2379	0.6189	-0.4531
(0.5,0.3)	1.1479	0.5732	-0.4214	1.1505	0.5752	-0.4211
(0.5,0.7)	0.9911	0.4949	-0.3648	0.9938	0.4969	-0.3638
(0.5,0.9)	0.9207	0.4590	-0.3391	0.9236	0.4618	-0.3381
160 segments						
(0.3,0.5)	0.9665	0.4822	-0.3542	0.9675	0.4838	-0.3541
(0.5,0.5)	1.0679	0.5333	-0.3911	1.0693	0.5346	-0.3914
(0.7,0.5)	1.1801	0.5897	-0.4324	1.1817	0.5909	-0.4325
(0.9,0.5)	1.3042	0.6520	-0.4782	1.3060	0.6530	-0.4780
(0.5,0.1)	1.2364	0.6191	-0.4537	1.2379	0.6189	-0.4531
(0.5,0.3)	1.1490	0.5740	-0.4204	1.1505	0.5752	-0.4211
(0.5,0.7)	0.9924	0.4958	-0.3641	0.9938	0.4969	-0.3638
(0.5,0.9)	0.9222	0.4604	-0.3387	0.9236	0.4618	-0.3381

the equation (48) may be written as

$$\frac{\partial^2\theta}{\partial X^2} + \frac{\partial^2\theta}{\partial Z^2} = 2\frac{\partial\theta}{\partial Z} \tag{49}$$

and the corresponding boundary conditions (see Figure 5) are

$$\begin{aligned}
&X = -(X_0 + X_1), \quad 0 \le Z \le \infty, \quad F = 0,\\
&-(X_0 + X_1) \le X \le -X_0, \quad Z = 0, \quad F = 0,\\
&-X_0 \le X \le X_0, \quad Z = 0, \quad F = -\pi/X_0,\\
&X_0 \le X \le X_0 + X_1, \quad Z = 0, \quad F = 0,\\
&X = X_0 + X_1, \quad 0 \le Z \le \infty, \quad F = 0,\\
&-(X_0 + X_1) \le X \le X_0 + X_1, \quad Z = \infty, \quad F = 2\theta.
\end{aligned}$$

Using a standard mathematical tool, Batu [2] derived an analytical solution to

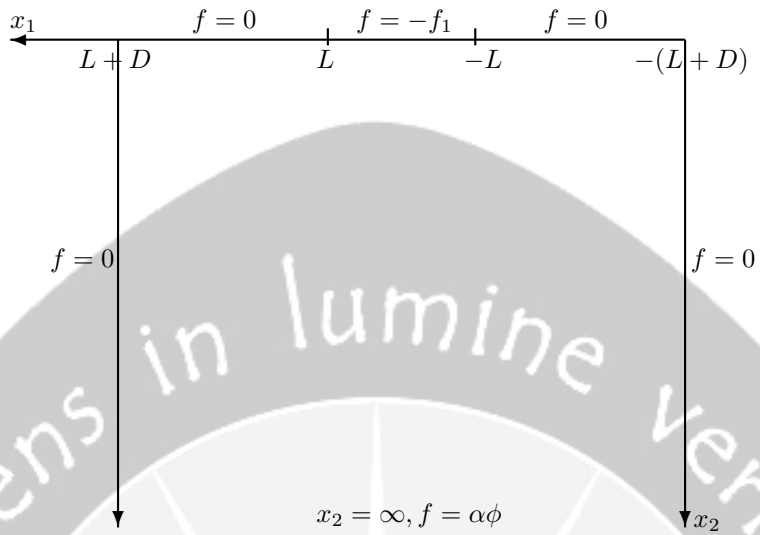


Figure 4: Boundary conditions for Problem 2.5

this problem as follows

$$\begin{aligned}
\theta &= \frac{\pi}{2(X_0+X_1)} + 2\sum_{n=1}^{\infty} \frac{I_n}{n} \sin\left(\frac{n\pi X_0}{X_0+X_1}\right) \cos\left(\frac{n\pi X}{X_0+X_1}\right) \\
U &= \frac{2\pi}{X_0+X_1} \sum_{n=1}^{\infty} I_n \sin\left(\frac{n\pi X_0}{X_0+X_1}\right) \sin\left(\frac{n\pi X}{X_0+X_1}\right) \qquad (50) \\
V &= \frac{\pi}{X_0+X_1} + 2\sum_{n=1}^{\infty} \frac{I_n}{n} \left\{1 + \left[\left(\frac{n\pi}{x_0+X_1}\right)^2 + 1\right]^{1/2}\right\} \\
&\quad \sin\left(\frac{n\pi X_0}{X_0+X_1}\right) \cos\left(\frac{n\pi X}{X_0+X_1}\right)
\end{aligned}$$

where

$$I_n = \frac{\exp\left\{Z - \left[\left(\frac{n\pi Z}{X_0+X_1}\right)^2 + Z^2\right]^{1/2}\right\}}{X_0\left\{1 + \left[\left(\frac{n\pi}{X_0+X_1}\right)^2 + 1\right]\right\}}.$$

We resolve this problem using the integral equation method. For this purpose we take the transformation

$$\theta = \exp(Z)\ \psi. \qquad (51)$$

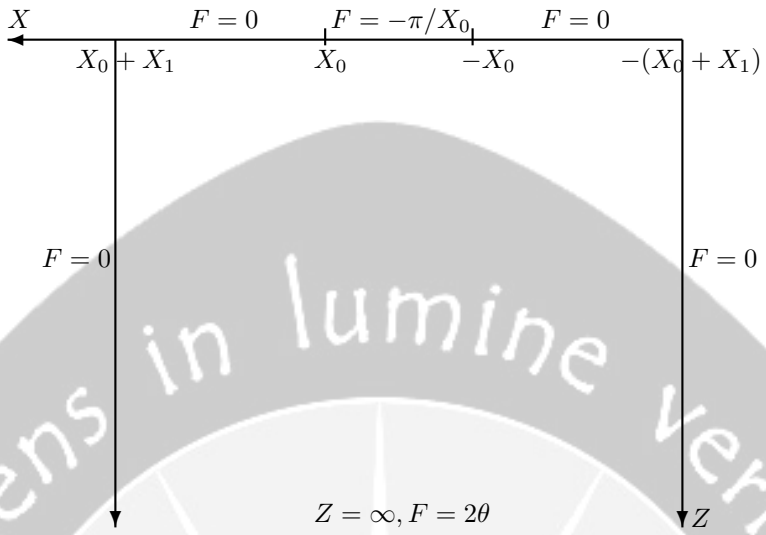


Figure 5: Boundary conditions of dimensionless variables for Problem 2.5

Substitution of (51) into (49) gives

$$\frac{\partial^2\psi}{\partial X^2} + \frac{\partial^2\psi}{\partial Z^2} - \psi = 0 \tag{52}$$

and

$$\begin{aligned} U &= -\exp(Z)\frac{\partial\psi}{\partial X} \\ V &= \exp(Z)\left(\psi - \frac{\partial\psi}{\partial Z}\right) \\ F &= \exp(Z)\left(\psi n_2 - P^{[\psi]}\right) \end{aligned}$$

where $P^{[\psi]}$ is given by (13). Equation (52) is a particular form of the constant coefficient equation (10) with $k = -1$ for the isotropic case ($\lambda_{11}^{(0)} = \lambda_{22}^{(0)} = 1$, $\lambda_{12}^{(0)} = \lambda_{21}^{(0)} = 0$). Therefore the boundary integral equation (14) may be used to find the solution of (52).

The corresponding boundary conditions, after the transformation (51), to be used in the integral equation method are (see Figure 6)

$$\begin{aligned} &X = -(X_0 + X_1), \quad 0 \le Z \le \infty, \quad P^{[\psi]} = 0, \\ &-(X_0 + X_1) \le X \le -X_0, \quad Z = 0, \quad P^{[\psi]} = -\psi, \\ &-X_0 \le X \le X_0, \quad Z = 0, \quad P^{[\psi]} = \pi/X_0 - \psi, \\ &X_0 \le X \le X_0 + X_1, \quad Z = 0, \quad P^{[\psi]} = -\psi, \\ &X = X_0 + X_1, \quad 0 \le Z \le \infty, \quad P^{[\psi]} = 0, \\ &-(X_0 + X_1) \le X \le X_0 + X_1, \quad Z = \infty, \quad P^{[\psi]} = -\psi. \end{aligned}$$

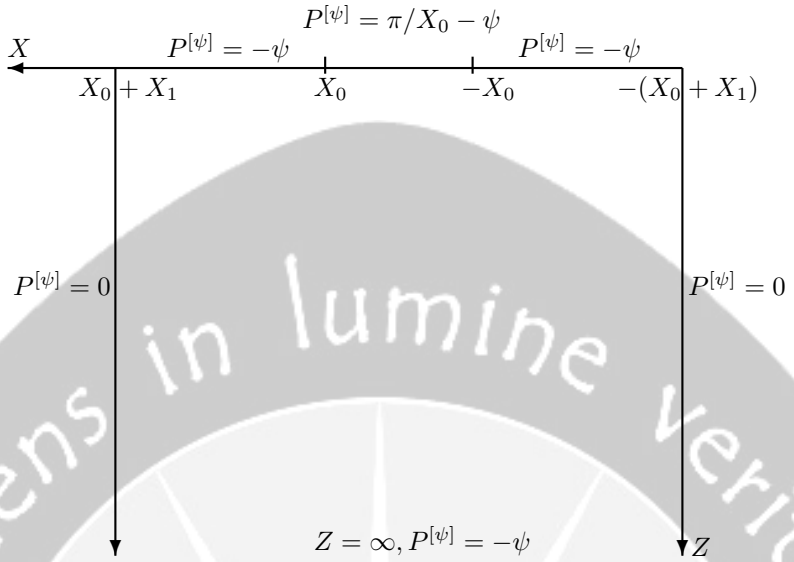


Figure 6: Boundary conditions for Problem 2.5 after the transformation (51)

Table 6 shows a comparison between the numerical and analytical solutions for the cases when $X_0 = 0.1, X_1 = 0.4$ and $X_0 = 0.25, X_1 = 0.25$. The values of the analytical solution are calculated by summing 100 first terms of the series solution (50). Whereas, when solving the problem using the integral equation method the boundary is divided into 600 segments of equal length. For the case when $X_0 = 0.1, X_1 = 0.4$, 200 segments are taken along each of the sides $X = -(X_0 + X_1), 0 \leq Z \leq 2$ and $X = X_0 + X_1, 0 \leq Z \leq 2$, 40 segments along each of the sides $-(X_0 + X_1) \leq X \leq -X_0, Z = 0$ and $X_0 \leq X \leq X_0 + X_1, Z = 0$, 20 segments along the side $-X_0 \leq X \leq X_0, Z = 0$ and 100 segments along the side $-(X_0 + X_1) \leq X \leq X_0 + X_1, Z = 2$. For the case when $X_0 = 0.25, X_1 = 0.25$, 200 segments are taken along each of the sides $X = -(X_0 + X_1), 0 \leq Z \leq 2$ and $X = X_0 + X_1, 0 \leq Z \leq 2$, 25 segments along each of the sides $-(X_0 + X_1) \leq X \leq -X_0, Z = 0$ and $X_0 \leq X \leq X_0 + X_1, Z = 0$, 50 segments along the side $-X_0 \leq X \leq X_0, Z = 0$ and 100 segments along the side $-(X_0 + X_1) \leq X \leq X_0 + X_1, Z = 2$. Moving down deeper than $Z = 2$ in the positive Z direction does not give more accurate results. It is observed from Table 6 that the numerical and the analytical solutions agree very well.

Problem 2.6 Consider Problem 2.5 again but with a different geometry of domain as for Problem 2.5. Referring to Figure 5, the section of straight line joining the points $(X, Z) = (-X_0, 0)$ and $(X, Z) = (X_0, 0)$ for Problem 2.5 is replaced by an arc of a circle. The circle is centered at point $(0, -\sqrt{r^2 - X_0^2})$ with radius r and positive values of Z. The value of the radius varies from $r/X_0 = 1.0$,

Table 6: Numerical and analytical solutions for Problem 2.5

Position	Numerical			Analytical		
(X, Z)	θ	U	V	θ	U	V
Case : $X_0 = 0.1, X_1 = 0.4$						
(.12,.08)	3.8494	9.2415	10.6176	3.8503	9.2754	10.6159
(.10,.09)	3.9849	8.0108	12.5756	3.9864	8.0419	12.5863
(.095,.1)	3.9757	7.0597	12.6935	3.9773	7.0863	12.7061
(.09,.11)	3.9611	6.2012	12.7113	3.9627	6.2239	12.7249
(.08,.12)	3.9689	5.2154	12.9674	3.9706	5.2342	12.9830
(.05,.13)	4.0411	3.2408	13.9517	4.0432	3.2524	13.9722
(.02,.14)	4.0479	1.2320	14.1372	4.0500	1.2363	14.1587
(.22,.05)	2.9706	6.6776	2.6034	2.9694	6.6800	2.5871
(.23,.06)	2.9388	6.1196	2.7832	2.9377	6.1225	2.7691
(.25,.12)	2.9753	4.5764	4.1086	2.9747	4.5814	4.1005
(.27,.20)	3.0111	3.0423	4.9654	3.0108	3.0467	4.9613
Case : $X_0 = 0.25, X_1 = 0.25$						
(.25,.12)	3.1416	3.9222	6.2830	3.1416	3.9316	6.2832
(.27,.20)	3.0936	2.3817	5.9111	3.0935	2.3866	5.9103
(.30,.30)	3.0740	1.3042	5.7791	3.0739	1.3066	5.7781
(.35,.37)	3.0537	.7571	5.6343	3.0536	.7585	5.6332
(.40,.42)	3.0493	.4200	5.6032	3.0491	.4207	5.6020
(.45,.44)	3.0441	.1982	5.5662	3.0440	.1985	5.5650
(.48,.46)	3.0503	.0723	5.6114	3.0501	.0724	5.6102
(.05,.30)	3.3477	.4121	7.7959	3.3481	.4129	7.7987
(.05,.35)	3.2993	.3185	7.4427	3.2996	.3191	7.4449
(.05,.40)	3.2623	.2450	7.1711	3.2625	.2454	7.1727
(.05,.45)	3.2339	.1879	6.9627	3.2341	.1883	6.9640

$r/X_0 = 1.9$, $r/X_0 = 2.8$ to $r/X_0 = 3.7$. In particular the bigger the value of the radius the flatter the arc of the circle will be. The geometry of the problem, after the transformation (51), is shown in Figure 7.

Table 7 and Table 8 show the results for the case when $X_0 = 0.1, X_1 = 0.4$. It is observed from Table 7 and Table 8 that as the value of the radius r/X_0 gets bigger the results converge to the corresponding results in Table 6 for Problem 2.5. Also, Table 9 and Table 10 show the results for the case when $X_0 = 0.25, X_1 = 0.25$. Again, Table 9 and Table 10 show that as the value of the radius r/X_0 gets bigger the results converge to the corresponding results in Table 6 for Problem 2.5.

7 Conclusion

Some boundary element methods are obtained for a class of two dimensional elliptic boundary value problems for inhomogeneous media. The methods can be applied

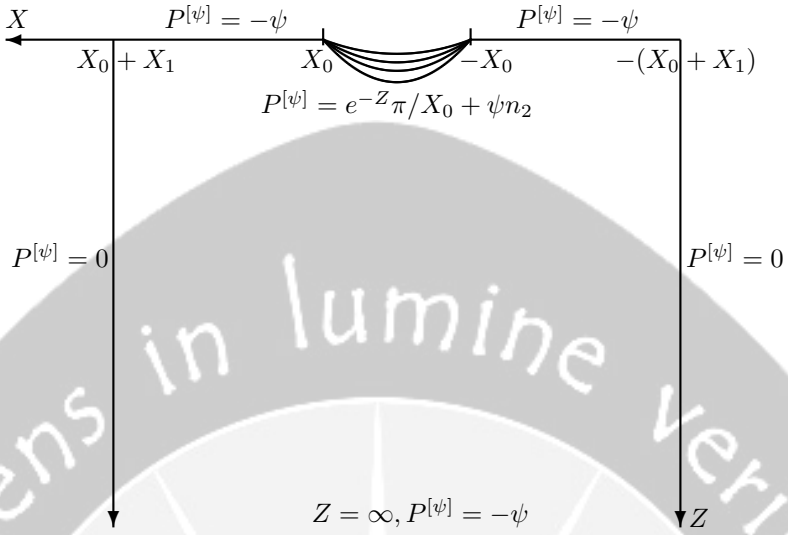


Figure 7: Boundary conditions for Problem 2.6 after the transformation (51)

to a variety of problems in such areas as antiplane strain in elastostatics and plane thermostatic problems for inhomogeneous anisotropic materials.

References

[1] Ang, W. T., Kusuma, J. and Clements, D. L. (1997), A boundary element method for a second order elliptic partial differential equation with variable coefficients. *Engng. Anal. Boundary Elements*, **18**, 311–316.

[2] Batu, V. (1978), Steady Infiltration from Single and Periodic Strip Sources. *Soil Sci. Soc. Am. J.*, **42**, 544–549.

[3] Cheng, A. H.-D. (1984), Darcy's Flow with Variable Permeability: A Boundary Integral Solution. *Water Resources Research*, **20**, 980–984.

[4] Cheng, A. H.-D. (1987), Heterogeneities in flows through porous media by boundary element method. *Topics in Boundary Element Research: Applications to Geomechanics*, **4**, 1291–1344.

[5] Clements, D. L. (1980), A boundary integral equation method for the numerical solution of a second order elliptic equation with variable coefficients. *J. Austrl. Math. Soc. (Series B)*, **22**, 218–228.

[6] D.L. Clements (1981), *Boundary Value Problems Governed by Second Order Elliptic Systems*, Pitman.

Table 7: Numerical solutions for Problem 2.6 when $X_0 = 0.1, X_1 = 0.4$

Position	$r/X_0 = 3.70$			$r/X_0 = 2.80$		
(X, Z)	θ	U	V	θ	U	V
(.12,.08)	3.8937	9.4892	10.5592	3.9276	9.6213	10.5934
(.10,.09)	4.0359	8.3293	12.6235	4.0729	8.4781	12.7031
(.095,.1)	4.0285	7.3550	12.7958	4.0661	7.4913	12.8940
(.09,.11)	4.0152	6.4675	12.8544	4.0530	6.5897	12.9665
(.08,.12)	4.0249	5.4522	13.1594	4.0635	5.5595	13.2894
(.05,.13)	4.1021	3.4123	14.2472	4.1428	3.4875	14.4175
(.02,.14)	4.1105	1.2993	14.4747	4.1518	1.3286	14.6605
(.22,.05)	3.0015	6.7145	2.5907	3.0263	6.7602	2.5966
(.23,.06)	2.9700	6.1605	2.7760	2.9947	6.2048	2.7850
(.25,.12)	3.0092	4.6285	4.1253	3.0351	4.6693	4.1500
(.27,.20)	3.0471	3.0876	5.0083	3.0740	3.1184	5.0466

[7] Clements, D. L. and Rogers, C. (1974), Wave propagation in inhomogeneous elastic media with (N + 1)-dimensional spherical symmetry. *Canadian J. Phys.*, **52**, 1246–1252.

[8] Gipson, G. S., Ortiz, J. C. & Shaw, R. P. (1995), Two-dimensional linearly layered potential flow by boundary elements, in *Boundary Elements XVII*, Editor: C.A. Brebbia, Springer-Verlag, Berlin.

[9] Rangogni, R. (1987), A solution of Darcy's flow with variable permeability by means of B.E.M and perturbation techniques, in *Boundary Elements IX Vol. 3*, Editor: C.A. Brebbia, Springer-Verlag, Berlin.

[10] Shaw, R. P. (1994), Green's functions for heterogeneous media potential problems. *Engng. Anal. Boundary Elements*, **13**, 219–221.

M.I. Azis: Department of Mathematics, Hasanuddin University, Jl. P. Kemerdekaan km. 10 Tamalanrea Makassar 90245, Indonesia.
Phone/Fax: +62 +411 585 643
E-mail: ivan@unhas.ac.id Website: www.unhas.ac.id\~ivan

Table 8: Numerical solutions for Problem 2.6 when $X_0 = 0.1, X_1 = 0.4$

Position	$r/X_0 = 1.90$			$r/X_0 = 1.00$		
(X, Z)	θ	U	V	θ	U	V
(.12,.08)	4.0450	10.0323	10.7655	5.9880	15.9985	14.3009
(.10,.09)	4.1990	8.9225	13.0049	6.2618	15.1858	18.2600
(.095,.1)	4.1936	7.8970	13.2449	6.2720	13.6085	19.1142
(.09,.11)	4.1814	6.9531	13.3532	6.2674	12.0663	19.6678
(.08,.12)	4.1937	5.8773	13.7244	6.3032	10.3320	20.6598
(.05,.13)	4.2789	3.7078	14.9651	6.4685	6.7509	23.3835
(.02,.14)	4.2895	1.4145	15.2503	6.4991	2.5942	24.1945
(.22,.05)	3.1132	6.9305	2.6306	4.5814	9.9127	3.5122
(.23,.06)	3.0813	6.3671	2.8289	4.5404	9.1683	3.8380
(.25,.12)	3.1250	4.8103	4.2451	4.6254	7.1175	6.0149
(.27,.20)	3.1665	3.2216	5.1836	4.7024	4.8616	7.5498

Table 9: Numerical solutions for Problem 2.6 when $X_0 = 0.25, X_1 = 0.25$

Position	$r/X_0 = 3.70$			$r/X_0 = 2.80$		
(X, Z)	θ	U	V	θ	U	V
(.25,.12)	3.1706	4.0655	6.2047	3.1983	4.1376	6.2173
(.27,.20)	3.1273	2.4714	5.9227	3.1564	2.5163	5.9601
(.30,.30)	3.1095	1.3547	5.8225	3.1391	1.3798	5.8701
(.35,.37)	3.0891	.7858	5.6816	3.1185	.8003	5.7296
(.40,.42)	3.0847	.4360	5.6523	3.1141	.4441	5.7007
(.45,.44)	3.0794	.2057	5.6148	3.1087	.2095	5.6628
(.48,.46)	3.0858	.0751	5.6619	3.1153	.0765	5.7108
(.05,.30)	3.3978	.4410	7.9592	3.4341	.4534	8.0601
(.05,.35)	3.3468	.3386	7.5834	3.3819	.3475	7.6744
(.05,.40)	3.3078	.2593	7.2957	3.3420	.2657	7.3795
(.05,.45)	3.2780	.1983	7.0758	3.3115	.2030	7.1543

Table 10: Numerical solutions for Problem 2.6 when $X_0 = 0.25, X_1 = 0.25$

Position	$r/X_0 = 1.90$			$r/X_0 = 1.00$		
(X, Z)	θ	U	V	θ	U	V
(.25,.12)	3.2865	4.3290	6.2955	4.9370	7.5490	9.1075
(.27,.20)	3.2469	2.6352	6.0942	4.8672	4.2778	9.2419
(.30,.30)	3.2306	1.4460	6.0246	4.8283	2.2297	9.0721
(.35,.37)	3.2094	.8383	5.8837	4.7916	1.2771	8.8077
(.40,.42)	3.2049	.4652	5.8553	4.7832	.7054	8.7482
(.45,.44)	3.1993	.2196	5.8162	4.7744	.3329	8.6834
(.48,.46)	3.2062	.0803	5.8667	4.7845	.1216	8.7584
(.05,.30)	3.5424	.4837	8.3476	5.2590	.4769	12.1521
(.05,.35)	3.4871	.3693	7.9373	5.1859	.4332	11.6690
(.05,.40)	3.4449	.2816	7.6246	5.1284	.3592	11.2708
(.05,.45)	3.4127	.2147	7.3863	5.0836	.2868	10.9533

Computation of Analysis Discrimination And Classification In Separating Two Classes Of Objects

Entin Hartini

P2TIK, BATAN, Serpong, Indonesia

Abstract: Discrimination and classification are multivariate techniques concerned with separating distinct set of objects and with allocating new objects to previously defined groups. Allocation or classification rule are usually developed from *learning* samples. Measured characteristics of randomly selected objects known to come from each of the two populations are examined for differences. A good classification procedure should result in few misclassifications. The chances or probabilities of misclassification and expected cost of misclassification (ECM) should be small. The estimated minimum ECM rule for two normal populations is tantamount to creating two univariate populations for the equal costs and equal priors discriminant function by taking appropriate linear combinations from two populations and then assigning a new observation to population 1 or otherwise. In sum for two populations, the maximum relative separation that can be obtained by considering linier combinations of the multivariate observations is equal to the sample squared distance between the two means. This is can be used in certain situations, to test whether two populations means differ significantly. A test for differences in mean vectors used an F- distribution.

HALF-SWEEP ARITHMETIC MEAN METHOD USING FINITE ELEMENT APPROXIMATION FOR POISSON'S EQUATION

J. Sulaiman[a], M.K. Hasan[b], M. Othman[c]

[a] Universiti Malaysia Sabah, Malaysia
[b] Universiti Kebangsaan Malaysia, Malaysia
[c] Universiti Putra Malaysia, Malaysia

Abstract. This paper discusses the application of the Half-Sweep Arithmetic Mean (HSAM) method by using the half-sweep linear finite element approximation equation based on the Galerkin scheme to solve one-dimensional Poisson's equation. Formulations of the full-sweep and half-sweep linear finite element approaches are also derived. Some numerical experiments are conducted to show that the Half-Sweep Arithmetic Mean (HSAM) method is superior to the Full-Sweep method.
Key-words: Half-Sweep Iterative, Arithmetic Mean Algorithm, Galerkin finite element scheme, Poisson's equation

1 Introduction

By using numerical techniques, there are many methods can be used by researchers to gain approximate solutions such as the finite difference, finite element, finite volume and boundary element methods. Those methods are one of the most efficient approximate techniques to solve any partial differential equation, which describes a certain problem in science and engineering.

Apart of those methods, the findings on the concept of the half-sweep iterative method and the two-stage iterative methods are definitely important in solving any system of linear equations. For instance, the half-sweep iteration is motivated by Abdullah [1] via the Explicit Decoupled Group (EDG) iterative method to solve two-dimensional Poisson's equations. Further discussions of the half-sweep iterative method, especially its applications on block iterative methods, are also reviewed, see in [2], [4], [14]. Secondly two-stage iterative methods are one of the efficient iterative methods in solving any system of linear equations. There are many two-stage iterative methods can be considered such as the Alternating Group Explicit (AGE) [3], the Iterative Alternating Decomposition Explicit (IADE) [7], the Reduced Iterative Alternating Decomposition Explicit (RIADE) [8], the Half-Sweep Iterative Alternating Decomposition Explicit (HSIADE) [9], the Quarter-Sweep Iterative Alternating Decomposition Explicit (QSIADE) [10], and the Arithmetic Mean (AM) [6] methods. In 2004, the standard AM method, however, has been modified in [11] by combining the concept of the half-sweep iteration and then called as the Half-Sweep Arithmetic Mean (HSAM) method. Moreover the effectiveness of the HSAM

method using the fourth-order standard finite difference approximation equation has also been shown in [12].

In this paper, we discuss the application of the Half-Sweep Arithmetic Mean (HSAM) method by using the half-sweep linear finite element approximation equation based on the Galerkin scheme to solve one-dimensional Poisson's equation. To show the efficiency of the HSAM method, let us consider one-dimensional Poisson's equation defined as

$$-\frac{d^2U}{dx^2} = f(x), \quad a_0 \le x \le b_0 \tag{1}$$

subject to the boundary conditions

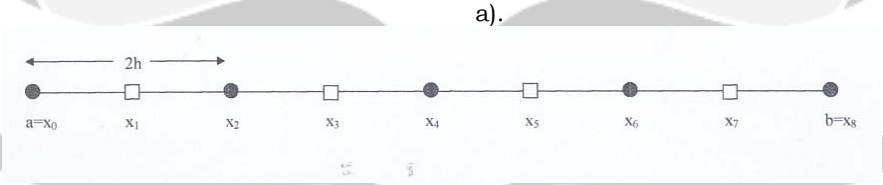
$$U(a_0) = \beta_0, \quad U(b_0) = \beta_1$$

and β_0, β_1, and $f(x)$ are constants and a continuous function**,** respectively.

To facilitate in formulating the full- and half-sweep linear finite element approximation equations for problem (1), we shall restrict our discussion onto uniform node points only as shown in Figure 1. Let assume the solution domain (1) can be uniformly divided into $m = 2^p,\ p \ge 2$ subinterval, which its distance, Δx defined as Equation (2)

$$\Delta x = \frac{(b_0 - a_0)}{m} = h,\ n = m - 1 \tag{2}$$

a).

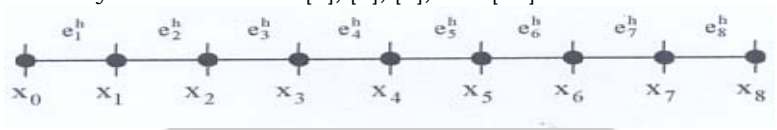


b).

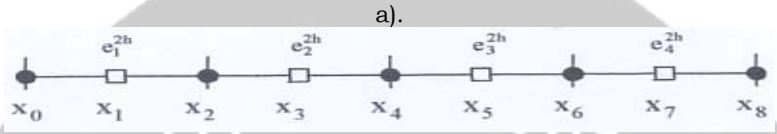
Figure 1. a). and b) show the distribution of uniform node points for the full- and half-sweep cases respectively.

Based on Figure 2, we need to build linear finite element networks in order to facilitate us to derive and implement in terms of the development of computational algorithms. The figure has shown that the linear finite element networks consist of two different sizes, which their element lengths are h and $2h$ respectively. However each element involves two node points only of type ●. For that reason, the implementation of the full- and half-sweep iterative algorithms will be applied onto the node points of the same type until the iterative convergence fixed is achieved.

Then other approximate solutions at remain points (points of the different type) are computed directly as discussed in [1], [2], [4], and [14].



a).



b).

Figure 2. The networks of linear elements with their length (a) h and (b) $2h$.

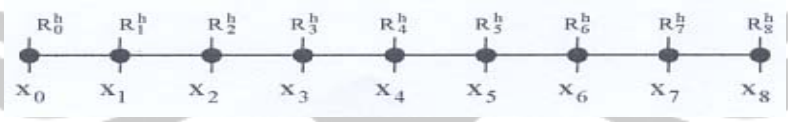
2 Formulation of the Half-Sweep Galerkin finite element approximation

As mentioned the previous section, we study the application of the Half-Sweep Arithmetic Mean (HSAM) method by using the half-sweep linear finite element approximation equation based on the Galerkin scheme to solve one-dimensional Poisson's equation. By considering all node points of type ● only, the general approximation of the function, $U(x)$ in the form of interpolation function for an arbitrary linear element e^{ph}, $p = 1,2$ is given by

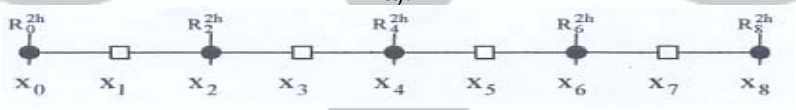
$$\underset{\sim}{U}\left[e^{ph}\right](x) = N_1^{ph}(x)U_i + N_2^{ph}(x)U_{i+p}, \qquad p = 1,2. \tag{3}$$

where,

$$h = x_{i+1} - x_i\,, \quad N_k^{ph}(x) = \frac{a_k}{ph}(x - b_k), \quad k = 1, 2\,, \quad \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}\,, \quad \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} x_{i+p} \\ x_i \end{bmatrix}.$$



a).



b).

Figure 3. a). and b). show the definition of the hat function, $R_j^{ph}(x)$ at the solution domain for the full- and half-sweep cases at m = 8.

To define the approximation of the functions, $U(x)$ and $f(x)$ for the entire domain, both functions will be approximated for the full- and half-sweep cases in the following form (Vichnevetsky [13])

$$\underset{\sim}{U}(x) = \sum_{j=0p,1p,2p}^{m} R_j^{ph}(x) U_j \tag{4}$$

and

$$\underset{\sim}{f}(x) = \sum_{j=0p,1p,2p}^{m} R_j^{ph}(x) f_j \tag{5}$$

for p = 1, 2. Equation (4) is an approximate solution for problem (1). In this paper, we shall consider the Galerkin finite element scheme to derive the full- and half-sweep linear finite element approximation equations for the problem. Thus, the choice of the weighted function, $W_j(x) = R_j^{ph}(x)$ is based on a set of the functions $\left\{ R_j^{ph}(x) \middle| j = 0,1p,2p, ..., m \right\}$, which form the approximation solution (4). Actually, each element of the function set is a first order piecewise polynomial functions. As a result, the definition for each hat function, $R_j^{ph}(x), \quad j = 0,1p,2p,...,m$ can be easily shown in Table 1.

Table 1. Definition of the hat function, $R_j^{ph}(x)$ in term of the shape function, $N_k^{ph}(x), k = 1,2$ for p = 1, 2, 4 in case of the networks of six linear elements.

Element (subinterval)	R_{0p}^{ph}	R_{1p}^{ph}	R_{2p}^{ph}	R_{3p}^{ph}	R_{4p}^{ph}	R_{5p}^{ph}	R_{6p}^{ph}
e_{1p}^{ph}	N_1^{ph}	N_2^{ph}	0	0	0	0	0
e_{2p}^{ph}	0	N_1^{ph}	N_2^{ph}	0	0	0	0
e_{3p}^{ph}	0	0	N_1^{ph}	N_2^{ph}	0	0	0
e_{4p}^{ph}	0	0	0	N_1^{ph}	N_2^{ph}	0	0
e_{5p}^{ph}	0	0	0	0	N_1^{ph}	N_2^{ph}	0
e_{6p}^{ph}	0	0	0	0	0	N_1^{ph}	N_2^{ph}

Let consider the Galerkin residual method [5] is defined as

$$\int_D R_k(x) E(x)\, dx = 0, \quad k = 0,1,2,\cdots,m \tag{6}$$

where $E(x) = \frac{d^2U}{dx^2} + f(x)$ is a residual function. By applying the differential formulate for the multiplication of two functions and substituting the boundary conditions into problem (1), Equation (6) can be rewritten as

$$\left[R_k(x) \frac{\partial U}{\partial x} \right]_a^b - \int_a^b \frac{\partial R_k}{\partial x} \frac{\partial U}{\partial x} \quad dx = F_k, \quad k = 0,1,2,\cdots,m \tag{7}$$

where,

$$F_k = \int_a^b \quad R_k(x) f(x) \quad dx$$

Beside this, the differential, $\underset{\sim}{U}(x)$ towards x is given by

$$\frac{d\underset{\sim}{U}}{dx} = \sum_{k=0}^{m} \frac{dR_k}{dx} U_k \tag{8}$$

By substituting Equation (8) into Equation (7), we can easily show the full- and half-sweep linear Galerkin finite element approximation equations generally stated in the following equation, see Vichnevetsky [13], Lewis and Ward [5] for the case, p = 1, which refers to the Full-Sweep iteration.

$$-U_{i-p} + 2U_i - U_{i+p} = b_i, \quad i = 1p, 2p, ..., m-p \tag{9}$$

where

$$b_i = \frac{(ph)^2}{6}\left(f_{i-p} + 4f_i + f_{i+p} \right)$$

The value of p, which corresponds to 1 and 2, represents the full- and half-sweep cases respectively.

To facilitate in developing a two-stage iterative method in the next section, we rewrite Equation (9) in a matrix form generally stated as

$$A\underset{\sim}{U} = \underset{\sim}{b} \tag{10}$$

where,

$$A = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} \left(\frac{m}{p} - p\right) x \left(\frac{m}{p} - p\right)$$

$$\underset{\sim}{U} = \begin{bmatrix} U_{1p} & U_{2p} & U_{3p} & \cdots & U_{m-p} \end{bmatrix}^T,$$

$$\underset{\sim}{b} = \begin{bmatrix} b_0 + U_0 & b_{1p} & b_{2p} & \cdots & b_{m-p} + U_m \end{bmatrix}^T.$$

3 Formulation of the HSAM method

In this section, we shall derive the formulation of the FSAM and HSAM schemes. As explained in previous section, the FSAM or HSAM method is one of two-stage iterative methods. It means that the iterative process for both methods involves two levels of virtual time such as $\underset{\sim}{U}^{(1)}$ and $\underset{\sim}{U}^{(2)}$. To derive the formulation of HSAM and FSAM methods, let assume the symmetry coefficient matrix, A in Equation (10) needs to be decomposed into

$$A = L + D + T \tag{11}$$

where L, D and T are lower triangular, diagonal and upper triangular matrices respectively. Generally, the scheme for both AM methods is given in [7], [11], [12] as follows

$$\left.\begin{aligned} (D+rL)\underset{\sim}{U}^{(1)} &= ((1-r)D - rT)\underset{\sim}{U}^{(k)} + r\underset{\sim}{f} \\ (D+rT)\underset{\sim}{U}^{(2)} &= ((1-r)D - rL)\underset{\sim}{U}^{(k)} + r\underset{\sim}{f} \\ \underset{\sim}{U}^{(k+1)} &= \frac{1}{2}\left(\underset{\sim}{U}^{(1)} + \underset{\sim}{U}^{(2)}\right) \end{aligned}\right\} \tag{12}$$

where r and $\underset{\sim}{U}^{(k)}$ represent as an acceleration parameter and an unknown vector at the k^{th} iteration respectively. By implementing some computer programs, we practically need to choose one optimal value of r, where its number of iterations is the smallest. The values of matrices L, D and T can be computed by using Equation (11). Then the general algorithm for FSAM and HSAM schemes in Equation (12) may be described in Algorithm I, see in [11], [12].

The FSAM and HSAM algorithms are explicitly performed by using all equations at level (1) and level (2) alternatively until the specified convergence criterion is satisfied. Then the Full-Sweep Gauss-Seidel (FGS) method acts as the control of comparison of numerical results.

Algorithm I. FSAM and HSAM schemes

i) at level (1)

a. Set $a \leftarrow 2.0, \quad b \leftarrow -1.0, \quad c \leftarrow -1.0,$

b. Set $w \leftarrow a(1-r), \quad v \leftarrow rb, \quad \lambda \leftarrow rc,$

c. For $i = 1p, 2p, 3p, \cdots, m-p,$

Calculate $$U_i^{(1)} \leftarrow \begin{cases} \left(wU_i^{(k)} - vU_{i+1}^{(k)} + rf_i\right)/a & ,i = 1p \\ \left(wU_i^{(k)} - \lambda U_{i-1}^{(1)} + rf_i\right)/a & ,i = m-p \\ \left(wU_i^{(k)} - vU_{i+1}^{(k)} - \lambda U_{i-1}^{(1)} + rf_i\right)/a & ,others \end{cases}$$

ii) at level (2)

d. For $i = m-p, m-2p, \cdots, 1,$

$$\text{Calculate } U_i^{(2)} \leftarrow \begin{cases} \left(wU_i^{(k)} - vU_{i+1}^{(2)} + rf_i\right)/a & ,i = 1p \\ \left(wU_i^{(k)} - \lambda U_{i-1}^{(k)} + rf_i\right)/a & ,i = m-p \\ \left(wU_i^{(k)} - vU_{i+1}^{(2)} - \lambda U_{i-1}^{(k)} + rf_i\right)/a & ,others \end{cases}$$

e. For $i = 1p, 2p, 3p, \cdots, m-p$,

$$\text{Calculate } U_i^{(k+1)} \leftarrow \frac{1}{2}\left(U_i^{(1)} + U_i^{(2)}\right)$$

4 Numerical experiments

To study the efficiency of the HSAM scheme by using the half-sweep linear finite element approximation equation (9) based on the Galerkin scheme, there are three items will be considered in comparison such as the number of iterations, execution time and maximum absolute error. Some numerical experiments were conducted in solving the following Poisson's equation

$$-\frac{d^2U}{dx^2} = 9\sin(3x), \qquad x \in [0,1] \tag{13}$$

Then boundary conditions and the exact solution of the problem (13) were defined by

$$U(x) = \sin(3x), \qquad 0 \le x \le 1, \tag{14}$$

All results of numerical experiments, obtained from implementation of the GS, FSAM and HSAM methods, have been recorded in Table 2. In the implementation mentioned above, the convergence test considered the tolerance error $\varepsilon = 10^{-10}$. Figures 4 and 5 show number of iterations and the execution time versus mesh size respectively.

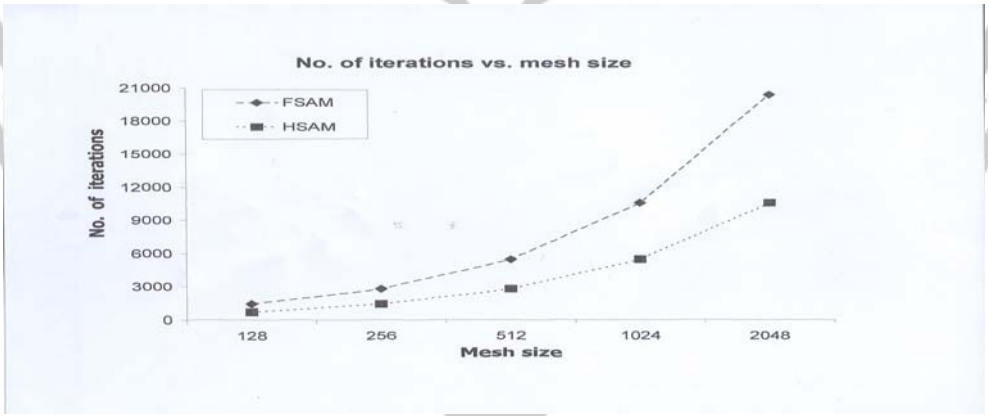


Figure 4. Number of iterations versus mesh size of the FSAM and HSAM methods.

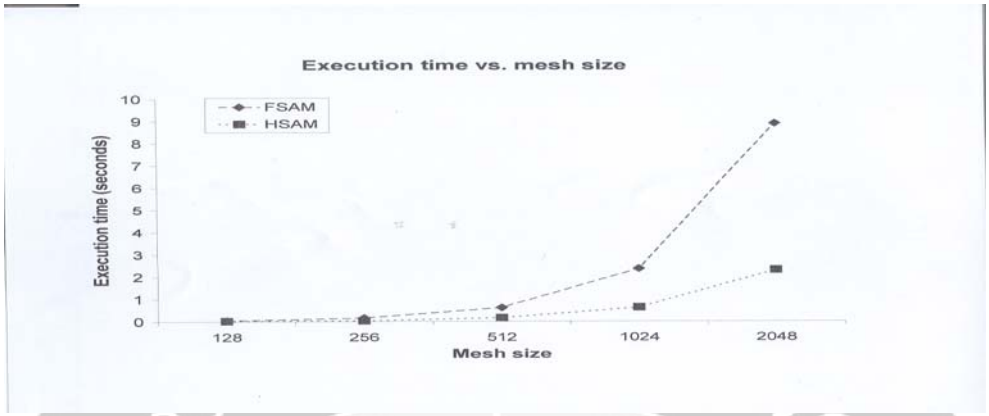


Figure 5. The execution time (seconds) versus mesh size of the FSAM and HSAM methods.

Table 2. Comparison of a number of iterations, the execution time (seconds) and maximum errors for the iterative methods.

No. of Iterations					
Methods (r optimum)	Mesh size				
	128	256	512	1024	2048
FGS	25950	94591	341534	1218827	4286118
FSAM	1429	2816	5467	10579	20383
HSAM	720	1429	2816	5467	10579
Execution time (Seconds)					
Methods (r optimum)	Mesh size				
	128	256	512	1024	2048
FGS	0.26	1.89	13.31	91.89	664.72
FSAM	0.04	0.16	0.62	2.37	8.90
HSAM	0.02	0.04	0.16	0.63	2.31
Maximum Absolute Errors					
Methods (r optimum)	Mesh size				
	128	256	512	1024	2048
FGS	4.2609e-5	1.1275e-5	5.3082e-6	1.1287e-5	4.2663e-5
FSAM	4.2451e-5	1.0627e-5	2.6844e-6	7.2707e-7	2.9438e-7
HSAM	1.6973e-4	4.2451e-5	1.0627e-5	2.6844e-6	7.2707e-7

5 Conclusion

In the previous section, it has shown that the full- and half-sweep linear finite element approximation equations based on the Galerkin scheme can be easily represented in the general form as shown in Equation (9). Through numerical results collected in Table 2, it seem that a number of iterations and the execution time for the HSAM have declined approximately 48.10 – 49.61% and 50.00 – 75.00% respectively as compared with the FSAM method, see in Figures 4 and 5.

Overall, it shows that the HSAM method is superior to the FSAM method in terms of a number of iterations and the execution time. This is because the computational complexity of the HSAM method is 50% less than the FSAM method. In terms of the accuracy, the approximate solutions of the HSAM method by using the half-sweep linear element approximation equation is in good agreement. Actually, it seems that the above conclusion of the efficiency of the HSAM method is obviously the same to the results obtained by using finite difference methods, see in [11], [12].

References

[1] Abdullah, A.R. [1991], The Four Point Explicit Decoupled Group (EDG) Method: A Fast Poisson Solver, *Int. J. Computer Maths.,* **38**, 61-70.

[2] Abdullah, A.R. & Ali, N.H.M. [1996], A comparative study of parallel strategies for the solution of elliptic pde's, *Parallel Algorithms and Applications*, **10**, 93-103.

[3] Evans, D.J. & Sahimi, M.S. [1988], The Alternating Group Explicit iterative method (AGE) to solve parabolic and hyperbolic partial differential equations. *Ann. Rev. Num. Fluid Mechanic and Heat Trans,* **2**, 283-389.

[4] Ibrahim, A.. 1993. *The Study of the Iterative Solution Of Boundary Value Problem by the Finite Difference Methods*. PhD Thesis. Universiti Kebangsaan Malaysia.

[5] Lewis, P.E. & Ward, J.P. 1991. The Finite Element Method: Principles and Applications. Wokingham: Addison-Wesley Publishing Company.

[6] Ruggiero, V. & Galligani, E. [1990], An iterative method for large sparse systems on a vector computer, *Computer Math. Applic.,* **20**, 25-28.

[7] Sahimi, M.S., Ahmad, A. & Bakar, A.A. [1993], The Iterative Alternating Decomposition Explicit (IADE) method to solve the heat conduction equation, *Int. J. Computer Maths.,* **47**, 219-229.

[8] Sahimi, M.S. & Khatim, M. [2001], The Reduced Iterative Alternating Decomposition Explicit (RIADE) Method for the Diffusion Equation, *Pertanika J. Sci. & Technol.* **9**(1), 13-20.

[9] Sulaiman, J., Hasan, M.K. & Othman, M. [2004], The Half-Sweep Iterative Alternating Decomposition Explicit (HSIADE) method for diffusion equations, In. J. Zhang, J.-H. He & Y. Fu (Eds). *Computational and Information Science 2004. Lecture Note on Computer Science (**LNCS 3314**): 57-63*. Berlin: Springer-Verlag.

[10] Sulaiman, J., Othman, M. & Hasan, M.K. [2004a], Quarter-Sweep Iterative Alternating Decomposition Explicit algorithm applied to diffusion equations, *Int. J. Computer Maths.,* **81**(12):1559-1565.

[11] Sulaiman, J., Othman, M. & Hasan, M.K. [2004b], A new Half-Sweep Arithmetic Mean (HSAM) algorithm for two-point boundary value problems. *Proceedings of the International Conference on Statistics and Mathematics and Its Application in the Development of Science and Technology*, Bandung, Indonesia, 169-173.

[12] Sulaiman, J., Othman, M. & Hasan, M.K. [2005], A fourth-order finite difference solver for Poisson's equations via the Half-Sweep Arithmetic Mean (HSAM) method. *Proceedings of the First IMT-GT Regional Conference Mathematics, Statistics and Their Applications*. Parapat, Indonesia, Editors: Mawengkang, H., Saib, S. & Sutarman, 139-146.

[13] Vichnevetsky, R. 1981. *Computer Methods for Partial Differential Equations, Vol I*. New Jersey: Prentice-Hall.

[14] Yousif, W.S. & Evans, D.J. [1995], Explicit De-coupled Group iterative methods and their implementations, *Parallel Algorithms and Applications*, **7**, 53-71.

J. SULAIMAN: Programme of Mathematics with Graphic Computer, School of Science & Technology, Universiti Malaysia Sabah, Locked Bag 2073, 88999 Kota Kinabalu, Sabah, Malaysia.
E-mail: jumat@ums.edu.my

M.K. HASAN: Department of Industrial Computing, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia
E-mail: khatim@ftsm.ukm.my

M. OTHMAN: Department of Communication Technology & Networks, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia.
URL: http://www.fsktm.upm.edu.my/~mothman/
E-mail: mothman@fsktm.upm.edu.my

The Application of Data Assimilation Method on Ground Water Pollution Problem

E. Apriliani, S. Sanjaya

Department of Mathematics, Institut Teknologi Sepuluh Nopember, Indonesia

Abstract: Data assimilation method is an estimation method which is combination between dynamical model with data measurement. There are two step what we must in data assimilation, the first is parameter identification and the second is variable estimation.

Data assimilation method has been applied in various problem such as meteorology, hydrology, air pollution and others. Here we applied data assimilation method to estimate the distribution of ground water pollution. The ground water pollution flow, for non reactive dissolved substance, can be write as advection - dispersion equation

$$D_x \frac{\partial^2 C}{\partial x^2} + D_y \frac{\partial^2 C}{\partial y^2} - v_x \frac{\partial C}{\partial x} v_y \frac{\partial C}{\partial y} = \frac{\partial C}{\partial t}$$

with boundary condition

$$C((x, y),0) = 0, \quad x \geq 0, \quad y \geq 0$$
$$C((0,0), t) = C_0, \quad t \geq 0$$
$$C((\infty, \infty), t) = 0, \quad t \geq 0$$

where x is the distance from the injection point.

In previous research we have done parameter identification, so that here just estimate the pollution concentrate. To get the pollution concentrate estimation, we must derive the mathematical model, discretize respect to time and space, applied into data assimilation algorithm by making computer program, make some simulation and finally analyze the result simulation.

Keyword: data assimilation, ground water pollution, concentrate distribution estimation

AN INVERSE THREE DIMENSIONAL ACOUSTIC PROBLEM SOLUTION FOR AXISYMMETRIC SOURCE IN FULL SPACE USING BOUNDARY ELEMENT METHOD AND GENERALIZED CROSS-VALIDATION

Ratnadewi[a] & Benjamin Soenarko[b]

[a]Department of Electrical Engineering,
Maranatha Christian University, Bandung, Indonesia

[b] Department of Engineering Physics, ITB,
Bandung, Indonesia

Abstract. The common problem found in acoustic is direct problem, in which for this type of problem, the acoustic source parameters are known and the acoustic pressure surrounding the source is sought. Another type of problem one may pose is such that the acoustic field pressures are known and the acoustic parameters are to be determined; this kind of problem is called an inverse problem.

This paper presents a solution of an inverse problem involving axisymmetric source in full space. The solution uses Boundary Element Method (BEM). The advantage of using the Boundary Element Method is the reduction of the dimension of the problem being solved, wherein three-dimensional problem is solved using two-dimensional treatment. For axisymmetric sources, the dimension of the problem can be further reduced and one deals only with one-dimensional computation.

In this paper Generalized Cross Validation (GCV) used to choice a parameter of regularization, this technique is used to correct the computational error. Test cases show that, the solution using regularization yields better result than the solution without regularization. The sources being evaluated include a radiating dome and a radiating cylinder respectively.

Key words: Inverse, BEM, Axisymmetric, SVD, GCV

1 Introduction

There are many problems in acoustics solved using the Boundary Element Method (BEM). In direct problem the pressure or normal velocity or the relation between them on the surface of vibrating object is used to determine the acoustic pressure at any field point [1-2].

In inverse problem the acoustics information at the acoustic field is known, and the acoustic information at the source is desired. The acoustic parameters at the source, such as acoustic pressure or impedance on the surface is to be determined [3].

For axisymmetric sources, the surface integral appearing in the calculation is reduced to a line integral. By taking advantage of the axisymmetric property of the

body in question, the modeling is simplified in that only the generator of the body needs to be discretized [4-5]
Kim and Ih [6] used interior BEM to determine the normal velocity and surface pressure of half-scaled automotive cabin. The difficulty arose was the ill-conditioned problem during the inversion caused by the singularity of the transfer matrix. They used regularization methods to stabilize the reconstructed field.

Combined Helmholtz Integral Equation Formulation (CHIEF) method was used to overcome the ill-conditioned nature of the surface matrix equation at certain characteristic frequencies [7]. In this method, CHIEF points were used to produce additional equations to obtain a unique solution.

Singular Value Decomposition (SVD) was employed to determined the number and location of the CHIEF points as proposed by Juhl [8-9].

P.A. Nelson [10], E.G. Williams [11] and Neumaier [15] proposed a technique for "regularizing" the solution of ill-posed inverse problems which, seems to be likely to occur more often than not when dealing with acoustic radiation.

The Generalized Cross Validation technique were developed for use primarily in statistical problems [12] and has been applied successfully in image processing [13] and in acoustic [10, 14] but not applied to problems with axisymmetric bodies.

2 Full Space Axisymmetric Boundary Element Method Formulation

The direct boundary element method is based on the Helmholtz integral equation [1,2]. The axisymmetric body in a full space is illustrated in Figure 1.

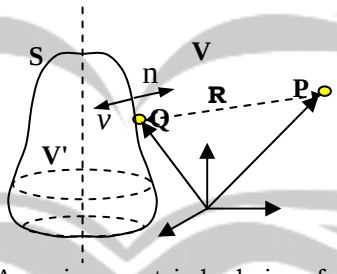


Figure 1. An axisymmetric body in a full space

Where Q is any point on S and P may be on V, S, or V' (the region inside the body bounded by (S), R ≡ |P-Q|, n is the unit normal on S directed away from the acoustic domain, v_n and p_s, are the surface normal velocity and sound pressure on S. C(P) is a coefficient. The value of C(P) depends on the location of point P. C(P) has the value 4π for P∈V, the value 2π for P∈S provided there is a unique tangent to S at such a P, and the value 0 for P∈V'. For P at an edge or corner of S, C(P) has the value [1-2]:

$$C(P) = 4\pi + \int_S \frac{\partial}{\partial v}\left(\frac{1}{R(P,Q)}\right) dS(Q) \tag{1}$$

The expression of C(P) in Eq. (1) is valid for all P for arbitrary nonsmooth S.
For axisymmetric body, a cylindrical coordinate system may be used (4,5), that is

$$C(P)\phi(P) = \int_L \left(\phi(Q) K^B(P,Q) - \frac{\partial \phi(Q)}{\partial v} K^A(P,Q) \right) \rho(Q)\, dL(Q) + 4\pi\phi^I(P) \qquad (2)$$

where

$$K^A(P,Q) = \int_0^{2\pi} \frac{e^{-ikR(P,Q)}}{R(P,Q)} d\theta(Q)$$

$$K^B(P,Q) = \int_0^{2\pi} \frac{\partial}{\partial v}\left(\frac{e^{-ikR(P,Q)}}{R(P,Q)} \right) d\theta(Q)$$

in which ϕ is the total scalar velocity potential satisfying the Helmholtz differential equation $\nabla^2 \phi + k^2 \phi = 0$, k is the wave number ω/c of the time harmonic waves present in the region V exterior to the surface S of an arbitrary body, ω is the circular frequency and c is the speed of sound, while ϕ^I is the incoming wave.

It can be shown that evaluations of $K^A(P,Q)$, $K^B(P,Q)$ and C(P) can performed partly numerically and partly analytically in terms of elliptic integrals. Equation (2) is the BEM formulation for axisymmetric bodies in a full space.

3 An Inverse Solution Formulation

The numerical implementation of Eq. (2) includes cases where P is on the surface S and P outside of S. For P on the surface, Eq. (2) can be written in matrix form:

$$[A]\Phi = [B]\,\Phi' + \Phi^I{}_S \qquad (3)$$

where [A] and [B] are NxN matrices, the element of vectors Φ, Φ', and $\Phi^I{}_S$ are values of ϕ, ϕ', and $-4\pi\phi^I$ on the surface. For P outside S (field point in V) the same procedure can be followed to obtain a matrix form of Eq. (2):

$$[C]\Phi + [D]\Phi' = \Phi_F + \Phi^I{}_F \qquad (4)$$

where [C] and [D] are NFxN matrices, with NF is the total number of field points, the element of vectors Φ_F and $\Phi^I{}_F$ are the value of $4\pi\phi$ and $-4\pi\phi^I$ on each field point. Eq. (3) and (4) are used in solving inverse problem, wherein $\Phi^I{}_S$, Φ_F and $\Phi^I{}_F$ are known while Φ and Φ' are to be solved.

In the direct BEM, either p (or ϕ) or ϕ' or the relation between them is known. Therefore by using Eq. (3), ϕ can be determined when ϕ' is known or vice versa. In the inverse BEM, no boundary condition is known, so there are more unknowns (2N) than the equations (N) in Eq. (3). To fully specify the problem, at least N more equations are needed, which are given by Eq. (4) with NF ≥ N.

To calculate ϕ and ϕ', rewrite Eq. (3) as

$$\Phi = [A]^{-1}[B]\,\Phi' + [A]^{-1}\Phi^{I}_{S} \tag{5}$$

Substituting Eq. (5) into Eq. (4) yields

$$([C][A]^{-1}[B] + [D])\,\Phi' = \Phi_F + \Phi^{I}_{F} + [C][A]^{-1}\,\Phi^{I}_{S} \tag{6}$$

which can be rewritten as

$$[G]\,\Phi' = Y \tag{7}$$

where Y is the right-hand side vector. Since the field points pressure and the incoming wave are known, vector Y can be determined. Then, Φ' and Φ can be found by solving Eqs. (7) and (5).

The solution of Eq. (5) may have a large error due to the ill-conditioned nature of matrices [A] and [B] at certain characteristic frequencies. One way to overcome this problem is by using CHIEF method. In CHIEF method, additional CHIEF points in V' are used to produce overdetermined equations to obtain an accurate inversion [7,8].

To determine the number and location of the CHIEF point, a simple method proposed by Juhl can be used. He used SVD of matrix [A], which is defined as [6,7,8]

$$[A] = [U][\,\Sigma][V]^{T} \tag{8}$$

where [U] and [V] are unitary matrices, satisfying the condition

$$[U][U]^{T} = [U]^{T}[U] = [I] \tag{9}$$
$$[V][V]^{T} = [V]^{T}[V] = [I] \tag{10}$$

with $[.]^T$ indicates a conjugate transpose matrix,

$$[\Sigma] = \begin{bmatrix} \sigma_1 & & & & \\ & \ddots & & & \\ & & \sigma_r & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & 0 \end{bmatrix} \tag{11}$$

in which σ_r is called the singular value of [A], and r is the rank of [A]. If [A] is singular then one or several of σ_i are zero, so that r < N. Juhl [7] pointed out that the number of CHIEF points needed to obtain good inversion equals the rank deficiency of [A], that is N-r.

The pseudo inverse of [A], denoted as $[A]^+$, can be calculated using the relation $[A]^+ = [V][\Sigma]^+[U]^T$, with

$$[\Sigma]^{+} = \begin{cases} \dfrac{1}{\sigma_i}, & \sigma_i \neq 0 \\ 0, & \sigma_i = 0 \end{cases} \tag{12}$$

where $[\Sigma]^{+}$ is the pseudo inverse of $[\Sigma]$.

4 Landweber Iteration

In many applications of linear algebra, the need arises to find a good approximation $\overline{\Phi'}$ to a vector $\Phi' \in R^n$ satisfying an approximate equation $[G]\,\Phi' \approx Y$ with ill-conditioned or singular $[G] \in R^{mxn}$, given $Y \in R^m$.

Usually, Y is a result of measurements the field points pressure contaminated by small error (noise), the measurement field pressures can be rewritten as [14]

$$\overline{Y} = [G]\,\Phi' + e \tag{13}$$

One can estimate the reconstructed velocity as

$$\overline{\Phi'} = [G^H G]^{-1}\, G^H\, \overline{Y} \tag{14a}$$
$$\overline{\Phi'} = [U]\, \mathrm{diag}[\sigma_i]^{-1}\, [V]^T\, \overline{Y} \tag{14b}$$

[G] is assumed to be known from the model. The vector $\overline{\Phi'} = [G]^{+}\,\overline{Y}$ is to be found, where $[G]^{+} = [U][\Xi]^{+}[V]^T$ is called pseudo inverse of [G].

By substituting Eq. [13] into [14a], the reconstruction error can be expressed as

$$\overline{\Phi'} - \Phi' = L\, e \tag{15}$$

where

$$L \equiv [G^H G]^{-1}\, G^H \tag{16}$$

The surface velocity in Eq. [6] and [7] can be estimated from the measured field pressure data by using the inverse iteration as

$$\Phi'^{\,j+1} = \Phi'^{\,j} + \beta\, G^H\, [\,Y - G\,\Phi'^{\,j}] \tag{17a}$$
$$= \beta\, G^H\, Y + [I_n - \beta\, G^H\, G]\,\Phi'^{\,j} \tag{17b}$$

where $\Phi'^{\,j}$ denotes the estimated source velocity at the jth step, I_n is the identity matrix with rank n, and β is a convergence parameter. Because Eq. (17b) can be considered as a geometric series, the necessary and sufficient condition for convergence as the iteration number j is increased is given by

$$|\,1 - \beta\sigma_i^2\,| < 1, \quad \text{for } i = 1,2,\ldots,n \tag{18}$$

If inverse of G is obtainable, the limit value of this geometric series for $j \to \infty$ is equivalent to the pseudo-inverse solution as

$$\overline{\Phi'}^{\,\infty} = [G^H G]^{-1}\, G^H\, \overline{Y}. \tag{19}$$

From this, it can be stated that the direct pseudo-inverse solution can be obtained by increasing the number of iterations sufficiently. However, if measurement errors are involved, the reconstruction error does not converge to a minimum value even if the number of iterations is infinitely increased, because the measurement errors can be amplified by nonradiating wave components during the backward reconstruction. Consequently, the iteration process should be terminated after some finite number of iterations chosen to achieve the minimum reconstruction error. It is particularly noteworthy that the high-order modes with small singular values possess a slower convergence rate than the lower modes, because the convergence rate depends on the geometric ratio of Eq. (17b). In other words, termination of the iteration process provides the low-pass filtering of the reconstruction field.

The direct implementation of the iterative filtering method can be written as

$$\overline{\Phi'}^{\,j} = \beta \sum_{i=0}^{j} [I-\beta G^H G]^i G^H \overline{Y} \qquad [20a]$$

$$= \beta \sum_{i=0}^{j} U\sigma[I-\beta\sigma^2]^i V^H \overline{Y} \qquad [20b]$$

$$= U\, \mathrm{diag}[(1-(1-\beta\sigma_i^2)^{j+1})/\sigma_i] V^H \overline{Y} \qquad [20c]$$

By comparing Eq. [20c] with Eq. [14b], the wave-vector filter matrix F_j can be defined as

$$\overline{\Phi'}^{\,j} = U\, F_j\, \mathrm{diag}[\sigma_i)^{-1}\, V^H \overline{Y} \qquad [21]$$

where

$$F_j = \mathrm{diag}\,[1-(1-\beta\sigma_i^2)^{j+1}) \qquad [22]$$

All the components of the filter matrix should be less than or equal to one. Divergence in the reconstruction field due to very small singular values can be suppressed by the filter. One can determine the optimal number of iterations that yields the minimum MSE.

5 Generalized Cross-Validation

When G is rank deficient or becomes increasingly ill-conditioned, this choice is impossible or increasingly useless. We may improve the condition of G^HG by modifying it. The simplest way to achieve this is by adding a small multiple of the identity. With this replacement, the formula [14a] turns into

$$\Phi' = [G^HG + n\lambda I\,]^{-1}\, G^H\, Y \qquad [23]$$

Generalized cross-validation offers a way to estimate appropriate values of parameters λ from the data. It is important since it does not require knowledge of the noise variance. The GCV estimate of λ in the ridge estimate Eq. [23] is the minimizer of $V(\lambda)$ given by

$$V(\lambda) = \frac{\frac{1}{n}\left\|(I - A(\lambda)Y)\right\|^2}{\left[\frac{1}{n}Tr\left(I - A(\lambda)\right)\right]^2} \qquad [24]$$

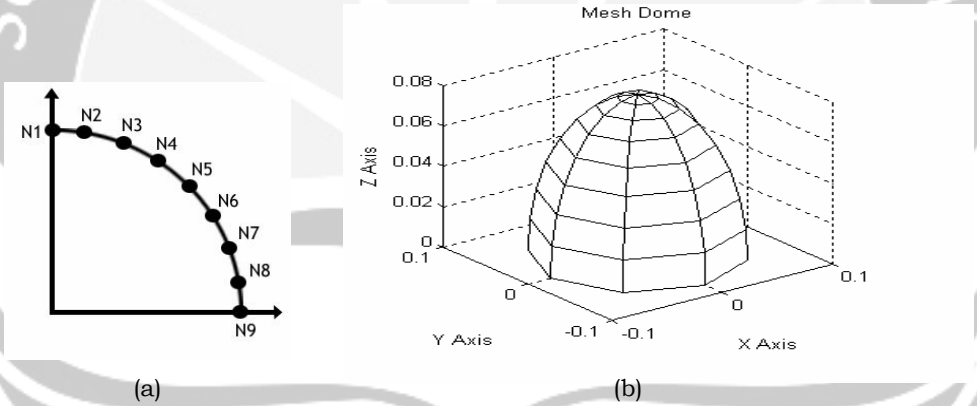
where

$$A(\lambda) = G(G^H G + n\lambda I\ ]^{-1}\ G^H \qquad [25]$$

Tr(.) denotes the trace of a matrix. Nhat Nguyen, Peyman Milanfar, Gene Golub [13] have used GCV for image restoration and resolution enhancement.

6 Test Cases

The axisymmetric source can be built from generator body as shown in Figure 2 for a radiating dome (radius a = 0.08m, 4 elements, 9 nodes) and a radiating cylinder (radius a = 0.04m, height t = 0.08m, 8 elements, 17 nodes) can be shown in Figure 3. The source can be discretised into a number of elements, which are assumed to radiate as point sources, and the sound field was measured.



(a) (b)

Figure 2. (a) Generator body (b) Axisymmetric source of a radiating dome

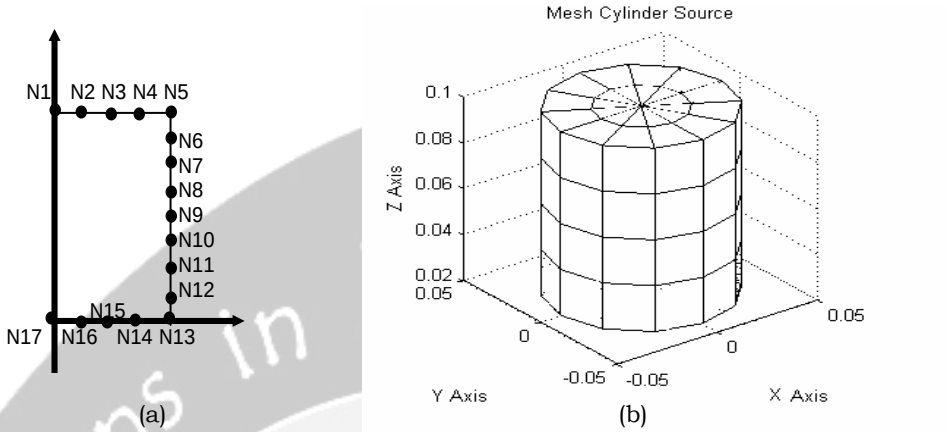


Figure 3. (a) Generator body (b) Axisymmetric source of a radiating cylinder

Figure 4. shows all of the field point distributions have the distance from the surface equal to 0.01m for radiating dome and 0.02m for radiating cylinder.

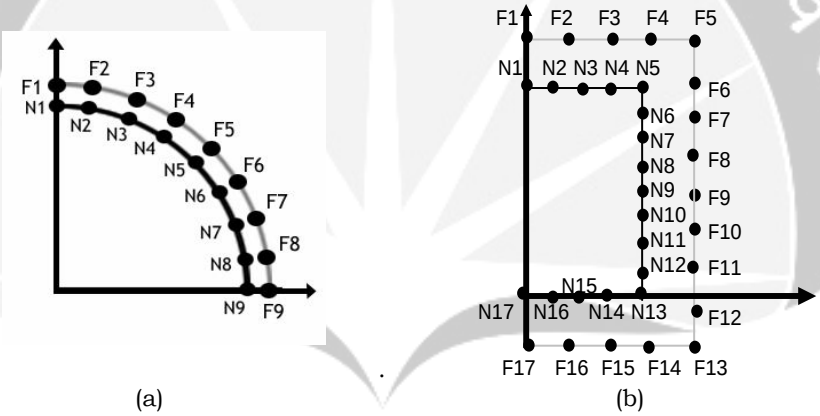


Figure 4. Fields point are distributed on the surface of imaginary space (a) dome (b) cylinder

The direct BEM calculation gives the surface pressure data. The prescribed surface data are referred to as the original data. Then the pressures at the selected field points are calculated using direct BEM. These pressures are then used as the given information to be employed in the present inverse formulation. Using this information as the input data, the present formulation calculates back the surface data. The results with regularization are compared with the result without regularization and the original data. Figure 5. shows the pressure graphic for dome. Figure 6. shows the original data, Figure 7. is the reconstruction without regularization and Figure 8. shows the reconstruction with regularization (β =0.0086 and l=964).

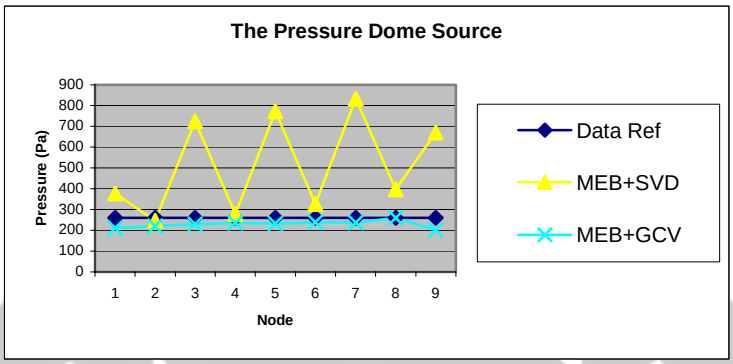


Figure 5. The pressure dome source

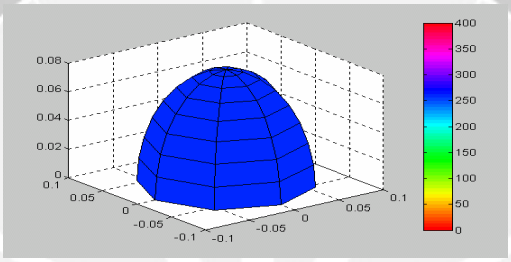


Figure 6. The original pressure data

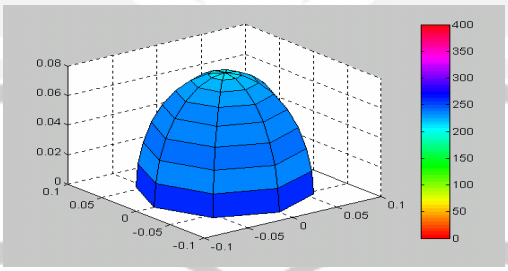


Figure 7. The reconstruction without regularization

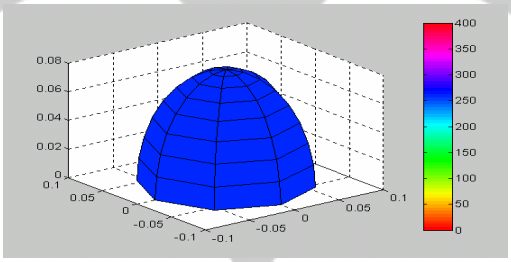


Figure 8. The reconstruction with regularization (β =0.0086 and l=964).

Figure 9. shows the pressure graphic for cylinder. Figure 10. shows the original data, Figure 11. is the reconstruction without regularization and Figure 12. shows the reconstruction with regularization (β =0.0086 and l=964).

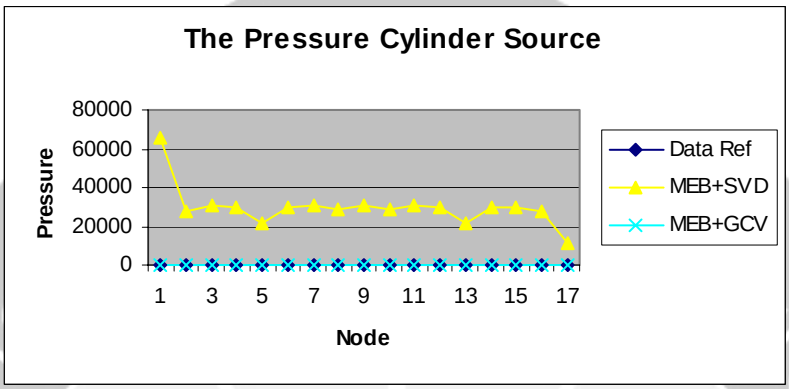


Figure 9. The pressure cylinder source

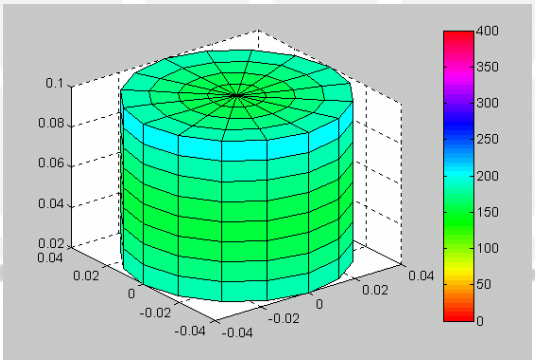


Figure 10. The original pressure data

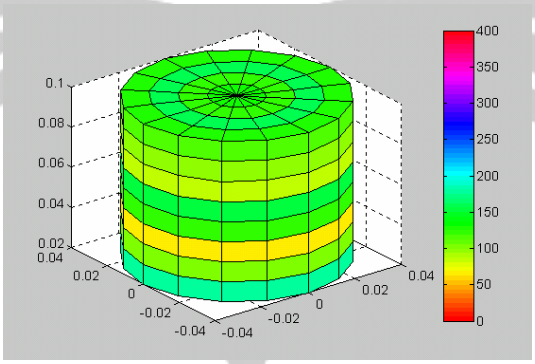


Figure 11. The reconstruction without regularization

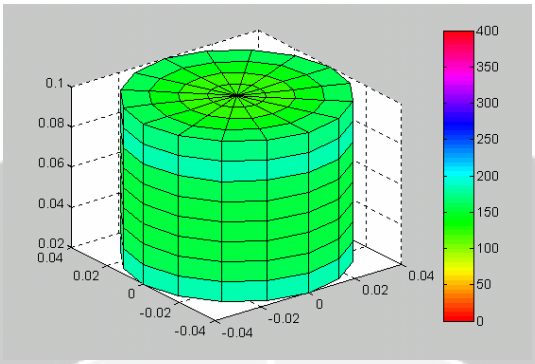


Figure 12. The reconstruction with regularization (β =0.0086 and l=964).

7 Conclusion

Inverse problems in acoustics may be solved with inverse BEM formulation using Singular Value Decomposition (SVD). Test cases were conducted dome and cylindrical geometry. The result obtained shows a good agreement with the available analytical values and apriorily known quantities. The Landweber iteration and Generalized Cross-Validation (GCV) can minimize error in the magnitude of the velocity potential.

Acknowledgement

The authors would like to acknowledge the support of Maranatha Christian University for this work.

References

[1] A.F. Seybert, B. Soenarko, F.J. Rizzo, and D.J. Shippy (1985), An Advanced Computational Method for Radiation and Scattering on Acoustic Waves in Three Dimensions, *J. Acoust. Soc. Am.*, **77 (2)**, 362 – 368.

[2] Benjamin Soenarko (1983), An Advanced Boundary Element Formulation for Acoustic Radiation and Scattering in Three Dimensions, *Ph.D Disertation at the University of Kentucky, America.*

[3] Ratnadewi, B. Soenarko [2005], On The Inverse Solution Using Boundary Element Method and Tikhonov Regularization, *Proceedings of the 4th International Conference on Numerical Analysis in Engineering*, **4**, B1.1.1-B1.1.8

[4] B. Soenarko, Dwi Urika (2003), An inverse BEM formulation in acoustics for axisymmetric bodies and boundary conditions in a half space, *Proceedings of the 3rd International Conference on Numerical Analysis in Engineering*, **3**, 3.17 – 3.21

[5] A.F. Seybert, B. Soenarko, F.J. Rizzo, and D.J. Shippy (1986), A Special integral equation formulation for acoustic radiation and scattering for axisymmetric bodies and boundary conditions, *J. Acoust. Soc. Am.*, **80 (4)**, 1241 – 1247

[6] B-K. Kim and J-G. Ih (1996), On the reconstruction of the vibro –acoustic field over surface enclosing an interior space using the boundary element method, *J. Acoust. Soc. Am.*, **100**, 3003 – 3016.

[7] Harry A. Schenck (1967), Improved Integral Formulation for Acoustic Radiation Problems, *J. Acoust. Soc. Am*, **44 (1)**, 41-58

[8] P.Juhl (1994), A Numerical Study of The Coefficient Matrix of The Boundary Element Method Near Characteristic Frequencies, *J. Sound and Vibration*, **175(1)**, 39-50.

[9] Gene H. Golub & Charles F. Van Loan (1996), Matrix Computations, *The John Hopkins University Press.*

[10] P.A. Nelson (2001), A Review of Some Inverse Problems in Acoustics, *International Journal of Acoustics and Vibration*, **6(3)**, 118-134.

[11] Earl G. Williams (2001), Regularization methods for near-field acoustical holography, *J. Acoust. Soc. Am.*, Vol. **110,** No. **4**, 1976 – 1988.

[12] Gene H. Golub, Michael Heath, and Grace Wahba (1978), Generalized Cross-Validation as a method for choosing a good ridge parameter, *Technometrics*, Vol. **21,** No. **2**, 215 – 223.

[13] Nhat Nguyen, Peyman Milanfar, Gene Golub (2001), Efficient Generalized Cross-Validation with Applications to Parametric Image Restoration and Resolution Enhancement, *IEEE Transactions on Image Processing*, Vol. 10, No. 9, 1299 – 1308.

[14] Bong-Ki Kim and Jeong-Guon Ih [2000], Design of an optimal wave-vector filter for enhancing the resolution of reconstructed source field by near-field acoustical holography (NAH), *J. Acoust. Soc. Am*, **107 (6)**, 3289 – 3297.

[15] Arnold Neumaier, Solving ill-conditioned and singular linear systems: A tutorial on regularization, 1-32

RATNADEWI: Ph D student at Department of Engineering Physics, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia.
Department of Electrical Engineering, Maranatha Christian University, Jl. Prof. Suria Sumantri no 65 Bandung 40164, Indonesia. Phone: +62 +22 2012186
E-mail: rdewi@bdg.centrin.net.id and ratnadewi@engineer.com

BENJAMIN SOENARKO: Department of Engineering Physics, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia. Phone/Fax: +62 +22 250 4424. E-mail: ben@tf.itb.ac.id and soenarko@bdg.centrin.net.id

EMERGENCE OF COMPLEX-VALUED SIGNAL PROCESSING

Andriyan Bayu Suksmono
Institut Teknologi Bandung, Indonesia

Abstract. This paper is a review on complex-valued signal processing, a class of new emerging signal processing techniques. This paradigm is based on a fact that various kinds of signals are naturally represented as (an N-dimensional) complex numbers, such as images acquired by coherent imaging (InSAR/Interferometric Synthetic Aperture Radar and MRI/Magnetic Resonance Imaging), and various digital modulation schemes. Most of current algorithms in signal/image processing consider only the magnitude (real-valued amplitude) and disregard the phase. In many cases, the later contains richer information. Recent research on various kinds of complex-valued neural network, which is mainly driven by the invention of Widrow's complex LMS algorithm, shows the advantages of using complex signal representation and consistent used of complex-valued algorithm. The discovery of CMRF (Complex-valued Markov Random Field) signifies a milestone in this emerging field. This paper will explain this paradigm-shift and provides various signal processing examples.

Key-words: complex signal representation, quadrature demodulation, signal processing, complex-valued neural network, CMRF, InSAR, MRI, LMS algorithm

1 Introduction

In the daily life, signals—including the two-dimensional one called image, are represented as a real number. Only in a few occasions, such as when the signal is Fourier transformed (FT), complex representation (real-imaginary or magnitude-phase) is needed. In many cases, the importance of phase information has been hidden. How important is the phase actually?

Oppenheim and Lim conducted a simple experiment [1], which is repeated and simplified here. Consider two different images, the "cameraman" shown in Fig.1 (a) and "rice" in Fig.1 (b). They are then Fourier-transformed, so that four arrays corresponding to the magnitude and phase of each images are obtained. Let f be the first image and g represent the second one, while $\mathbf{F} \equiv |\mathbf{F}| \angle \mathbf{F}$ and $\mathbf{G} \equiv |\mathbf{G}| \angle \mathbf{G}$ are their corresponding FT. First, discard the magnitude of the first image. Then reconstruct the image by inverse FT

$$\hat{f}_1 = FT^{-1}\left(e^{j\angle \mathbf{F}}\right) \tag{1}$$

The result is displayed in Fig.1(c). Second, combine the amplitude of **F** with the phase of **G** and then reconstruct a "combined" image

$$\hat{f}_2 = FT^{-1}\left(|\mathbf{F}|e^{j\angle\mathbf{G}}\right) \tag{2}$$

The result is displayed in Fig.1 (d). Observation on the reconstructed images shows that, even when the magnitude is unknown, phase information can still be used to reconstruct the image, although with a lower quality. On the other hand, the second reconstructed image shows the opposite case with the magnitude.

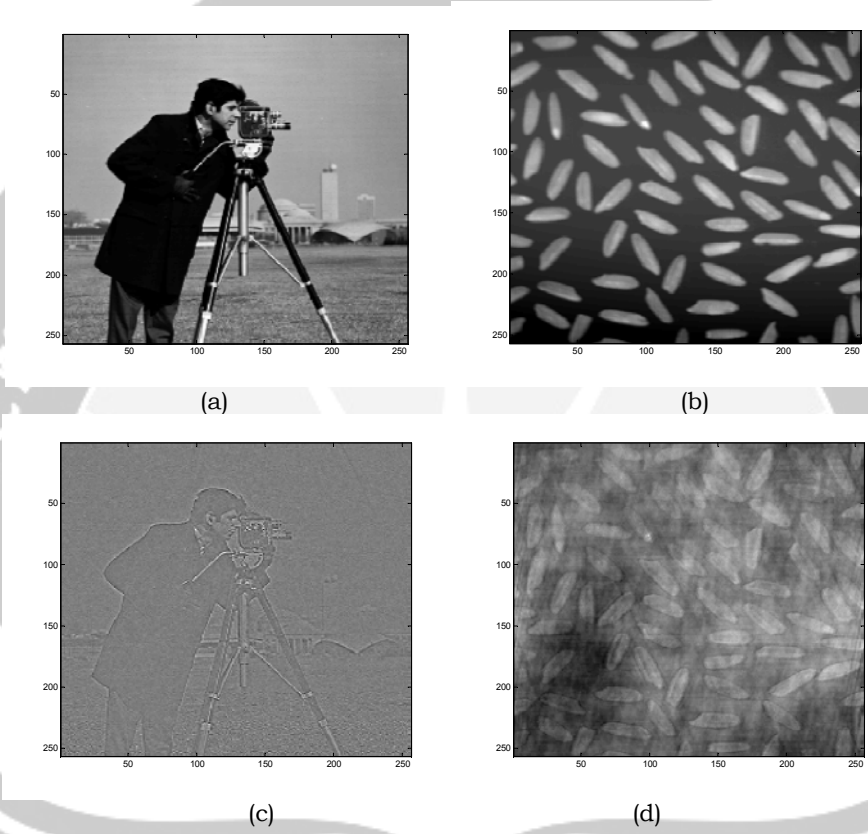


Fig.1 Demonstration of the important of phase information: (a) "cameraman", (b) "rice", (c) "cameraman" without magnitude information and (d) "cameraman" magnitude combined with "rice" phase.

Actually, there are more kinds of complex-valued signals than just a Fourier transformed ones. In fact, some imaging devices-such as MRI (Magnetic Resonance Imaging) and InSAR (Interferometric Synthetic Aperture Radar), communication devices such as quadrature demodulators, provides complex-valued signals. And it is very reasonable to state that the proper way to manipulate these signals is by complex-valued processing algorithm. The purpose of this paper is to review development of such algorithms, accompanied with comparisons to their real-valued counterparts when applicable.

The rest of the paper is organized as follows. Section 2 describes various kinds of complex-valued signals and its generation. Various complex-valued processing algorithms are explained in Section 3. The paper is concluded in Section 4.

2 Complex Signals and Its Generation

An electronic component responsible for generating complex-valued data is the IQ (Inphase-Quadrature) demodulator shown as a block diagram in Fig.2. It is a part of an electronic imaging system (MRI, InSAR, and GPR-the Ground Penetrating Radar) and communication devices. The input of this device is a modulated signal coming from a sensor, such as antennas or microphones, which is a mixture of information and a carrier wave. The signal is then split into two branches; the first one is demodulated by multiplying with a locally generated sinusoid whose frequency equal to the carrier, while the other one is multiplied by a $\pi/2$-shifted sinusoid, i.e. a *cosine* signal. After a low-pass filtering process (LPF), two signals called the inphase (I) and quadrature (Q) are obtained. In a digital processing, both of them are then converted to a digital format by an ADC (Analog to Digital Converter). The output of this demodulator is expressed as

$$\mathbf{Z} = \mathbf{I} + j\mathbf{Q} \tag{3}$$

where , $j = \sqrt{-1}$ is the imaginary number.

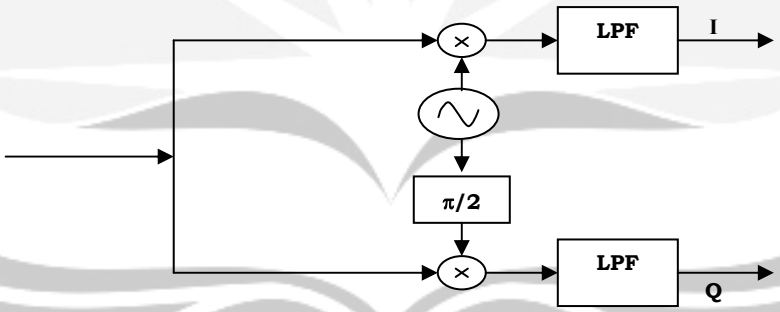
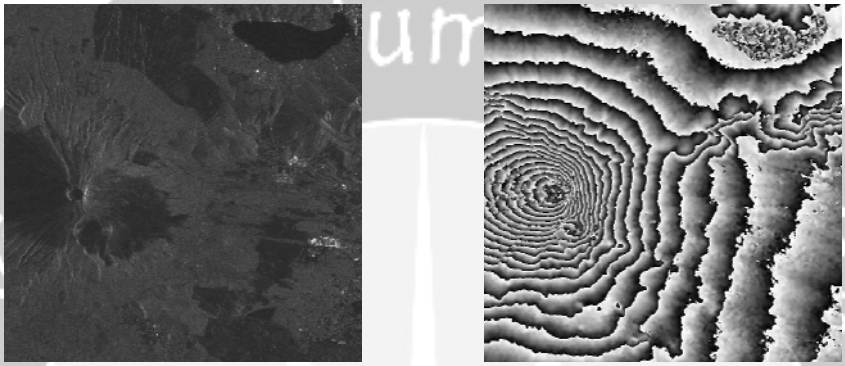


Fig. 2 Block diagram of a quadrature demodulator. Output of this circuit can be expressed as a complex number Z = I + *j*Q

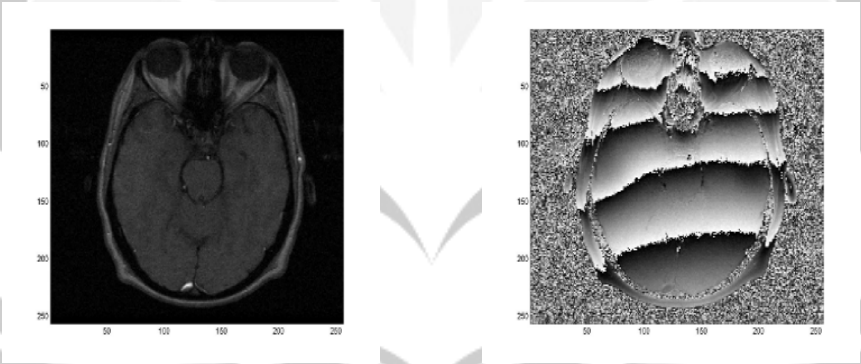
In an imaging system, magnitude part of the signal has a direct meaning, i.e. it is similar to the image captured by human eyes. Therefore, most of image processing algorithm on photographic image can be applied here, but some assumptions on noise behavior should be taken into account. In coherent imaging systems, such as InSAR, the noise is multiplicative instead of additive one. On the other hand, phase image has no immediate meaning by merely observing it directly. In the InSAR, it corresponds to terrain elevation, while in the MRI, it corresponds to water-fat distribution, temperature or bio-fluid flows.

Two kinds of complex-valued image examples are presented here. Fig.3 displays (a) a magnitude and (b) a phase of an InSAR image around Mt. Fuji, Japan, captured by JERS-1 imaging satellite. While Fig.4 is MRI images of a human head section, with (a) the magnitude and (b) the phase. In both of these cases, fringes are found to appear in the phase images, which is wrapping effects caused by *arctan* usage in phase retrieval. In this case, phase unwrapping (PU) is an important process to obtain an unwrapped phase image, to get a valid interpretation about the image.



(a) (b)

Fig.3 Complex InSAR images of Mt. Fuji: (a) magnitude and (b) phase image



(a) (b)

Fig. 4 Complex MRI images of a head section: (a) magnitude and (b) phase image

3 Processing the Complex Signals

Signal processing assumes a particular underlying model of the signal. The most popular one is the statistical model. In the first discussion, a two dimensional signals-- the image, is considered. Array of number representing an image is also called field. Since a meaningful image does not totally random, a "structural' model, such as Markov Random Field (MRF), is introduced. According to the MRF model, although a point in a field is random, there is a certain degree of

relationship between a pixel and its neighbor expressed as parameters. Since most of images are real numbers, it will be named as the real-valued MRF. This concept is extended to image whose pixel contains both magnitude and phase and it is called complex-valued MRF (CMRF). In the second part, processing algorithm of complex-valued signals captured by an array is discussed.

3.1 MRF and CMRF Model for Images

In an MRF, a site (pixel position) is denoted by *s*, while its intensity (grey scale) value is denoted by x_s. The *colour* of a pixel corresponds to its brightness. Let a finite rectangular $M \times N$ lattice L represents an image. A G-levels *colouring* of lattice L denoted by x_s is a function assigning the site-s in L to the set of {0, 1, …, G-1}. Subset of sites $t \in N_s$ are called neighbours of site s if the conditional probability of the colour x_s depends only on the sites $t \in N_s$, i.e.,

$$P\left(x_s \middle| x_1, x_2, \ldots, x_{s-1}, x_{s+1}, \ldots, x_{M \times N}\right) = P\left(x_s \middle| x_t; t \in N_s\right) \tag{4}$$

Markov random field (MRF) is defined such that the field has a joint probability density on the set of all possible colourings x_s for all sites-s of the lattice L subject to the conditions of: positivity, Markovianity and homogeneity.

According to the Ising model, the probability value is determined by the pixel's energy $E_I(x_s)$ and environment's temperature T. The energy depends on the neighbourhood configuration, order of the model and interaction strength between sites. The probability of a site-s, whose intensity is x_s, with a neighbourhood-t is expressed as:

$$P\left(x_s\right) = \frac{1}{Z} e^{\frac{-E_I\left(x_s\right)}{T}} \tag{5}$$

$$E_I\left(x_s\right) = -\left(\alpha_s x_s + \sum_{t \in N_s} \Lambda_{I\,st} x_s x_t\right) \tag{6}$$

where Z is the partition function for normalization, α_s and Λ_{Ist} are MRF model parameters, and N_s is the neighborhood set of site-s. The model parameter Λ_{Ist} is the interaction strength between sites s and t, while α_s corresponds to the strength of external field. These parameters can be determined by least square method.

The complex-valued MRF (CMRF) model can also be constructed in a similar fashion. Instead of the real-valued pixel x_s, CMRF uses the complex-valued one z_s. The energy of the CMRF is defined as

$$E\left(z_s\right) \equiv \frac{1}{2\sigma^2}\left(\left| z_s - \sum_{t \in N_s} \Lambda_{st}^* z_t \right|^2\right) \tag{7}$$

CMRF parameters Λ_{st} that corresponds to Λ_{Ist} in equation (6) generally have complex values. Based on MSE metric, the estimated value of (non-causal) CMRF parameters Λ_{st} and variance σ^2 are:

$$\hat{\boldsymbol{\Lambda}}^* \equiv \left[\sum_{s\in L} z_s \boldsymbol{Q}_s^*\right]\left[\sum_{s\in L} \boldsymbol{Q}_s \boldsymbol{Q}_s^*\right]^{-1} \tag{8}$$

$$\hat{\sigma}^2 \equiv \frac{1}{M\times N}\sum_{s\in L}\left|z_s - \hat{\boldsymbol{\Lambda}}^* \boldsymbol{Q}_s\right|^2 \tag{9}$$

where

$$\boldsymbol{Q}_s = \begin{bmatrix} z_{s+\tau_1} & z_{s+\tau_{-1}} & \cdots & z_{s+\tau_{12}} & z_{s+\tau_{-12}} \end{bmatrix}^T \tag{10}$$

$\boldsymbol{Q}_s$ is neighborhood vector and $()^*$ and $\hat{}$ mean the complex-conjugate transpose and estimated value, respectively. A 5th order neighborhood N_s has been considered.

3.2 Adaptive Complex-Amplitude Classifier for InSAR Images

As a first complex-valued algorithm example, an adaptive texture classification system that deals with the height and the reflectance is presented. Combination of reflectance--the magnitude of the image, and height--the phase image, yields a complex-valued image. The system generates and processes slope-direction-insensitive complex-amplitude texture features. It consists of two-stage preprocessor and a self-organizing map (SOM) neural network.

First, an input complex-amplitude image is decomposed into small blocks of a size $L\times L$. Each image block is locally unwrapped so that the data is converted into amplitude and height information. The unwrapped data is fed to the feature extractor. The extractor calculates stochastic characteristics of the block to generate a feature vector (complex mean and covariances).

The generated feature vectors are fed to the SOM. The SOM yields adaptive templates (references) for classification. After a set of templates has been generated, each feature vector is fed again to the SOM to be classified into one of the classes represented by the templates. Accordingly the image is segmented depending on the statistical characteristics of the local complex-amplitude data. The feature parameter vector that is insensitive to the slope is constructed as

$$K' \equiv [M, K(0,0), K'(0,1), K'(1,0), K'(1,1)] \tag{11}$$

where

$$M = \frac{1}{L^2}\sum_{i=0}^{L-1}\sum_{j=0}^{L-1} z(i,j) \tag{12}$$

$$K'(\xi,\eta) = \left|K(\xi,\eta)\right| e^{j\left|\varphi(\xi,\eta)\right|} \tag{13}$$

$$K(\xi,\eta) = \left|K(\xi,\eta)\right| e^{j\arg(K(\xi,\eta))} = \frac{1}{L^2}\sum_{i=0}^{L-1}\sum_{j=0}^{L-1} z^*(i,j) z(i-\xi, j-\eta) \tag{14}$$

By using this feature, the InSAR image is assumed to be a 2nd order (causal) CMRF. By using a 16×16 block size and 16 classes, the results are displayed in Fig.5.

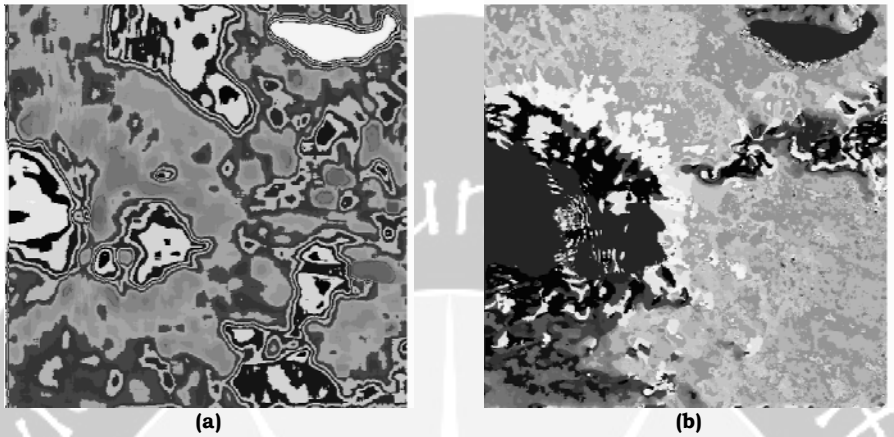


Fig. 5 Classification result: (a) using magnitude only and (b) magnitude-phase

Figure 5(a) is the classification result when the phase information is omitted. This is the conventional way to classify images using only amplitude (intensity) information. The lake is classified into a single class completely. However the mountain region contains so many classes that one cannot find a mountain-shaped cluster.

Figure 5 (b) shows the classification result obtained when complex-amplitude representation is used. In this case, the lake is classified perfectly. Moreover, the mountain is covered almost by a single class so that a mountain shape can be observed clearly. In the south direction of the lake (Lake Yamanaka), one can also find a cluster. Actually, it is low mountains (e.g. Mt. Ohora), which does not appear in conventional classification result (Fig.5 (a)). Accordingly, the lake and the mountains are segmented successfully by the complex-amplitude scheme.

3.3. Adaptive Phase Filtering and InSAR Image Restoration

The second application of complex-valued algorithm to be discussed in this Section is the adaptive residue-noise filtering of InSAR images. In an InSAR image, each pixel at a certain location of the amplitude image is associated with its counterpart in the phase image. They are both combined to yield a complex-valued image. During the restoration process, the complex-valued image is divided into small windows where stationary statistical condition is assumedly be satisfied, which is 64×64 in this study.

In each window, residues; which appear in the presence of phase noise, are detected and it as well as its neighbors is marked. The CMRF parameter is estimated by choosing the unmarked area whenever possible, or one can use them

anyway if the image is greatly corrupted. Then restoration algorithm is applied. Two filtering methods are proposed: the steepest descent and the Monte Carlo Metropolis method.

3.3.1. CMRF Filtering by Steepest Descent Method

After the parameters are estimated, the value of marked pixels are updated by using steepest-descent method as follows

$$z_s(k+1) = z_s(k) + \mu \Delta z_s(k) \tag{15}$$

$$\Delta z_s(k) = z_s(k) - \hat{\Lambda}^*(k) \boldsymbol{Q}_s(k) \tag{16}$$

where μ is the learning rate ($0<\mu<1$). This algorithm decreases the energy exponentially. The process is repeated by feeding the estimation output of the network to the input recurrently, until a certain amount of residues number is achieved.

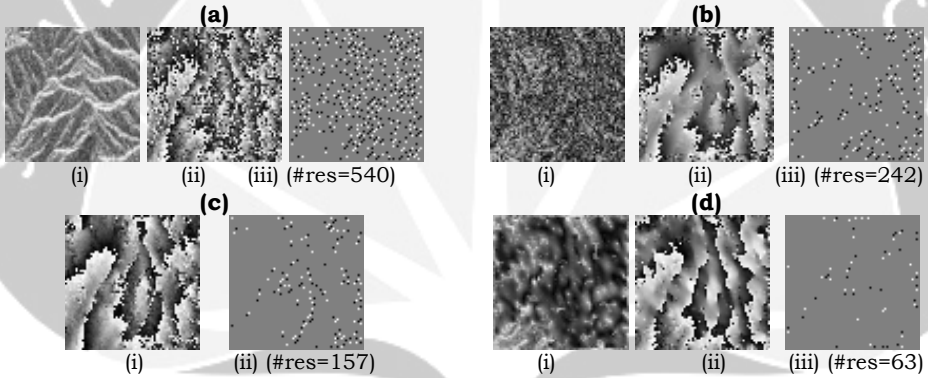


Fig.6 Performance comparison of the complex-valued method with the Lee and G-W filters: (a) Unfiltered ((i) amplitude, (ii) phase, (iii) residue map); (b) Lee filtered: (i) coherence, (ii) phase, (iii) residue map; (c) G-W filtered: ((i) phase, (ii) residue map; and (d) Complex- valued method: (i) amplitude, (ii) phase, (iii) residue map.

In the experiment, an InSAR-unfiltered image is used. The performance is compared with standard Lee filter's and the Goldstein-Werner (G-W) filter. Figure 6 (a) shows the unfiltered original image: (i) amplitude, (ii) phase and (iii) residue map. Before the processing, 540 residues are detected. Firstly, the Lee filter is applied to the original image. The result is depicted in Fig. 6(b): (i) coherence map, (ii) filtered phase, and (iii) residue map. It is found that the residue number is reduced to 242. Secondly, the G-W filter is applied. The result is shown in Fig. 6(c): (i) phase image and (ii) residue map. The residue is reduced to 157 after a strong filtering (by setting filter's parameter α=1). At last, the complex-valued method is applied. The result is depicted in Fig. 6(d): (i) amplitude, (ii) phase and (iii) residue map. It is found that the residue number has been reduced to 63.

It is observed that the proposed method not only reduce the residue more than others, but also give a better phase image result as follows. In the middle area of the Lee-filtered image, a smearing effect that joins fringes together and erases some detail of objects is observed. The similar behaviour is also found in the G-W filtered: the smearing effect occurs in the left and right parts of the image. In contrast, the propose method gives a lower smearing effect to the fringe, so that the detail is more preserved.

3.3.2. CMRF Filtering by Monte Carlo Metropolis Algorithm

A measure of the distance between a current pixel values z_s with the estimated one $\hat{z}_s$; i.e. the error energy; is defined as

$$E(z_s) = \frac{1}{2\sigma^2(T)} \| z_s - \hat{z}_s \|^2 \tag{17}$$

$$\hat{z}_s \equiv \sum_{t \in N_s} \Lambda_{st} z_t \tag{18}$$

In this method, the estimation criterion is the minimum mean square error (MMSE) evaluated to the entire complex-valued image in an M×N block, that is to say, it has to minimize the following function:

$$E_\varepsilon = \frac{1}{M \times N} \sum_{s \in L} \| z_s - \hat{z}_s \|^2 \tag{19}$$

The restoration process is achieved by decreasing the energy function (19) by using the Monte Carlo Metropolis (MM) algorithm. In the process, a corrupted pixel is randomly chosen and updated by adding (or subtracting) a small (complex) random value. If the update brings the system to a lower energy state, then it is accepted. But if the energy increases, it is accepted it some probability. The procedure is iterated until a certain convergence level is achieved. In the cycle of the restoration algorithm, some residues are eliminated due to combination of –1 and +1 residues or it disappears by itself.

In the experiment, simulated complex-valued image is used. The phase image represents a wrapped linear slope in vertical direction. Homogeneous amplitude (unity) is assumed. An area in the complex image is then multiplied by zero mean complex Gaussian multiplicative noise (variance=0.5). The effect of complex multiplicative noise is multiplicative in the amplitude, while additive in the phase.

Fig. 7(a) shows the simulated phase with noise (left part) and its corresponding residue distribution (right part). The number of detected residue is 115. The MM algorithm is applied, where the initial chosen temperature T_0 = 0.26 and final temperature T_{final} = 0.00026. The detected current temperature $T_{current}$=0.11, is equal to the estimated variance ($\hat{\sigma}^2(T)$). The system is evolved for 7500 cycles. The restored phase image is displayed in Fig.7 (b). For a comparison, a restoration/

filtering result using complex boxcar filter—a mere averaging in complex domain, is provided in Fig. 7 (c).

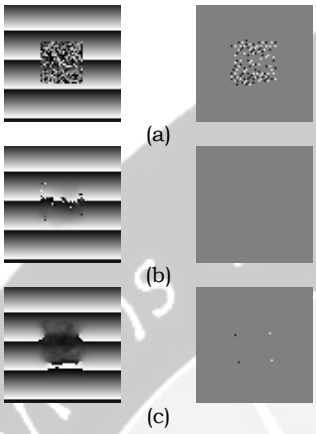


Fig.7 Simulated Phase (left part) and its corresponding residue distribution (right part): (a) original simulated phase degraded by noise (residue=115) (b) restoration by the proposed method (residue=0), and (c) restoration by complex boxcar method (residue=4).

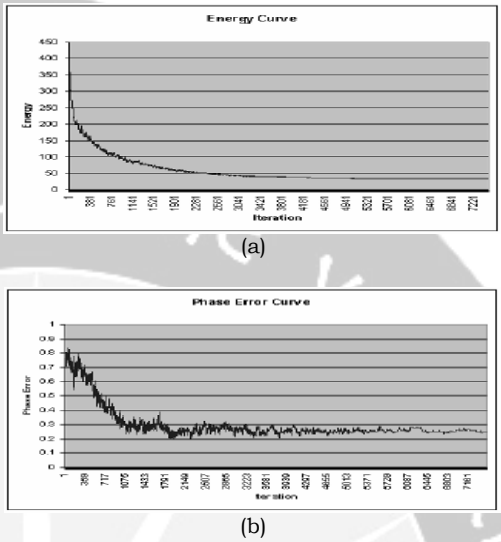


Fig.8 Curves of energy (a) and phase error (b) as a function of iteration number.

It is observed that the result of the MM algorithm is better than the boxcar filter. Because of the averaging process in the boxcar filter, an area that contains noise is smeared out such that important fringe detail is lost as it is observed in Fig.7 (c). On the other hand, the complex-valued method restores the fringe detail as well as slopes in the corrupted areas (Fig. 7(b)). Additionally, while boxcar method reduces the residues from 115 to 4, the complex-valued method eliminates all the residues.

Figure 8 displays the evolution of energy (a) and phase error (b). The phase error is calculated as the averaged absolute value of the phase difference between the estimated phase image and the original noiseless image. It is observed that the energy as well as the phase error decreases monotonically. In both of the curves, a high fluctuation in the initial iterations occurs and it fades away afterwards.

3.4 Beamforming By a Complex-Valued Neural Network

In this Section, neuro-beamforming by an intelligent complex-valued algorithm is explained. Beamforming is a capability of a sensor-array to direct its spatial response to a certain direction. The response means both transmission and reception, but in this discussion, the later is assumed. Beamforming capability is important in various engineering and scientific applications, such as, to increase channel capacity in communication through spatial multiplexing.

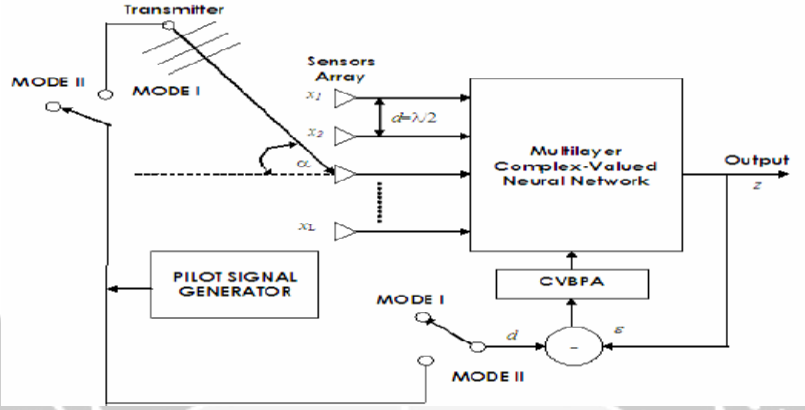


Fig. 9. Diagram block of the CVNN-based intelligent beamforming system based on Widrow's two-mode adaptation. During mode-I, the processor is connected to the sensors array, while in mode-II it is connected to pilot (teacher) signal.

It is assumed that the signal is narrowband and a quadrature scheme is used; therefore a complex signal is received by the sensor array. Widrow's 2-mode adaptation is adopted here, as displayed in Fig.9. Each element--the sensor, receives transmitted waves from the vicinity, with different delay (and noise) according to sensor's position in the array. The array should receive waves from given directions, and to reject interferences from another ones. The adaptation is performed by a complex-valued multilayer perceptron (CVMLP).

In the CVMLP, let $x_1 \ldots x_L$ denote the network input, $v_{11} \ldots v_{ML}$ denote the weight of the hidden layer, $w_{11} \ldots w_{NM}$ denote the weight in output layer, $y_1 \ldots y_N$ are the output of the hidden layer and $z_1 \ldots z_N$ are the network output. During the learning process, the weights are updated by using CVBPA (Complex-Valued Back Propagation Algorithm) as follows:

Output units:

$$w_{kj}^{new} = w_{kj}^{old} + \Delta w_{kj} \tag{20}$$

$$\Delta w_{kj} = \eta \left(t_k - z_k\right) \overline{f}'\left(O_k\right) \overline{y}_j \tag{21}$$

Hidden units:

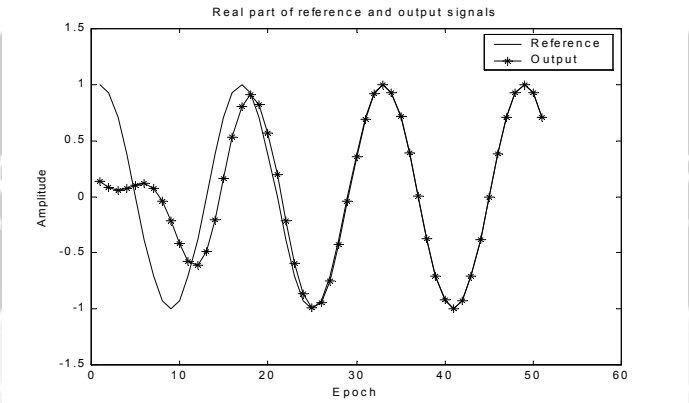
$$v_{ji}^{new} = v_{ji}^{old} + \Delta v_{ji} \tag{22}$$

$$\Delta v_{ji} = \eta \overline{x}_i \overline{f}'\left(H_j\right) \sum_k \delta_k w_{kj} \tag{23}$$

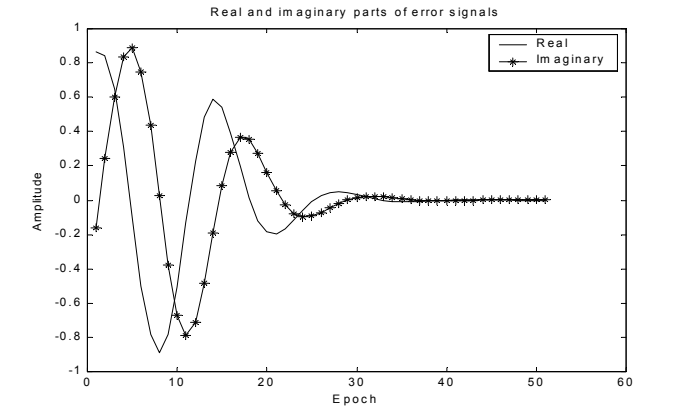
$$\delta_k = \left(t_k - z_k\right) \tag{24}$$

where η is the learning speed, t_k is the target signal, f' is the derivative of the activation function, and the bar above variables (or functions) indicates complex conjugate operation.

Fig.10 shows learning process during a neuro-beamforming session: (a) comparison between the output and reference---both are real parts, and (b) error decays of both real and imaginary parts, showing that the CVMLP works properly. Next, the performance of the neuro-beamformer is compared with CLMS (Complex-Valued Least Mean Square) algorithm with exact matrix method as a reference. Note that this exact solution cannot be obtained in the real situation.



(a)



(b)

Fig.10 Learning process of in the intelligent beam pointing by CVNN. The array adapts incoming signal from DOA +15^o: (a) Comparison between the output and the reference signals showing adaptation of the waveform. (b) Shows that both of the real and imaginary parts of the errors are decreasing during the adaptation process.

The simulation beamforming result is displayed in Fig.11: (a) initial array beampattern, (b) array pattern obtained by exact matrix method used as reference, (c) beampattern by CLMS method and (d) beampattern after CVNN method. It is obvious that the CVNN perform better than the CLMS.

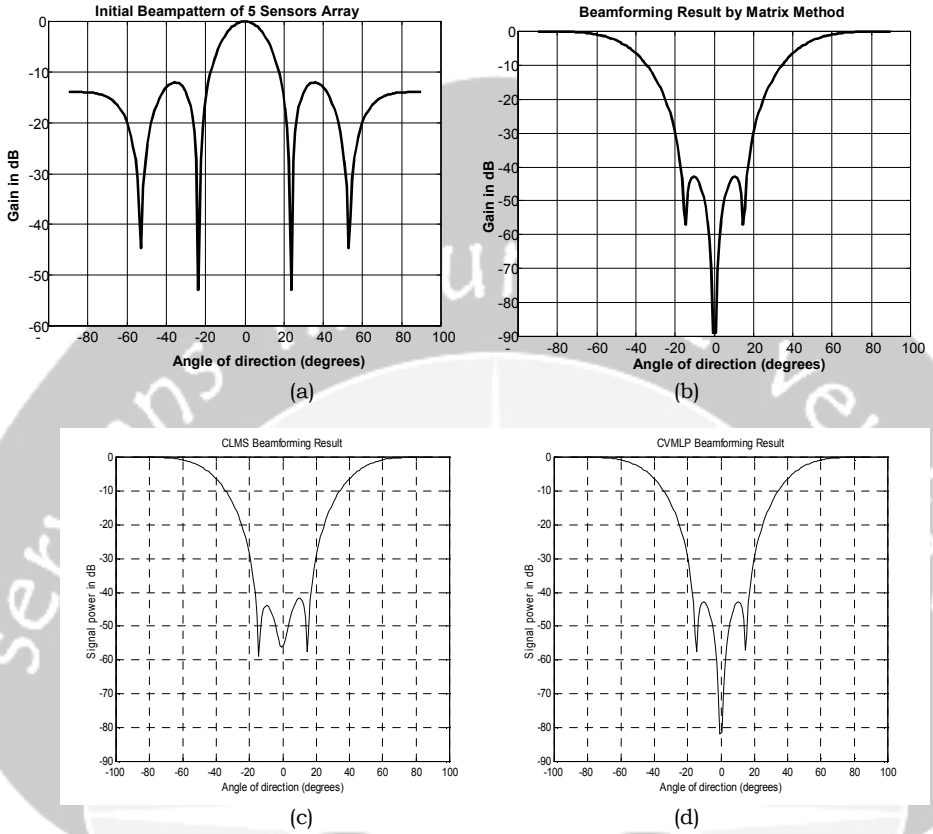


Fig.11:((a) initial array beampattern, (b) pattern obtained by exact matrix method, (c) beampattern by CLMS method, and (d) beampattern after CVNN method.

4 Conclusions

Various application of complex-valued signal processing has been discussed: InSAR image segmentation, complex-valued image denoising, and intelligent beamforming. The advantages of using the complex-valued algorithm are also shown.

Acknowledgments

The author would like to acknowledge Dr. Masanobu Shimada of JAXA (previously EORC-NASDA), Japan for supplying the InSAR image and Prof. J. Pauly of Information System Laboratory, Stanford University, USA, for a permission to use the MRI image.

References

[1] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol 69, No. 5, pp. 529-541, May 1981.

[2] A. Hirose Ed., *Complex-Valued Neural Networks – Theories and Applications*, Series on Innovative Intelligence, Vol. 5, World Scientific, 2003.

[3] B. Widrow, J.J. McCool, and M. Ball. "The complex LMS algorithm", *Proceeding of the IEEE*, pp. 719-720.

[4] A.B. Suksmono and A. Hirose, "Adaptive noise reduction of InSAR image based on Complex-MRF model and its application to phase unwrapping problem", *IEEE Trans. on Geoscience and Remote Sensing*, Vol. 40, No. 3, March 2002, pp. 699-709.

[5] A.B. Suksmono and A. Hirose, "Adaptive complex-amplitude texture classifier that deals with both height and reflectance for interferometric SAR images", *IEICE Transaction on Electronics*, ISSN 0916-8524, Vol. E83-C, no. 12, pp. 1905-1911, Dec. 2000, pp. 1912-1916.

[6] A.B. Suksmono and A. Hirose, "Interferometric SAR image restoration using Monte-Carlo Metropolis method", *IEEE Transaction on Signal Processing*, Vol. 50, No.2, Feb. 2002, pp. 290-298.

[7] A.B. Suksmono and A. Hirose, "Intelligent beamforming using a complex-valued neural network," *Journal of Fuzzy and Intelligent Systems*, Vol. 15, No. 3-4 (2004), IOS Press, pp.139-147.

[8] A.B. Suksmono and A. Hirose, "Beamforming of Ultra-Wideband Pulses by A Complex-Valued Spatio-Temporal Multilayer Neural Networks" *International Journal of Neural Systems*, Vol. 15, No. 1&2, April 2005, pp. 85-92.

[9] V. Tarokh, N. Seshadari, and A.R. Calderbank, "Space-time codes for high data rate wireless communication: Performance criterion and code construction," *IEEE Trans. On Information Theory*, Vol. 44, No.2, March 1998, pp.744-765.

ANDRIYAN B. SUKSMONO: Imaging & Image Processing Research Group and
The Radio Communication & Microwave Laboratory, Department of Electrical Engineering, Institut Teknologi Bandung, Jl. Ganesha No. 10, Bandung, Indonesia, Phone:+62-22-2501661, Fax:+62-22-2534133
Email: suksmono@ltrgm.ee.itb.ac.id, suksmono@yahoo.com

A STUDY OF SOME ASPECTS OF SPACE-TIME MODELS

Darwis S[a], Ruchjana B N[b],

[a] Institut Teknologi Bandung, Bandung, Indonesia
[b] Universitas Pajajaran, Bandung, Indonesia

Abstract. Studies on well placement consist of time series observations from several spatial locations. A correlated production time series are best considered as components of some vector stochastic process whose specification includes not only the serial dependence of each component but also the interdependent between component series. For example, in a reservoir performance study, the measurements might include production volume of two or three nearest wells, porosity, permeability and reservoir geometry. This paper proposes a space time approach t0 the problem of well placement optimization. A class of vector time series models is introduced to describe relationship among several wells production and proposes a simulation approach to forecast the performance of a new well. After an introduction of vector process, least squares parameter estimation procedures are discussed. Some useful notions such as spatial weight, time lag, and spatial lag are introduced. An empirical study is presented and some descriptions of production simulation is investigated.
Key-words: space-time models, least squares

1. Introduction

Optimal prediction of the location of new wells is a multidimensional problem that depends on fluid flow, reservoir geometry and economic criteria. The ideal procedure is the evaluation of well productivity which location is changed in each simulation. The approach would provide more reliable results, but demands computational efforts. Jatibarang volcanic oil reservoir is very complex having heterogeneities in horizontal and verical directions. These heterogeneities have to be taking into account in order to make reliable forecasts of future performance of a new well. Production strategy becomes more complex when horizontal wells are considered due to their interaction with reservoir. Aanonsen et al [1] proposed a method to optimized well locations based on expected reservoir performance. Oil production is used as a measure of reservoir response. Let R denote the reservoir response (e.g. oil production). The objective was to optimize the expected value for R. Regression and kriging were used to reduce the number of simulation run. Nakajima and Schiozer [4] propose quality map approach to guide horizontal well placement. The cumulative oil production of a well placed at a certain location is chose as the objective of the optimization process. Flow rate, viscosity, formation factor, oil density, oil mobility are the significant factors that affect well location. Pfeifer and Deutsch [6] illustrate the space-time modeling approach for Boston arrest data from the 14 areas and 72 periods. The forecasting model for the number of arrests was $Z(t) - Z(t-1) = -.812\varepsilon(t-1) + .092W\varepsilon(t-1) + \varepsilon(t)$. The forecasting of space time process was investigated by Giagomini and Granger [3]. Improved forecasting performance can be obtained by imposing spatial correlation.

A space time models for climate systems was proposed for assessing the consistency of numerical models with field observation [7]. The application of the forecasting methodology can be used for short-term prediction on time direction. This study proposes a hybrid of vector time series process and spatial simulation to forecast, in spatial direction, the production profile of a new well. The space-time parameters were used as a measure of well performance.

2. Vector time series

Consider N interrelated well response time series $\{Z(s_i,t), i=1,\ldots,N, t=1,\ldots,T\}$ with $E\,Z(s_i,t)^2<\infty$. If all $Z(s_i,t)$ were multivariate normal, then the distributional properties completely determined by the means $\mu(s,t)$ and covariance $\Gamma_{ij}(t,t+h)$

$$\mu(s_i,t)=EZ(s_i,t) \qquad \Gamma_{ij}(t,t+h)=E\left[\left(Z(s_i,t)-\mu(s_i,t)\right)\left(Z(s_j,t+h)-\mu(s_j,t+h)\right)\right]$$

Even when the observations are nonnormal, the means and covariance specify the second-order properties, the covariance measure the dependence between different spatial series and the dependence within the same series. When dealing with N interrelated series , it is more convenient to use vector notation

$$\mathbf{Z}(t)=(Z(s_1,t),\ldots,Z(s_N,t))^t,\ \boldsymbol{\mu}(t)=E\mathbf{Z}(t),$$
$$\boldsymbol{\Gamma}(t,t+h)=E\left[\left(\mathbf{Z}(t)-\boldsymbol{\mu}(t)\right)\left(\mathbf{Z}(t+h)-\boldsymbol{\mu}(t+h)\right)^t\right]=\left(\Gamma_{s_is_j}(t,t+h)\right)_{i,j=1}^{N}$$

A particularly important role is played by the class of stationary vector process. The series $\{\mathbf{Z}(t), t=0,1\ldots\}$ is to be stationary if $\boldsymbol{\mu}(t)$ and $\boldsymbol{\Gamma}(t,t+h)$ are independent of t. A stationary series will be denoted by $(\mathbf{Z}(t),\boldsymbol{\mu},\boldsymbol{\Gamma}(h))$. The simplest vector time series is white noise. The N-vector series $\boldsymbol{\varepsilon}(t)$ is said to be white noise with mean 0 and covariance Σ, $\boldsymbol{\varepsilon}(t)\sim \text{iid}(\mathbf{0},\boldsymbol{\Sigma})$, if and only if $\boldsymbol{\varepsilon}(t)$ is stationary with mean 0 and covariance Σ. Vector white noise is used as a building block from which can be constructed a variety of vector time series. The linear process is those of form

$$\mathbf{Z}(t)=\sum_{j=-\infty}^{\infty}\boldsymbol{\psi}_j\boldsymbol{\varepsilon}(t-j),\boldsymbol{\varepsilon}(t)\sim \text{iid}(\mathbf{0},\boldsymbol{\Sigma}),\boldsymbol{\Gamma}(h)=\sum_{j=-\infty}^{\infty}\boldsymbol{\psi}_{j-h}\boldsymbol{\Sigma}\boldsymbol{\psi}_j^t$$

The Vector process $\text{VAR}_N(1)$ process for a system of N wells is given by

$$\mathbf{Z}(t)=\boldsymbol{\Phi}\mathbf{Z}(t-1)+\boldsymbol{\varepsilon}(t),\boldsymbol{\varepsilon}(t)\sim \text{iid}(\mathbf{0},\boldsymbol{\Sigma})$$

and can be expressed as vector $\text{MA}_N(\infty)$

$$\mathbf{Z}(t)=\sum_{j=0}^{\infty}\boldsymbol{\Phi}^j\boldsymbol{\varepsilon}(t-j)$$

provided all the eigenvalues of $\boldsymbol{\Phi}$ are less than 1 in absolute value; i.e

$|\mathbf{I} - B\boldsymbol{\Phi}| \neq 0$ for all $B \in C$ such that $|B| \leq 1$. The VAR models represent all correlations and cross-correlation within and between the N time series. The vector autoregressive process for N = 3 wells located at s_1, s_2, s_3, $VAR_3(1)$, is given by

$$\begin{pmatrix} Z(s_1,t) \\ Z(s_2,t) \\ Z(s_3,t) \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{21} & \phi_{22} & \phi_{23} \\ \phi_{31} & \phi_{32} & \phi_{33} \end{pmatrix} \begin{pmatrix} Z(s_1,t-1) \\ Z(s_2,t-1) \\ Z(s_3,t-1) \end{pmatrix} + \begin{pmatrix} \varepsilon(s_1,t) \\ \varepsilon(s_2,t) \\ \varepsilon(s_3,t) \end{pmatrix}$$

or

$$\mathbf{Z}(t) = \boldsymbol{\Phi}\mathbf{Z}(t-1) + \boldsymbol{\varepsilon}(t)$$

The $VAR_3(1)$ model implies that current flow depend not only on the previous flow but also on the nearest wells [9]. There is a feedback relationship between the three processes; current flow will also be influenced by the flow performance at the previous period. The estimation of the parameter of a stationary vector process plays an important part in describing and modeling the dependence structure between the components of the process. Let $\boldsymbol{\Gamma}(k)$ be a covariance matrix for the vector $AR_N(1)$, note that $E\mathbf{Z}(t-k)\boldsymbol{\varepsilon}(t)^t = 0$ for k=1,2,...

$$\boldsymbol{\Gamma}(k) = E\mathbf{Z}(t-k)\mathbf{Z}(t)^t = \begin{cases} \boldsymbol{\Gamma}(-1)\boldsymbol{\Phi}^t + \boldsymbol{\Sigma} & k = 0 \\ \boldsymbol{\Gamma}(0)\left(\boldsymbol{\Phi}^t\right)^k & k = 1,2,\ldots \end{cases}$$

For k = 1, the parameter Φ, and error covariance, Σ, can be determined by solving the Yule-Walker equations

$$\boldsymbol{\Gamma}(0)\boldsymbol{\Phi}^t = \boldsymbol{\Gamma}(1) \Leftrightarrow \Phi = \Gamma^t(1)\Gamma^{-1}(0)$$

$$\Sigma = \Gamma(0) - \Phi\Gamma(0)\Phi^t$$

The STARMA model represent an extension of univariate process into the spatial domain and has proven useful in modeling flow histories of spatially located process when the relative spatial locations of the wells in the system can be used to help explain the correlative structure [5]. Consider the $STAR_N(1_1)$ model

$$\mathbf{Z}(t) = \phi_{10}\mathbf{Z}(t-1) + \phi_{11}\mathbf{W}\mathbf{Z}(t-1) + \boldsymbol{\varepsilon}(t),$$
$$E\boldsymbol{\varepsilon}(t) = 0, E\boldsymbol{\varepsilon}(t)\boldsymbol{\varepsilon}(t)^t = \sigma^2\mathbf{I}, E\mathbf{Z}(t)\boldsymbol{\varepsilon}(t+s)^t = \mathbf{0}, s = 1,2,\ldots$$

where $\mathbf{W}$ is the $N \times N$ matrix of spatial weights. The $STAR_N(1_1)$ can be written as linear model

$$\mathbf{Z}(t) = (\mathbf{Z}(t-1) \quad \mathbf{WZ}(t-1))\mathbf{\Phi} + \boldsymbol{\varepsilon}(t)$$

where $\Phi = (\phi_{10} \quad \phi_{11})^t$ and t=1,2,....,T. The least squares estimate of Φ is

$$\hat{\mathbf{\Phi}} = \begin{pmatrix} \sum_{t=1}^{T} \mathbf{Z}(t-1)^t \mathbf{Z}(t-1) & \sum_{t=1}^{T} \mathbf{Z}(t-1)^t \mathbf{WZ}(t-1) \\ \sum_{t=1}^{T} \mathbf{Z}(t-1)^t \mathbf{WZ}(t-1) & \sum_{t=1}^{T} \mathbf{Z}(t-1)^t \mathbf{W}^t \mathbf{WZ}(t-1) \end{pmatrix}^{-1} \begin{pmatrix} \sum_{t=1}^{T} \mathbf{Z}(t-1)^t \mathbf{Z}(t) \\ \sum_{t=1}^{T} Z(t-1)^t \mathbf{W}^t \mathbf{Z}(t) \end{pmatrix}$$
$$\cong \begin{pmatrix} \hat{\gamma}_{00}(0) & \hat{\gamma}_{10}(0) \\ \hat{\gamma}_{10}(0) & \hat{\gamma}_{11}(0) \end{pmatrix}^{-1} \begin{pmatrix} \hat{\gamma}_{00}(1) \\ \hat{\gamma}_{10}(1) \end{pmatrix}$$

where $\hat{\gamma}_{\ell k}(s)$ is the space-time autocorrelation defined as

$$\hat{\gamma}_{\ell k}(s) = \frac{1}{N(T-s)} \sum_{t=1}^{T-s} \mathbf{Z}(t)^t \mathbf{W}^t \mathbf{WZ}(t+s)$$

In STAR models, the parameters are independent of spatial locations. Ruchjana (2002) proposed a generalization of STAR with spatial varying parameters, denoted by GSTAR models. The vector $GSTAR_N(1_1)$ process for a system of N wells is given by

$$\mathbf{Z}(t) = \mathbf{\Phi}_{10}\mathbf{Z}(t-1) + \mathbf{\Phi}_{11}\mathbf{WZ}(t-1) + \boldsymbol{\varepsilon}(t)$$

where $\Phi_{10} = \text{diag}(\phi_{10}(s_1),\ldots,\phi_{10}(s_N))$ is the $N \times N$ diagonal matrix of autoregressive parameters, and $\Phi_{11} = \text{diag}(\phi_{11}(s_1),\ldots,\phi_{11}(s_N))$ is the $N \times N$ diagonal matrix of spatial dependence parameters. For N = 2 wells, $GSTAR_2(1_1)$ can be written as

$$\begin{cases} Z(s_1,t) = \phi_{10}(s_1)Z(s_1,t-1) + \phi_{11}(s_1)Z(s_2,t-1) + \varepsilon(s_1,t) \\ Z(s_2,t) = \phi_{11}(s_2)Z(s_1,t) + \phi_{10}(s_2)Z(s_2,t-1) + \varepsilon(s_2,t) \end{cases}$$

or

$$\mathbf{Z}(t) = \mathbf{\Phi}\mathbf{Z}(t-1) + \boldsymbol{\varepsilon}(t)$$

For N=3, and weight matrix

$$W = \begin{pmatrix} 0 & .6 & .4 \\ .7 & 0 & .3 \\ .6 & .4 & \end{pmatrix}$$

$GSTAR_3(1_1)$ is represented by

$$\begin{pmatrix} Z(s_1,t) \\ Z(s_2,t) \\ Z(s_3,t) \end{pmatrix} = \left[\begin{pmatrix} \phi_{10}(s_1) & 0 & 0 \\ 0 & \phi_{10}(s_2) & 0 \\ 0 & 0 & \phi_{10}(s_3) \end{pmatrix} + \begin{pmatrix} \phi_{11}(s_1) & 0 & 0 \\ 0 & \phi_{11}(s_2) & 0 \\ 0 & 0 & \phi_{11}(s_3) \end{pmatrix} \begin{pmatrix} 0 & .6 & .4 \\ .7 & 0 & .3 \\ .6 & .4 & 0 \end{pmatrix} \right] \begin{pmatrix} Z(s_1,t-1) \\ Z(s_2,t-1) \\ Z(s_3,t-1) \end{pmatrix}$$

$$= \left[\begin{pmatrix} \phi_{10}(s_1) & 0 & 0 \\ 0 & \phi_{10}(s_2) & 0 \\ 0 & 0 & \phi_{10}(s_3) \end{pmatrix} + \begin{pmatrix} 0 & .6\phi_{11}(s_1) & .4\phi_{11}(s_1) \\ .7\phi_{11}(s_2) & 0 & .3\phi_{11}(s_2) \\ .6\phi_{11}(s_3) & .4\phi_{11}(s_3) & 0 \end{pmatrix} \right] \begin{pmatrix} Z(s_1,t-1) \\ Z(s_2,t-1) \\ Z(s_3,t-1) \end{pmatrix}$$

or

$$\begin{cases} Z(s_1,t) = \phi_{10}(s_1)Z(s_1,t-1) + .6\phi_{11}(s_1)Z(s_2,t-1) + .4\phi_{11}(s_1)Z(s_3,t-1) \\ Z(s_2,t) = .7\phi_{11}(s_2)Z(s_1,t-1) + \phi_{10}(s_2)Z(s_2,t-1) + .3\phi_{11}(s_2)Z(s_3,t-1) \\ Z(s_3,t) = .6\phi_{11}(s_3)Z(s_1,t-1) + .4\phi_{11}(s_3)Z(s_2,t-1) + \phi_{10}(s_3)Z(s_3,t-1) \end{cases}$$

Alternatively $\text{GSTAR}_3(1_1)$ can be expressed in constrained $\text{VAR}_3(1)$ as

$$\mathbf{Z}(t) = \Phi\mathbf{Z}(t) + \boldsymbol{\varepsilon}(t) \qquad (1)$$

The estimate of $\boldsymbol{\Phi}$ is obtained by left-multiplying $\mathbf{Z}(t-1)$ on both sides of the transposed equation of (1) and then taking expectations

$$\boldsymbol{\Gamma}(0)\boldsymbol{\Phi}^t = \boldsymbol{\Gamma}(1)$$

3. Simulated annealing

Kriging interpolation involves the generation of images of the reservoir properties and commonly used to visualize reservoir heterogeneities. The method is designed for stationary spatial process and a large number of experience data used for modeling spatial correlation. Therefore, kriging techniques not well suited for reproducing geological reservoir patterns where the number of data are very limited. Simulated annealing (Deutsch, Journel, 1991) can reproduce reservoir attribute patterns, similar to those generated by other stochastic simulation methods, and developed in the area of optimization combinatorial problems. In the context of this paper, the components are the GSTAR parameters. The objective is to reproduce the pattern of spatial correlation of the model parameters based on experience production data modeled as GSTAR models. The essential components of simulated annealing are the objective function O, a procedure to update the objective function, a perturbation, and a procedure for reducing the system parameter. The procedure is consist of the following steps: generate an initial image, establish a schedule for lowering the control parameters, perturb the image, compute a new objective function, and establish the acceptance probability

$$p = \begin{cases} 1 & O_{new} \le O_{old} \\ e^{\left(\frac{O_{old} - O_{new}}{c}\right)} & O_{new} > O_{old} \end{cases}$$

If the perturbation is accepted then update the image and reset the objective to O_{new}. The objective is to reproduce an image that honors correlation model, and the data values at known locations, i.e., a match of realization correlation γ with correlation model γ_{model}

$$\sum_{h}\left[\frac{\gamma(h) - \gamma_{model}(h)}{\gamma_{model}(h)}\right]^2$$

4. Application

The application proposes a combination of vector time series process and spatial simulation to forecast the GSTAR parameters of new well locations. It is supposed that the experience wells locations are fixed, while the new well varied in the xy-plane. Let s = (x, y) denotes the location of new well. GSTAR and simulated annealing are used to describe production profile of a new well. Table 1 shows the parameters of GSTAR from three production wells. Table 2 shows the simulated annealing results on $nx \times ny = 7 \times 10 = 70$ grids of siz $xsiz \times ysiz = 100 \times 200 = 20\,000$. The initial spatial correlation was sph(a = 1700 , c = .0035) for ϕ_{10} and sph(a = 1700, c = .0037). The spatial image γ_{model} was obtained by minimizing the objective function

$$\sum_{h}\left[\frac{\gamma(h) - \gamma_{model}(h)}{\gamma_{model}(h)}\right]^2$$

The initial image is modified by swapping pairs of grids s_i, s_j chosen at random.

The patterns of simulated annealing were quite similar to the nearest to the experience wells. The autoregressive parameters ϕ_{10} exhibit fairly stable behaviour. The spatial parameters ϕ_{11} were of larger variation due to large variation in experience data.

Assuming constant flowing pressure, the relationship between flow rate and time for producing wells is

$$dq(t) = -\beta q(t)^{b+1} dt$$

where b and β , 0 < b < 1 , are empirically determined constants. Solutions to flow rate equations show the expected decline in flow rate as the production time increases. Three declines curves have been identified based on the value of b. The exponential decline curve corresponds to b = 0 has the solution

$$q(t) = q(0)e^{-\beta t}$$

where q(0) is initial rate and β is a factor that is determined from field data. The hyperbolic decline curve corresponds to a value b in the range 0 < b < 1 has the the form

$$q(t)^{-b} = b\beta t + q(0)^{-b}$$

The harmonic decline curve corresponds to b = 1 has the form

$$q(t)^{-1} = b\beta t + q(0)^{-1}$$

The exponential decline $\ell n\, q(t) = \ell n\, q(0) - \beta t$ has the form $Y - \overline{Y} = \beta(X - \overline{X}) + \varepsilon$ for a straight line with slope β. Cumulative production decline curve is the integral of the rate from the initial rate q(0) at t = 0 to the rate q at time t. The cumulative production for the exponential decline rate is

$$\int_0^t q(t)\, dt = \frac{q(t) - q(0)}{\beta}$$

The exponential decline factor is found from equation

$$\beta = -\frac{1}{t} \ln \frac{q(t)}{q(0)}$$

Assuming constant flowing pressure and relationship of the cummulative production and decline rate, the optimization of well locations reduces to

$$s_{opt} = \arg\max_{s} \phi_{11}(s)$$

The proposed solution for well placement problem is $s_{opt} = (4,6)$.

Table 1. Input for simulated annealing, the GSTAR parameters $\phi_{10}(s), \phi_{11}(s)$ from three well locations

Well s_i	Coordinate (x_i, y_i)	$\phi_{10}(s_i)$	$\phi_{11}(s_i)$
1	(4,2)	.963	.078
2	(7,5)	.852	.005
3	(1,10)	.943	.078

Table 2. Simulated annealing of GSTAR parameters $\phi_{10}(s), \phi_{11}(s)$, nx = 7, ny = 10, gridsize = 70, xsiz = 100, ysiz = 200, blocksize = 20 000. The ϕ_{10} parameters are smoother compared to the ϕ_{11} parameters

.929,.081	.949,.093	.918,.061	.877,.038	.907,.073	.945,.098	.944,.073
.935,.093	.955,.106	.894,.045	.860,.022	.878,.058	.949,.108	.952,.083
.936,.118	.958,.119	.875,.025	.852,.005	.870,.032	.955,.125	.957,.100
.955,.114	.963,.125	.962,.006	.852,.005	.853,.016	.963,.125	.960,.110
.956,.118	.963,.852	.963,.006	**.852,.963**	.854,.955	.963,.935	.961,.113
.956,.117	.963,.125	.963,.006	.852,.005	.853,.011	.963,.886	.960,.113
.953,.112	.963,.125	.853,.005	.852,.005	.852,.007	.961,.125	.958,.112
.946,.108	.958,.123	.876,.026	.852,.005	.873,.025	.958,.119	.955,.104
.939,.092	.956,.107	.901,.049	.865,.023	.889,.052	.954,.108	.951,.085
.937,.079	.952,.098	.922,.067	.886,.034	.910,.072	.949,.097	.943,.075

5. Summary

This paper presents the spatial prediction of space-time models and its application in the determination of optimal well locations. A GSTAR model was proposed as a model for oil production time series data. The parameter was estimated using least squares approach. The production profile of a new well was interpolated using simulated annealing technique.

Acknowledgment

The authors thank Jose Rizal for providing simulated annealing computation of GSTAR parameters

References

[1]. Aanonsen S I, Eide A L, Holden, Aasen J O, 1995, Optimizing reservoir performance under uncertainty with application to well location, *Society of Petroleum Engineers*, SPE 30710

[2]. Deutsch C V, Journel A G, 1991, The application of simulated annealing to stochastic reservoir modeling, *SPE Advanced Technology Series*, 2:2

[3]. Giacomini R, Granger C W, 2002, *Aggregation of space-time process*, University of California

[4]. Nakajima L and Schiozer D J, (2003), *Horizontal well placement optimization using quality map definition,* Canadian International Petroleum Conference

[5]. Pfeifer, P.E and S. J. Deutsch. (1980), A Three-Stage Iterative Procedure for Space-Time Modeling, *Technometrics*, 22:1, 35-47

[6]. Pfeifer, P.E and S. J. Deutsch (1980), Identification and interpretation of first order space time ARMA models, *Technometrics.*,22:3, 397-408
[7]. Niu X F, McKeague I W, Elsner J B, [1999], *Seasonal space-time for climate system*, Florida State University
[8]. Ruchjana B N. [2002], *Generalized space-time autoregressive models for well production*, Institut Teknologi Bandung
[9]. Wei W.W.S. (1990), *Time Series Analysis*, Addison-Wesley

Sutawanir DARWIS: Department of Mathematics, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia. e-mail:sdarwis@dns.math.itb.ac.id

Budi Nurani Ruchjana: Department of Mathematics, Padjadjaran University, Jl Raya Bandung-Sumedang Km 21, Jatinangor, Indonesia, email: bnurani@yahoo.com

COINTEGRATION AND CAUSALITY BETWEEN ECONOMIC VARIABLES AND ELECTRONIC OUTPUT

Z. Arsad, Y. Y. Au, W. S. Tham

Universiti Sains Malaysia, Penang, Malaysia

Abstract. This paper investigates the short and long run dynamics between macroeconomic variables and electronic industry output in Malaysia during the period of 1991-2004. Therefore, this study attempts to look at factors affecting the performance of electronic industry. The methodology employed is the standard cointegration analysis and the vector error correction model. The findings of the study suggest that in the long-run macroeconomic variables such as money supplies M1, M2 and M3; interest rates on loan and deposit, foreign currency exchanges, consumer price index and industrial index share long run equilibrium with output of electronic industry. Therefore, government policies that take into account all these variables may be able to predict future performance of electronic industry in Malaysia. In the short run there is significant evidence for unidirectional causality from Japanese Yen, Singapore Dollars and M2 to the electronic output. In addition, average lending rate may also play a small role in the performance of electronic industry.
Key-words: macroeconomic variables, unit root test, cointegration

1 Introduction

Electronic industry is the largest component of the overall manufacturing sector in Malaysia. In general, Malaysia is the main exporter of electronic products in the world. The identification of various factors that affect a certain industry is a major concern in theory and practice. This assessment is particularly relevant in the electronic sector, given the importance and size of this industry in Malaysia, as well as its relationship with other macroeconomic variables.

During the period 1998-2004, the percentage of exported electronic components from the overall exported manufactured products has been consistently high, in the range of 65% to 70%. Over the past few years the electronic industry has experienced some drastic changes due to changes in economic environment, in particular the Asian Economic Crisis, the drastic fall of interest rates, the change of consumer demands and the evolution of high technology on the electronic products. In addition, the manufacturing sector has been one of the main targets of the Malaysian government for fiscal and monetary policy aimed at achieving low inflation and unemployment and also to be less dependent on the agricultural sector.

It is widely known that economic environments have a profound effect on the growth of many industries including the electronic industry. History has also shown that performance of the electronic industry is closely linked to macroeconomic variables. There are many literature reviews on the effect of macroeconomic variables on industries. Studies on the effect of macroeconomic variables on various industries include [12] and [1] for the housing industry, [8]

and [3] for the Cotton, and Pulp and Paper industry respectively; and [2], [10] and [11] on the life insurance industry.

Other than works on the life insurance industry as documented, for example, by [11], econometric studies on other industries in Malaysia are rather limited. Therefore, the goal of this paper is to empirically investigate the short- and long-run relationship of macroeconomic variables on the total electronic output. Vector error correction model (VECM) and Granger-causality analysis are used to investigate the relationship. The remainder of the paper is organized as follows: Section 2 presents the methodology and empirical framework, Section 3 describes the data used, results of the analysis and discussion, while Section 4 provides some concluding remarks.

2 Empirical Framework

The empirical framework of this study is based on the standard methods of cointegration and vector autoregressions (VAR). The analysis begins with exploring the integration and cointegration properties of the variables before employing the unrestricted vector autoregression model. The results from cointegration test enable us to model short run dynamic interactions among the variables within the VAR system. If the variables are found to be non-stationary and non-cointegrated, the dynamic interactions among the variables are assessed according to the standard VAR model with variables expressed in first difference. Conversely, if the variables are found to be cointegrated, error correction models should be employed.

Since the influential work of [9], it has been widely accepted that most macroeconomic variables are non-stationary or are integrated series. A variable is said to be integrated of order d (written as I(d)) if it requires differencing d times to achieve stationarity. Therefore, any variable that is integrated of order 1 or higher is non-stationary. The findings that the variables are individually non-stationary integrated of the same order is a necessary condition for establishing the presence of cointegration among the variables. A set of variables is said to be cointegrated if they are non-stationary integrated of the same order and yet their linear combination is stationary. The evidence for cointegration suggests that the variables cannot drift farther away from each other arbitrarily. Any deviations of a variable from the long run relationship will result in some variables adjusting to return back to the long run path.

The order of integration of the series can be determined by conduction the Augmented Dickey-Fuller (ADF) unit root test. The ADF test make allowance for higher order correlation by assuming that the series follows an AR(p) process. The test is carried out based on the regression:

$$\Delta y_t = \alpha_0 + \alpha_1 t + \alpha_2 y_{t-1} + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + \varepsilon_t$$

where y_t is the series under investigation. The regression with a drift and time trend is normally used to account the possibility of non-zero means and the presence of deterministic trend in the series. The null hypothesis for a unit root test can be written as:

$$H_0 : \alpha_2 = 0 \qquad \text{vs.} \qquad H_1 : \alpha_2 < 0$$

If the null hypothesis of unit roots is rejected, then the level of the series is said to be integrated of order one, I(1). The maximum lag length p required for serial correlation correction in the ADF regression is determined on the basis of evidence provided by Akaike Information Criterion (AIC).

Having established that each of the series is non-stationary, analysis proceeds to examine whether there exists some long run equilibrium relationship between electronic industry products and variables of interests. A vector autoregressive approach (VAR-based) of [5,6] and [7] is used to test for cointegration. The technique uses maximum likelihood procedure to determine the number of cointegrating vectors among a vector of time series, and derive a likelihood ratio test for the hypothesis that there is a given number of these cointegrating vectors. Consider a k-dimensional vector model:

$$y_t = \phi_0 + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t$$

and this equation can be rewritten into a vector error correction model (VECM) as follows:

$$\Delta y_t = \kappa_0 + \sum_{i=1}^{p-1} \kappa_i \Delta y_{t-i} + \vartheta_k y_{t-k} + \upsilon_t \qquad (1)$$

where κ is the short run dynamics that reflects the immediate response of the dependent variable to a change in explanatory variables. The matrix ϑ represents the long run relationship between the variables. The number of cointegrating vectors is determined by the rank of the ϑ matrix. The matrices ε_t and υ_t are a vector of normal distributed error with zero mean and constant variance.

Two types of likelihood ratio tests are applied to determine the number of cointegrating vectors, namely maximum eigenvalue and trace test statistics. The trace test is a likelihood ratio test for maximum r cointegrating vectors against the alternative equals to p, the number of variables:

$$\lambda_{\text{trace}} = -N \sum_{i=r+1}^{k} \ln\left(1 - \hat{\lambda}_i\right)$$

where N is the number of observations and $\hat{\lambda}_i$ are estimated eigenvalues of ϑ matrix. The maximum eigenvalue test has an identical null hypothesis as trace test, with the alternative hypothesis of (r+1):

$$\lambda_{\max} = -N \ln\left(1 - \hat{\lambda}_{r+1}\right)$$

As noted by [4], if two series are individually I(1) and cointegrated, a causal relationship exist in at least direction. The authors suggest the use of VECM to examine the short run dynamic, long-run equilibrium and Granger causality relationship between variables. The Granger reparameterization of (1) leads to:

$$\Delta y_t = \kappa_0 + \sum_{i=1}^{p-1} \kappa_i \Delta y_{t-i} + \varphi \varepsilon_{t-1} + \upsilon_t$$

ε_{t-1} is called the error correction term and is obtained from the cointegration equation (1) above. This term represents the response of the dependent variable for any departure from equilibrium. That is φ measures the extent to which any disequilibrium at time (t-1) is compensated for at time t. The matrix κ_i (k x k) is the short run adjustment coefficient and is used to investigate short run interactions between the variables. The null hypothesis of no causation from j-th explanatory variable to dependent variable i, i.e $\kappa_{1j} = \kappa_{2j} = \ldots = \kappa_{p-1,j} = 0$ can be investigated using an F-test. Failing to reject the null hypothesis implies that the j-th variable does not Granger cause the dependent variable.

3 Data, Analysis and Results

The data used in this paper are monthly Electronic Component produced (EC), Monetary Aggregates (M1, M2 and M3), Consumer Price Index (CPI), Index of Industrial Production (IPP), Interest Rates on Saving Deposits (DEP), Average Lending Rate (ALR), Exchange Rates of Singapore Dollars, British Pound Sterling and Japanese Yen (SGD, STL and YEN). The data was obtained from the Monthly Statistical Bulletin of Bank Negara Malaysia and Department of Statistics Malaysia. The data covers the period from January 1991 to June 2004, a total of 162 observations. All the data series are expressed in logarithmic form for the analysis.

The analysis and findings of ADF Unit Root test is presented in Table 1. Generally, the results indicate that most variables are non-stationary at level, both with and without trend, with the exception of electronic output (Log EC), monetary supply (Log M3) and consumer price index (Log CPI). However, all the series depict a stationary pattern after first-differencing. In other words, these variables are integrated of order 1 or I(1) and therefore we can proceed to the Johansen and Juselius cointegration test to investigate long run dynamic among the variables.

Table 2 reports the Johansen and Juselius cointegration test statistics. As can be seen in the table, there are three cointegration vectors, at 1% significant level) between the electronic output (EC) and explanatory variables, suggesting the electronic output in Malaysia is cointegrated with Monetary aggregates, interest rates, Consumer and Industrial Indices and foreign currency exchanges. Therefore the results suggest a long run equilibrium between electronic output and the explanatory variables.

Since all the variables are cointegrated, the vector error correction model can be established. The vector error correction model (VECM) approach allows us to distinguish between short run and long run dynamics between the variables. The estimated model is reported in Table 3. The result clearly shows significant negative error correction term at 5% level, implying that the dependent variable, Log EC, has the tendency to adjust to any deviations from long run equilibrium as represented by the cointegration relationship. The estimated coefficient of the error

term indicates that 48.3% of the system disequilibrium is corrected within a single month.

Table 1. Results for ADF Unit Root test

Variable	Level		First Difference	
	Without trend	With trend	Without trend	With trend
Log EC	-0.0810 [2]	-3.3555*[2]	-14.1838*[1]	-14.1640*[1]
Log M1	-1.0764 [2]	-2.2289 [2]	-9.2604* [1]	-9.2498* [1]
Log M2	-2.7562 [2]	-0.8336 [3]	-7.4928* [1]	-7.9406* [1]
Log M3	-3.2782*[3]	-1.0594 [3]	-5.2795* [2]	-6.2379* [2]
Log DEP	-0.4517 [2]	-1.3744 [2]	-8.7444* [1]	-8.8494* [1]
Log ALR	-0.7049 [2]	-2.1755 [3]	-5.5421* [1]	-5.9157* [1]
Log CPI	-3.2554*[1]	-0.5013 [1]	-7.6139* [1]	-8.4072* [1]
Log IPP	-1.4704 [3]	-2.6910 [3]	-8.7592* [2]	-8.7907* [2]
Log SDG	-1.4440 [3]	-2.2027 [3]	-5.7930* [2]	-5.7926* [2]
Log STL	-0.7666 [1]	-2.5256 [1]	-8.3237* [1]	-8.3978* [1]
Log YEN	-1.3470 [1]	-2.8417 [3]	-6.1686* [3]	-6.1536* [3]

Note: * indicates significance at 5% level, [ ] represents optimal lag

Table 2. Johansen and Juselius Cointegration Test

H_0	Eigenvalue	Trace-stats
$r = 0$	0.4455	368.457**
$r \leq 1$	0.3306	274.710**
$r \leq 2$	0.2777	210.898**
$r \leq 3$	0.2541	159.180*
$r \leq 4$	0.1866	112.564
$r \leq 5$	0.1319	79.722
$r \leq 6$	0.1241	57.239

Note: ** and * indicate rejection of null hypothesisat the 1% and 5% levels of significance

The Granger causality test was carried out to investigate the short run dynamics between the variables. The p-values of the F-test in the final column of Table 3 suggest that electronic output is Granger-caused by Monetary aggregate M2, Singapore Dollars and Japanese Yen. The p-value also provide little evidence that average lending rate may play a small role in influencing the output of electronic industry in Malaysia.

Table 3. Estimation of Error Correction Model and Granger Causality Test

Variable	Coefficient	t-stats	Lag	p(F-stats)
Constant	0.0254	1.471	-	-
Log EC	-0.1089	-1.267	2	0.398
Log M1	0.2253	0.821	4	0.403
Log M2	2.1009	1.485	3	0.032
Log M3	2.1340	2.732*	5	0.101
Log DEP	-0.0178	-0.116	4	0.670
Log ALR	0.6053	1.518	5	0.061
Log CPI	-2.3640	-0.774	4	0.740
Log IPP	0.0432	0.209	4	0.212
Log SGD	-0.3220	-0.486	3	0.034
Log STL	0.1035	0.453	4	0.605
Log YEN	0.2747	1.071	5	0.004
ErrorCT	-0.4830	-3.915*	-	-

Note: * indicates significance at 5% level

4 Conclusion

This paper attempts to identify factors that influence the performance of electronic industry in Malaysia during the period 1991-2004. Both financial and economic variables are considered and their long- and short-run dynamics are investigated using cointegration techniques. Financial variables consist of interest rates on deposit account of commercial bank. Meanwhile, average lending rate, Monetary aggregates, consumer price index, industrial production index and foreign currency exchanges are the economic variables.

The empirical results suggest significant long-run interaction between electronic output and both of the financial and economic variables. This suggests that these explanatory variables cannot drift too far apart and therefore may be considered to be dependent of each other. The finding from VECM model indicates a reasonably fast speed of adjustment for financial and economic variables to fully capture the shock in the output of electronic industry. The results of causality analysis indicate that the electronic output is influenced by money supply, Singapore Dollars, Japanese Yen and to a certain extent by the average lending rate.

References

[1] Baffoe-Bonnie, J. (1998). The dynamic impact of macroeconomic aggregates on housing prices and stock of houses: A national and regional analysis. *The Journal of Real Estate Finance and Economics*, **17**, 179-197.

[2] Campbell. R. A. (1980). The demand for life insurance: an application of the economic of uncertainty. *Journal of Finance*, **35**, 1155-1172.

[3] Chao, W. S. Brongiorno, J. (2002). Exports and growth: a causality analysis for the pulp and paper industries based on international panel data. *Applied Economics*, **34**, 1-13.
[4] Engle, R. F. Granger, C. W. J. (1987). Cointegration and error-correction representation, estimation and testing. *Econometrica*, **55**, 251-276.
[5] Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, **12**, 231-254.
[6] Johansen, S. (1991). Estimation and hypothesis testing of cointegrating vectors in Gaussian vector autoregressive models. *Econometrica*, **59**. 1151-1580.
[7] Johansen, S. Juselius, K. (1990). Maximum likelihood estimation and inference on cointegration with application to the demand for money. *Oxford Bulletin of Economics and Statistics*, **52**, 169-211.
[8] Meyer, L. (1999). An economic analysis of U.S total fiber demand and cotton mill demand. *Cotton and Wool Situation and Outlook Yearbook*, 23-28.
[9] Nelson, C. R. Plosser, C. I (1982). Trends and random walks in macroeconomic time series: some evidence and implications. *Journal of Monetary Economics*, **10**, 139-162.
[10] Proper, C. (1989). An econometric analysis of the demand for private health insurance in England and Wales. *Applied Economic*, **21**, 777-792.
[11] Redzuan, H. Yaakob, R. (2004). Factors affecting the life insurance demand in Malaysia. *Proceedings of the Malaysian Finance Association 6th Annual Symposium*, Langkawi, 358-367.
[12] Schwab, R. (1983). Real and nominal interest rates and the demand for housing. *Journal of Urban Economics*, **13**, 181-195.

Appendix A: Details of the Authors

Zainudin Arsad: School of Mathematical Sciences, Universiti Sains Malaysia, 11800 Minden, Penang, Malaysia. Fax: 00-60-4-657-0910.
E-mail: zainudin@cs.usm.my

Au Yuen Yee: BSc. student at School of Mathematical Sciences, Universiti Sains Malaysia, 11800 Minden, Penang, Malaysia. (Graduating Aug. 2005)
E-mail: mindy_214@yahoo.com

Tham Wai See: BSc. Student at School of Mathematical Sciences, Universiti Sains Malaysia, 11800 Minden, Penang, Malaysia. (Graduating Aug. 2005)
E-mail: jctham@yahoo.com.hk

Supperreplication Methods for Single-Asset Barrier Options with Uncertain Volatility

Komang Dharmawan

Dept. of Mathematics, Udayana University, Bali, Indonesia

Abstract. The aim of this paper is to study single-asset barrier options, where the volatility of the stocks is assumed to define an interval-valued bounded stochastic process. The bounds on the volatility may represent, for instance, the extreme values of the volatility of traded options. As the volatility is not known exactly, the value of the option can not be determined. Nevertheless, it is possible to calculate extreme values. We show that these values correspond to the best and the worst case scenarios of the future volatility for short positions and long positions in the portfolio of the options. Our main tools are the equivalence of the option pricing, a certain stochastic control problem, and the resulting concept of super-hedging. This concept has been well known for some time but never applied to barrier options. Using rather standard arguments, we derive the Hamilton-Jacobi-Bellman equation for the value function. Then we define the super price and super-hedging strategy for the barrier options and show equivalence with the control problem studied above. The superprice price can be found by solving the nonlinear Hamilton-Jacobi-Equation studied above. It is called sometimes the Black-Scholes-Barenblatt (BSB) equation. This is the Hamilton-Jacobi-Bellman equation of the exit control problem. We include an example: pricing and hedging of a single-asset barrier option and its numerical solution using the finite difference method.

Key-words: Single-asset Barrier Option, Stochastic Volatility, Super-replication Method, Stochastic Optimal Control, Hamilton-Jacobi-Bellman equation

1 Introduction

Barrier option have become increasingly popular in the financial markets as they are less expensive compare to standard options like in Black and Scholes [3] and they are also valuable tools for trading and hedging in many situations (see [4], [16] or [26]). The pricing of barrier options is well understood when the volatility is assumed to be constant. The closed-form solutions for down-and-out options have been provided by Rich [23] and Rubinstein and Reiner [25].

The motivation of this work comes from the study of Mathematical Finance. In this paper we describe the financial strategy (super-replicating strategy) that a trader can follow in order to manage his/her model risk for barrier options. Suppose that the trader precisely knows the model followed by the real market, and that this model is given by a stochastic differential equation. Then she/he is able to construct a strategy which perfectly hedges the option. This seems to be unrealistic

for some reasons, for example: choice of the modelling stochastic processes, the assumption of constant volatilities, etc.

When one has a rather precise information on the model of the market, then one can take advantage of the robustness of formula of Black and Scholes type (see El Karoui, Jeanblanc-Picque and Shreve [11] and Romagnoli and Vargiolu [24]). When one has only a vague information on the model of the market, what strategies that he/she can use to protect their positions? To answer this questions, Avellaneda, Levy, and Paras [1] and Avellaneda and Paras [2], Romagnoli and Vargiolu [24], Gozzi and Vargiolu [14] propose super-replication methods to solve the problems.

In the case of barrier options or other options where the payoff function are non-convex, the method proposed by El-Karoui, Jeanblanc-Picque, and Shreve [11], is not applicable. This is due to the fact that the barrier option price may not increase monotonically with volatility. Moreover, the value function of the option is neither convex nor concave. To sell barrier options, one generally trades them above their theoretical Black-Scholes price. Another method used to hedge and price barrier options is by static hedging. This strategy does not involve continuous re-balancing as in dynamic hedging. Such static hedging normally involves setting up a portfolio at the beginning of the contract that is guaranteed to match the payout of the options to be hedged(see Derman, Ergener, and Kani [8]).

In this paper, we analysis the robustness of European single-asset barrier options. Our work is motivated by Avellaneda, Levy, and Paras [1] and Avellaneda and Paras [2], Romagnoli and Vargiolu [24], Gozzi and Vargiolu [14], but we discuss hedging strategy of a single-asset barrier option, an option governed by a one dimensional diffusion process. We also assume that the volatility is not known, so the market model is incomplete. This reflects that there are many choices for derivatives asset prices that can exist in an uncertainty market. The source of the uncertainty mainly comes from unpredictable volatility.

By assuming that the value function of the exit control problem v is continuous with respect to time t and space of price x, and is regular enough to apply the Ito formula, we show that the pair (v, Δ) is a superstrategy, Theorem 3.5. Furthermore, the exit control problem is a 'bang-bang' solution, see Theorem 4.3. By Lemma7.4 in [5], $\Delta_t = \frac{\partial}{\partial S} v(t, S_t)$ is bounded. Therefore, choosing Δ as a superstrategy is valid and makes sense.

2 Uncertain Volatility

According to Arbitrage Pricing Theory, if the market presents no arbitrage opportunities, there exists a probability measure on future scenario such that the price of any security is the expectation of its discounted cash-flows, Duffie [10]. Such a probability is known as a *martingale measure*, Harisson and Pliska [15]. It is true that pricing measure is often difficult to calculate precisely and there may exist more than one measure which is consistent with a given market, Avellaneda, Levy,

and Paras [1]. Based on this fact, it is useful to view incomplete markets as they are reflecting the many choices for derivatives asset prices that can exist in an uncertainty market. The source of the uncertainty mainly comes from unpredictable volatility. Avellaneda, Levy, and Paras [1] and Avellaneda and Paras [2], assume that the underlying asset S_t follows a diffusion process with non-constant interest rate and volatility

$$dS_t = r_t S_t dt + \sigma_t S_t dW_t. \tag{1}$$

The volatility function σ_t is known and fluctuates within an interval

$$0 < \sigma_{min} \leq \sigma_t \leq \sigma_{max} \tag{2}$$

The volatility process (σ_t) that satisfies (1) induces a unique probability measure $\mathbb{Q} = \mathbb{Q}_t^\sigma$ on the space of prices S_t. Let Σ denote the set of all measures that can be induced within the constraint (2).

Now we define a contingent claim for the barrier option.

Definition 2.1. A price process $\{V_t; 0 \leq t \leq T\}$ for the barrier option is any adapted process satisfying

$$V_T = h(S_T)\mathbf{1}_{\{\tau > T\}}$$

where $h : \mathbb{R}_+ \to [0, \infty)$ is a given function and τ is the *first hitting time* of the process S_t on the barrier H defined by

$$\tau = \inf\{t > 0; \quad S_t \geq H\},$$

where S_t is the solution of (1).

The payoff function h is not necessarily a convex function. It can be mixed between convex or concave function.

Definition 2.2. The option price at time $T - t$ with initial stock price $S_t = x$ is given by

$$v(t, x) = \mathbb{E}\left[h(S_T)\mathbf{1}_{\{\tau > T\}} \,|S_t = x\right], \quad 0 \leq t \leq T$$

with terminal and boundary condition

$$\begin{aligned} v(T, x) &= h(x), \qquad x < H \\ v(t, x) &= 0, \qquad\quad x \geq H \end{aligned}$$

Then, by Ito's theorem, the problem can be converted into a nonlinear partial differential equation (DPP), which is a version of the Black-Scholes-Barenblatt (BSB) equation (see [19]), where $\mathbb{E}^{\mathbb{Q}}$ is the expectation operator with respect to the measure $\mathbb{Q}$ and the dynamic price process (1).

3 Stochastic Controls and replicating strategy

We assume that the volatilities are stochastic, but restricted to move within an admissible interval Σ. In the real situation the agent does not know the true volatilities, instead he uses another model, that is,

$$dS_t = r_t S_t dt + \gamma_t S_t dW_t.$$

to hedge the contingent claim, where $\gamma \in \Sigma$ is a certain admissible volatility, $S_t = x$ is the initial condition. The volatility can be interpreted as a control to find the worst or best case price of the single asset barrier option. In this model, there are two sources of uncertainty, that is W_t and the volatility γ. Since the agent does not know these two objects, he will estimate the fair price of the claim within the interval price, which is known as the interval of admissible prices. The arbitrage free price of the barrier option is given by

$$v_t^\gamma = \mathbb{E}_{\mathbb{Q}} \left[h(S_{T-t}^{t,x,\gamma}) \mathbf{1}_{\{\tau > (T-t)\}} \right].$$

Since we do not know yet whether our contingent claim is attainable or not. Therefore, we expect that the arbitrage free price of the claim lies in the interval

$$v^-(t,x) \leq v_t^\gamma \leq v^+(t,x), \quad 0 \leq t \leq T,$$

where

$$v^+(t,x) = \sup_{\gamma \in \Sigma} \mathbb{E} \left[h(S_{T-t}^{t,x,\gamma}) \mathbf{1}_{\{\tau > (T-t)\}} \right]$$

and

$$v^-(t,x) = \inf_{\gamma \in \Sigma} \mathbb{E} \left[h(S_{T-t}^{t,x,\gamma}) \mathbf{1}_{\{\tau > (T-t)\}} \right].$$

Now we adopt the definition of a replicating strategy for unspecified volatilities given in Touzi [28], Karatzas [18], or Frey [13].

Definition 3.1. A super-replicating price for a contingent claim h at time t is given by

$$\begin{aligned} \bar{v}(t,x) \quad = \quad & \inf \{ x \geq 0 \ \mid \ \forall \gamma \in \Sigma \ \exists \Delta \text{ admissible, such that} \\ & X_T^{t,x,\Delta} \geq h(S_{T-t}^{t,x,\gamma}) \mathbf{1}_{\{\tau > (T-t)\}}; \mathbb{P} - a.e., \Big\}. \end{aligned}$$

This is the minimum price that the agent can accept in order to super-replicate the claim. Any such process Δ, which may depend on γ, is called a super-replicating strategy or superstrategy.

Definition 3.2. A sub-replicating price for the contingent claim h at time t is given by

$$\begin{aligned} \underline{v}(t,x) \quad = \quad & \sup \{ x \geq 0 \ \mid \ \forall \gamma \in A(\Sigma) \ \exists \Delta \text{ admissible, such that} \\ & X_T^{t,x,\Delta} \leq h(S_{T-t}^{t,x,\gamma}) \mathbf{1}_{\{\tau > (T-t)\}}; \mathbb{P} - a.e., \Big\}. \end{aligned}$$

Any such process Δ, which may depend on γ, is called a sub-replicating strategy or substrategy.

Remark 3.3. Another version of super-replicating and sub-replicating strategy is also given by El-Kouri *et al.* [11] or Romagnoli and Vargiolu [24](see also [28], [17], and [7]. If $\bar{v} = \underline{v} = v_t$, then v_t is the arbitrage free price of the contingent claim h.

Theorem 3.4. *The process*

$$X_t^{t,x,\Delta} \equiv X_t = \sup_{\gamma \in \Sigma} \mathbb{E}\left[h(S_{T-t}^{t,x,\gamma})\mathbf{1}_{\{\tau > (T-t)\}}\right]$$

is a supermartingale

Proof. The portfolio is self-financing, and X_t is bounded from below, hence by Theorem 3.5 in Krylov [20], p.149, X_t is a supermartingale. □

As we noticed in Definition 3.1 the super-replicating depends on the choice of volatility process (γ_t). This choice can create arbitrage opportunities. Therefore, v_t^+ (respectively v_t^-) may be considered as a stochastic control problem where the lower bound and the upper bound of its solution can be interpreted as a sub 'arbitrage' price and super 'arbitrage' price, respectively. Before we convert our problem into a stochastic control problem in the HJB equation, the following theorem gives an idea that with a super-replicating strategy, one can have $X_T^{t,x,\Delta} \geq h(S_{T-t}^{t,x,\gamma})\mathbf{1}_{\{\tau > (T-t)\}}$. This means the portfolio overhedges the contingent claim. The following theorem is the significant result of this paper.

Theorem 3.5. *Let v be a price process for a contingent claim and let Δ be a portfolio process. If $\bar{v}$ is the super-hedging price as defined by Definition 3.1, then there exists a pair (v, Δ) such that*

$$\bar{v}(t,x) = v(t,x) = \sup_{\gamma \in \Sigma} \mathbb{E}\left[h(S_{T-t}^{t,x,\gamma})\mathbf{1}_{\{\tau > (T-t)\}}\right].$$

In particular,

$$\bar{v}(T,x) \geq h(S_{T-t}^{t,x,\gamma})\mathbf{1}_{\{\tau > (T-t)\}} \quad a.s.$$

Proof. Take a super-replicating strategy Δ associated with an upper hedging price as defined in Definition 3.1. Then

$$X_T^{x,\Delta,v} \geq \mathbb{E}(S_{T-t}^{t,x,\gamma})\mathbf{1}_{\{\tau > (T-t)\}}$$

for every admissible control γ, but by Proposition 4.4.2 in Krylov [20], $X_T^{x,\Delta}$ is a supermartingale. This implies that

$$\bar{v}(t,x) = X_t^{x,\Delta} \geq \mathbb{E}\left[X_T^{t,x,\Delta} \,|\mathcal{F}_t\right] \geq \mathbb{E}\left[h(S_{T-t}^{t,x,\gamma})\mathbf{1}_{\{\tau > (T-t)\}}\right] \quad \forall \gamma \text{ admissible}.$$

Hence, $\bar{v}(t,x) \geq v(t,x)$.
To prove that $\bar{v}(t,x) \leq v(t,x)$, we apply Ito's formula to the process $S_{\tau \wedge T}^{x,\gamma}$, giving

$$\begin{aligned} v(\tau \wedge T, S_{\tau \wedge T}^{x,\gamma}) &= v(t,x) + \int_t^{\tau \wedge T} \left(\frac{\partial}{\partial t} v(r, S_r^{t,x,\gamma}) + \frac{\partial^2}{\partial x^2} v(r, S_r^{x,\gamma})\right) dr \\ &\quad + \int_t^{\tau \wedge T} x\gamma \frac{\partial}{\partial x} v(r, S_r^{t,x,\gamma}) dW_r. \end{aligned}$$

Taking expectation of both sides, we have

$$\mathbb{E}(S_T^{t,x,\gamma})\mathbf{1}_{\{\tau>T\}} = v(t,x) + \mathbb{E}\int_t^{\tau\wedge T}\left(\frac{\partial}{\partial t}v(r,S_r^{x,\gamma}) + \frac{\partial^2}{\partial x^2}v(r,S_r^{x,\gamma})\right)dr.$$

Now we take the supremum of both sides, giving

$$\bar{v}(t,x) \le v(t,x) + \sup_{\gamma\in\Sigma}\mathbb{E}\int_t^{\tau\wedge T}\left(\frac{\partial}{\partial t}v(r,S_r) + \frac{\partial^2}{\partial x^2}v(r,S_r^{x,\gamma})\right)dr. \tag{3}$$

Since the expectation in (3) is zero, we have

$$\bar{v}(t,x) \le v(t,x).$$

Therefore,

$$\bar{v}(t,x) = v(t,x) = \sup_{\gamma\in\Sigma}\mathbb{E}\left[(S_{T-t}^{t,x,\gamma})\mathbf{1}_{\{\tau>(T-t)\}}\right].$$

□

4 Pricing and Hedging

Let us consider a very common example of barrier option, that is the knock out and up call of the European type. If $0 \le t \le T$, $S_t = x$ and the call has not knocked out prior to time t, then the price process for this option is given by an adapted process, $\{v_t; 0 \le t \le T\}$, satisfying

$$v_T = (S_T - K)\mathbf{1}_{\{\tau>T\}}.$$

Here K is the strike price of the option and τ is the first moment of time when the process S_t hits the barrier H, defined by

$$\tau = \inf\{t \ge 0;\quad S_t \ge H\}.$$

Assuming that $\mathbb{P}$ is already the risk neutral measure and $0 < K < H$, the value of the knock-out barrier option at time t with initial stock price x is given by

$$J(t,x;\sigma) = \mathbb{E}\left[(S_{T-t}^{t,x,\sigma} - K)^+\mathbf{1}_{\{\tau>(T-t)\}}\right], \quad 0 \le t \le T. \tag{4}$$

For a constant volatility, $\sigma_t = \sigma$, the explicit solution of (4) can be derived by the method of reflection principle of Brownian motion, Rich [23]. One may refer to Rich [23] for the closed form solution of (4). The value function (4) is also the solution of partial differential equation

$$\frac{\partial}{\partial t}v(t,x) + \frac{1}{2}\sigma^2x^2\frac{\partial^2}{\partial x^2}v(t,x) = 0, \quad 0 \le t < T, \quad 0 \le x < H \tag{5}$$

with terminal and boundary conditions

$$v(T,x) = (x-K)^+, \quad 0 \le x < H \tag{6}$$

$$v(t,x) = 0, \qquad x \ge H, \quad 0 \le t \le T. \tag{7}$$

Now assume that the true volatility is limited to move in a certain interval, i.e.

$$\sigma_t \in I = [\sigma_{min}, \sigma_{max}].$$

The assumption that (σ_t) is adapted to $\mathbb{F}$ makes it functional of the Brownian paths $\{W_t, 0 \leq t \leq T\}$, so that it is dependent on the past of the Brownian motion or stock price. This volatility can be interpreted as a control to find the worst and the best case price of the barrier option. Since the seller does not know the true volatility, he will estimate the fair price of the claim within an interval of prices, which is known as the interval of admissible prices. Therefore, we expect that the price of the claim lies in the interval

$$v^-(t, S_t^{t,x,\sigma}) \leq v_t^\sigma \leq v^+(t, S_t^{t,x,\sigma}),$$

where

$$v^-(t, S_t^{t,x,\sigma}) = \inf_{\sigma \in I} \mathbb{E}\left[(S_{T-t}^{t,x,\sigma} - K)^+ \mathbf{1}_{\{\tau > (T-t)\}}\right]$$

and

$$v^+(t, S_t^{t,x,\sigma}) = \sup_{\sigma \in I} \mathbb{E}\left[(S_{T-t}^{t,x,\sigma} - K)^+ \mathbf{1}_{\{\tau > (T-t)\}}\right].$$

As already discussed in the previous section, in order to have superstrategy, we fix the price v_t of the option and the quantities Δ_t of the risky asset S_t in the hedging portfolio $X_t^{x,\Delta,v}$ as

$$v_t = v(t, S_t^{t,x,\sigma}), \qquad \Delta_t = \frac{\partial}{\partial x} v(t, S_t^{t,x,\sigma}), \quad 0 \leq t \leq \tau \wedge T$$

where v is the solution of the HJB equation

$$\frac{\partial}{\partial t} v(t,x) + \frac{1}{2} \sup_{\sigma \in I} \sigma_t^2 x^2 \frac{\partial^2}{\partial x^2} v(t,x) = 0, \quad 0 \leq t < T, \quad 0 \leq x < H \tag{8}$$

with terminal and boundary conditions

$$v(T,x) = (x-K)^+, \quad 0 \leq x < H \tag{9}$$

$$v(t,x) = 0, \quad x \geq H, \quad 0 \leq t \leq T. \tag{10}$$

Therefore, the portfolio process satisfies

$$d(X_t^{x,\Delta,v}) = \Delta_t \sigma S_t^{t,x,\sigma} dW_t.$$

Initially, at $t = 0$, take

$$X_0^{x,\Delta,v} = v(0, S_0).$$

Then

$$X_{\tau \wedge T}^{x,\Delta,v} = v(\tau \wedge T, S_{\tau \wedge T}^{t,x,\sigma})$$

with terminal and boundary conditions

$$v(T,x) = (x-K)^+, \quad \text{if } \tau > T$$

$$v(t,H) = 0, \quad \text{if } \tau \leq T.$$

Remark 4.1. In order to have delta hedging admissible, we have to impose the condition (see Lemma 7.4 in [5])

$$\mathbb{E}\int_0^T \Delta_t^2 dt < \infty.$$

Moreover, when

$$\Delta_t = \frac{\partial}{\partial x} v(t, S_t^{t,x,\sigma}), \quad 0 \le t \le \tau \wedge T,$$

then $X_t^{x,\Delta,v}$ is a supermartingale.

Theorem 4.2. *Suppose that v is a solution of (8)-(10) for any convex payoff function. Then v is not convex or concave in x for any $t > 0$.*

Proof. We prove by contradiction. Suppose that v is convex or concave for all t and x. Note that v is positive for $x < H$ and v approaches zero when $x \to H$ and $x \to 0$ for every fixed time t. Therefore, it must be concave. However, if t approaches T, $v(T, x) = h(x)$ which is a convex function. This produces a contradiction. □

Theorem 4.3. *Let v be a solution of the HJB equation (8) with terminal condition (9) and boundary condition (10), and define*

$$\sigma_t(x) \quad = \quad \begin{cases} \sigma_{max} & \text{if} \quad \dfrac{\partial^2}{\partial x^2} v(t,x) \ge 0 \\ \sigma_{min} & \text{if} \quad \dfrac{\partial^2}{\partial x^2} v(t,x) \le 0. \end{cases}$$

Then σ_t is an optimal 'bang-bang' control, v is the superprice and Δ is the super-strategy.

Proof. Since v is a unique solution of the HJB equation (8)-(10), then v is the optimal price. Then, clearly σ_t is an optimal control, because HJB is computed with supremum. □

5 Practical Issues

The contingent claim $h(x) = (x - K)^+ \mathbf{1}_{\{\tau > T\}}$ is discontinuous at the barrier H. This results in an unbounded delta hedging at the maturity of the barrier option. The large delta hedging may cause instability in the hedging strategy (See figure 1). The delta hedging becomes very negative near the barrier, $\Delta_t = \frac{\partial}{\partial x} v(t, x) \to -\infty$ as $t \to T$.

If our portfolio consists of a non-risky asset invested in a money market and risky assets in a stock, then in the case where the stock price does not cross the barrier, the seller covers this short position with funds shares in the stock. If the stock price hits the barrier and the option is knocked out, the hedging strategy is in

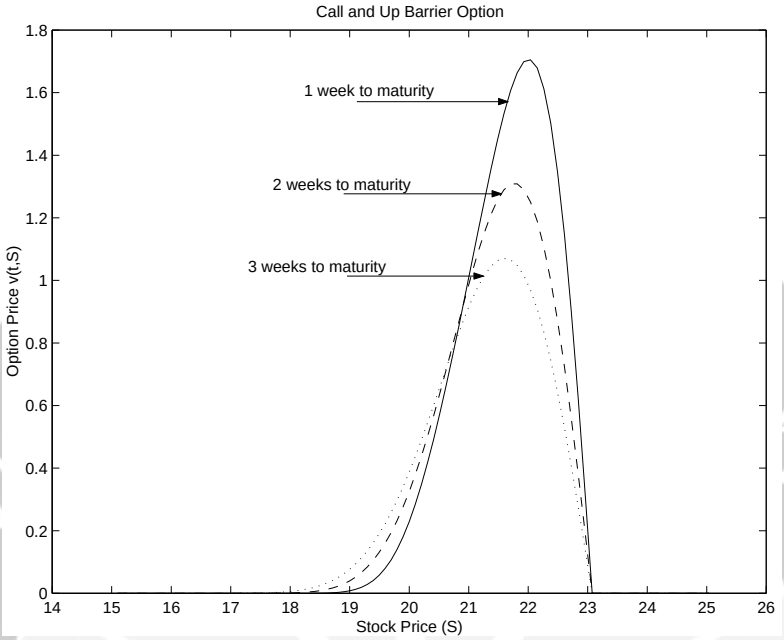


Figure 1: The barrier option price given by (8)- (10) with K=20, H=23, $\sigma = 0.20$

the region where Δ_t is large and negative. In this case, the seller covers his short position with the money market.

To avoid the large delta being taken, one can put a constraint on the hedging portfolio and then use this constraint to bound the super-replication strategy. This approach has been suggested by Schmock, Shreve, and Wystup [26]. They impose constraints on the delta and show that the cheapest super-replicating claim that satisfies this constraint can be found as the solution of a dual problem of a stochastic control problem. Another method to avoid instability in the hedging strategy is proposed by Shreve [27], Chap.20, p.218. He imposes the boundary condition

$$v(t,x) + \alpha H \frac{\partial}{\partial x} v(t,x) = 0, \quad x \geq H, \quad 0 \leq t \leq T,$$

instead of

$$v(t,H) = 0 \quad x \geq H, \quad 0 \leq t \leq T$$

where α is a 'tolerance parameter'. This approach guarantees that the $H\Delta_t$ remains bounded and the value of the portfolio is always sufficient to cover a hedging error within $\alpha H \Delta_t$ of the short position.

6 Numerical simulation

In this subsection, we consider a numerical example which illustrates the previous discussion. In particular, we generate a Call and Up barrier option of European type with strike price K =\$20 and barrier H =\$23. Since the true volatility is not known, we expect the volatility to be moving within interval $[\sigma_{min}, \sigma_{max}] = [0.10, 0.20]$ and option expiration at $T = 0.25$ year. Then we use the HJB equation (8)- (10) to calculate the superprice. We also assume that we initially can buy or sell the option at the mid volatility, $(\sigma_{max} + \sigma_{min})/2$. Here we report in Figure 2 the subprice and superprice barrier option computed using explicit schemes.

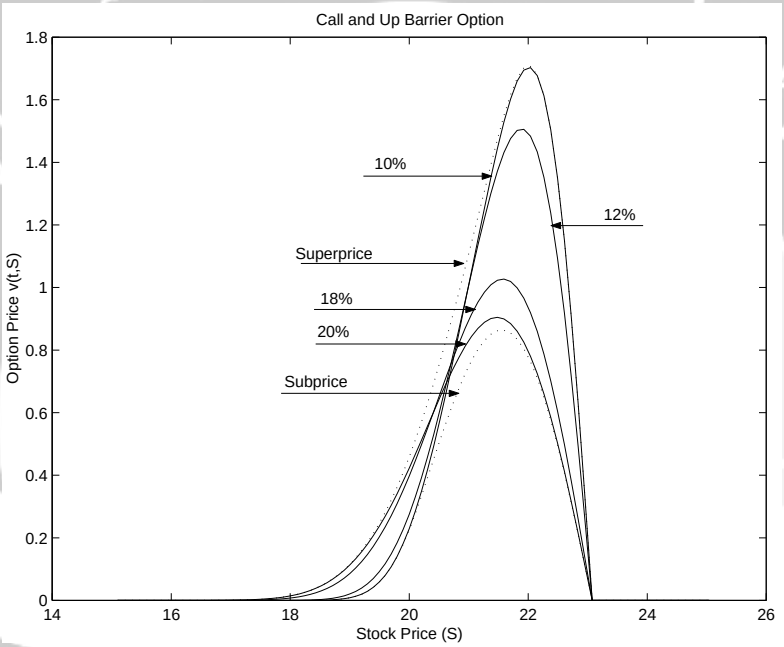


Figure 2: The dotted lines represent the superprice and subprice of the barrier option computed by (8)- (10) and the solid lines represent the extreme value of the option computed by the linear equation (5)-(7).

Figure 2 illustrates a comparison between the extreme prices that are obtained by pricing with a constant volatility, linear PDEs and those obtained from the BSB equation. Since the extreme prices for options are obtained by using the two extreme volatilities, one might believe that the extreme price for the portfolios would be given by the Black-Scholes prices with some constant volatility σ in the range $\sigma_{min} \leq \sigma \leq \sigma_{max}$. As shown in Figure 2, the theoretical price calculated by the Black-Scholes formula is too low to enter into a delta-hedging strategy that protects against the worst case situation. The superhedging strategy obtained from the BSB equation would protect the hedger against the movement of the volatilities within the band.

The following figure shows superstrategy Δ with 1-3 months to maturity, computed using equation (4). It shows how delta of the portfolio superstrategy varies as the option gets closer to maturity.

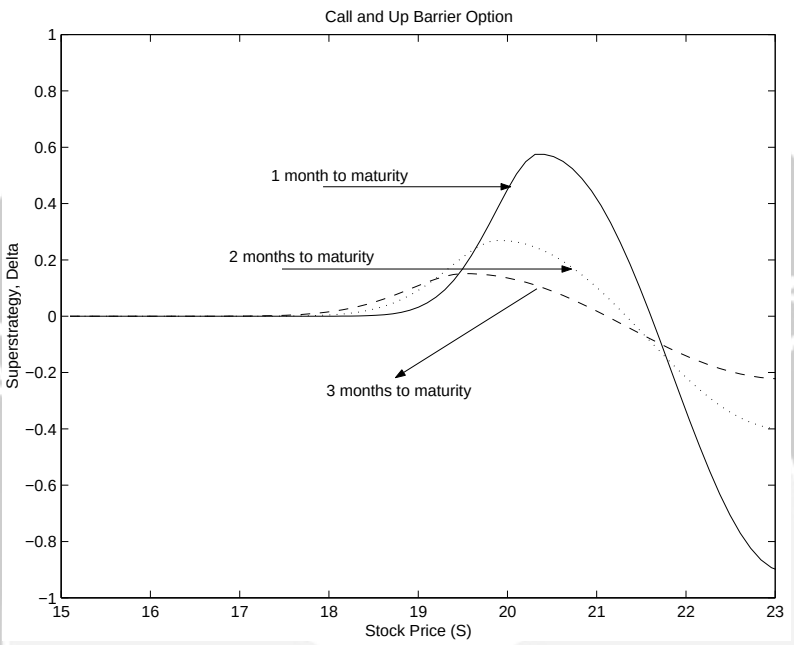


Figure 3: Delta superstrategy computed from the superprice given in figure 2

7 Concluding Remark

Our results in this work can be summarized as follows.

- If $v \in C^{2,1}([0,\infty))$ then the Black-Scholes-Barenblatt (BSB) equation (8)-(10) is easily derived from the dynamic programming principles via Ito's formula and there exists a unique solution in the classical sense.

- The value function v is not convex or concave for all x for each time t (see Theorem 4.2). Therefore, the superprice is determined dynamically: it is either the upper bound or the lower bound of the volatility matrix, according to the convexity or concavity of the value function with respect to the stock prices.

- The finite difference method with explicit schemes gives quite stable solutions

Acknowledgment

I would like to thank to Dr. Ben Goldys of The School of Mathematics, The University of New South Wales, Sydney Australia for his supervision during the author studied in Australia.

References

[1] Avellaneda, M., Levy, A., and Paras, A. 1995. *Pricing and hedging derivatives in markets with stochastic volatilities*. Applied Mathematical Finance. 2, p.73-88.

[2] Avellaneda, M. and A. Paras. 1996. *Managing the volatility risk of portfolios of derivatives securities: the Lagrangian uncertain volatility model*. Applied Mathematical Finance 3, p.21-52.

[3] Black, F. and M. Schole. 1973. *Pricing of Options and Corporate Liabilities*. Journal of Political Economy, 81,p637-659.

[4] Brown, H., D. Hobson, and L. Rogers. 2001. *Robust hedging of barrier options*. Mathematical Finance. 11(3), p.285-314.

[5] Crandall, M. G. M. Kocan, and A. Swiech. 2000.L^P-*Theory for fully nolinear uniformly parabolic equations*. Communication in Partial Differential Equations, 25(11&12), p.1997-2053.

[6] Cvitanic, J. and I. Karatzas. 1993. *Hedging contingent claim with constrained portfolio*. Annal Applied Probability. 3. p.652-681.

[7] Davis M.H.A. and J.M.C. Clark. 1994. *A note on super-replicating strategies*. Phil. Trans. R. Soc. London A 347, p.485-494.

[8] Derman, E., D Ergener, and I. Kani. 1995. *Static options replication*. Journal of Derivatives, 2(4). p.78-95

[9]

[10] Duffie D. 1996. *Dynamic Asset Pricing Theory*. Standford University, Academic Press, California

[11] El-Karoui N., M. Jeanblanc-Pique, S.E. Shreve. 1998. *Robustness of the Black-Schole Formula*. Math. Finance 8(2), p.93-126.

[12] Frey, R. and C.A. Sin. 1999. *Bounds on European option prices under stochastic volatility*, Mathematical Finance. 9(2), p.97-116.

[13] Frey R. 2000. *Superreplication in stochastic volatility models and optimal stopping*. Finance and Stochastics. 4. p.161-187.

[14] Gozzi, F and T. Vargiolu. 2002. *Supperreplication of European Multiasset Derivatves with Bounded Stochastic Volatility*. Math. Methods of Oper. Research.

[15] Harrison, J.M. and S.R Pliska. 1981. *Matingales and stochastic integrals in the theory of continuous trading*. Stochastic processes and Applications. 11. p.215-260.

[16] Heynen R. and Harry Kat. 1994. *Crossing Barrier*. Risk, 7(6). p.46-49

[17] Hobson, David G. 1998. *Volatility misspecification, option pricing and superreplication via coupling*. The Annals of Applied Probability 8(1), p.193-205

[18] Karatzas, I. 1996. *Lectures on mathematics of finance*. CRM monograph series, ISSN 1065-8599; v.8

[19] Karatzas, I. and S.E. Shreve. 1991. *Brownian Motion and Stochastic Calculus*. Springer-Verlag, New York.

[20] Krylov, N.V. 1990. *Control Diffusion Processes*. Springer-Verlag, New York.

[21] Lyons, T. 1993. *Uncertain volatility and the risk-free synthesis of derivatives*. Applied Mathematical Finance 6, p.1969-1984.

[22] Merton R.C. 1973. *The theory of rational option pricing*. Bell Journal of Economics and Management Scince. 4. p.141-183.

[23] Rich D. R. 1994. *The mathematical foundations of barrier option-pricing theory*. Advances in Futures and Operation Research, 7,p.267-311.

[24] Romagnoli S. and T. Vargiolu. 2000. *Robustness of the Black-Scholes approach in the case of option on several assets*. Finance and Stochastics, 4, p.325-341.

[25] Rubinstein M. and E.S. Reiner. 1991. *Breaking down the barrier*. Risk 4(8), p.28-35.

[26] Schmock U., S. E. Shreve and U. Wystup. 2001. *Valuation of exotic under shortselling constraints*. Working Paper. http://www.math.ethz.ch/ schomock

[27] Shreve, E.S.*Lecture on Stochastic Calculus and Finance* . http://www-2.cs.cmu.edu/ chal/shreve.html

[28] Touzi Nizar. 1999. *Super-replication under proportional transaction costs: From discrete to continuous-time models*. Mathematical Methods of Operations Reserach. 50, p.297-320.

Komang Dharmawan: Department of Mathematics, Universitas Udayana, Kampus Bukit Jimbaran, Bali, Indonesia.
Phone/Fax: +62 +361 703137.
E-mail: Komang_Dharmawan@yahoo.com.au

OPTIMISING PARAMETER ESTIMATION FOR DOUBLE SAMPLING CONTROL CHART

Dradjad Irianto

Department of Industrial Engineering
Bandung Institute of Technology, Indonesia.

Abstract. The double sampling control chart is aimed at improving the ability to detect any out-ofcontrol condition by observing the second sample without any interruption. This paper briefly discusses the advantage of double sampling procedures proposed Daudin (1990). Instead of minimizing the expected average sample size as in Daudin (1990), we propose an optimizing parameter estimation to maximize the power to detect a small shift of process' mean value. This optimization leads to a revised procedure of double sampling control chart.

Keywords: Double Sampling Control Chart, Optimization, Power of Control Chart

1. Introduction

Statistical process control is a well-known method to understand the process variability and to improve the quality of process. Among the statistical process control tools, control chart is aimed at monitoring the process. A control chart is designed to identify variation in process, either as a result of unassignable causes, or as a result of assignable (or special) causes. In this respect, the standard Shewhart $\overline{X}$ control chart has been widely used, but it is slow in detecting small shift. A number of alternatives have been proposed to improve the sensitivity of control chart, e.g. employing warning limits.

Reynolds et al. (1988) proposed a variable sampling interval (VSI) control chart in accordance to an out-of-control warning or signal. If a signal occurs, next sampling is taken in a shorter sampling interval; otherwise it is reasonable to take a longer sampling interval. Costa (1992) proposed a variable sample size (VSS) control chart using the same idea of VSI. The double sampling procedure (DS) uses the same idea as in VSI and VSS, but the second sample is observed with zero time intervals. The double sampling (DS) control chart was firstly proposed by Croasdale (1974). Daudin, Duby and Trecourt (1990) and Daudin (1992) proposed DS control chart that utilizes the information from both samples at the second stage. Irianto and Shinozaki (1998) discussed both DS procedures and proposed the advantage of Daudin's procedure compared to Croasdale's. Instead of minimizing the expected sample size, Irianto and Shinozaki (1998) maximized the power to detect a small shift of mean value.

2. The DS Control Chart Procedure

With DS control chart, the second sample will be observed only if the first sample is signaling a warning of deviation of the mean value. The procedures is described as follows (and is exhibited in Figure 1):

1.Take a sample of size n_1, X_{1i}, $i = 1,2,\cdots.n_1$ from a population with mean value μ_0 and a known standard deviation σ.

2.Calculate the sample mean $\overline{X}_1 = \sum_{i=1}^{n_1} X_{1i} / n_1$.

3.If $(\overline{X}_1 - \mu_0) / (\sigma / \sqrt{n_1})$ is in I_1, the process is considered to be under control.

4.If $(\overline{X}_1 - \mu_0) / (\sigma / \sqrt{n_1})$ is in I_3, the process is considered to be out of control.

5.If $(\overline{X}_1 - \mu_0) / (\sigma / \sqrt{n_1})$ is in I_2, take a second sample of size n_2, X_{2i}, $i = 1,2,\cdots.n_2$.

6.Calculate the sample mean $\overline{X}_2 = \sum_{i=1}^{n_2} X_{2i} / n_2$.

7.Calculate the total sample mean $\overline{X} = (n_1 \overline{X}_1 + n_2 \overline{X}_2) / (n_1 + n_2)$.

8.If $-L < \dfrac{\overline{X}_1 - \mu_0}{\sigma / \sqrt{n_1}} < -L_1$ or $L_1 < \dfrac{\overline{X}_1 - \mu_0}{\sigma / \sqrt{n_1}} < L$ and if $\dfrac{\overline{X} - \mu_0}{\sigma / \sqrt{n_1 + n_2}} < -L_2$ or $\dfrac{\overline{X} - \mu_0}{\sigma / \sqrt{n_1 + n_2}} > L_2$, then the process is considered to be out of control, otherwise the process is considered to be under control.

Let $\overline{Z}_1 = (\overline{X}_1 - \mu_0) / (\sigma / \sqrt{n_1})$ and $\overline{Z} = (\overline{X} - \mu_0) / (\sigma / \sqrt{n_1 + n_2})$, then the probabilities that the process is considered to be under control by the first sample and after taking the second sample can be formulated as $P_{a1} = \Pr[\overline{Z}_1 \in I_1]$ and $P_{a2} = \Pr[\overline{Z}_1 \in I_2 \text{ and } \overline{Z} \in I_4]$ respectively, and the probability that process under control is $P_a = P_{a1} + P_{a2}$. We assume the characteristic of output of process follows a normal distribution $N(\mu, \sigma^2)$.

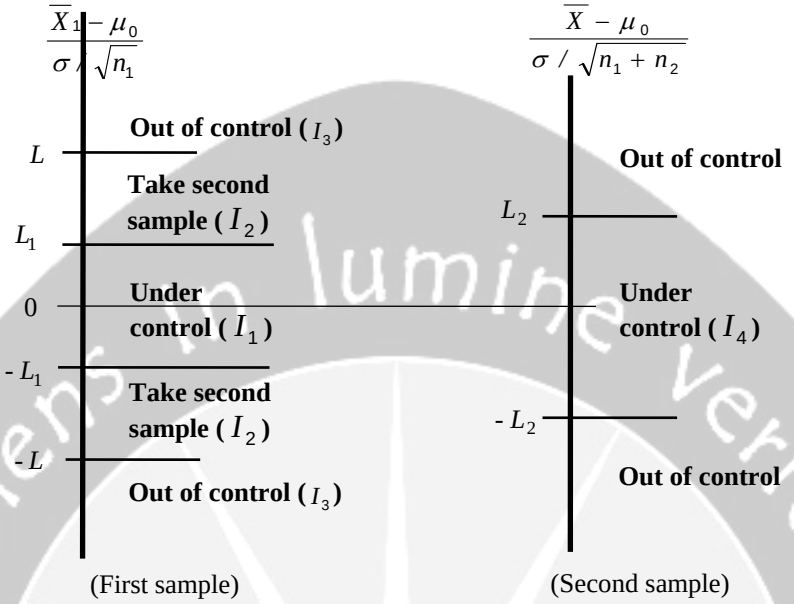


Figure 1. The Double Sampling Control Chart.

For a shift from the mean value $\delta = (\mu_0 - \mu)/\sigma$, the probability that the process is considered to be under control becomes:

$$P_a = \Phi[L_1 + \delta\sqrt{n_1}] - \Phi[-L_1 + \delta\sqrt{n_1}] + \int_{z \in I_2^*} \{\Phi[cL_2 + rc\delta - z\sqrt{n_1/n_2}] - \Phi[-cL_2 + rc\delta - z\sqrt{n_1/n_2}]\}\phi(z)dz \quad (1)$$

where, $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and cumulative distribution functions of standard normal distribution respectively, $r = \sqrt{n_1 + n_2}$, $c = r/\sqrt{n_2}$ and $I_2^* = [-L + \delta\sqrt{n_1}, -L_1 + \delta\sqrt{n_1}) \cup (L_1 + \delta\sqrt{n_1}, L + \delta\sqrt{n_1}]$. The average run length and the expected total sample size are given in equation (2) and (3), respectively:

$$\text{ARL} = 1/(1 - P_a), \text{ and} \quad (2)$$

$$n_1 + n_2 \cdot \Pr[\overline{Z}_1 \in I_2]. \quad (3)$$

3. Optimising Control Limits

There are five parameters required to specify the DS control charts, i.e. L_1, L_2, L, n_1 and n_2. Daudin et al. (1990) suggested an optimization procedure as follows :

$$\underset{n_1, n_2, L, L_1, L_2}{\text{Min}} \quad n_1 + n_2 \cdot \Pr[\overline{X}_1 \in I_2 | \mu = \mu_0] \quad (4)$$

Subject to:

(i) Pr[Out of Control | $\mu = \mu_0$] = α , that is

$$1 - \{\Phi[L_1] - \Phi[-L_1]\} - \int_{z \in I_2} \{\Phi[cL_2 - z\sqrt{n_1/n_2}\,] - \Phi[-cL_2 - z\sqrt{n_1/n_2}\,]\}\phi(z)dz = \alpha .$$

(ii) Pr[Out of control | $\mu = \mu_1$] = β (for a given intended shift $\delta = |\mu_1 - \mu_0|$), that is

$$1 - \{\Phi[L_1 + \delta\sqrt{n_1}\,] - \Phi[-L_1 + \delta\sqrt{n_1}\,]\}$$
$$- \int_{z \in I_2^*} \{\Phi[cL_2 + rc\delta - z\sqrt{n_1/n_2}\,] - \Phi[-cL_2 + rc\delta - z\sqrt{n_1/n_2}\,]\}\phi(z)dz = \beta .$$

To find the solution, they proposed an algorithm as follows :

(i) Determine n_1 and n_2 .

(ii) For a given value of L (suggested to be large, 4 or 5 times of standard deviation), both constraints are used to determine the values of L_1 and L_2 .

(iii) Find the optimal composition of L_1, L_2 and L that minimize the objective function for all possible pairs of (n_1, n_2).

The optimum result is given by the minimum size of (n_1, n_2). This optimization procedure is mainly motivated by the inspection cost. Differently, Irianto and Shinozaki (1998) considered the power to detect any deviation of the process' mean. It means that knowing the ARL of the chart is more important than cost consideration. Therefore the motivation is to maximize the power while setting sample sizes n_1 and n_2 so that the expected total sample size is fixed when $\mu = \mu_0$. The optimization is formulated as follows:

$$\underset{L,L_1,L_2}{\text{Max}}\ 1 - \{\Phi[L_1 + \delta\sqrt{n_1}\,] - \Phi[-L_1 + \delta\sqrt{n_1}\,]\}$$

$$- \int_{z \in I_2^*} \{\Phi[cL_2 + rc\delta - z\sqrt{n_1/n_2}\,] - \Phi[-cL_2 + rc\delta - z\sqrt{n_1/n_2}\,]\}\phi(z)dz . \qquad (5)$$

Subject to:

(i) E[total sample size | $\mu = \mu_0$] = n, that is

$$n_1 + n_2 \cdot \Pr[\bar{Z}_1 \in I_2 | \mu = \mu_0] = n \quad \Leftrightarrow \quad L = \Phi^{-1}\left[\frac{n - n_1}{2n_2} + \Phi[L_1]\right].$$

(ii) Pr[Out of Control | $\mu = \mu_0$] = α , that is

$$1 - \{\Phi[L_1] - \Phi[-L_1]\} - \int_{z \in I_2^*} \{\Phi[cL_2 - \sqrt{n_1/n_2}\ z] - \Phi[-cL_2 - \sqrt{n_1/n_2}\ z]\}\phi(z)dz = \alpha$$

From the first constraint, L is expressed in terms of L_1 , it reduces the number of parameter. Since the left hand of the second constraint is an increasing function of L_2 , L_2 is uniquely determined for fixed L_1 and L . For a given value of L_1 , the values of L and L_2 are uniquely determined. Since $\Phi(L_1) = \Phi(L) - (n - n_1)/(2n_2)$

and $(1-\alpha/2) \le \Phi(L) \le 1$, then $\Phi^{-1}\left[1-\frac{n-n_1}{2n_2}-\frac{\alpha}{2}\right] \le L_1 \le \Phi^{-1}\left[1-\frac{n-n_1}{2n_2}\right]$. This range of L_1 is quite small if α is small.

4. Numerical Results

Usually, standard Shewhart chart is used as the basis for comparison. The P_a =Pr[Out of Control $\mid \mu = \mu_1$] and the average run length (ARL) of the standard Shewhart $\bar{X}$ control chart (for *n*=5 and *L*=3) is shown in Table 1. The P_a =Pr[Out of Control $\mid \mu = \mu_1$] and the average sample size of DS (as in equation 3) for some shift are presented in Table 2, where α is set at 0.0027 and sample sizes $n_1 = n_2$ =4 and n =5.

Table 1. ARL and P_a of Shewhart Chart (for n =5 and *L*=3)

Shift δ	P_a	ARL
0	0.0027	370.4
0.5	0.0064	155.2
1	0.0228	43.9

Table 2. P_a and average sample size of DS Chart (for $n_1 = n_2$ =4 and n =5)

Limits			Power					
L_1	L	L_2	δ =0.2	0.4	0.5	0.6	0.8	1
1.15	3.8014	2.9924	0.0076	0.0303	0.055	0.094	0.2253	0.4229
			Average sample size					
			5.1474	5.5496	5.814	6.099	6.6511	7.0693

Clearly, the numerical result of the DS control chart gives better performance shown by higher power. Accordingly, the out of control signal occurs in a shorter interval than the standard Shewhart control chart, thus further action can be performed sooner. However, as shown in Table 2, it should be noted that the average sample size increases as the shift of process mean gets larger. In general, this average sample size affects the economic performance of control chart.

5. Investigating Control Limits

Table 5 shows control limits of DS charts for some pairs of n_1 and n_2 that give an expected sampling number *n*=5. This result shows that maximizing power leads to higher value of L_1 and lower value of L_2. Based on the first constraint, the higher value of L_1 is related to higher value of *L*, which is limited to $L = \infty$. It means that L is no longer necessary, and the double sampling control chart can be simplified as in Figure 2. The first sample is similar with one proposed by Croasdale (1974).

Table 5. P_a and Average Sample Size of DS Chart for Some Pairs of Sample Sizes.

DDS Charts	L_1	L	L_2	Power		Average sample size	
				$\delta = 0.5$	$\delta = 1.0$	$\delta = 0.5$	$\delta = 1.0$
$n_1=4$	0.671	3.0590	3.3435	0.0288	0.2273	5.3130	5.5341
$n_2=2$	0.672	3.1589	3.1652	0.0329	0.2582	5.3207	5.5769
$n=5$	0.673	3.3057	3.0720	0.0357	0.2766	5.3295	5.6314
	0.674	3.6057	3.0149	0.0375	0.2882	5.3405	5.7143
	0.67449	∞	2.9999	0.0379	0.2910	5.3492	5.8224
$n_1=4$	0.963	3.0599	3.3813	0.0321	0.2696	5.5596	6.1212
$n_2=3$	0.964	3.1360	3.2175	0.0373	0.3059	5.5683	6.1703
$n=5$	0.965	3.2362	3.1218	0.0410	0.3292	5.5780	6.2289
	0.966	3.3854	3.0557	0.0440	0.3459	5.5890	6.3039
	0.967	3.7058	3.0087	0.0461	0.3577	5.6030	6.4200
	0.96742	∞	2.9961	0.0467	0.3606	5.6127	6.5517
$n_1=4$	1.275	3.0473	3.4577	0.0376	0.3467	5.9138	7.0941
$n_2=5$	1.276	3.0969	3.2942	0.0446	0.3897	5.9234	7.1483
$n=5$	1.277	3.1555	3.1924	0.0499	0.4175	5.9335	7.2087
	1.278	3.2274	3.1184	0.0543	0.4378	5.9445	7.2776
	1.279	3.3210	3.0605	0.0580	0.4536	5.9565	7.3591
	1.280	3.4575	3.0135	0.0611	0.4662	5.9702	7.4614
	1.281	3.7271	2.9754	0.0637	0.4762	5.9872	7.6119
	1.28155	∞	2.9593	0.0647	0.4801	6.0020	7.8212
$n_1=4$	1.376	3.0680	3.3660	0.0453	0.4115	6.0572	7.5478
$n_2=6$	1.377	3.1140	3.2452	0.0517	0.4434	6.0672	7.6065
$n=5$	1.378	3.1677	3.1602	0.0569	0.4657	6.0779	7.6716
	1.379	3.2323	3.0945	0.0613	0.4826	6.0893	7.7448
	1.380	3.3139	3.0412	0.0651	0.4960	6.1017	7.8298
	1.381	3.4261	2.9966	0.0683	0.5069	6.1156	7.9329
	1.382	3.6110	2.9590	0.0711	0.5158	6.1319	8.0709
	1.38299	∞	2.9292	0.0733	0.5225	6.1567	8.3903

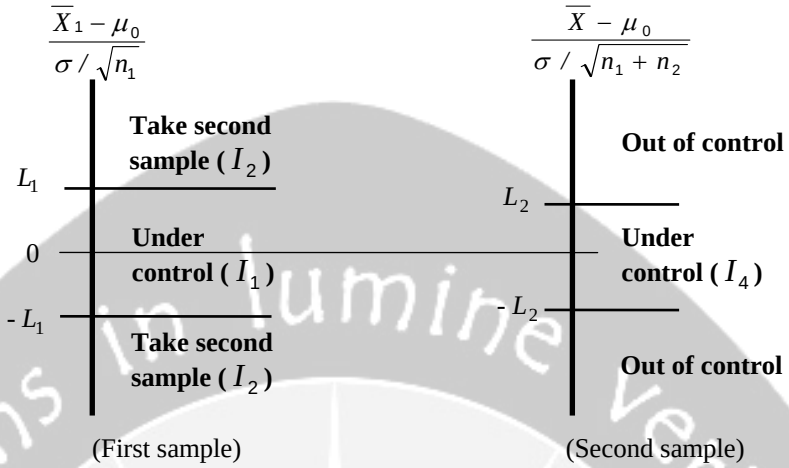


Figure 2. Revised DS control chart.

6. Concluding Remarks

Despite of the advantage of giving sooner signal of mean shift, the DS procedure needs a complicated calculation. Efficiency of the calculation is improved by changing the optimization problem (4) into (5). Since the value of L_1 is limited by the first constraint, the optimization procedure is less complicated. In this paper, the numerical result shows that the optimum limits lead to $L = \infty$. This revised control chart is simpler since it has lesser parameter.

References

[1] Costa, F.B.A. (1994), $\overline{X}$ Charts with Variable Sample Size, *Journal of Quality Technology*, **26(3)**, 155-163.

[2] Croasdale, R. (1974), Control Charts for a Double-Sampling Scheme Based On Average Production Run Lengths, *International Journal of Production Research*, **12(5)**, 585-592.

[3] Daudin, J.J., C. Duby, and P. Trecourt (1990), Plans de Controle Double Optimaux (Maitrise des Procedes et Controle de Reception), *Rev. Statistique Appliquee*, **38(4)**, 45-59.

[4] Daudin, J.J. (1992), Double Sampling $\overline{X}$ Charts, *Journal of Quality Technology*, **24(2)**, 78-87.

[5] Irianto, D., and N. Shinozaki (1998), An Optimal Double Sampling $\overline{X}$ Control Chart, *International Journal of Industrial Engineering – Theory, Applications and Practice*, **5(3)**, 226-234.

[6] Reynolds, M.R. Jr., R.W. Amin, J.C. Arnold, and J.A. Nachlas (1988), $\overline{X}$ Charts with Variable Sampling Interval, *Technometrics*, **30(2)**, 181-192.

DRADJAD IRIANTO: Department of Industrial Engineering, Institut Teknologi Bandung, Gedung Labtek III, Jl. Ganesha 10, Bandung 40132, Indonesia. Tel./fax.: (022) 250 6449.
E-mail: irianto@lspitb.org

CONFIDENCE BANDS FOR AIR POLLUTANT (CARBON MONOXIDA) UNDER DOUBLE TYPE-II CENSORING WITH BOOTSTRAP PERCENTILE

Akhmad Fauzy[a], Noor Akma Ibrahim[b], Isa Daud[b], Mohd. Rizam Abu Bakar[b]

[a] Universitas Islam Indonesia, Jogjakarta, Indonesia
[b] Universiti Putra Malaysia, Serdang, Malaysia

Abstract. This paper describes existing methods and develops new methods for constructing confidence bands for survivor function of two parameters exponential distribution under double type-II censoring. Our results are built on extensions of previous work by [11] and Balakrishnan [1]. They use maximum likelihood estimator to construct interval estimation under double type-II censoring. The confidence bands are developed for survivor function using the confidence region about survivor function. We will use another method, known as the bootstrap percentile from [4]. This method gives shorter confidence bands compared to the traditional method.
Key words: air pollutant, bootstrap percentile, confidence bands, double type-II, survivor function

1. Introduction

The survivor function or reliability function is a property of any random variable that maps a set of events, usually associated with mortality or failure of some system, onto time. It captures the probability that the system will survive beyond a specified time. The term reliability function is common in engineering while the term survivor function is used in a broader range of applications, including human mortality.

The exponential distribution is often proposed as a model in life testing and reliability because of its simplicity and the ease with which estimates can be calculated. [2] deals with inference procedures for the one-parameter exponential model. Inference for the two-parameter exponential model has been studied by [9], [10], [12] and many others, based on complete samples and type-II censored data.

In reliability studies, due to time limitations and/or other restrictions on data collection, several lifetimes of units put on test may not be observed. In addition, sometimes the lowest and/or highest few observations in a sample could be due to some negligence or some other extraordinary reasons. In such cases, it is convenient to remove those outlying observations. Type-II censored samples are considered here, whereby, in an ordered sample of size n, a known number of observations is missing at either end (single censoring) or both ends (double censoring). Doubly censored samples have been considered, by authors, like [1] and Raqab [11]. They used maximum likelihood estimator to construct interval estimation for survivor function of two parameters exponential distribution under double type-II censoring. Using the intervals estimation for survivor functions at

every lifetime develops confidence bands for survivor function. This band allows us to see the region in which the survivor function lies.

Bootstrap method is a computer-based method for assigning measures of accuracy to statistical estimates, especially to calculate the confidence interval. The aim of using bootstrap method is to gain the best estimation from minimal data [5].

[6] used bootstrap method to construct interval estimation for two parameters exponential distribution under double type-II censoring. In [7] bootstrap method was utilised to construct the interval estimation for survivor function for two parameters exponential distribution under double type-II censoring. In this paper the focus is to make comparison of the confidence bands for survivor function obtained by the conventional method and bootstrap percentile method.

2. Methodology

An example of real data is analysed to illustrate the procedure. The data is an air quality date extracted from the Malaysian Data Report 2000 obtained from the Department of Environment, Ministry of Science, Technology and Environment. The confidence band for the survivor function was first constructed by the traditional approach. From the bootstrap's repeated samples, the convergence condition is determined. This will be fallowed by developing the confidence band for the survivor function. The S-Plus software was used in the development of the programme.

3. Theory

The actual survival time of an individual, t, can be regarded as the value of a variable T, which can take any non-negative value. The survivor function, $S(t)$, is defined to be the probability that the survival time is greater than or equal to t, and so:

$$S(t) = P(T \geq t) = 1 - F(t). \tag{1}$$

The survivor function can therefore be used to represent the probability that an individual survives from the time origin to some time beyond t [3].

The essential element in lifetime data analysis is the presence of a nonnegative response, t, with appreciable dispersion and often with censoring. Due to sampling methods or the occurrence of some competing risk of removal from the study, several lifetimes of individuals may be censored. By censored data we mean that, in a potential sample of size n, a known number of observations is missing at either end (single censoring) or both ends (double censoring). The type of censoring just described is often called type-II censoring [8].

Suppose some initial observations are censored in addition to some final observations being censored. Out of the n components put to test, suppose the experimenter fails to observe the first r and the last s, observations are then said to be double type II censoring.

$$t_{r+1:n} \leq t_{r+2:n} \leq \ldots \leq t_{n-s:n}. \tag{2}$$

Two Parameters Exponential Distribution

The two parameters exponential distribution has probability density function [10]:

$$f(t;\mu,\theta)=\frac{1}{\theta}\exp\left(-\frac{t-\mu}{\theta}\right)\ ;\ t\ge\mu,\ \mu\ge 0,\ \theta>0. \tag{3}$$

Here μ is the warranty time and θ is the residual mean life. Once again, it is simple exercise to derive the maximum likelihood estimation of the θ as [1]:

$$\hat{\mu}=t_{r+1:n}+\hat{\theta}\ln\left(\frac{n-r}{n}\right), \tag{4}$$

$$\hat{\theta}=\frac{\sum_{i=r+1}^{n-s}t_{i:n}+s\,t_{n-s:n}-(n-r)t_{r+1:n}}{n-s-r}. \tag{5}$$

The following quantities are independent, with exact sampling distributions:

$$\frac{2n(t_{r+1:n}-\mu)}{\theta}\sim\chi^2_2 \quad\text{and}\quad \frac{2(n-s-r)\hat{\theta}}{\theta}\sim\chi^2_{(2(n-s-r)-2)}, \tag{6}$$

where $\chi^2_{(2(n-s-r)-2)}$ is the chi-squared distribution with $(2(n-s-r)-2)$ degrees of freedom. It follows that the ratio of these quantities divided by the ratio of their degrees of freedom is a F variable:

$$\frac{2n(t_{1:n}-\mu)}{2\theta}\Big/\frac{2(n-s-r)\hat{\theta}}{\theta(2(n-s-r)-2)}=\frac{n(n-s-r-1)(t_{1:n}-\mu)}{(n-s-r)\hat{\theta}}\sim F_{(2,2(n-s-r)-2)}, \tag{7}$$

with $(2,2(n-s-r)-2)$ degrees of freedom.

A two-sided, equal-tail, $(1-\alpha)$ level confidence interval on μ is constructed from the probability statement:

$$\Pr\left(F_{(2,2(n-s-r)-2;\alpha/2)}\le\frac{n(n-s-r-1)(t_{1:n}-\mu)}{(n-s-r)\hat{\theta}}\le F_{(2,2(n-s-r)-2;1-\alpha/2)}\right)=1-\alpha. \tag{8}$$

The $(1-\alpha)$ confidence intervals for μ is:

$$\left(\hat{\mu}-\frac{(n-s-r)\hat{\theta}\,F_{(2,2(n-s-r)-2;1-\alpha/2)}}{n(n-s-r-1)}\right)=\hat{\mu}_{\min}\le\mu\le\left(\hat{\mu}-\frac{(n-s-r)\hat{\theta}\,F_{(2,2(n-s-r)-2;\alpha/2)}}{n(n-s-r-1)}\right)=\hat{\mu}_{\max}. \tag{9}$$

A confidence interval on θ similarly constructed from the probability statement:

$$\Pr\left(\chi^2_{(2(n-s-r)-2;\alpha/2)} \le \frac{2(n-s-r)\hat{\theta}}{\theta} \le \chi^2_{(2(n-s-r)-2;1-\alpha/2)}\right) = 1-\alpha. \quad (10)$$

The $(1-\alpha)$ confidence intervals for θ is:

$$\frac{2(n-s-r)\hat{\theta}}{\chi^2_{(2(n-s-r)-2;1-\alpha/2)}} = \hat{\theta}_{\min} \le \theta \le \frac{2(n-s-r)\hat{\theta}}{\chi^2_{(2(n-s-r)-2;\alpha/2)}} = \hat{\theta}_{\max}. \quad (11)$$

Survivor function on two parameters exponential distribution is:

$$S(t) = \int_t^\infty f(t)\,dt = \int_t^\infty \theta^{-1}\exp(-(t-\mu)/\theta)\,dt = \exp(-(t-\mu)/\theta). \quad (12)$$

The $(1-\alpha)$ confidence for survivor function is:

$$\exp\left(-\frac{(t-\hat{\mu}_{\min})}{\hat{\theta}_{\min}}\right) < S(t) < \exp\left(-\frac{(t-\hat{\mu}_{\max})}{\hat{\theta}_{\max}}\right). \quad (13)$$

Bootstrap Percentile Method

Bootstrap method is a computer-based method for assigning measures of accuracy to statistical estimates, especially to calculate the confidence interval. Bootstrap itself comes from the phrase "*pull oneself up by one's Bootstrap*" which means to stand up by one's own feet and do with minimal resources. The minimal resource is a minimum data, data that are free from certain assumption or data with no assumption at all about the population distribution. The aim of using bootstrap method is to gain the best estimation from minimal observation. The Bootstrap's percentile procedures for the confidence bands for survivor function on two parameters exponential distribution under double type-II censoring are as follows:

1. give an equal opportunity $1/(n-s-r)$ to every observation of type-II censoring,
2. take $(n-s-r)$ sample with replication,
3. do step 2 until *B* times in order to get an "*independent bootstrap replications*", $\hat{\beta}^{*1}, \hat{\beta}^{*2}, \ldots, \hat{\beta}^{*B}$, and search for convergence condition. Calculate:

$$S(t)^{*b} = \exp\left(-\frac{(t^{*b}-\mu^{*b})}{\theta^{*b}}\right) \quad \text{with } \mu^{*b} = t^{*b}_{r+1:n} + \hat{\theta}^{*b}\ln\left(\frac{n-r}{n}\right) \text{ and}$$

$$\hat{\theta}^{*b} = \frac{\sum_{i=r+1}^{n-s} t_{i:n}^{*b} + s\, t_{n-s:n}^{*b} - (n-r)\, t_{r+1:n}^{*b}}{n-s-r}, \quad (14)$$

4. define the confidence interval at the level $(1-\alpha)$ of the bootstrap percentile for survivor function of one and two parameters exponential distribution under double type-II censoring as:

$$\left[S(t)^{*b(\alpha/2)}, S(t)^{*b(1-\alpha/2)}\right], \quad (15)$$

5. confidence bands for survivor function are developed using the intervals estimation for survivor functions at every lifetime.

4. Results And Discussion

The data presented in air pollutant (special case: carbon monoxide) data report 2000 Malaysia from Department of Environment, Ministry of Science, Technology and Environment. The first 22 observations in a random sample of 28 lifetimes from carbon monoxide (ppm/part per million)) on 1st December 2000 are as follow:

- , - , - , 0.1100, 0.4950, 0.5338, 0.6075,
0.6150, 0.7029, 0.7350, 0.7871, 0.8650, 0.8925, 0.8938,
0.9429, 0.9543, 0.9629, 1.0186, 1.0500, 1.0514, 1.0625,
1.2171, 2.3050, 2.8038, 2.9275, - , - , -
.

These data are of double type-II censoring. Using Lilliefors test, the data are exponentially distributed. As an illustration we will use these data to construct confidence bands for the survivor function.

Based on the two parameters exponential distribution under double type-II, the intervals estimation for survivor functions at every lifetime, are tabulated in Table 1.

Table 1. The floor (F) and ceiling (C) for survivor functions to every lifetime at the level of 99% and 95 % with traditional method and bootstrap percentile method

Life time	Traditional method				Bootstrap percentile method			
	99%		95%		99%		95%	
	F	C	F	C	F	C	F	C
0.1100	0.58631	0.94449	0.69007	0.93471	0.89286	1.00000	0.89286	1.00000
0.4950	0.37331	0.81770	0.46154	0.78915	0.54638	1.00000	0.57809	1.00000
0.5338	0.35671	0.80591	0.44321	0.77580	0.51999	1.00000	0.55332	0.97144
0.6075	0.32718	0.78398	0.41037	0.75107	0.47333	0.90161	0.50914	0.89286
0.6150	0.32431	0.78178	0.40716	0.74860	0.46883	0.89286	0.50484	0.88509
0.7029	0.29255	0.75647	0.37144	0.72022	0.41910	0.82269	0.45715	0.79586
0.7350	0.28174	0.74743	0.35919	0.71012	0.40228	0.79941	0.44088	0.76524
0.7871	0.26505	0.73300	0.34016	0.69404	0.37642	0.76298	0.41395	0.72364

0.8650	0.24191	0.71193	0.31357	0.67067	0.33921	0.70671	0.37316	0.67121
0.8925	0.23423	0.70464	0.30469	0.66261	0.32555	0.68877	0.35955	0.65420
0.8938	0.23388	0.70429	0.30428	0.66223	0.32501	0.68793	0.35901	0.65347
0.9429	0.22079	0.69146	0.28906	0.64809	0.29753	0.65688	0.33429	0.62463
0.9543	0.21786	0.68852	0.28564	0.64485	0.29205	0.65006	0.32938	0.61866
0.9629	0.21567	0.68631	0.28309	0.64242	0.28681	0.64488	0.32566	0.61362
1.0186	0.20204	0.67214	0.26708	0.62688	0.25797	0.61102	0.30017	0.58057
1.0500	0.19473	0.66429	0.25846	0.61828	0.23951	0.59429	0.28761	0.56416
1.0514	0.19442	0.66394	0.25809	0.61790	0.23904	0.59353	0.28712	0.56351
1.0625	0.19190	0.66119	0.25511	0.61489	0.23442	0.58821	0.28321	0.55775
1.2171	0.16008	0.62400	0.21706	0.57448	0.16171	0.51874	0.22914	0.48607
2.3050	0.04470	0.41524	0.06966	0.35608	0.01097	0.24709	0.05386	0.22642
2.8038	0.02491	0.34451	0.04137	0.28596	0.00324	0.18453	0.02610	0.16437
2.9275	0.02155	0.32891	0.03635	0.27082	0.00240	0.17164	0.02171	0.15208

Connecting the intervals estimation of every lifetime develops confidence bands for survivor function with traditional method and bootstrap percentile method.

Comparison of Confidence Bands

Figure 1 and 2 give the confidence bands for survivor function on two parameters exponential distribution under double type-II censoring using the traditional method and the bootstrap percentile method.

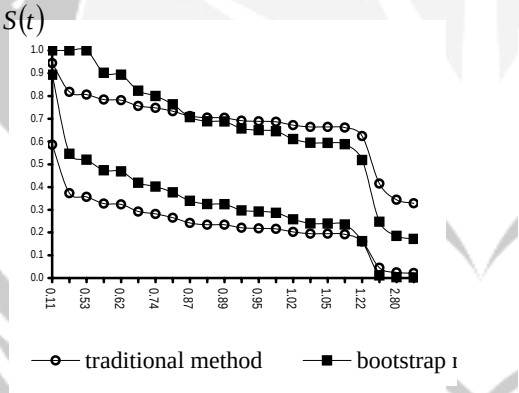


Figure 1.
99% Confidence bands for survivor function

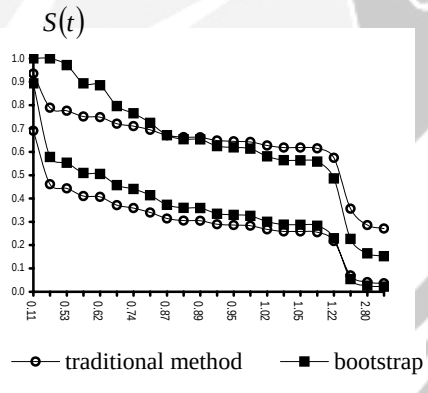


Figure 2.
95% Confidence bands for survivor function

From these figures 1 and 2 with 99% and 95% respectively, the width of the confidence regions for the survivor function with bootstrap percentile are narrower compared to the traditional method.

5. Conclusion

Using the intervals estimation for survivor functions at every lifetime develops the confidence bands for survivor function. Bootstrap percentile method proved to be

more potential in constructing confidence bands for survivor function on two parameters exponential distribution under double type-II censoring than the traditional method. Therefore, bootstrap method can be used as an alternative method in constructing confidence bands.

6. References

[1] Balakrishnan. (1990). On the Maximum Likelihood Estimation of the Location and Scale Parameters of the Exponential Distribution Based on Multiply Type-II Censored Samples. *Journal of Applied Statistics*. 17:55-61.

[2] Bartholomew, D. J. (1963). The Sampling Distribution of an Estimate Arising in Life Testing. *Technometrics* 5. 361-374.

[3] Collett, D. (1996). *Modelling Data In Medical Research*. London: Chapman & Hall.

[4] Efron, B. (1979). Bootstrap Method: Another Look at the Jackknife. *The Annals of Statistics*, 7(1), 1-26.

[5] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

[6] Fauzy, A., Ibrahim, N. A., Daud, I. & Abu Bakar, M. R. (2003a). Interval Estimation for One and Two Parameters Exponential Distribution under Double Type-II Censoring with Bootstrap Percentile. *Weekly Seminar to Institute for Mathematical Research*. UPM.

[7] Fauzy, A., Ibrahim, N. A., Daud, I. & Abu Bakar, M. R. (2003b). Interval Estimation for Survivor Function of Exponential Distribution under Double Type-II Censoring with Bootstrap Percentile. *International Conference on Research and Education in Mathematics*, Institute for Mathematical Research, UPM.

[8] Fernandez, A. J. (2000). Estimation and Hypothesis Testing for Exponential Lifetime Models with Double Censoring and Prior Information. *Journal of Economic and Social Research*, 2(2), 1-17.

[9] Hsieh, H. K. (1981). On Testing the Quality of Two Exponential Distributions. *Technometrics* 23. 263-269.

[10] Lawless. (1982). *Statistical Models and Methods for Lifetime Data*. New York: John Wiley & Sons.

[11] Raqab, M. Z. (1995). On the Maximum Likelihood Prediction of the Exponential Distribution Based on Double Type-II Censored Samples. *Pakistan Journal of Statistics*, 11, 1-10.

[12] Singh, N. (1985). A Simple and Asymptotically Optimal Test for the Equality of K Exponential Distribution Based on Type II Censored Samples. Comm. Statist. Theory Methods. 14.1615-1625.

AKHMAD FAUZY: Department of Statistics, Universitas Islam Indonesia, Jl. Kaliurang Km 14,4 Sleman, Jogjakarta, 55501, +62 +274 895920, afauzy@fmipa.uii.ac.id

NOOR AKMA IBRAHIM, ISA DAUD, MOHD RIZAM ABUBAKAR, Institute for Mathematical Research and Department of Mathematics, Universiti Putra Malaysia, Serdang, Selangor 43400, nakma@fsas.upm.edu.my, isa@fsas.upm.edu.my, mrizam@fsas.upm.edu.my

GENETICS PROBABILITY DISTRIBUTION IN DISCRETE-TIME MARKOV CHAIN

Mike Susmikanti

P2TIK-BATAN, Serpong-Tangerang, Indonesia

Abstract. Markov processes represent the simplest generalization of independent processes by permitting the outcome at any instant to depend only on the outcome that preceed it. A special kind of Markov process is a Markov chain. Both Markov chains and Markov processes can be discrete-time or continuous-time, depending on whether the time index set is discrete or continuous. The writer is mostly concerned with the state limiting behavior of genetic model, the behavior of various occupation times and their probability distribution are of special interest. in discrete-time Markov chain. The probability that a new cell has moved into state e_k is given by the hypergeometric distribution. In another genetic model, the transition probability that the next generation has moved into state e_k (k genes of type A and N-k genes of type B) from state e_j is given by the binomial distribution.
Key-words: Probability, Genetics

1. Introduction

Markov processes represent the simplest generalization of independent processes by permitting the outcome at any instant to depend only on the outcome that preceeds it and none before that.

A special kind of Markov process is a Markov chain where the system can ocuppy a finite or countably infinite number of states e_1, e_2, ..., e_j, ... such that the future evolution of the process, once it is in a given state, depends only on the present state and not on how it arrived at the state.

Both Markov chain and Markov processes can be discrete-time or continuous-time, depending on the time index set is discrete or continuous. In this case is mostly concerned with the transient and steady state limiting behavior of genetic model in discrete-time Markov chain. In addition, the behavior of various occupation time, first passage times, state occupancy times, and their probability distribution are of special interest.

The probability that a new cell has moved into state e_k is given by the hypergeometric distribution. In another genetic model, the transition probability that the next generation has moved into state e_k from state e_j is given by the binomial distribution. It will be interesting to study the limiting behavior of the population based on these models after several generations.

Markov processes are named after A. A. Markov (1856-1922), who introduced this concept for the discrete parameter systems with a finite number of states (1907). The theory for denumerable (countably infinite) chains was initiated by Kolmogorov (1936) followed by Doeblin (1937), Doob (1942), Levy (1951), and many others.

2. Markov Process

In a Markov process $\mathbf{x}(t)$, the past has no influence on the future if the present is specified. This means that if $t_{n-1} < t_n$, then

$$P[\mathbf{x}(t_n) \le x_n \mid \mathbf{x}(t),\ t \le t_{n-1}] = P[\mathbf{x}(t_n) \le x_n \mid \mathbf{x}(t_{n-1})] \quad (1)$$

From (1) it follows that if $t_1 < t_2 < \ldots\ldots < t_n$, then

$$P[x(t_n) \le x_n \mid \mathbf{x}(t_{n-1}),\ \ldots\ldots,\ \mathbf{x}(t_1)] = P[\mathbf{x}(t_n) \le x_n \mid \mathbf{x}(t_{n-1})] \quad (2)$$

3. Transition probabilities

In a discrete-time Markov chain $\{\mathbf{x}_n\}$ with a finite or infinite set of states $e_1, e_2, \ldots, e_i, \ldots$, let $\mathbf{x}_n = \mathbf{x}(t_n)$ represent the state of the system at $t = t_n$. If $t_n = n\boldsymbol{T}$, then for $n \ge m \ge 0$, the sequence $\mathbf{x}_m \to \mathbf{x}_{m+1} \to \ldots \mathbf{x}_n, \ldots$ represents the evolution of the system. Let

$$p_i(m) = P\{\mathbf{x}_m = e_i\} \quad (3)$$

represent the probability that at time $t = t_m$ the system occupies the state e_i, and

$$p_{ij}(m, n) \cong P\{\mathbf{x}_n = e_j \mid \mathbf{x}_m = e_i\} \quad (4)$$

The probability that the system goes into state e_j at $t = t_n$ given that it was in state e_i at $t = t_m$. The numbers $p_{ij}(m, n)$ represent the transition probabilities of the Markov chain from state e_i at t_m *to* e_j at t_n. Equation (3) and (4) completely determine the system, since $m < n < r$,

$$\begin{aligned} &P\{\mathbf{x}_r = e_i, \mathbf{x}_n = e_j, \mathbf{x}_m = e_k\} \\ &= P\{\mathbf{x}_r = e_i \mid \mathbf{x}_n = e_j\}\ P\{\mathbf{x}_n = e_j \mid \mathbf{x}_m = e_k\}\ P\{\mathbf{x}_m = e_k\} \\ &= p_{ji}(n,r)\ p_{kj}(m,n)\ p_k(m) \end{aligned} \quad (5)$$

4. Mixed type population of constant size

Consider two-population type A and B each multiplying independently according to branching processes $[\mathbf{x}_n]$ and $[\mathbf{y}_n]$ given by

$$\mathbf{x}_{n+1} = \sum_{i=1}^{x_n} \xi_i \qquad \mathbf{y}_{n+1} = \sum_{j=1}^{y_n} \eta_j \quad (6)$$

Let

$$P(\xi_i = k) = a_k \geq 0 \quad P(\eta_j = k) = b_k \geq 0 \qquad k = 0, 1, 2, \ldots \qquad (7)$$

represents the respective progency distributions for single individuals in each population. Then

$$A(z) = \sum_{k=0}^{\infty} a_k z^k \qquad B(z) = \sum_{k=0}^{\infty} b_k z^k \qquad (8)$$

represent their respective moment generating functions.

To compute the transition probabilities

$$p_{jk} = P\{\mathbf{x}_{n+1} = k \mid \mathbf{x}_n = j\} \qquad (9)$$

We can use the conditional moment generating function

$$\sum_{k=0}^{\infty} p_{jk} z^k = \sum_{k=0}^{\infty} z^k \; P\{\mathbf{x}_{n+1} = k \mid \mathbf{x}_n = j\} \cong E[z^{x_{n+1}} \mid x_n = j]$$

$$= E[z^{\sum_{i=1}^{j} y_1} \mid x_n = j] = [E\{z^{y_1}\}]^j = P^j(z) \qquad (10)$$

Thus the one-step transition probability p_{jk} is given by the coefficient of z^k in the expansion of $P^j(z)$ is

$$p_{jk} = \{P^j(z)\}_k \qquad (11)$$

$A^i(z)$ gives the generating function for the number of offspring of i individuals for the type-A population, that is,

$$P\{\mathbf{x}_{n+1} = j \mid \mathbf{x}_n = i\} = \{A^i(z)\}_j \qquad (12)$$

The two dimensional process evolves as a sequence of pairs of random variables $(\mathbf{x}_n, y_n)$ composed of the independent branching processes $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_n\}$ so that

$$P[\mathbf{x}_{n+1} = j_1, \mathbf{y}_{n+1} = j_2 \mid \mathbf{x}_n = i_1, \mathbf{y}_n = i_2]$$

$$= P\{\mathbf{x}_{n+1} = j_1 \mid \mathbf{x}_n = i_1\} \; P\{\mathbf{y}_{n+1} = j_2, \mathbf{y}_n = i_2]$$

$$= \{A^{i_1}(z)\}_{j_1} \{B^{i_2}(z)\}_{j_2} \qquad (13)$$

Consider the special situation, where the combined population remains fixed over all generations. Thus

$$x_n + y_n = N \qquad n = 0, 1, 2, \ldots. \qquad (14)$$

In that case if $\{\mathbf{x}_n = i\}$, then necessarily $\{\mathbf{y}_n = N - i\}$, so that the one-step transition probability for the event $\{\mathbf{x}_{n+1} = j\}$ given $\{\mathbf{x}_n = i\}$;

$$p_{ij} = P\{x_{n+1} = j \mid x_n = i, x_n + y_n = x_{n+1} + y_{n+1} = N\}$$

$$= \frac{P\{x_{n+1} = j \mid x_n = i, x_n + y_n = x_{n+1} + y_{n+1} = N\}}{P\{x_{n+1} + y_{n+1} = N,\ x_n = i, x_n + y_n = N\}}$$

$$= \frac{P\{x_{n+1} = j \mid x_{n+1} + y_{n+1} = N \mid x_n = i, x_n + y_n = N\}}{P\{x_{n+1} + y_{n+1} = N \mid x_n = i, x_n + y_n = N\}}$$

$$= \frac{P\{x_{n+1} = j \mid y_{n+1} = N - j \mid x_n = i,\ y_n = N - i\}}{P\{x_{n+1} + y_{n+1} = N \mid x_n = i, y_n = N - i\}}$$

$$p_{ij} = \frac{(A^i(z))_j (B^{N-i}(z))_{N-j}}{(A^i(z)B^{N-i}(z))_N} \qquad i, j = 0,1, \ldots, N \qquad (15)$$

Here we used equation (13) in simplifying the numerator, and the denominator expression follows, since moment generating function for the sum random variable

$$z_{n+1} = x_{n+1} + y_{n+1} = \sum_{m=1}^{x_n} \xi_m + \sum_{m=1}^{y_n} \eta_m \qquad (16)$$

under the condition $x_n = i$, $y_n = N - i$, is given by $A^i(z)B^{N-i}(z)$.

5. Second order binomial model

Suppose the individuals in either population A or B can have at most two progency with common probabilities (for ξ_i or η_i)

$$P\{\xi_i = 0\} = q^2 \qquad P\{\xi_i = 1\} = 2pq \qquad P\{\xi_i = 2\} = p^2 \qquad (17)$$

where $q = 1 - p$ $\qquad 0 < p < 1$

So that their common moment generating function is given by

$$A(z) = B(z) = (q + pz)^2$$

In this case, equation (15) reduces to

$$P_{ij} = \frac{\binom{2i}{j}\binom{2(N-i)}{N-j}}{\binom{2N}{N}} \qquad i,j = 0,1,\ldots\ldots,N \qquad (18)$$

This coincides with the *hypergeometric* genetics model.

6. Poisson population model

Suppose the two branching processes A and B follow independent Poisson progency distribution with mean values λ and μ respectively. Then,

$$A(z) = e^{\lambda(z-1)} \qquad B(z) = e^{\mu(z-1)} \tag{19}$$

and hence from (15) we obtain

$$p_{ij} = \frac{(e^{-i\lambda}(i\lambda)^j / j!)(e^{-(N-i)\mu}[(N-i)\mu]^{N-j}/(N-j)!)}{e^{-(i\lambda+(N-i)\eta)}([i\lambda+(N-i)\mu]^N / N!)}$$

$$p_{ij} = \binom{N}{j}\left(\frac{i\lambda}{i\lambda+(N-i)\mu}\right)^j\left(\frac{(N-i)\mu}{i\lambda+(N-i)\mu}\right)^{N-j} \tag{20}$$

for $i, j = 0, 1, 2. \ldots, N$

This represents a *binomial* model.

In the special case when λ = μ, equation (20) simplifies to

$$p_{ij} = \binom{N}{j}\left(\frac{i}{N}\right)^j\left(1-\frac{i}{N}\right)^{N-j} \qquad i, j = 0, 1, \ldots., N \tag{21}$$

In addition, it coincides with the *binomial* sampling model in genetics model.

7. Genetics

Consider a population that is able to produce new offspring of like kind. For each member let p_k, k = 0, 1, 2, ... represent the probability of creating k new members. The direct descendents of the *n*th generation form the (*n+1*) generation. The members of each generation are independent of each other. Suppose x_n represents the size of *n*th generation. It is clear that x_n depends only on x_{n-1} since $x_n = \sum_{i=1}^{x_{n-1}} y_i$, where y_i represents the number of offspring of the i member of the (n-1) generation, and the manner in which the value of x_{n-1} was reached is of no consequence. Thus, x_n represents a Markov chain.

Suppose each cell of an organism contains two types of genes A and B, the total number of genes in each cell adds up to N. The cell is an state e_j, j = 0, 1, 2, …. , *N*, if it contains exatly j genes of type A and N – j genes of type B. Prior to cell division each gene duplicates itself so that the two new cells in the next generation each inherit N genes chosen at random from the pol of 2j genes of type A and 2N-2j genes of type B. The probability that a new cell has moved into state e_k is given by the hypergeometric distribution

$$p_{jk} = \frac{\binom{2j}{k}\binom{2N-2j}{N-k}}{\binom{2N}{N}} \qquad (22)$$

where $j, k = 0, 1, \ldots, \max(0, 2j - N) \le k \le \min(2j, N)$

In another genetic model, let e_j represent the present state as defined above so that the probability of selecting a gene of type A in the next generation is simply $p = j/N$. Suppose the N genes in he next generation are determined by random selection resulting from N Bernoulli trails with the A-gene probability equal to p. In that case, the transition probability that the next generation has moved into state e_k (k genes of type A and N-k genes of type B) from state e_j is given by the binomial distribution with

$$p_{jk} = \binom{N}{k}\left(\frac{j}{N}\right)^k\left(1-\frac{j}{N}\right)^{N-k} \quad j, k = 0, 1, \ldots, N \qquad (23)$$

It will be interesting to study the limiting behavior of the population based on these models after several generations.

8. Conclusion

Markov processes represent the simplest generalization of independent processes by permitting the outcome at any instant to depend only on the outcome that precedes it. A special kind of Markov process is a Markov chain. Genetics process can specify in a Markov chain, the probability that a new cell has moved from the cell of an organism contains two types of genes A and B was give by the hypergeometric distribution. In another genetic model, the transition probability that the next generation has moved into state e_k from state e_j is given by the binomial distribution. We can study the characteristics of the population based on these models after several generations.

References

[1] Dudewicz, Edward J., Syracuse University; Mishra, Satya N.; University of South Alabama**;** ***Modern Mathematical Statistics***; John Wiley & Sons Inc., 1988

[2] Papoulis, Athanasios; Pillai, S. Unnikrishna; Polytechnic University; ***Probability, Random Variables, and Stochastic Processes***; McGraw-Hill Companies Inc., fourth edition, 2002.

MIKE SUSMIKANTI: P2TIK-BATAN, Serpong-Tangerang, Indonesia

E-mail: mike@batan.go.id

AVERAGE PRICE OPTIONS IN ENERGY MARKETS

S.A.Borovkova[a], F.J.Permana[b]

[a] Delft University of Technology, the Netherlands
[b] Delft University of Technology, the Netherlands

Abstract. In recent years, commodity markets have experienced significant growth, in terms of volumes of trades and financial instruments available to market participants. Commodity futures, swaps and options are routinely traded on exchanges as well as over-the counter. However, commodity option markets are not as developed yet as markets for financial options. Characteristic features of commodities present significant challenges for commodity option pricing and hedging.

Although exchange-traded energy options are not very liquid, over-the-counter (OTC) energy options are becoming a common risk management tool. However, OTC options are exotic, non-European style contracts, designed to suit the needs of a particular customer. Hence, their pricing and hedging requires more sophisticated tools than standard option pricing models, such as Black-Scholes.

Most options in energy markets are Asian-style, i.e. based on the average price of an underlying asset over a certain period. Basket options (for which the underlying value is a weighted sum of a number of assets) are also very common to energy markets, because portfolios of energy companies usually consist of more than one product. In both Asian and basket options, a sum or an average of asset prices play the key role.

In this paper we give an overview of several methods for pricing and hedging Asian and basket options, and derive the corresponding greeks. We compare option prices and hedge costs obtained by these methods on the basis of simulations and historical prices from energy markets.
Key-words: *Asian option, basket option, hedging, the greeks.*

1 Introduction

Since the introduction of Black and Scholes option pricing model, options have become one of the most popular financial instruments. While option markets in equity, foreign exchange or fixed income markets are very well developed, they are relatively new and not as widespread for commodities. In commodity, and in particular, energy markets, exchange-traded options are often illiquid. However, during the past ten years, OTC energy options are becoming a common risk management tool.

Most OTC options in energy are exotic, non-European style contracts. These options are not standardized products, but tailor made instruments designed to suit risk management needs of particular customer (e.g. an oil producing company or an airline). This, and particular features of energy prices make energy options

pricing and hedging more involved, but also more interesting and challenging topic for researchers.

The volatility of energy prices is very high (for example, the crude oil price volatility is always over 30 %), compared to volatilities of stocks or financial indices (which are mostly in the range 10 – 15 %). Consequently, European-style options would be very expensive in energy markets. Market participants hence look for other types of options, which would be more attractive tools for hedging market risk. A typical exotic option which is very common in energy markets is an Asian option. It is cheaper than a European option, because its payoff is based on the average price of an underlying asset (e.g. oil futures contract) over a period of time. Moreover, most delivery contracts in energy are priced on the basis of an average price over a certain period. Companies are interested in hedging price risk in these contracts and hence, they perceive Asian options as a more attractive alternative to European options, better suited to their needs.

Most energy companies have several products in their portfolio (e.g. crude oil, gasoline, natural gas), and are interested in managing the risk of the entire portfolio. So a basket option, whose underlying is the value of a portfolio of assets, is also a typical risk management tool in energy markets.

The underlying of both Asian and basket options is an *average price,* for an Asian option it is the average price of a certain asset over a time period, and for a basket option it is a weighted average of prices on a number of different assets. So option pricing methods are similar for both types of options.

A common model for an asset price process (a stock or index price, a price of a futures or forward contract) is the Geometric Brownian motion. This model assumes that the price distribution is log-normal. Most popular option price models (Black and Scholes (1973), Black (1975)) are heavily dependent on this assumption. However, sums (or averages) of log-normal random variables are not log-normal, which immediately makes option pricing and hedging for Asian and basket options much more complicated. In fact, there exists no closed form solution for the price of an Asian option, even under a simple assumption of log-normality of an asset price.

Wakeman and Turnbull (1991, [8) introduced a quick way to price Asian option, by simply assuming that arithmetic average of log-normally distributed random variables is also log-normally distributed, and matching the first two moments. Vorst (1992, [9)) offered an analytic approach that approximates the arithmetic average option by the geometric average option. Curran (1994, [2) introduced an arithmetic average option pricing model by conditioning on the geometric mean price. All three approaches are widely used in practice, and are incorporated in various option pricing software (e.g. FEA). However, neither of the above papers discusses the greeks, which are vital for option risk management. Vorst (1992,(9)) the expression for hedge ratios, but neither derives it nor discusses it in detail.

The situation is quite similar for basket options. Wakeman again offered the option pricing method based in the log-normal distribution approximation and Milevsky and Posner (1998, [6]) proposed reciprocal-gamma distribution approach for valuing basket option. But again, none of the papers discussed the greeks.

Developing the option hedging strategy is an important part of option pricing. For delta-hedging, the expression for the option's delta is needed. Calculation of other greeks is also essential for risk monitoring and management of an option portfolio. Deriving the greeks is challenging for Asian and basket options because no closed-form solution for the option price is available, and the best one can hope for is some form of approximation. This is probably the reason most papers on Asian or basket option do not discuss the greeks at all or such discussion is very limited. In this paper we shall derive all the greeks for various average price option pricing methods. We shall also compare option prices and hedge costs obtained by the methods mentioned above on the basis of simulations and for historical prices from energy markets.

Most energy options are written on futures or forward contracts or on swaps and not on the physical commodity. Recall that, if the interest rate is constant, then futures and forward prices are equivalent. Also swaps can be reduced to futures, at least if expiry dates of swaps and the corresponding futures coincide. Hence, in this paper we only consider options on futures. Because futures contract does not require initial investment, under the risk-adjusted probability measure, futures price follows the Geometric Brownian Motion (GBM) with zero drift:

$$dF(t,T) = \sigma(t,T)F(t,T)dW(t) \tag{1}$$

where $F(t,T)$ is the futures price at time t for expiry at time T, $t_0 < t$ is the current time, $\sigma(t,T)$ is the futures price volatility, and $W(t)$ is the Wiener process. Under this assumption, $F(t,T)$ is log-normally distributed with mean and variance given by

$$\hat{E}(\log(F(t,T))) = \log(F(t_0,T)) - 0.5\sigma(t,T)(t-t_0) \tag{2}$$

$$\hat{\mathrm{var}}(\log(F(t,T))) = \sigma(t,T)^2(t-t_0). \tag{3}$$

Based on log-normal distribution property, we also have

$$\hat{E}(F(t,T)) = F(t_0,T) \tag{4}$$

The expressions for the volatilities of the futures prices for different expiries $(\sigma(t,T))_T$ (the so-called *term structure of volatilities*) can be obtained from the spot price volatility, by assuming a certain spot price process and deriving the corresponding futures prices by means of risk-neutral valuation, using the fact that, under the risk-neutral probability measure, the futures price is the expected spot price. However, this is not the subject of this paper, and here we assume the futures price volatility to be given and fixed at the time of option's issue. It can be estimated from historical futures prices or implied from liquidly traded options, if those are available.

We shall also assume everywhere that the futures contract and the option written on it expire at the same date *T* (in reality, energy options expire a few days prior to the corresponding futures or forward contract). So everywhere we shall omit the

index *T* denoting the futures expiry date and denote *F(t)* the futures price, which is the asset price underlying the option, and σ - its volatility.

The paper is organized as follows. Section 2 of this paper discusses Asian option pricing by analytical approximation approaches (log-normal distribution approximation, Curran method and Vorst method). Also there we shall derive all the greeks. Section 3 discusses basket option pricing by the log-normal and reciprocal-gamma distribution approximation and the corresponding greeks. Section 4 compares the option prices and hedge costs obtained from various methods on the basis of simulations and for historical data from energy markets. Finally we present conclusions, comments and suggestions for future work in Section 5. Most calculations are presented in the Appendix.

2 Asian Options

Asian options are options whose payoffs are based on the average asset price over a period of time. The averaging period can start at the inception of the contract or at some later date. Here we will be concerned with arithmetic average strike option. This is the typical Asian option whose payoff depends on the arithmetic average price and the strike price. The payoff of this option is:

$$payoff = \max(\eta(F_A - X), 0),$$

where X is the strike price, F_A is the arithmetic average of the futures price over period $[t_0, T]$, η =1 for call option and η = -1 for put option.

The payoff of a European option is based on the outright asset price, while the payoff of an Asian option is based on the average price over some period. Kemna and Vorst (1992,(9)) have shown that the volatility of geometric average price is by the factor $\sqrt{3}$ lower than the volatility of the asset price, when the averaging is done continuously. Since the Asian option can be thought of as a European option whose underlying value is the average asset price, we can rephrase this theoretical result in terms of volatilities behind Asian and European options: the "Asian" volatility (i.e. the one behind Asian option) is approximately by the factor 0.57 lower than the "European" volatility, i.e. the volatility behind the corresponding European option (on the same asset and with the same maturity). This theoretical result implies that the average price (i.e. Asian) option is cheaper than the European option.

As we mentioned earlier, the closed form price formula doesn't exist for options based on the arithmetic average, because the arithmetic average of log-normally distributed variables is not log-normally distributed. One way to derive some kind of closed form solution is to use an analytical approximation approach. We shall review here three analytical approximation methods: Vorst method, Wakeman method and Curran method.

The hedging of Asian option should be considered in two cases: the averaging period starts today or at some later date (newly issued option), or the averaging period has already started (already issued option). In fact, the analysis for newly issued options can be extended for already issued options, so we shall consider

newly issued options first. We shall derive the greeks (delta, gamma, rho, theta and vega) for the Wakeman and Vorst methods. Curran method does not allow for analytical expressions for greeks, but requires time-consuming numerical calculations, so greeks for Curran method will not be considered here.

2.1. Vorst Method

For $t_0 \le t \le T$, the continuous form of geometric average of the futures price over interval $[t_0, T]$ is defined as:

$$G(T) = \exp\left(\frac{1}{T - t_0} \int_{t_0}^{T} \log(F(\tau)).d\tau \right),$$

where T is the maturity date.

Assuming the futures price $F(t)$ follows GBM, Kemna and Vorst (1992,[9]) have shown that the variable $G(T)$ is also log-normally distributed with mean and variance

$$\hat{E}(\log(G(T))) = \log(F(t_0)) + 0.5.(-0.5.\sigma).(T - t_0),$$

$$\hat{Var}(\log(G(T))) = \frac{\sigma^2}{3}.(T - t_0).$$

This implies that the volatility of geometric average of futures price is equal to $\frac{1}{\sqrt{3}}$ (0.57735) of the volatility of futures price.

Vorst approach approximates the value of the arithmetic average strike call by the geometric average strike call. The geometric average strike call price is the lower bound of the arithmetic average strike call price. The upper bound of that can be defined in terms of the arithmetic average strike call price and the expectations of arithmetic average and geometric average. Vorst approximation takes a value between the lower bound and the upper bound.

Assume that $t_i - t_{i-1} = \Delta t, i - 1,2,...,n$ and the average period begins at t_m and ends up at $t_n = T$, where T denotes the maturity date. The option is valued at time t_0. The discrete form of equations (2)-(4) are

$$\hat{E}(\log(F(t_k))) = \log(F(t_0)) - \frac{1}{2}.\sigma^2.(t_k - t_0),$$

$$\hat{\text{var}}(\log(F(t_k))) = \sigma^2.(t_k - t_0),$$

$$\hat{E}(F(t_k)) = F(t_0), \qquad k = m, m+1,...,n.$$

Now we define the discrete form of geometric average as

$$G(T) = \sqrt[N]{\prod_{j=m}^{n} F(t_j)}.$$

Product of lognormal random variables is itself lognormal. Consequently, $G(T)$ is also lognormal. It can be shown (see *Appendix A*) that the parameters of this distribution are

$$M = \hat{E}(\log(G(T))) = \log(F(t_0)) - \tfrac{1}{2}\sigma^2\left\{(t_m - t_0) + \tfrac{1}{2}(T - t_m)\right\},$$

$$V = \hat{\mathrm{var}}(\log(G(T))) = \sigma^2\left\{(t_m - t_0) + \frac{(2N-1)}{6N}(t_n - t_m)\right\},$$

$$\hat{E}(G(T)) = e^{M + \frac{1}{2}V}.$$

The closed form formula of the geometric average strike call price is obtained by applying Black formula:

$$c_G = e^{-rn.\Delta t}.\left(e^{M+\frac{1}{2}V}.N(d_1) - X.N(d_2)\right)$$

where $$d_1 = \frac{\log\left(e^{M+\frac{1}{2}.V}\right) - \log(X) + \frac{1}{2}V}{\sqrt{V}} = \frac{M - \log(X) + V}{\sqrt{V}}$$

$$d_2 = d_1 - \sqrt{V}.$$

The geometric average is always less than arithmetic average. Consequently ,

$$\max(G(T) - X, 0) \le \max(A(T) - X, 0) \le \max(G(T) - X, 0) + A(T) - G(T)$$

.

Taking the expectation to the second inequality, we obtain

$$e^{-r(T-t_0)}\hat{E}(\max(A(T) - X, 0)) \le e^{-r(T-t_0)}\hat{E}(\max(G(T) - X, 0)) + e^{-r(T-t_0)}\left(\hat{E}(A(T)) - E(G(T))\right)$$

It means

$$c_G \le c_A \le c_G + e^{-r(T-t_0)}\left(\hat{E}(A(T)) - \hat{E}(G(T))\right).$$

In other words, we can say that :

- Geometric average strike call price c_G is lower bound of arithmetic average strike call price c_A
- $c_G + e^{-r(T-t_0)}\left(\hat{E}(A(T)) - \hat{E}(G(T))\right)$ is upper bound of arithmetic average strike call price c_A,

where $\hat{E}(A(T)) = \hat{E}\left(\frac{1}{N}\sum_{k=m}^{n} F(t_k)\right) = \frac{1}{N}\sum_{k=m}^{n} \hat{E}(F(t_k)) = F(t_0)$

Vorst approach takes a value between lower bound and upper bound to approximate the call price, and the call price is

$$\hat{c}_A = e^{-r(T-t_0)}\left(e^{M+\frac{1}{2}V}.N(d_1^*) - X^*.N(d_2^*)\right) \quad (5)$$

,where $M = \hat{E}(\log(G(T))) = \log(F(t_0)) - \frac{1}{2}\sigma^2\{(t_m - t_0) + \frac{1}{2}(T - t_m)\}$

$$V = Cov(\log(G(T)), \log(G(T))) = \sigma^2(t_m - t_0) + \frac{\sigma^2.(2N-1)}{6.N}(T - t_m)$$

$$\hat{E}(A(T)) = F(t_0)$$

$$\hat{E}(G(T)) = e^{M+\frac{1}{2}V}$$

$$X^* = X - (E(A(T)) - E(G(T)))$$

$$d_1^* = \frac{M - \log(X^*) + V}{\sqrt{V}}$$

$$d_2^* = d_1^* - \sqrt{V}$$

The value of $\hat{c}_A$ is located between the lower bound c_G and the upper bound $c_G + e^{-r(T-t_0)}\left(\hat{E}(A(T)) - \hat{E}(G(T))\right)$. The error of approximation is

$$|\hat{c}_A - c_A| \le e^{-r(T-t_0)}\left(\hat{E}(A(T)) - \hat{E}(G(T))\right).$$

Call price valued at time t is obtained by substituting t_0 with t in formula (5). The greeks are derived from the call price formula by taking the first and second derivatives. They are simplified as (see *appendix B* for details) :

$$\Delta = \frac{e^{-r(T-t)}}{F(t)}\left(e^{M+\frac{1}{2}V}.N(d_1^*) - (-E(A(T)) + E(G(T))).N(d_2^*)\right).$$

$$\Pi = \frac{-1}{F(t)}.\Delta + E(G(T))\left(N(d_1^*) - N(d_2^*)\right) +$$

$$\frac{e^{-r(T-t)}}{F(t)^2\sqrt{V}}.E(G(T))\left(n(d_1^*) - n(d_2^*)\right) + \frac{e^{-r(T-t)}}{F(t)}\left(N(d_2^*) - n(d_2*)\right).$$

$$\rho = -(T-t).\hat{c}_A.$$

$$\Theta = r.\hat{c} + \frac{\partial \hat{c}_A}{\partial M}.\frac{\partial M}{\partial t} + \frac{\partial \hat{c}_A}{\partial V}.\frac{\partial V}{\partial t}.$$

$$\upsilon = \frac{\partial \hat{c}_A}{\partial M}.\frac{\partial M}{\partial \sigma}. + \frac{\partial \hat{c}_A}{\partial V}\frac{\partial V}{\partial \sigma}$$

where $\frac{\partial \hat{c}_A}{\partial M} = e^{-r(T-t)}.E(G(T)).\left(N(d_1^*) - N(d_2^*)\right);$

$$\frac{\partial \hat{c}_A}{\partial V} = \frac{1}{2}.\frac{\partial \hat{c}_A}{\partial M} + \frac{e^{-r(T-t)}}{2\sqrt{V}}X^*.n(d_2^*);$$

$$\frac{\partial M}{\partial t} = \frac{1}{2}.\sigma^2;\ \frac{\partial V}{\partial t} = -\sigma^2;$$

$$\frac{\partial M}{\partial \sigma} = -\sigma\left\{(t_m - t) + \tfrac{1}{2}(T - t_m)\right\};$$

$$\frac{\partial V}{\partial \sigma} = 2.\sigma.\left\{(t_m - t) + \frac{(2N-1)}{6.N}(T - t_m)\right\}.$$

2.2. Wakeman Method

Wakeman approach is based on the assumption that arithmetic average of lognormal random variables is also lognormal. Under this assumption, we can apply Black formula to value an Asian option by replacing the futures price F_0 with M_1 and the square of volatility σ^2 with $\frac{1}{T-t_0}\log\left(\frac{M_2}{(M_1)^2}\right)$, where M_1 and M_2 are the first and second central moment of arithmetic average variable $A(T)$.

Taking the same assumptions and notations used in Vorst method, we assume that the averaging period starts at t_m, ends up at $t_n = T$ and the option is valued at time t_0. It can be shown (see *appendix C*) that

$$M_1 = F(t_0)$$

$$M_2 = \frac{F(t_0)^2.e^{\sigma^2(t_m - t_0)}}{N^2\left(e^{\sigma^2.\Delta t} - 1\right)}\left\{\left(e^{N\sigma^2\Delta t} - 1\right) + 2\left(\frac{e^{\sigma^2\Delta t}\left(e^{(N-1)\sigma^2\Delta t}\right)}{e^{\sigma^2.\Delta t} - 1} - N + 1\right)\right\}.$$

Suppose that $A(T)$ is lognormally distributed with parameters η and V. From lognormal distribution properties, we know that

$$M_1 = \hat{E}(A(T)) = \exp\left(\eta + \frac{1}{2}V\right) \text{ and } M_2 = \hat{E}\left(A(T)^2\right) = \exp(2\eta + 2V).$$

From those, we obtain

$$V = \log\left(\frac{M_2}{M_1^{\,2}}\right).$$

The closed form formula can be obtained by applying Black formula, replacing F_0 with M_1 and σ with σ^*, where $(\sigma^*)^2 = \frac{1}{T-t_0}\log\left(\frac{M_2}{(M_1)^2}\right)$. The Asian call price at time t_0 is

$$c = e^{-r(T-t_0)}\{M_1.N(d_1) - X.N(d_2)\} \tag{6}$$

where:

$$N = n - m + 1$$

$$M_1 = F(t_0)$$

$$M_2 = \hat{E}\left(A(T)^2\right) = \frac{1}{N^2}\left(M_{21} + 2M_{22}\right)$$

$$M_{21} = F(t_0)^2.e^{\sigma^2(t_m - t_0)}.\left(\frac{e^{N.\sigma^2.\Delta t} - 1}{e^{\sigma^2.\Delta t} - 1}\right)$$

$$M_{22} = \frac{F(t_0)^2.e^{\sigma^2(t_m - t_0)}}{e^{\sigma^2.\Delta t} - 1}\left(e^{\sigma^2.\Delta t}\frac{\left(e^{(N-1)\sigma^2.\Delta t} - 1\right)}{\left(e^{\sigma^2.\Delta t} - 1\right)} - (N-1)\right)$$

$$(\sigma^*)^2 = \frac{1}{T-t_0}\log\left(\frac{M_2}{(M_1)^2}\right)$$

$$d_1 = \frac{\log(M_1) - \log(X) + \frac{1}{2}(\sigma^*)^2(T-t_0)}{\sigma^*\sqrt{T-t_0}}$$

$$d_2 = d_1 - \sigma^*\sqrt{T-t_0}$$

$N(.)$ is cumulative standard normal distribution function.

By substituting t_0 with t in formula (6), we obtain the call price at time t. The greeks are derived by taking the first and second derivatives of call price formula. Assume the greeks are calculated at discrete times $t = t_i$, where $i = 0,1,2,\ldots,m-1$ (newly issued option in which the averaging period start at time t_m). They are simplified as follows (see *Appendix* D for details)

$$\Delta = e^{-r(T-t)}N(d_1)$$

$$\Gamma = e^{-r(T-t)}.n(d_1).\frac{1}{F(t).\sigma^{*}\sqrt{T-t}}.$$

$$\Theta = r.c + e^{-r(T-t)}\left\{M_1.n(d_1).\frac{\partial d_1}{\partial t} - X.n(d_2).\frac{\partial d_2}{\partial t}\right\}$$

$$\rho = -(T-t).c$$

$$\upsilon = e^{-r(T-t)}\left\{F(t).n(d_1).\frac{\partial d_1}{\partial \sigma} - X.n(d_2).\frac{\partial d_2}{\partial \sigma}\right\}, \text{ where}$$

$$\frac{\partial d_1}{\partial t} = -\sigma^2\left[\frac{-1}{2}\left(\log\left(\frac{F(t)}{X}\right)\right)\left\{\log\left(\frac{M_2}{(M_1)^2}\right)\right\}^{-3/2} - \frac{1}{4}\left\{\log\left(\frac{M_2}{(M_1)^2}\right)\right\}^{-1/2}\right]$$

$$\frac{\partial d_2}{\partial t} = -\sigma^2\left[\frac{-1}{2}\left(\log\left(\frac{F(t)}{X}\right)\right)\left\{\log\left(\frac{M_2}{(M_1)^2}\right)\right\}^{-3/2} + \frac{1}{4}\left\{\log\left(\frac{M_2}{(M_1)^2}\right)\right\}^{-1/2}\right]$$

$$\frac{\partial d_1}{\partial M_2} = -\sigma^2\left[\frac{-1}{2M_2}\left(\log\left(\frac{F(t)}{X}\right)\right)\left\{\log\left(\frac{M_2}{M_1^{\,2}}\right)\right\}^{-3/2} - \frac{1}{4M_2}\left\{\log\left(\frac{M_2}{M_1^{\,2}}\right)\right\}^{-1/2}\right]$$

$$\frac{\partial d_2}{\partial M_2} = -\sigma^2\left[\frac{-1}{2M_2}\left(\log\left(\frac{F(t)}{X}\right)\right)\left\{\log\left(\frac{M_2}{M_1^{\,2}}\right)\right\}^{-3/2} + \frac{1}{4M_2}\left\{\log\left(\frac{M_2}{M_1^{\,2}}\right)\right\}^{-1/2}\right]$$

$$U = e^{\sigma^2.\Delta t}.$$

$$\frac{\partial U}{\partial \sigma} = 2.\sigma.\Delta t.e^{\sigma^2.\Delta t} = 2.\sigma.\Delta t.U$$

$$U^{*} = \frac{U^{m-i-1}}{(U-1)^2}\left[(m-i-1)U - m + i\right]$$

$$U^{**} = \frac{U^{m-i}}{(U-1)^3}\left[(m-i-1)U - m + i - 1\right]$$

$$M_{21} = F^2.e^{\sigma^2(m-i)\Delta t}.\left(\frac{e^{N.\sigma^2.\Delta t}-1}{e^{\sigma^2.\Delta t}-1}\right) = F^2.\frac{U^{(m-i)}}{U-1}.\left(U^N - 1\right).$$

$$M_{22} = \frac{F^2.U^{(m-i+1)}}{U-1}\left(U^{(N-1)} - 1\right) - F^2.(N-1)\frac{U^{(m-i)}}{U-1}$$

$$\frac{\partial M_{21}}{\partial U} = F(t)^2.U^*.\left(U^N - 1\right) + F(t)^2.\frac{U^{m-i}}{U-1}.N.U^{(N-1)}$$

$$\frac{\partial M_{22}}{\partial U} = F(t)^2.U^{**}.\left(U^{(N-1)} - 1\right) + F(t)^2.\frac{U^{m+1-i}}{(U-1)^2}.(N-1).U^{(N-2)} - F^2.(N-1).U^*$$

$$\frac{\partial M_2}{\partial U} = \frac{1}{N^2}\left[\frac{\partial M_{21}}{\partial U} + 2.\frac{\partial M_{22}}{\partial U}\right]$$

$$\frac{\partial d_1}{\partial \sigma} = \frac{\partial d_1}{\partial M_2}.\frac{\partial M_2}{\partial U}.\frac{\partial U}{\partial \sigma}$$

$$\frac{\partial d_2}{\partial \sigma} = \frac{\partial d_2}{\partial M_2}.\frac{\partial M_2}{\partial U}.\frac{\partial U}{\partial \sigma}.$$

2.3. Curran Method

If we assume that the futures price $F(t_i)$ is lognormally distributed, then the two variables, $\log(F(t_i))$ and geometric average $\log(G)$ are jointly normally distributed. Curran approach is based on the standard properties of the joint normal distribution. In particular, conditional on $\log(G)$, the log-price $\log(F(t_i))|\log(G) = x$ is normally distributed. Then the call price is derived from this conditional random variable.

Assume that the averaging period starts at time t_m and ends up at time $t_n = T$. From Vorst method we obtained that $\log(F(t_i))$ and $\log(G(T))$ are normally distributed with parameters:

$$\hat{E}(\log(F(t_i))) = \log(F(t_0)) - \tfrac{1}{2}\sigma.(t_i - t_0)$$

$$\sigma_i^2 = \hat{Var}(F(t_j)) = \tfrac{1}{2}\sigma(t_i - t_0)$$

$$\mu_G = M = \hat{E}(\log(G(T))) = \log(F(t_0)) - \tfrac{1}{2}\sigma^2\left\{(t_m - t_0) + \tfrac{1}{2}(T - t_m)\right\}$$

$$\sigma_G^2 = Cov(\log(G(T)), \log(G(T))) = \sigma^2\left\{(t_m - t_0) + \frac{(2N-1)}{6N}(t_n - t_m)\right\}$$

$$\hat{E}(A(T)) = F(t_0)$$

Denote coefficient correlation between $\log(F(t_i))$ and $G(T)$ as ρ_i. Based on joint normal distribution property (see *Appendix E*), $\log(F(t_i))$, conditional on $\log(G(T)) = x$, is normally distributed with mean and variance :

$$E\left[\left(\log\left(F\left(t_i\right)\right)\middle|\log\left(G\left(T\right)\right)=x\right)\right]\log\left(F_0\right)-\frac{1}{2}.\sigma^2\left(t_i-t_0\right)+\left(x-\mu_G\right).\rho_i.\frac{\sigma_i}{\sigma_G}$$

$$Var\left[\left(\log\left(F\left(t_i\right)\right)\middle|\log\left(G\left(T\right)\right)=x\right)\right]=\left(1-\rho_i^{\,2}\right)\sigma_i^{\,2}.$$

Using lognormal distribution property, we obtain

$$E\left[F\left(t_i\right)\middle|\log\left(G\left(T\right)=x\right)\right]$$

$$=F_0\exp\left[-\frac{1}{2}\sigma^2\left(t_i-t_0\right)+\left(x-\mu_G\right)\rho_i\frac{\sigma_i}{\sigma_G}+\frac{1}{2}\left(1-\rho_i^{\,2}\right)\sigma_i^{\,2}\right]$$

Consequently,

$$E\left[A\left(T\right)\middle|G\left(T\right)=x\right]=\frac{1}{N}\sum_{i=m}^{n}F_0.\exp\left[\left(\log\left(x\right)-\mu_G\right).\rho_i.\frac{\sigma_i}{\sigma_G}-\rho_i^{\,2}.\sigma_i^{\,2}\right]$$

It can be shown *(see appendix E)* that the call price can be approximated in simple form as follows

$$\widehat{c}=e^{-r.n.dt}.\left\{\frac{1}{N}.\sum_{i=m}^{n}F_0.N\left(\frac{\mu_G+\sigma_i.\sigma_G.\rho_i-\log\left(LB\right)}{\sigma_G}\right)+X.N\left(\frac{\mu_G-\log\left(LB\right)}{\sigma_G}\right)\right\}$$

where $f\left(x\right)$ is probability density function of $\log\left(G\left(T\right)\right)$.

Lower bound *LB* is given by

$$LB=\arg\min\left[x\middle|E\left(A\left(T\right)\middle|G\left(T\right)=x\right)=X\right].$$

It is clear that $LB<G\left(T\right)<X$.

LB can be calculated using a numerical method (e.g. Newton-Raphson method).

2.4. Already Issued Option

As we mention earlier, the analysis for newly issued options can be extended for already issued options. Suppose that the averaging period is $\left[t_m,T\right]$, where t_n=T, and the option price will be valued at discrete time t_i , $m\le i<n$. In this case, the futures prices $F\left(t_m\right),F\left(t_{m+1}\right),...,F\left(t_i\right)$ have been already observed. Define

$$N=n-m+1,$$

$$N^*=n-i$$

$$B\left(t_i\right)=\frac{1}{i-m+1}\sum_{k=m}^{i}F\left(t_i\right)=\frac{1}{N-N^*}\sum_{k=m}^{i}F\left(t_i\right)$$,and

$$D(t_i,T)=\frac{1}{n-i}\sum_{k=i+1}^{n}F(t_k)=\frac{1}{N^*}\sum_{k=i+1}^{n}F(t_k).$$

It can be shown that

$$A(T)=\frac{N-N^*}{N}B(t)+\frac{N^*}{N}D(t_i,T).$$

Payoff of the call price is :

$$\text{Pay off} = \max(A(T)-X,0)$$

$$= \max\left(\frac{N-N^*}{N}B(t)+\frac{N^*}{N}D(t_i,T)-X,0\right)$$

$$= \frac{N^*}{N}\max(D(t_i,T)-X^*,0), \text{where}$$

$$X^* = \frac{N}{N^*}\left(X-\frac{N-N^*}{N}B(t_i)\right).$$

If $X^* \geq 0$, the option value can be approximated using the method developed for newly issued option where the averaging period starts at t_{i+1} and the strike price is equal to X^*. Afterwards, we have to multiply the result by $\frac{N^*}{N}$.

If $X^* < 0$, $\max(A(T)-X,0)=\frac{N^*}{N}\max(D(t_i,T)-X^*,0)$.

Call price $c = \frac{N^*}{N}.e^{-r.(n-i).\Delta t}.\hat{E}(D(t_i,T)-X^*,0)$

$$= \frac{N^*}{N}.e^{-r.(n-i).\Delta t}.\hat{E}(D(t_i,T)-X^*)$$

$$= \frac{N^*}{N}.e^{-r.(n-i).\Delta t}\left[\hat{E}(D(t_i,T))-X^*\right],$$

where $\hat{E}(D(t_i,T))=\hat{E}\left(\frac{1}{N^*}\sum_{k=i+1}^{n}F(t_k)\right)=F(t_i)$.

Next, we shall derive the call price and the greeks using Vorst method in case of $X^* \geq 0$. Let us define

$$\overline{G} = \prod_{k=i+1}^{n}F(t_k)$$

$$\overline{A} = \frac{1}{N^*}\sum_{k=i+1}^{n} F(t_i) = D(t_i, T).$$

We apply the call price formula for newly issued option, replacing N by N^* and m with $i+1$. The call price at time t_i, $i = m, m+1, m+2, ..., n-1$ is

$$\hat{c} = \frac{N^*}{N} e^{-r(T-t_i)} \left(e^{M+\frac{1}{2}V}.N(d_1^*) - K^*.N(d_2^*) \right),$$

where $M = \hat{E}(\log(\overline{G})) = \log(F(t_i)) - 0.5 * \sigma^2 \{\Delta t + \frac{1}{2}(T - t_{i+1})\}$

$$V = Cov(\log(\overline{G}), \log(\overline{G})) = \sigma^2.\Delta t + \sigma^2 \left(\Delta t + \frac{(2N^* - 1)}{6.N^*}(T - t_{i+1}) \right),$$

$$\hat{E}(\overline{A}) = F(t_i)$$

$$\hat{E}(\overline{G}) = e^{M+\frac{1}{2}V}$$

$$B(t_i) = \frac{1}{m - i + 1}\sum_{k=m}^{i} F(t_k)$$

$$X^* = \frac{N}{N^*}\left(X - \frac{N - N^*}{N} B(t_i) \right)$$

$$K^* = X^* - \left(E(\overline{A}) - E(\overline{G}) \right)$$

$$d_1^* = \frac{M - \log(K^*) + V}{\sqrt{V}}$$

$$d_2^* = d_1^* - \sqrt{V}$$

By taking the first and second derivatives, we derive the greeks.. The greeks formulae at time t_i can be simplified as follows:

$$\Delta = \frac{N^*}{N}.\frac{e^{-r(T-t_i)}}{F(t_i)} \left(e^{M+\frac{1}{2}V}.N(d_1^*) - \left(-E(\overline{A}) + E(\overline{G}) \right).N(d_2^*) \right)$$

$$\Pi = \frac{-1}{F(t_i)}.\Delta + \frac{N^*}{N}.E(\overline{G})\left(N(d_1^*) - N(d_2^*) \right) +$$

$$\frac{N^*}{N}\left\{ \frac{e^{-r(T-t_i)}}{F(t_i)^2\sqrt{V}}.E(\overline{G})\left(n(d_1^*) - n(d_2^*) \right) + \frac{e^{-r(T-t_i)}}{F(t_i)}\left(N(d_2^*) - n(d_2 *) \right) \right\}$$

$$\rho = -(T - t_i).\hat{c}$$

$$\Theta = r.\hat{c} + \frac{\partial \hat{c}}{\partial M}.\frac{\partial M}{\partial t} + \frac{\partial \hat{c}}{\partial V}.\frac{\partial V}{\partial t}$$

$$\upsilon = \frac{\partial \hat{c}}{\partial M}.\frac{\partial M}{\partial \sigma} + \frac{\partial \hat{c}}{\partial V}.\frac{\partial V}{\partial \sigma}$$

where $\frac{\partial \hat{c}}{\partial M} = \frac{N^*}{N}.e^{-r(T-t_i)}.E(\overline{G}).\left(N(d_1^*) - N(d_2^*)\right)$

$$\frac{\partial \hat{c}}{\partial V} = \frac{N^*}{N}\left\{\frac{1}{2}.\frac{\partial \hat{c}}{\partial M} + \frac{e^{-r(T-t_i)}}{2\sqrt{V}}X^*.n(d_2^*)\right\}$$

$$\frac{\partial M}{\partial t} = \frac{1}{2}.\sigma^2;\ \frac{\partial V}{\partial t} = -\sigma^2.$$

$$\frac{\partial M}{\partial \sigma} = -\sigma\left\{\Delta t + \tfrac{1}{2}(T - t_{i+1})\right\};$$

$$\frac{\partial V}{\partial \sigma} = 2.\sigma.\left\{\Delta t + \frac{(2N^* - 1)}{6.N^*}(T - t_{i+1})\right\}.$$

In case of $X^* < 0$,

$$\hat{E}\left(D(t_i, T) - X^*, 0\right) = \hat{E}\left(D(t_i, T) - X^*\right) = \hat{E}\left(D(t_i, T)\right) - X^*.$$

The call price at time t_i, $i = m, m+1, m+2, \ldots, n-1$ is

$$\hat{c}_A = \frac{N^*}{N}e^{-r(T-t_i)}\left(F(t_i) - K^*\right),$$

where K^* is defined in the same way as in case of $X^* \geq 0$.

Using the same notations used in the calculation for already issued options by Vorst method, Wakeman method approximates the call price at time t_i, $m \leq i < n$ for $X^* \geq 0$ by the following formula:

$$c_i = \frac{N^*}{N}e^{-r(T-t_i)}\left\{M_1.N(d_1) - X^*.N(d_2)\right\}$$

where: $M_1 = \hat{E}(\overline{A}) = F(t_i)$

$$B(t_i) = \frac{1}{m - i + 1}\sum_{k=m}^{i} F(t_k)$$

$$M_2 = \hat{E}\left(\overline{A}^2\right) = \frac{1}{\left(N^*\right)^2}\left(M_{21} + 2M_{22}\right)$$

$$M_{21} = F\left(t_i\right)^2 . e^{\sigma^2 . \Delta t} . \left(\frac{e^{N^* . \sigma^2 . \Delta t} - 1}{e^{\sigma^2 . \Delta t} - 1}\right)$$

$$M_{22} = \frac{F\left(t_i\right)^2 . e^{\sigma^2 . \Delta t}}{e^{\sigma^2 . \Delta t} - 1}\left(e^{\sigma^2 . \Delta t}\frac{\left(e^{\left(N^* - 1\right)\sigma^2 . \Delta t} - 1\right)}{\left(e^{\sigma^2 . \Delta t} - 1\right)} - \left(N^* - 1\right)\right)$$

$$\left(\sigma^*\right)^2 = \frac{1}{T - t_i}\log\left(\frac{M_2}{\left(M_1\right)^2}\right)$$

$$d_1 = \frac{\log\left(M_1\right) - \log\left(X\right) + \frac{1}{2}\left(\sigma^*\right)^2\left(T - t_i\right)}{\sigma^*\sqrt{T - t_i}}$$

$$d_2 = d_1 - \sigma^*\sqrt{T - t_i}$$

The greeks formulas can be simplified as follows *(* see *appendix F* for details*):*

$$\Delta = \frac{N^*}{N} . e^{-r\left(T - t_i\right)} N\left(d_1\right)$$

$$\Gamma = \frac{N^*}{N} e^{-r\left(T - t_i\right)} . n\left(d_1\right) . \frac{1}{F\left(t_i\right) . \sigma^*\sqrt{T - t_i}} .$$

$$\rho = -\frac{N^*}{N}\left(T - t_i\right) . c$$

$$\Theta = r . c + \frac{N^*}{N} . e^{-r\left(T - t_i\right)}\left\{M_1 . n\left(d_1\right) . \frac{\partial d_1}{\partial t} - X^* . n\left(d_2\right) . \frac{\partial d_2}{\partial t}\right\}$$

$$\upsilon = \frac{N^*}{N} . e^{-r\left(T - t_i\right)}\left\{F\left(t_i\right) . n\left(d_1\right) . \frac{\partial d_1}{\partial \sigma} - X^* . n\left(d_2\right) . \frac{\partial d_2}{\partial \sigma}\right\},$$

where

$$\frac{\partial d_1}{\partial t} = -\sigma^2\left[\frac{-1}{2}\left(\log\left(\frac{F\left(t_i\right)}{X^*}\right)\right)\left\{\log\left(\frac{M_2}{\left(M_1\right)^2}\right)\right\}^{-3/2} - \frac{1}{4M_2}\left\{\log\left(\frac{M_2}{{M_1}^2}\right)\right\}^{-1/2}\right]$$

$$\frac{\partial d_2}{\partial t} = -\sigma^2\left[\frac{-1}{2}\left(\log\left(\frac{F(t_i)}{X^*}\right)\right)\left\{\log\left(\frac{M_2}{(M_1)^2}\right)\right\}^{-3/2} + \frac{1}{4M_2}\left\{\log\left(\frac{M_2}{M_1^{\,2}}\right)\right\}^{-1/2}\right]$$

$$\frac{\partial d_1}{\partial M_2} = -\sigma^2\left[\frac{-1}{2M_2}\left(\log\left(\frac{F(t_i)}{X^*}\right)\right)\left\{\log\left(\frac{M_2}{M_1^{\,2}}\right)\right\}^{-3/2} - \frac{1}{4M_2}\left\{\log\left(\frac{M_2}{M_1^{\,2}}\right)\right\}^{-1/2}\right]$$

$$\frac{\partial d_2}{\partial M_2} = -\sigma^2\left[\frac{-1}{2M_2}\left(\log\left(\frac{F(t_i)}{X^*}\right)\right)\left\{\log\left(\frac{M_2}{M_1^{\,2}}\right)\right\}^{-3/2} + \frac{1}{4M_2}\left\{\log\left(\frac{M_2}{M_1^{\,2}}\right)\right\}^{-1/2}\right]$$

$$U = e^{\sigma^2.\Delta t}.$$

$$\frac{\partial U}{\partial \sigma} = 2.\sigma.\Delta t.e^{\sigma^2.\Delta t} = 2.\sigma.\Delta t.U$$

$$U^* = \frac{-1}{(U-1)^2}$$

$$U^{**} = \frac{-2U}{(U-1)^3}$$

$$\frac{\partial M_{21}}{\partial U} = F(t_i)^2.U^*.\left(U^{N^*} - 1\right) + F(t_i)^2.\frac{U}{U-1}.N^*.U^{(N^*-1)}$$

$$\frac{\partial M_{22}}{\partial U} =$$

$$F(t_i)^2.U^{**}.\left(U^{(N^*-1)} - 1\right) + F(t_i)^2.\frac{U^2}{(U-1)^2}.(N^* - 1)U^{(N^*-2)} - F^2.(N^* - 1)U^*$$

$$\frac{\partial M_2}{\partial U} = \frac{1}{(N^*)^2}\left[\frac{\partial M_{21}}{\partial U} + 2.\frac{\partial M_{22}}{\partial U}\right]$$

$$\frac{\partial d_1}{\partial \sigma} = \frac{\partial d_1}{\partial M_2}.\frac{\partial M_2}{\partial U}.\frac{\partial U}{\partial \sigma}$$

$$\frac{\partial d_2}{\partial \sigma} = \frac{\partial d_2}{\partial M_2}.\frac{\partial M_2}{\partial U}.\frac{\partial U}{\partial \sigma}$$

3 Basket Options

A basket option is an option whose payoff depends on the value of a portfolio (basket) of underlying assets. Although the weight of underlying assets price in the basket can be negative (e.g portfolio of an oil company which buys crude oil and

sells gasoline), in this paper we only consider basket option with all weights being positive. We shall discuss two analytical approximation approaches for basket option valuation: the lognormal and reciprocal-gamma distribution approaches.

For basket option with N assets, not only volatilities of asset prices but also correlations between assets play an important role. Also, when delta-hedging a basket option, we need to calculate exposures of the option to all the underlying assets. So we define delta (Δ) as a vector of length N consisting of $\Delta_i = \frac{\partial c}{\partial F_i(t)}$, *i=1,2,…,N.* Gamma (Γ) is defined as matrix (NxN), where $\Gamma_{i,j} = \frac{\partial^2 c}{\partial F_j(t).\partial F_i(t)}$.

Vega (ν) of a basket option measures the sensitivity of the call price with respect to volatility (ν_i) and correlation $(\rho_{i,j})$. It is defined as a matrix (NxN), where $\nu_{i,j} = \frac{\partial c}{\partial \sigma_i}$ if $i = j$, and $\nu_{i,j} = \frac{\partial c}{\partial \rho_{i,j}}$ if $i \neq j$.

3.1. Log-normal Distribution Approach

This approach is introduced by Wakeman and is essentially the same as that for Asian options. It is again based on the assumption that the sum of lognormals is also lognormal. We calculate the first two central moments of the basket at the maturity of the option in a risk-neutral world, and then we assume that the value of the basket is lognormally distributed at that time. The option can be regarded as an option on a futures contract, since we already assumed that all underlying assets are futures. It means Black model can be applied to calculate the option price.

Suppose that there are N futures (assets) in the basket. Each asset price $F_i(t)$ follows GBM with the volatility of asset i denoted by σ_i and the log-expected return correlation between two different assets, indexed by i and j, is denoted by $\rho_{i,j}$. The basket value is given by:

$$B(t) = \sum_{i=1}^{N} a_i . F_i(t),$$

where a_i is weight of the asset i and a_i's satisfy $\sum_{i=1}^{N} a_i = 1$. The first two central moments are :

$$M_1 = \hat{E}(B(T)) = \hat{E}\left(\sum_{i=1}^{N} a_i . F_i(T)\right) = \sum_{i=1}^{N} a_i . \hat{E}(F_i(T)) = \sum_{i=1}^{N} a_i . F_i(t_0)$$

$$M_2 = \hat{E}\left(B(T)^2\right) = \sum_{j=1}^{N}\sum_{i=1}^{N} a_i.a_j F_i(t_o).F_j(t_o).\exp\left(\rho_{ij}.\sigma_i.\sigma_j.(T-t_0)\right)$$

Assume that the basket value $B(T)$ is lognormally distributed with parameters μ and $\sqrt{V}$. It means that

$$M_1 = \hat{E}(B(T)) = \exp(M + 0.5.V) \text{ and}$$
$$M_2 = \hat{E}\left(B(T)^2\right) = \exp(2.M + 2.V).$$

It can be shown that $V = \hat{Var}(\log(B(T))) = \log\left(\frac{M_2}{M_1^{\,2}}\right)$.

Regarding this option as an option on futures contract, we can apply Black formula to value the call by substituting F_0 with M_1 and $\sigma^2.(T-t_0)$ with $\sqrt{V}$. The closed-form formula for the call price is :

$$c = \exp(-r.(T-t_0))[M_1.N(d_1) - X.N(d_2)], \tag{7}$$

where $d_1 = \dfrac{\log(M_1/X) + 0.5.V}{\sqrt{V}}$

$d_2 = d_1 - \sqrt{V}$.

When the call price is valued at time t , formula (7) can be applied by substituting t_0 with t . By taking the first and the second derivatives of call price formula, the final expression of the greeks is simplified to *(*see *appendix G* for details*)*

$$\Delta_i = \exp(-r(T-t))\left[\frac{\partial M_1}{\partial F_i(t)}.N(d_1) + M_1.n(d_1).\frac{\partial d_1}{\partial F_i(t)} - X.n(d_2).\frac{\partial d_2}{\partial F_i(t)}\right]$$

$$\Gamma_{i,j} = e^{-r(T-t)}\left[a_i n(d_1)\frac{\partial d_1}{\partial F_j(t)} + Xn(d_2)\frac{\partial^2\sqrt{V}}{\partial F_j(t).\partial F_i(t)} + X\frac{\partial n(d_2)}{\partial F_j(t)}\frac{\partial\sqrt{V}}{\partial F_i(t)}\right]$$

$$\rho = -(r(T-t)).c.$$

$\Theta = r.c + \exp(-r(T-t)).X.n(d_2).\dfrac{\partial\sqrt{V}}{\partial t}$, since $M_1.n(d_1) = X.n(d_2)$.

$$v_{i,j} = \begin{cases} \exp(-r(T-t)).X.n(d_2).\dfrac{\partial\sqrt{V}}{\partial\rho_{i,j}}, & i \neq j \\ \exp(-r(T-t)).X.n(d_2).\dfrac{\partial\sqrt{V}}{\partial\sigma_i}, & i = j \end{cases},$$

where $\frac{\partial M_1}{\partial F_i(t)} = a_i$

$$\frac{\partial M_2}{\partial F_i(t)} = 2a_i \sum_{j=1}^{N} a_j . F_i(t) . F_j(t) . \exp\left(\sigma_i . \sigma_j . \rho_{i,j} (T-t)\right)$$

$$\frac{\partial \sqrt{V}}{\partial F_i(t)} = \frac{1}{2M_1 . M_2 \sqrt{V}} . \left[\frac{\partial M_2}{\partial F_i(t)} . M_1 - 2. \frac{\partial M_1}{\partial F_i(t)} . M_2 \right]$$

$$\frac{\partial d_2}{\partial F_i(t)} = \frac{\partial d_1}{\partial F_i(t)} - \frac{\partial \sqrt{V}}{\partial F_i(t)} \text{ since } d_2 = d_1 - \sqrt{V} .$$

$$\frac{\partial d_1}{\partial F_i(t)} = \frac{\frac{1}{M_1} . \frac{\partial M_1}{\partial F_i(t)} . \sqrt{V} - \log(M_1 / X) . \frac{\partial \sqrt{V}}{\partial F_i(t)}}{V} + \frac{1}{2\sqrt{V}} \frac{\partial \sqrt{V}}{\partial F_i(t)}$$

$$\frac{\partial^2 M_2}{\partial F_j(t) . \partial F_i(t)} = 2 . a_i . a_j . \exp\left(\sigma_i . \sigma_j . \rho_{i,j} (T-t)\right)$$

$$w_{1,i} = . \frac{\partial M_2}{\partial F_i(t)} . M_1 - 2. \frac{\partial M_1}{\partial F_i(t)} . M_2$$

$$w_2 = 2 . M_1 . M_2 . \sqrt{V}$$

$$\frac{\partial^2 \sqrt{V}}{\partial F_j(t) . \partial F_i(t)} = \frac{\frac{\partial w_{1,i}}{\partial F_j(t)} . w_2 - w_{1,i} . \frac{\partial w_2}{\partial F_j(t)}}{{w_2}^2}$$

$$\frac{\partial w_{1,i}}{\partial F_j(t)} = a_j . \frac{\partial M_2}{\partial F_i(t)} + M_1 . \frac{\partial^2 M_2}{\partial F_j(t) . \partial F_i(t)} - 2 . a_i . \frac{\partial M_2}{\partial F_j(t)} \text{ and}$$

$$\frac{\partial w_2}{\partial F_j(t)} = 2 . M_2 . \sqrt{V} . \frac{\partial M_1}{\partial F_j(t)} + 2 . M_1 . \sqrt{V} . \frac{\partial M_2}{\partial F_j(t)} + 2 . M_1 . M_2 . \frac{\partial \sqrt{V}}{\partial F_j(t)} .$$

$$\frac{\partial n(d_2)}{\partial F_i(t)} = - d_2 . n(d_2) . \frac{\partial d_2}{\partial F_i(t)}$$

$$\frac{\partial \sqrt{V}}{\partial t} = \frac{\partial \sqrt{V}}{\partial M_2} . \frac{\partial M_2}{\partial t}$$

$$\frac{\partial \sqrt{V}}{\partial M_2} = \frac{1}{2 . M_2 . \sqrt{V}}$$

$$\frac{\partial M_2}{\partial t} = \sum_{j=1}^{N}\sum_{i=1}^{N} -\rho_{i,j}.\sigma_i.\sigma_j.a_i.a_j.F_i(t).F_j(t).\exp\left(\sigma_i.\sigma_j.\rho_{i,j}(T-t)\right).$$

$$\frac{\partial\sqrt{V}}{\partial\rho_{i,j}} = \frac{1}{2.M_2.\sqrt{V}}.\frac{\partial M_2}{\partial\rho_{i,j}}$$

$$\frac{\partial M_2}{\partial\rho_{i,j}} = 2.a_i.a_j.F_i(t).F_j(t).\sigma_i.\sigma_j.(T-t).\exp\left(\sigma_i.\sigma_j.\rho_{i,j}(T-t)\right)$$

$$\frac{\partial M_2}{\partial\sigma_i} = 2\sum_{j=1}^{N}\rho_{i,j}.\sigma_j.(T-t).a_i.a_j.F_i(t).F_j(t).\exp\left(\sigma_i.\sigma_j.\rho_{i,j}(T-t)\right).$$

$$\frac{\partial\sqrt{V}}{\partial\sigma_i} = \frac{1}{2.M_2.\sqrt{V}}.\frac{\partial M_2}{\partial\sigma_i}.$$

3.2. Reciprocal- Gamma Distribution Approach.

This approach was introduced by Milevsky and Posner (1998, [6]). It is also moment matching approach, based on the assumption that the sum of lognormal random variables is reciprocal-gamma distributed. This is based on the theoretical result stating that, if the number of lognormal random variables tends to infinity, the distribution of their sum tends to reciprocal Gamma distribution. Of course, in a typical application to basket options there are fewer than 10 assets in a basket, but Milevsky and Posner (1998, [7]) argue that reciprocal Gamma distribution still approximates the sum of lognormals better than any other distribution, e.g. another lognormal (as in Wakeman method).

If random variable X is gamma distributed with parameters of α and β ($Gamma(\alpha,\beta)$), then the probability density function of X is

$$g(x,\alpha,\beta) = \frac{e^{-x/\beta}.\left(\frac{x}{\beta}\right)^{\alpha-1}}{\beta.\Gamma(\alpha)},\ x\geq 0,\ \alpha,\beta>0.$$

A random variable Y, which is reciprocal-gamma distributed, is defined by the relationship of distribution functions between gamma and reciprocal-gamma :

$$G_R(y,\alpha,\beta) = 1 - G(1/y,\alpha,\beta),\ \forall y>0,\ \alpha,\beta>0.$$

Consequently, the probability density functions of those are related as:

$$g_R(x,\alpha,\beta) = \frac{g(1/y,\alpha,\beta)}{y^2},\ x\geq 0,\ \alpha,\beta>0$$

(8)

It can be shown that

$$E\left(Y^i\right)=\frac{1}{\beta^i(\alpha-1)(\alpha-2)...(\alpha-i)},\ \text{i=1,2,3,...} \tag{9}$$

Milevsky and Posner (1998, [6]) have shown that the distribution of the arithmetic average of stocks prices $\frac{1}{N}\sum_{i=1}^{N}S(t_i)$ converges to reciprocal gamma distribution when both $N\to\infty$ and $T\to\infty$, and $r-q<\frac{1}{2}\sigma^2$. (It is assumed that the stock price $S(t_i)$ follows GBM with the expected return μ (in risk neutral world the expected return μ is equal to the risk-free interest rate r), the continuous dividend yield rate q, volatility σ and time to maturity *T*). In practice, both N and T are, of course, finite, and in fact, rather small. But Milevsky and Posner (1998, (7)) has shown that the reciprocal gamma distribution can approximate the sum of lognormals better than lognormal distribution in the Kolmogorov-Smirnov sense.

Using the same notation as for lognormal distribution approach, we define $H=\sum_{i=1}^{N}a_i.F_i(t_0)$. If $B(T)$ is divided by H (we denote this ratio $B^*(T)$), the basket value is normalized to have mean 1. We assume that $B(T)$ is reciprocal gamma distributed. Consequently $B^*(T)$ is also reciprocal gamma distributed. Taking into account the first two central moments of $B^*(T)$ we have

$$M_1=\hat{E}\left(B^*(T)\right)=1$$

$$M_2=\hat{E}\left(B^*(T)^2\right)=\hat{E}\left(\frac{1}{H^2}B(T)^2\right)=\frac{1}{H^2}\sum_{j=1}^{N}\sum_{i=1}^{N}a_i.a_j.F_i(t_0).F_j(t_0).\exp\left(\rho_{ij}.\sigma_i.\sigma_j.T\right).$$

Suppose that $B^*(T)$ is reciprocal-gamma distributed with parameters α and β. From eq. (9) we have

$$\hat{E}\left(B^*(T)\right)=\frac{1}{\beta.(\alpha-1)}$$

$$\hat{E}\left(B^*(T)^2\right)=\frac{1}{\beta^2(\alpha-1)(\alpha-2)}.$$

It means $\beta=\frac{1}{\alpha-1}$.

In term of M_2, the parameters α and β can be written as

$$\alpha = \frac{2.M_2 - 1}{M_2 - 1}$$

$$\beta = \frac{1}{\alpha - 1} = 1 - \frac{1}{M_2} \text{ since } \beta = \frac{1}{\alpha - 1}.$$

Using the reciprocal gamma distribution with these parameters and using eq. (8), the call price formula is *(see appendix H)*

$$c = \exp(-r(T - t_0))[H.G(u, \alpha - 1, \beta) - X.G(u, \alpha, \beta)] \tag{10}$$

Cumulative distribution function of Gamma distribution $G(d)$ in the closed form formula plays the same role to cumulative normal distribution function $N(d)$ in Black's formula for the call price.

The call price valued at time t can be calculated by substituting t_0 with t in eq. (10). The greeks are derived from the call price formula by taking the first and second derivatives of the call price.

Parameters α and β are functions of the random variable M_2, and M_2 in turn is a function of random variables $F_i(t)$, t and parameters σ_i and $\rho_{i,j}$. Consequently, parameters α and β in the call price formula in eq.(10) are also some functions of $F_i(t)$, t, σ_i and $\rho_{i,j}$. This implies that reciprocal-gamma approximation approach does not allow for analytical expression of the greeks, but requires numerical calculations. One alternative is to regard the parameters of α and β as constants. Then we can derive the closed form formula for delta under this assumption:

$$\Delta_i = \frac{\partial c}{\partial F_i(t)}$$

$$= e^{-r(T-t)}\left[\frac{\partial H}{\partial F_i(t)} G\left(\frac{H}{X}\middle|\alpha - 1, \beta\right) + \frac{H}{X} g\left(\frac{H}{X}\middle|\alpha - 1, \beta\right) - \frac{X}{X} g\left(\frac{H}{X}\middle|\alpha, \beta\right)\right]$$

$$= e^{-r(T-t)}.a_i.G\left(\frac{H}{X}\middle|\alpha - 1, \beta\right), \text{ since } H.g\left(\frac{H}{X}\middle|\alpha - 1, \beta\right) = X.g\left(\frac{H}{X}\middle|\alpha, \beta\right)$$

where $g(.,\alpha,\beta)$ is the probability density function of Gamma distribution with parameters α and β.

4 Simulation Study

We compare approximation values for the arithmetic average call option price using three approaches described above, with the values obtained by the Monte Carlo simulations. Table 1 summarizes the results.

In general, the prices obtained by all the methods (Monte Carlo, Wakeman, Curran and Vorst) are relatively close, but lognormal approximation values are closest to the values obtained by Monte Carlo. This table also shows that the geometric average price is indeed the lower bound for the call price, as the theoretical arguments imply.

Table 1: Asian call option price

Strike price	***36***	***39***	***40***	***41***	***43***
Monte Carlo	5.0674	3.2112	2.7065	2.2667	1.5487
Wakeman	4.9811	3.1383	2.6450	2.2106	1.5068
Curran	4.9807	3.1381	2.6448	2.2105	1.5069
Vorst	4.9794	3.1364	2.6429	2.2086	1.5049
Geometric Average	4.9590	3.1212	2.6295	2.1968	1.4962

Note: Fo=40; r=2.5 %; σ =35 %; Δt =1/365 (1 year = 365 trading days); Averaging period :[75,100]; number of replication for MC = 10,000

Option issuer needs to hedge the option and monitor his risk exposure. For this, the hedge ratio (i.e. the option's delta) and other greeks are important. Performance of delta hedge is investigated by calculating the hedge cost of an option and the hedge error. Using price paths generated by Monte Carlo simulation we calculate hedge costs and hedge errors. Tables 2 and 3 summarize our results.

Hedge error is defined as (hedge cost – call price) and % hedge error is defined as (total cost / call price)x100 %. Hedge errors are relatively low for all methods. It indicates that the delta hedging strategy performs well. Although total costs and hedge errors obtained by Wakeman method and Vorst method are relatively close, Wakeman method gives a better performance: its hedge errors are lower than those obtain by Vorst method. This conclusion is supported by histograms 1-6 and Table 4. We generated 1000 price paths, and the distribution of hedge errors are represented by these histograms. For both methods the hedge errors are in the range 3.6-6 % and are approximately normally distributed.

In an ideal world, hedging should be done continuously. In practice, hedging is done at discrete time intervals (e.g. daily). The hedge frequency affects the hedge error. The hedge error is smaller if the hedge frequency is higher. We can observe

the distribution of hedge error when the hedge frequency varies. Table 5 and histograms 7-12 present the results. In this case the hedge error is around 5 – 6 %. Hedge performance of both Wakeman and Vorst methods are relatively the same. Table 5 also shows that the hedge error tends to be smaller both in the mean and the standard deviation if the hedge frequency is higher. Histograms 7-12 show that the hedge error is approximately normally distributed.

How well the greeks are approximated can be checked plotting the greeks versus the underlying price (futures price). These plots are displayed in Figures 1-10. The behavior of all the greeks is very similar to those for European options.

Table 2: In the money (A(T) > X)

Method	*Wakeman*	*Vorst*
Hedge cost	3.1625	3.1629
Call price	3.1383	3.1364
Hedge error	0.0242	0.0266
	(0.77 %)	(0.85 %)

Note:Fo=40; A(T)=42.36; X=39;
T= 100 days; r=2.5 %; σ =35 %

Table 3: Out of the money A(T) < X

Method	*Wakeman*	*Vorst*
Hedge cost	1.5644	1.5633
Call price	1.5068	1.5049
Hedge error	0.0577	0.0584
	(3.83 %)	3.88 (%)

Note: Fo=40; A(T)=35.88; X=43 ;
T= 100 days; r=2.5 %; σ =35 %

Table 4: Distribution of hedge error

strike price	*36*		*40*		*43*	
method	*Vorst*	*Wakeman*	*Vorst*	*Wakeman*	*Vorst*	*Wakeman*
Mean	0.1822 3.66 %	0.1805 3.62 %	0.1352 5.12 %	0.1333 5.04 %	0.0916 6.09 %	0.0898 5.96 %
standard deviation	0.1992	0.1993	0.2267	0.2266	0.2468	0.2466
Minimum	-0.5581	-0.5609	-0.5439	-0.5470	-0.8019	-0.8052
Maximum	1.3405	1.3375	0.9319	0.9283	1.3150	1.3111

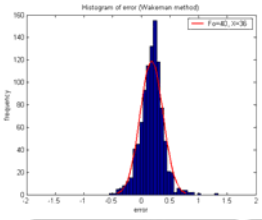
Note: Fo=40; r=2.5 %; σ =35 %; Δt =1/365 (365 trading days in a year)
Averaging period [75,100];

Table 5: Distribution of hedge error for various hedge frequency

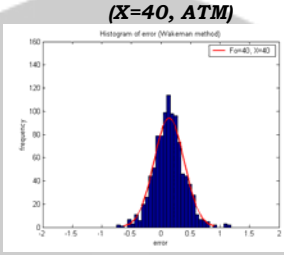
hedge frequency	*10 days*		*5 days*		*1 days*	
Method	*Vorst*	*Wakeman*	*Vorst*	*Wakeman*	*Vorst*	*Wakeman*
call price	2.6429	2.645	2.6429	2.6450	2.6429	2.6450
Mean	0.1491 5.64 %	0.1471 5.56 %	0.1456 5.51 %	0.1436 5.43 %	0.1403 5.31 %	0.1383 5.23 %
standard deviation	0.9148	0.9148	0.5988	0.5988	0.2486	0.2486
minimum	-2.6398	-2.6412	-1.6288	-1.6305	-0.7236	-0.7264
maximum	4.4303	4.4287	2.4658	2.4616	1.1578	1.1547

Note: Fo=40; X=40; r=2.5 %; σ =35 %; Δt =1/365 (1 year = 365 trading days).
Averaging period: [75,100];Number of replication = 1000.

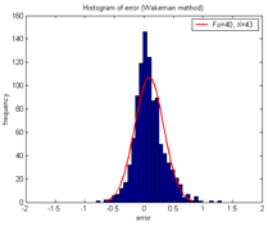
Histogram1: Wakeman method (X=36,OTM)



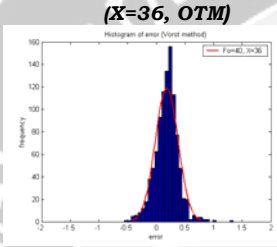
Histogram 2: Wakeman method (X=40, ATM)



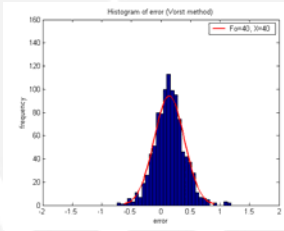
Histogram 3: Wakeman method (X=43, ITM)



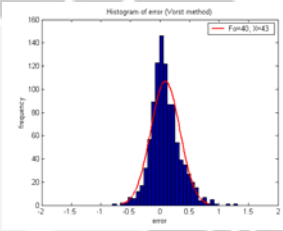
Histogram 4: Vorst method (X=36, OTM)



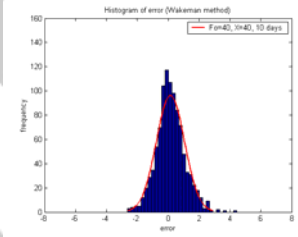
Histogram 5: Vorst method (X=40, ATM)



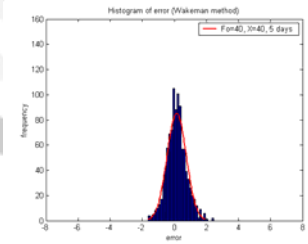
Histogram 6: Vorst method (X=43, ITM)



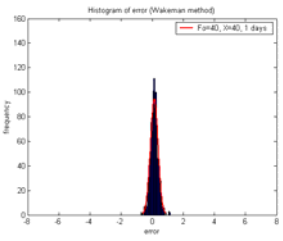
Histogram 7 : Wakeman method (10 days)



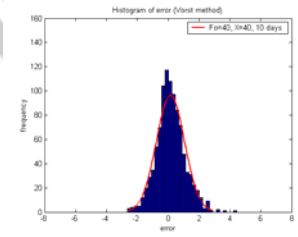
Histogram 8: Wakeman method (5 days):



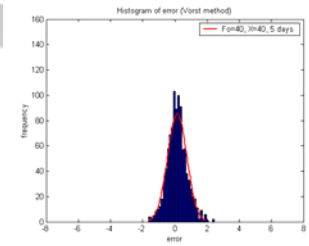
Histogram 9: Wakeman method (1 day)



Histogram 10: Vorst method (10 day)



Histogram 11: Vorst method (5 days)



Histogram 12: Vorst method (1 day)

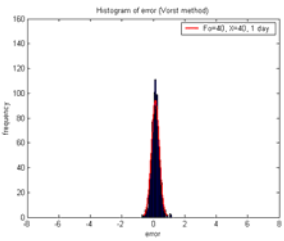


Figure 1:
Delta (Wakeman method)

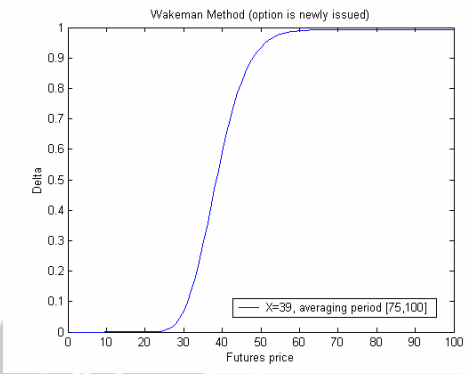


Figure 2:
Delta (Vorst method)

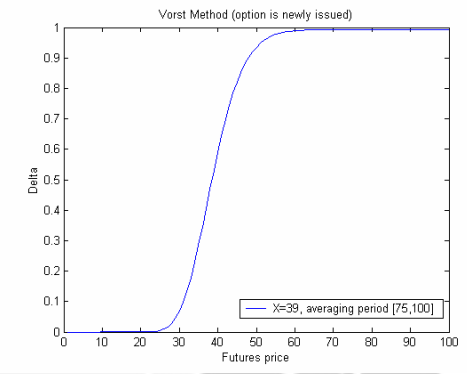


Figure 3:
Gamma (Wakeman method)

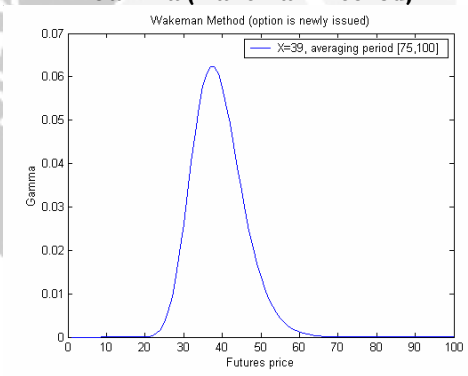


Figure 4:
Gamma (Vorst method)

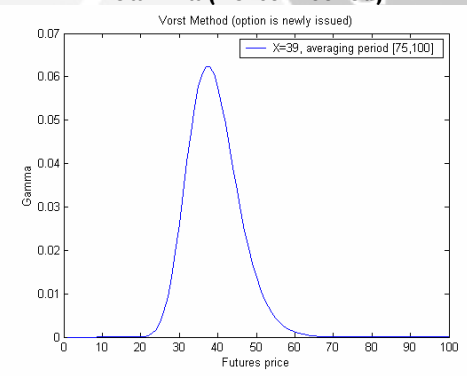


Figure 5:
Rho (Wakeman method)

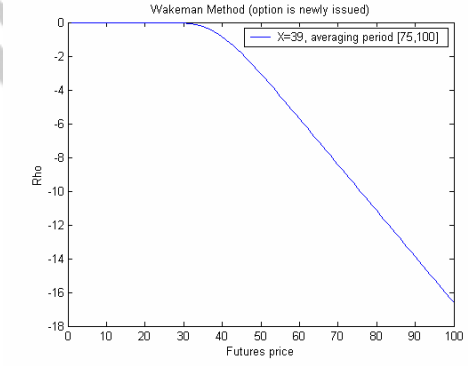


Figure 6:
Rho (Vorst method)

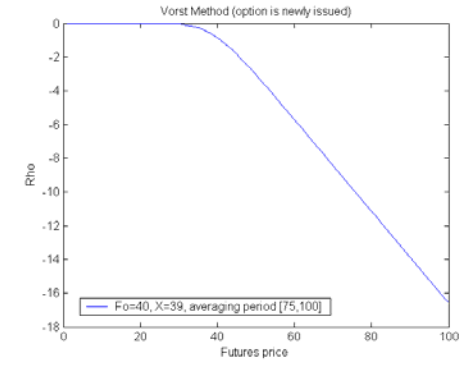


Figure 7:
Theta (Wakeman method)

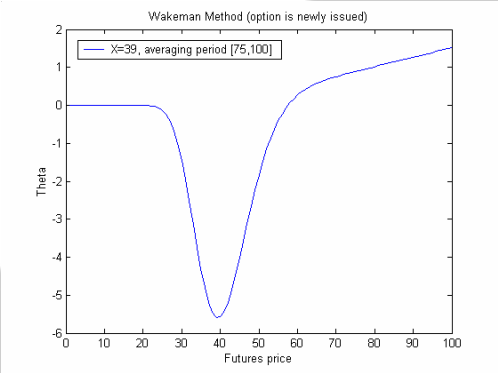


Figure 8:
Theta (Vorst method)

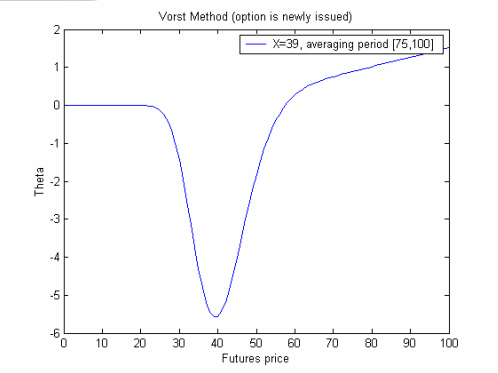


Figure 9:
Vega (Wakeman method)

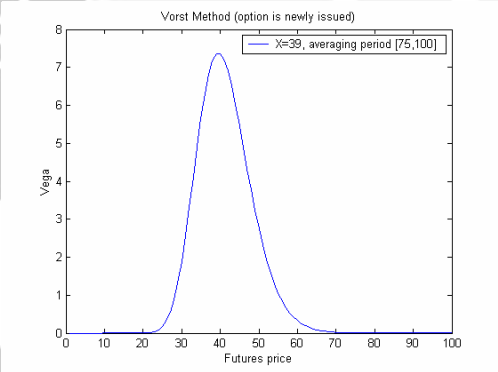
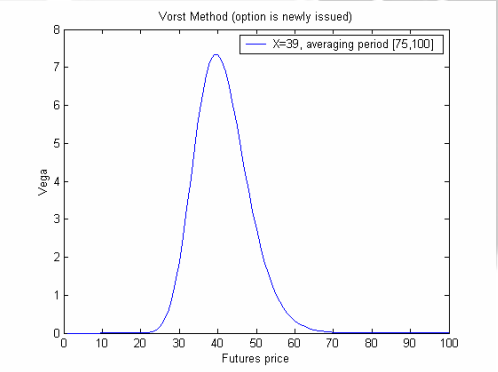


Figure 10:
Vega (Vorst method)



We can also check approximation of the greeks using a relationship between theta and gamma (involving call price, futures price, interest rate and volatilities), derived from Ito's lemma. This relationship is

$$\frac{dc}{\partial t}+\frac{1}{2}\frac{\partial^2 c}{\partial F^2}.\sigma^2.F^2 = r.c\,.$$

In terms of gamma and theta,

$$\Theta+\frac{1}{2}.\Gamma.\sigma^2.F^2 = r.c.$$

To check whether this relationship holds for our approximated values, we calculate the discrepancy as:

$$discrepancy = abs\left[r.c - \left(\Theta + \frac{1}{2}.\Gamma.\sigma^2.F_2 \right) \right]$$

If the model is correct, this theoretical relationship should hold exactly. Using the same paths used to calculate hedge costs and hedge error, we find that the discrepancy is zero for Wakeman method and is less than 0.009344 for Vorst method , when the calculation is done until 6 digits. It means that the theoretical relationship between theta and gamma is approximately satisfied by both methods. The values of the greeks obtained by both methods are very close as well. These are shown in Table 6. Still, the discrepancy obtained by Wakeman method is lower than that obtained by Vorst method. In other words , Wakeman method shows a better performance in terms of the greeks.

Table 6: Greeks values

Strike price	*36*		*39*		*41*	
Method	*Wakeman*	*Vorst*	*Wakeman*	*Vorst*	*Wakeman*	*Vorst*
call price	4.9811	4.9794	3.1383	3.1364	2.2106	2.2086
put price	1.0084	1.0067	2.1452	2.1432	3.2037	3.2017
Delta	0.7572	0.7573	0.5889	0.5889	0.4711	0.4711
Gamma	0.0459	0.0459	0.0577	0.0577	0.0592	0.0592
Rho	-1.3647	-1.3642	-0.8598	-0.8593	-0.6056	-0.6051
Theta	-4.3777	-4.3731	-5.5732	-5.6476	-5.7427	-5.7939
Vega	5.8634	5.8474	7.3602	7.343	7.5508	7.5336

Note:Fo=40; r=2.5 %; σ =35 %; Δt =1/365; averaging period {75,100).

Next, we performed the simulation study for basket options. Table 7 compares the call option prices on a basket consisting 4 assets obtained by the lognormal and reciprocal gamma distribution approaches and Monte Carlo simulation. We applied those methods to three data sets. The datasets are:

The futures prices : $F_1(t_0)$=100; $F_2(t_0)$=90; $F_3(t_0)$=95; $F_4(t_0)$=100;

Volatilities : σ_1=20 %; σ_2 =25 %; σ_3 =30 %; σ_4 =25 % per-annum.

The weights : a_1 =20 %; a_2 =15 %; a_3 =25 %; a_4 =40 %.

The risk free interest rate (r) : 5 % per-annum.

The time to expiry (T) :365 days (365 trading days per-year).

Dataset 1: uncorrelated assets ($\rho_{i,j}$ =0, $i \neq j$, $i, j = 1,2,3,4$).

Dataset 2: the assets have high and positive correlation, the correlation matrix:

$$\begin{bmatrix} 1 & 0.80 & 0.90 & 0.75 \\ 0.80 & 1 & 0.85 & 0.70 \\ 0.90 & 0.85 & 1 & 0.90 \\ 0.75 & 0.70 & 0.90 & 1 \end{bmatrix}$$

Dataset 3: the assets have low and positive correlation, the correlation matrix:

$$\begin{bmatrix} 1 & 0.100 & 0.050 & 0.035 \\ 0.100 & 1 & 0.020 & 0.001 \\ 0.050 & 0.020 & 1 & 0.040 \\ 0.035 & 0.001 & 0.040 & 1 \end{bmatrix}$$

The number of replication for Monte Carlo simulation method is 100x1000. It means we generate 1000 paths to obtain the call price, and this simulation is replicated 100 times.

The discrepancies among the two methods are relatively small. It is around 1 cent when the call price is around 4 $ (datasets 1 and 3) and 10 cents when the call price is around 7.5 $. In general we can conclude that the accuracy of log-normal and reciprocal-gamma distribution approximation is the same.

Table 7: Call price of basket option

Method	*data set 1*	*data set 2*	*data set 3*
log-normal	3.9927	7.5793	4.1949
Reciprocal-gamma	3.9877	7.4790	4.1872
Monte Carlo	3.9956 *(0.2289*)*	7.5556 *(0.4135*)*	4.1886 *(0.2409*)*

** standard deviation*

If the model works well, the hedge error should be small. Table 8,9 and 10 presents the numerical results came for a basket consisting of two assets (i.e. two kinds of futures). The parameters of this basket are

The futures prices : $F_1(t_0)$=95; $F_2(t_0)$=105.

Volatilities : σ_1=20 %; σ_2 =30 % per-annum.

The weights : a_1 =70 %; a_2 =0.3 %.

The risk free interest rate r : 5 % per-annum.

The time to expiry T :100 days (365 trading days per-year).

We have three data sets with the different log-expected return correlation.

Data set A: $\rho_{1,2}$ =0 (uncorrelated assets)

Data set B: $\rho_{1,2}$ =0.9 (positive and highly correlated assets)

Data set C: $\rho_{1,2}$ =0.1 (positive and low correlated assets).

For each data set we generate 3 different paths.

Table 8: Hedge errors of paths generated from dataset A

Path	*A1*		*A2*		*A3*	
method	*log-normal*	*r-gamma*	*log-normal*	*r-gamma*	*log-normal*	*r-gamma*
hedge cost	3.4154	3.3855	2.8801	2.8958	2.5499	2.5357
call price	2.5047	2.5085	2.5047	2.5085	2.5047	2.5085
hedge error	0.9107 (36.36 %)	0.8770 (34.96 %)	0.3754 (14.99 %)	0.3874 (15.44 %)	0.0452 (1.81 %)	0.0272 (1.08 %)

Table 9: Hedge errors of paths generated from dataset B

Path	*B1*		*B2*		*B3*	
method	*log-normal*	*r-gamma*	*log-normal*	*r-gamma*	*log-normal*	*r-gamma*
hedge cost	5.1751	5.1249	4.7077	4.7958	4.0175	4.0500
call price	3.6964	3.6930	3.6964	3.6930	3.6964	3.6930
hedge error	1.4787 (40.00 %)	1.4319 (38.77 %)	1.0113 (27.36 %)	1.1028 (29.86 %)	0.3212 (8.69 %)	0.3570 (9.67 %)

Table 10: Hedge errors of paths generated from dataset C

Path	*C1*		*C2*		*C3*	
method	*log-normal*	*r-gamma*	*log-normal*	*r-gamma*	*log-normal*	*r-gamma*
hedge cost	2.7016	2.6842	3.2000	3.1852	4.1165	4.1844
call price	2.5201	2.5238	2.5201	2.5238	2.5201	2.5238
hedge error	0.1814 (7.20 %)	0.1603 (6.36 %)	0.6799 (26.98 %)	0.6614 (26.21 %)	1.5964 (63.34 %)	1.6606 (65.80 %(

These tables show that the performances of both log-normal and reciprocal-gamma distribution approaches are approximately equal. It is difficult to conclude which method is better than another one because hedge errors of both are relatively the same. Interestingly, the range of hedge error is very wide. For instance, in Table 8 we have three paths generated for each dataset. The hedge errors vary from 1 % through 36 %.

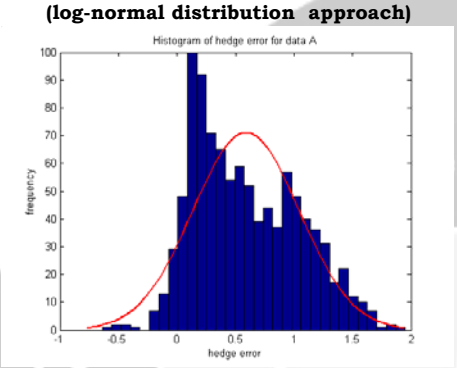
The further investigation is done by replicating the paths from dataset A. We plot the empirical distribution of hedge errors from this collection of paths. The results are shown in Table 11 and histograms 13 and 14. The hedge errors vary from 0.02 % to 83.68 % for reciprocal-gamma distribution approach, and in the range 0.04-77.85 % for the lognormal approach. The means of hedge errors are also high for both, around 23-24 %. Moreover the distribution of hedge errors is not normal.

Table 11: Distribution of hedge errors from dataset A

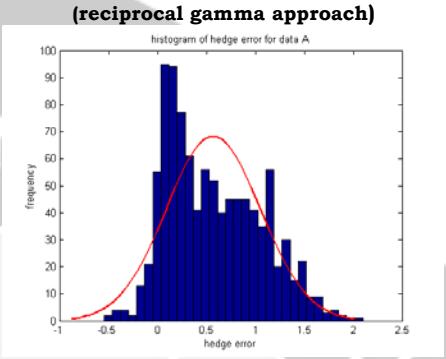
Statistics	*log-normal (2.5047*)*			*reciprocal-gamma (2.5085*)*		
Distribution	*hedge cost*	*hedge error*	*% hedge error*	*hedge cost*	*hedge error*	*% hedge error*
Mean	3.0938	0.5891	24.01	3.0792	0.5707	23.47
standard deviation	0.4530	0.4530	17.43	0.4819	0.4819	18.32
Minimum	1.8654	-0.6393	0.04	1.9658	-0.5426	0.02
Maximum	4.4545	1.9498	77.85	4.6075	2.0990	83.68

**) call price*

Histogram 13
(log-normal distribution approach)



Histogram 14:
(reciprocal gamma approach)



We also investigate the hedge performance of model by observing the hedge errors versus hedge frequencies as well. The hedge error should be lower if the hedge frequency is higher. But Tables 12 and 13 fail to support this theoretical result.

Table 12:
Simulation on path A3 using log-normal distribution approach

Hedge frequency	*1 day*	*3 days*	*5 days*	*10 days*	*15 days*
Hedge cost	2.5500	2.4112	2.5694	2.8054	2.8404
call price	2.5047	2.5047	2.5047	2.5047	2.5047
Hedge error	0.0453 (1.81 %)	-0.0935 (3.73 %)	0.0647 (2.58 %)	0.3007 (12.01 %)	0.3357 (13.40 %)

Table 13:
Simulation on path A3 using reciprocal-gamma distribution approach

Hedge frequency	*1 day*	*3 days*	*5 days*	*10 days*	*15 days*
Hedge cost	2.5357	2.3990	2.5578	2.7933	2.8251
call price	2.5085	2.5085	2.5085	2.5085	2.5085
Hedge error	0.0272	-0.1095	0.0493	0.2848	0.3166
% hedge error	1.08	4.36	1.97	11.35	12.62

Why does this happen? Let us look again at the results for the paths C1-C3. These paths, each of length 100 days, are generated by Monte Carlo simulation using the same parameters. When the correlation between the assets is estimated on the basis of all 100 days, the estimates are relatively close to the given correlation (path C1: $\widehat{\rho}_{1,2}$ = 0.2069, path C2: $\widehat{\rho}_{1,2}$ =0.2093, path C3: $\widehat{\rho}_{1,2}$ =0.1944 and $\rho_{1,2}$ =0.2). However, if we estimate the correlation on the basis of fewer days, the differences in estimated values are striking, as shown in Table 14.

Both methods assume the correlation between assets in the basket is constant and given. But Table 14 shows that empirical correlations do not satisfy this assumption, even when theoretical correlation is constant. The estimated correlations are rather erratic, and they change over the observation period. At glance we can see that the path C1 is the best in the sense that the estimated correlation values tend to be more constant and closer to the given correlation compared to paths C2 and C3. And the results also show the lowest hedge error obtained for the path C1, and much lower than those obtained for the paths C2 and C3. This indicates that the estimation of the correlation between asset prices plays crucial role in basket option pricing and hedging. To get a good performance from the model (e.g. hedge performance), the estimation of correlation should be accurate. Moreover the constant correlation assumption over the averaging period should hold.

Table 14:
The evolution of the expected correlation values of paths generated from dataset C

observation period	*observed expected return correlation coeeficient*		
	path C1	*path C2*	*path C3*
the first 20 days	0.1246	0.4657	0.2386
the first 40 days	0.1775	0.2923	0.0909
the first 60 days	0.2485	0.1908	0.1675
the first 80 days	0.2471	0.1988	0.1299
the first 100 days	0.2069	0.2093	0.1944

4 Conclusions and Further work

Over-the-counter energy options, which are becoming increasingly popular, are mostly exotic, Asian style contracts. Basket options are also typical for energy markets, where portfolios usually consist of several energy products. The similarity between Asian and basket options is that their payoffs depend on the average price (or a sum of prices) of underlying assets. The main difficulty in valuing these options is the same: the average (or sum) of lognormal random variables is not lognormally distributed. Numerical or analytical approximation approaches are needed to price these options.

In this paper we reviewed three analytical approximation approaches for Asian options: Wakeman, Vorst and Curran methods. We derived the greeks for Vorst and Wakeman methods and tested all the methods on a simulation study. All methods produce close results in terms of valuing Asian options. However, we recommend Wakeman method for pricing, hedging and calculation of greeks: the option prices are closest to those obtained by the Monte Carlo simulation, the method gives (approximate) closed-form solutions for the option prices and the greeks, and it produces the lowest hedge errors. The greeks for Asian options behave similarly to those for European options and the theoretical relationship between Theta and Gamma (derived from the Ito calculus) is approximately

satisfied by both Wakeman and Vorst method. Again, the discrepancy in this relationship is lower for Wakeman method.

Pricing and hedging basket option is complicated by the additional problem of correlations within the basket. Basket option with N correlated underlying assets involves N sources of uncertainty. Moreover, these are correlated with each others. The greatest risk in basket option pricing is the model risk, more precisely, the estimation of correlation coefficients between asset prices. Basket option can be significantly mispriced if the correlation coefficients are not specified correctly.

We reviewed here two approaches for valuing basket options: the lognormal and the reciprocal-gamma distribution approximation. Both methods produce close results for option valuation. Delta hedging derived by both these methods is rather unreliable, except when the estimates of the correlation coefficients are accurate and the assumption of a constant correlation during the averaging period holds. Otherwise the hedge errors turn out to be very high, and also with high variance.

A number of issues deserve further investigation. First, the Asian-basket option is also typical energy options. Combining an Asian option and a basket option pricing approaches to develop an Asian-basket option model is still a very challenging task.

Second, it would be useful to develop a basket option model in which the portfolio weights can be negative. In practice, a portfolio of an energy oil company consist of long positions on a commodity, e.g. crude oil or natural gas, and short positions on a "product", e.g. gasoline, or electricity. In such cases, basket options with possibly negative weights are needed to hedge market risk of such portfolios. Option pricing for such baskets can be developed by the shifted-log normal distribution approximation approach.

Third issue also arises from our analysis of basket options, especially when considering real-life prices. In reality, the constant correlation assumption used in valuing a basket option is often unrealistic, since correlations between assets in a basket often change in time. To deal with this issue, the basket option pricing model with dynamic correlation should be developed.

References

[1] Arnold, L (1974), *Stochastic diferential equations, theory and applications,* Willey, New York.

[2] Curran,M. (1994), Valuing Asian and Portfolio Options by Conditioning on the Geometric Mean Price, *Management Science*, Vol. 40, No. 12, pp. 1705-1711.

[3] Eydeland, A. and Wolyniec, K. (2003). Energy and Power Risk Management: New Developments in Modeling, Pricing and Hedging, John Willey and Sons.Akhmediev & A. Ankiewicz (1997), *Solitons, nonlinear pulses and beams*, Chapman & Hall, London.

[4] Hull,J.D. (2002), Option, Futures and Other Derivatives, 5th ed., Prentice Hall.

[5] Kemna, A.G.Z., and Vorst, A.C.F. (1990), Pricing method for Options Based on Average Asset Values, *Journal of Banking and Finance*, Vol. 14, pp. 113-129.

[6] Milevsky, M.A. and Posner, S.E (1998), A Closed-Form Approximation for Valuing Basket Options, *The Journal of Derivatives*, pp. 54-61.

[7] Milevsky, M.A. and Posner , S.E. (1998), Asian Options, the sum of lognormals and the Reciprocal Gamma Distribution, *Journal of Financial and Quantitative Analysis*, Vol. 33, pp. 409-422.

[8] Turnbull,S.M. and Wakeman,L.M. (1991), A Quick Algorithm for Pricing European Average Options, *Journal of Financial and Quantitative Analysis*, Vol. 26, pp. 377-389.

[9] Vorst, T (1992), Prices and Hedge Ratios of Average Exchange Rate Options, *International Review of Financial Analysis*, Vol.1, No.3, pp. 179-193.

[10] Wilmott, P. (2000), Derivatives: The Theory and Practice of Financial Engineering, John Willey and Sons.

A Mean and variance of the discrete form of Geometric Average

The discrete form of geometric average is defined as

$$G(T) = \sqrt[N]{\prod_{k=m}^{n} F(t_k)},$$

where $N = n - m + 1$.

$$\begin{aligned} M &= \hat{E}(\log(G(T))) \\ &= \tfrac{1}{N}\sum_{k=m}^{n} \hat{E}(\log(F(t_k))) \\ &= \tfrac{1}{N}\left\{ N.\log(F(t_i)) - \tfrac{1}{2}.\sigma^2 \sum_{k=m}^{n}(t_k - t_i)\right\} \\ &= \log(F(t_i)) - \tfrac{1}{2}\sigma^2\left\{(t_m - t_i) + \tfrac{1}{2}(T - t_m)\right\}, \end{aligned}$$

since $\sum_{k=m}^{n}(t_k - t_i) = N(t_m - t_i) + \sum_{k=m}^{n}(t_k - t_m)$

$$= N.(t_m - t_i) + (0 + \Delta t + 2.\Delta t + ... + (n-m).\Delta t)$$
$$= N.(t_m - t_i) + \frac{1}{2}.(n-m).(n-m+1).\Delta t$$
$$= N.(t_m - t_i) + \frac{1}{2}.(t_n - t_m).N$$
$$= N\left\{(t_m - t_i) + \frac{1}{2}.(t_n - t_m)\right\}.$$

For $j \le k$, $Cov(\log(F(t_j)), \log(F(t_k))) = \sigma^2(t_j - t_i)$.

$$Cov\left(\log(F(t_j)), \sum_{k=m}^{n} \log(F(t_k))\right) = \sum_{k=m}^{j-1} Cov(\log(F(t_j)), \log(F(t_k))) + \sum_{k=j}^{n} Cov(\log(F(t_j)), \log(F(t_k)))$$

$$= \sigma^2 \sum_{k=m}^{j-1}(t_k - t_i) + \sigma^2 \sum_{k=j}^{n}(t_j - t_i)$$
$$= N.\sigma^2(t_m - t_i) + \sigma^2 \sum_{k=m}^{j-1}(t_k - t_m) + \sigma^2 \sum_{k=j}^{n}(t_j - t_m)$$
$$= N.\sigma^2(t_m - t_i) + \sigma^2\{0 + \Delta t + 2.\Delta t + ... + (j-m-1)\Delta t + (n-j+1)(j-m)\Delta t\}$$
$$= N.\sigma^2(t_m - t_i) + \sigma^2.\Delta t\{\tfrac{1}{2}.(j-m).(j-m-1) + (n-j+1).(j-m)\}$$
$$= N.\sigma^2(t_m - t_i) + \sigma^2.\Delta t.\frac{1}{2}.(j-m).(2n-m-j+1).$$

$$V = Var(\log(G(T)))$$
$$= Cov(\log(G(T)), \log(G(T)))$$

$$= Cov\left(\frac{1}{N}\sum_{j=m}^{n} \log(F(t_j)), \frac{1}{N}\sum_{k=m}^{n} \log(F(t_k))\right)$$
$$= \frac{1}{N^2}\sum_{j=m}^{n} Cov\left(\log(F(t_j)), \sum_{k=m}^{n} \log(F(t_k))\right)$$
$$= \frac{1}{N^2}\sum_{j=m}^{n}\left[N.\sigma^2.(t_m - t_i) + \sigma^2.\Delta t.\frac{1}{2}.(j-m).(2n-m-j+1)\right]$$

$$= \sigma^2 (t_m - t_i) + \frac{\sigma^2 . \Delta t}{2.N^2} \sum_{j=m}^{n} (j-m).(2n-m-j+1)$$

$$= \sigma^2 (t_m - t_i) + \frac{\sigma^2 . \Delta t}{2.N^2} \sum_{j=1}^{n-m-1} (j-1).(2n-2m+2-j)$$

$$= \sigma^2 (t_m - t_i) + \frac{\sigma^2 . \Delta t}{2.N^2} \sum_{j=1}^{N} \{(2N+1)j - j^2 - 2N\}$$

$$= \sigma^2 (t_m - t_i) + \frac{\sigma^2 . \Delta t}{N^2} \left[(2N+1).\frac{1}{2}.N.(N+1) - \frac{1}{6}.N.(N+1).(2N+1) - 2N^2 \right]$$

$$= \sigma^2 (t_m - t_i) + \frac{\sigma^2 . \Delta t}{6.N} (2N-1)(N-1)$$

$$= \sigma^2 (t_m - t_i) + \frac{\sigma^2}{6.N} (2N-1)(n-m).\Delta t$$

$$= \sigma^2 \left\{ (t_m - t_i) + \frac{(2N-1)}{6N} (t_n - t_m) \right\}.$$

B The Greeks of Asian option using Vorst method

For newly issued option (option valued at time t, $t = t_0, t_1, ..., t_{m-1}$), Vorst formula for Asian call on futures based on arithmetic average is

$$\hat{c}(F, r, \sigma, t) = e^{-r(T-t)} \left(e^{M + \frac{1}{2}V} . N(d_1^*) - X^* . N(d_2^*) \right),$$

where $N = n - m + 1$

$$M = \hat{E}(\log(G(T))) = \log(F(t)) - 0.5.\sigma^2 \{(t_m - t) + \tfrac{1}{2}(T - t_m)\}$$

$$V = Cov(\log(G(T)), \log(G(T))) = \sigma^2 \left\{ (t_m - t) + \frac{(2N-1)}{6.N} (T - t_m) \right\}$$

$$\hat{E}(A(T)) = F_0$$

$$\hat{E}(G(T)) = e^{M + \frac{1}{2}V}$$

$$X^* = X - (E(A(T)) - E(G(T)))$$

$$d_1^* = \frac{M - \log(X^*) + V}{\sqrt{V}}$$

$d_2^* = d_1^* - \sqrt{V}$

$N(.)$ is standard normal distribution function.

Let $n(.)$ be standard normal probability density function. We first will derive some expression which is used to obtain the greeks. Since $d_2^* = d_1^* - \sqrt{V}$, and V and M are independent of F, we have

$$\frac{\partial d_1^*}{\partial F} = \frac{\partial d_2^*}{\partial F} \text{ and } \frac{\partial d_1^*}{\partial M} = \frac{\partial d_2^*}{\partial M}$$

$$\frac{\partial d_2^*}{\partial V} = \frac{\partial d_1^*}{\partial V} - \frac{1}{2\sqrt{V}}.$$

Consequently,

- $$e^{M+\frac{1}{2}V}.n(d_1).\frac{\partial d_1^*}{\partial F} = e^{M+\frac{1}{2}V}.\frac{1}{\sqrt{2\pi}}.\exp\left(-\frac{1}{2}(d_1^*)^2\right).\frac{\partial d_2^*}{\partial F}$$

$$= e^{M+\frac{1}{2}V}.\frac{1}{\sqrt{2\pi}}.\exp\left(-\frac{1}{2}(d_2^*)^2 - d_2^*.\sqrt{V} - \frac{1}{2}V\right).\frac{\partial d_2^*}{\partial F}$$

$$= X^*.n(d_2).\frac{\partial d_2^*}{\partial F},$$

since $\exp\left(-d_2^*.\sqrt{V}\right) = \exp\left(-M + \log X^*\right) = X^*.\exp(-M)$

- $e^{M+\frac{1}{2}V}.n(d_1).\frac{\partial d_1^*}{\partial M} = X^*.n(d_2).\frac{\partial d_2^*}{\partial M}$, and

- $$e^{M+\frac{1}{2}V}.n(d_1).\frac{\partial d_1^*}{\partial V} = X^*.n(d_2).\left(\frac{\partial d_1^*}{\partial V} - \frac{1}{2\sqrt{V}}\right)$$

$$e^{M+\frac{1}{2}V}.n(d_1).\frac{\partial d_1^*}{\partial V} - X^*.n(d_2).\frac{\partial d_1^*}{\partial F} = \frac{1}{2\sqrt{V}}\,X^*.n(d_2).$$

$$\frac{\partial E(G(T))}{\partial F} = e^{M+\frac{1}{2}V}.\frac{\partial M}{\partial F} = E(G(T)).\frac{1}{F}.$$

$$\frac{\partial E(A(T))}{\partial F} = 1 = \frac{F}{F} = \frac{E(A(T))}{F}.$$

$$\frac{dX^*}{\partial F} = -\frac{\partial E(A(T))}{\partial F} + \frac{\partial E(G(T))}{\partial F} = \frac{-E(A(T)) + E(G(T))}{F}.$$

$$\frac{dX^*}{\partial M} = \frac{\partial E(G(T))}{\partial M} = E(G(T)).$$

$$\frac{\partial \hat{c}}{\partial M} = e^{-r(T-t)}\left(e^{M+\frac{1}{2}V}.N\left(d_1^*\right) + e^{M+\frac{1}{2}V}.n\left(d_1^*\right).\frac{\partial d_1}{\partial M} - \frac{\partial X^*}{\partial M}.N\left(d_2^*\right) - X^*.n\left(d_2^*\right).\frac{\partial d_2}{\partial M}\right)$$

$$= e^{-r(T-t)}\left(e^{M+\frac{1}{2}V}.N\left(d_1^*\right) - \frac{\partial X^*}{\partial M}.N\left(d_2^*\right)\right)$$

$$= e^{-r(T-t)}.E(G(T)).\left(N\left(d_1^*\right) - N\left(d_2^*\right)\right).$$

$$\frac{\partial \hat{c}}{\partial V} = e^{-r(T-t)}\left(\frac{1}{2}.e^{M+\frac{1}{2}V}.N\left(d_1^*\right) + e^{M+\frac{1}{2}V}.n\left(d_1^*\right).\frac{\partial d_1}{\partial V} - \frac{\partial X^*}{\partial V}.N\left(d_2^*\right) - X^*.n\left(d_2^*\right).\frac{\partial d_2}{\partial V}\right)$$

$$= e^{-r(T-t)}\left(\frac{1}{2}.E(G(T)).N\left(d_1^*\right) - E(G(T)).N\left(d_2^*\right) + \frac{1}{2\sqrt{V}}X^*.n\left(d_2^*\right)\right)$$

$$= \frac{1}{2}.\frac{\partial \hat{c}}{\partial M} + \frac{e^{-r(T-t)}}{2\sqrt{V}}X^*.n\left(d_2^*\right).$$

$$\frac{\partial M}{\partial t} = \frac{1}{2}.\sigma^2.$$

$$\frac{\partial V}{\partial t} = -\sigma^2.$$

$$\frac{\partial M}{\partial \sigma} = -\sigma\left\{\left(t_m - t\right) + \tfrac{1}{2}\left(T - t_m\right)\right\}.$$

$$\frac{\partial V}{\partial \sigma} = 2.\sigma.\left\{\left(t_m - t\right) + \frac{(2N-1)}{6.N}\left(T - t_m\right)\right\}.$$

$$\Delta = \frac{\partial \hat{c}}{dF} =$$

$$e^{-r(T-t)}\left(e^{M+\frac{1}{2}V}.\frac{\partial M}{\partial F}.N\left(d_1^{\,*}\right) + e^{M+\frac{1}{2}V}.n\left(d_1^{\,*}\right).\frac{\partial d_1}{\partial F} - \frac{\partial X^*}{\partial F}.N\left(d_2^{\,*}\right) - X^*.n\left(d_2^{\,*}\right).\frac{\partial d_2}{\partial F}\right)$$

$$= e^{-r(T-t)}\left(e^{M+\frac{1}{2}V}.\frac{1}{F}.N\left(d_1^{\,*}\right) - \frac{\partial X^*}{\partial F}.N\left(d_2^{\,*}\right)\right)$$

$$= \frac{e^{-r(T-t)}}{F}\left(e^{M+\frac{1}{2}V}.N\left(d_1^{\,*}\right) - \left(-E\left(A(T)\right) + E\left(G(T)\right)\right).N\left(d_2^{\,*}\right)\right).$$

$$\Pi = \frac{\partial \hat{c}^{\,2}}{d^2F} = \frac{\partial \Delta}{dF}$$

$$= \frac{-e^{-r(T-t)}}{F^2}\left(e^{M+\frac{1}{2}V}.N\left(d_1^{\,*}\right) - \left(-E\left(A(T)\right) + E\left(G(T)\right)\right).N\left(d_2^{\,*}\right)\right) +$$

$$\frac{e^{-r(T-t)}}{F}\frac{\partial}{\partial F}\left(e^{M+\frac{1}{2}V}.N\left(d_1^{\,*}\right) - \left(-E\left(A(T)\right) + E\left(G(T)\right)\right).N\left(d_2^{\,*}\right)\right)$$

$$= \frac{-1}{F}.\Delta + \frac{e^{-r(T-t)}}{F}\frac{\partial}{\partial F}\left(e^{M+\frac{1}{2}V}.N\left(d_1^{\,*}\right) - \left(-E\left(A(T)\right) + E\left(G(T)\right)\right).N\left(d_2^{\,*}\right)\right)$$

$$= \frac{-1}{F}.\Delta + \frac{e^{-r(T-t)}}{F}\left(e^{M+\frac{1}{2}V}.\frac{\partial M}{\partial F}.N\left(d_1^{\,*}\right) + e^{M+\frac{1}{2}V}.n\left(d_1^{\,*}\right).\frac{\partial d_1^{\,*}}{\partial M}.\frac{\partial M}{\partial F} + N\left(d_2^{\,*}\right) + \right.$$

$$\left. F.n\left(d_2^{\,*}\right).\frac{\partial d_2^{\,*}}{\partial M}.\frac{\partial M}{\partial F} - e^{M+\frac{1}{2}V}.N\left(d_2^{\,*}\right).\frac{\partial M}{\partial F} - e^{M+\frac{1}{2}V}.n\left(d_2^{\,*}\right).\frac{\partial d_2^{\,*}}{\partial M}.\frac{\partial M}{\partial F}\right)$$

$$= \frac{-1}{F}.\Delta + \frac{e^{-r(T-t)}}{F}\left(e^{M+\frac{1}{2}V}.\frac{N\left(d_1^{\,*}\right)}{F} + e^{M+\frac{1}{2}V}.\frac{n\left(d_1^{\,*}\right)}{F.\sqrt{V}} + N\left(d_2^{\,*}\right) + \frac{F.n\left(d_2^{\,*}\right)}{F.\sqrt{V}} - \right.$$

$$\left. e^{M+\frac{1}{2}V}.\frac{N\left(d_2^{\,*}\right)}{F} - e^{M+\frac{1}{2}V}.\frac{n\left(d_2^{\,*}\right)}{F.\sqrt{V}}\right)$$

$$= \frac{-1}{F}.\Delta + \frac{e^{-r(T-t)}}{F^2}E\left(G(T)\right)\left(N\left(d_1^{\,*}\right) - N\left(d_2^{\,*}\right)\right) +$$

$$\frac{e^{-r(T-t)}}{F^2\sqrt{V}}.E(G(T))\left(n\left(d_1^*\right)-n\left(d_2^*\right)\right)+\ \frac{e^{-r(T-t)}}{F}\left(N\left(d_2^*\right)-n(d_2*)\right)$$

$$=\frac{-1}{F}.\Delta+\frac{1}{F^2}.\frac{\partial\hat{c}}{\partial M}+\frac{e^{-r(T-t)}}{F^2\sqrt{V}}.E(G(T))\left(n\left(d_1^*\right)-n\left(d_2^*\right)\right)+$$

$$\frac{e^{-r(T-t)}}{F}\left(N\left(d_2^*\right)-n(d_2*)\right).$$

$$\rho=\frac{\partial\hat{c}}{dr}=-(T-t).e^{-r(T-t)}\left(e^{M+\frac{1}{2}V}.N\left(d_1^*\right)-X^*.N\left(d_2^*\right)\right)=-(T-t).\hat{c}\ .$$

$$\Theta=\frac{\partial\hat{c}}{dt}$$

$$=\text{r}.e^{-r(T-t)}\left(e^{M+\frac{1}{2}V}.N\left(d_1^*\right)-X^*.N\left(d_2^*\right)\right)+\ e^{-r(T-t)}\left(\frac{\partial c^*}{\partial M}.\frac{\partial M}{\partial t}+\frac{\partial c^*}{\partial V}.\frac{\partial V}{\partial t}\right)$$

$$,\text{where } c^*=e^{M+\frac{1}{2}V}.N\left(d_1^*\right)-X^*.N\left(d_2^*\right).$$

$$=r.\hat{c}+\frac{\partial\hat{c}}{\partial M}.\frac{\partial M}{\partial t}+\frac{\partial\hat{c}}{\partial V}.\frac{\partial V}{\partial t}\ \text{, since } e^{-r(T-t)}.\frac{\partial c^*}{\partial M}=\frac{\partial\hat{c}}{\partial M}\ \text{and}\ e^{-r(T-t)}.\frac{\partial c^*}{\partial V}=\frac{\partial\hat{c}}{\partial V}$$

$$\upsilon=\frac{\partial\hat{c}}{d\sigma}=\frac{\partial\hat{c}}{\partial M}.\frac{\partial M}{\partial\sigma}+\frac{\partial\hat{c}}{\partial V}.\frac{\partial V}{\partial\sigma}.$$

C The first and second central moment of arithmetic average using Wakeman method

$$M_1=\hat{E}(A(T))$$

$$=\hat{E}\left(\frac{1}{N}\sum_{k=m}^{n}F(t_k)\right)$$

$$=\frac{1}{N}\sum_{k=m}^{n}\hat{E}(F(t_k))$$

$$=F(t_0).$$

Denote $\rho_{j,k}$ as coefficient correlation between $F(t_j)$ and $F(t_k)$.

For $j < k$,

$$\hat{E}\left(F(t_j).F(t_k)\right) = F(t_0)^2.\exp\left(\rho_{j,k}.\sigma^2.\sqrt{t_j - t_0}.\sqrt{t_k - t_0}\right) = F(t_0)^2.e^{\sigma^2(t_j - t_0)},$$

since $\rho_{j,k} = \dfrac{Cov\left(F(t_j), F(t_k)\right)}{\sigma_{F(t_j)}.\sigma_{F(t_k)}} = \dfrac{\sigma^2(t_j - t_0)}{\sigma\sqrt{t_j - t_0}.\sigma.\sqrt{t_k - t_0}} = \sqrt{\dfrac{t_j - t_0}{t_k - t_0}}$, $j < k$.

$$M_2 = \hat{E}\left(A(T)^2\right)$$

$$= \hat{E}\left(\frac{1}{N^2}\sum_{j=m}^{n} F(t_j)\sum_{k=m}^{n} F(t_k)\right)$$

$$= \frac{1}{N^2}\left\{\sum_{j=m}^{n}\hat{E}\left(F(t_j)^2\right) + 2\sum_{k=m+1}^{n}\sum_{j=m}^{k-1}\hat{E}\left(F(t_j).F(t_k)\right)\right\}.$$

$$\sum_{j=m}^{n}\hat{E}\left(F(t_j)^2\right) = \sum_{j=m}^{n} F(t_0)^2.e^{\sigma^2(t_j - t_0)}$$

$$= F(t_0)^2.e^{\sigma^2(t_m - t_0)}\sum_{j=m}^{n}.e^{\sigma^2(t_j - t_m)}$$

$$= F(t_0)^2.e^{\sigma^2(t_m - t_0)}.\left(\frac{e^{N.\sigma^2.\Delta t} - 1}{e^{\sigma^2.\Delta t} - 1}\right).$$

$$\sum_{k=m+1}^{n}\sum_{j=m}^{k-1}\hat{E}\left(F(t_j).F(t_k)\right) = \sum_{k=m+1}^{n}\sum_{j=m}^{j-1} F(t_0)^2.e^{\sigma^2(t_j - t_0)}$$

$$= F(t_0)^2.e^{\sigma^2(t_m - t_0)}\sum_{k=m+1}^{n}\sum_{j=m}^{j-1} e^{\sigma^2(t_j - t_m)}.$$

$$= F(t_0)^2.e^{\sigma^2(t_m - t_0)}\sum_{k=m+1}^{n}\left(\frac{e^{(j-m)\sigma^2.\Delta t} - 1}{e^{\sigma^2.\Delta t} - 1}\right)$$

$$= \frac{F(t_0)^2.e^{\sigma^2(t_m - t_0)}}{e^{\sigma^2.\Delta t} - 1}\left(\sum_{k=m+1}^{n} e^{(j-m)\sigma^2.\Delta t} - \sum_{k=m+1}^{n} 1\right).$$

The first part in the bracket can be simplified as $e^{\sigma^2.\Delta t}\dfrac{\left(e^{(N-1).\sigma^2.\Delta t} - 1\right)}{\left(e^{\sigma^2.\Delta t} - 1\right)}$, and the second part in the bracket is equal to $N - 1$.

So, we have

$$\sum_{k=m+1}^{n}\sum_{j=m}^{k-1}\hat{E}\left(F\left(t_j\right).F\left(t_k\right)\right)=$$

$$\frac{F\left(t_0\right)^2.e^{\sigma^2\left(t_m-t_0\right)}}{e^{\sigma^2.\Delta t}-1}\left(e^{\sigma^2.\Delta t}\frac{\left(e^{(N-1)\sigma^2.\Delta t}-1\right)}{\left(e^{\sigma^2.\Delta t}-1\right)}-\left(N-1\right)\right).$$

Now the second central moment M_2 can be simplified as:

$$M_2=\frac{F\left(t_0\right)^2.e^{\sigma^2\left(t_m-t_0\right)}}{N^2\left(e^{\sigma^2.\Delta t}-1\right)}\left\{\left(e^{N\sigma^2\Delta t}-1\right)+2\left(\frac{e^{\sigma^2\Delta t}\left(e^{(N-1)\sigma^2\Delta t}\right)}{e^{\sigma^2.\Delta t}-1}-N+1\right)\right\}$$

D The Greeks of Asian option using Wakeman method

For newly issued option (option is valued at time t, $t=t_0,t_1,...,t_{m-1}$), Wakeman formula for Asian call on futures based on arithmetic average is

$$c=e^{-r(T-t)}\left\{M_1.N\left(d_1\right)-X.N\left(d_2\right)\right\}$$

where:

$$M_1=F$$

$$M_2=\hat{E}\left(A(T)^2\right)=\frac{1}{N^2}\left(M_{21}+2M_{22}\right)$$

$$M_{21}=F^2.e^{\sigma^2(m-i)\Delta t}.\left(\frac{e^{N.\sigma^2.\Delta t}-1}{e^{\sigma^2.\Delta t}-1}\right)$$

$$M_{22}=\frac{F^2.e^{\sigma^2\left(t_m-t\right)}}{e^{\sigma^2.\Delta t}-1}\left(e^{\sigma^2.\Delta t}\frac{\left(e^{(N-1)\sigma^2.\Delta t}-1\right)}{\left(e^{\sigma^2.\Delta t}-1\right)}-\left(N-1\right)\right)$$

$$\left(\sigma^*\right)^2=\frac{1}{T-t}\log\left(\frac{M_2}{\left(M_1\right)^2}\right)$$

$$d_1=\frac{\log\left(M_1\right)-\log\left(X\right)+\frac{1}{2}\left(\sigma^*\right)^2\left(T-t\right)}{\sigma^*\sqrt{T-t}}$$

$$=\frac{\log\left(F\right)-\log\left(X\right)+\frac{1}{2}\left(\sigma^*\right)^2\left(T-t\right)}{\sigma^*\sqrt{T-t}}$$

$$d_2=d_1-\sigma^*\sqrt{T-t}$$

$$\sigma^* \sqrt{T-t} = \frac{1}{\sqrt{T-t}} \cdot \left\{ \log\left(\frac{M_2}{(M_1)^2} \right) \right\}^{1/2} \cdot \sqrt{T-t} = \left\{ \log\left(\frac{M_2}{(M_1)^2} \right) \right\}^{1/2}.$$

Using this result we obtain

$$d_1 = \log\left(\frac{F(t)}{X} \right) \left\{ \log\left(\frac{M_2}{(M_1)^2} \right) \right\}^{-1/2} + \frac{1}{2} \left\{ \log\left(\frac{M_2}{(M_1)^2} \right) \right\}^{1/2}$$

$$d_2 = \log\left(\frac{F(t)}{X} \right) \left\{ \log\left(\frac{M_2}{(M_1)^2} \right) \right\}^{-1/2} - \frac{1}{2} \left\{ \log\left(\frac{M_2}{(M_1)^2} \right) \right\}^{-1/2}$$

Consequently,

$$\frac{\partial d_1}{\partial M_2} = -\sigma^2 \left[\frac{-1}{2M_2} \log \frac{F(t)}{X} \left\{ \log \frac{M_2}{M_1^2} \right\}^{-3/2} - \frac{1}{4.M_2} \cdot \left\{ \log\left(\frac{M_2}{(M_1)^2} \right) \right\}^{-1/2} \right].$$

$$\frac{\partial M_2}{\partial t} = \frac{1}{N^2} \left(\frac{\partial M_{21}}{\partial t} + 2 \frac{\partial M_{22}}{\partial t} \right) = -\sigma^2 . M_2 \text{ ,since } \frac{\partial M_{21}}{\partial t} = -\sigma^2 . M_{21};$$

$$\frac{\partial M_{22}}{\partial t} = -\sigma^2 . M_{22}$$

$$\frac{\partial d_1}{\partial t} = \frac{\partial d_1}{\partial M_2} \cdot \frac{\partial M_2}{\partial t}$$

$$= -\sigma^2 \left[\frac{-1}{2} \log\left(\frac{F(t)}{X} \right) \left\{ \log\left(\frac{M_2}{(M_1)^2} \right) \right\}^{-3/2} - \frac{1}{4} \left\{ \log\left(\frac{M_2}{(M_1)^2} \right) \right\}^{-1/2} \right]$$

$$\frac{\partial d_2}{\partial M_2} = -\sigma^2 \left[\frac{-1}{2M_2} \log\left(\frac{F(t)}{X} \right) \left\{ \log\left(\frac{M_2}{(M_1)^2} \right) \right\}^{-3/2} + \frac{1}{4M_2} \left\{ \log\left(\frac{M_2}{(M_1)^2} \right) \right\}^{-1/2} \right]$$

$$\frac{\partial d_2}{\partial t} = \frac{\partial d_2}{\partial M_2} \cdot \frac{\partial M_2}{\partial t}$$

$$= -\sigma^2 . \left[\frac{-1}{2} \log\left(\frac{F(t)}{X} \right) \left\{ \log\left(\frac{M_2}{(M_1)^2} \right) \right\}^{-3/2} + \frac{1}{4} \left\{ \log\left(\frac{M_2}{(M_1)^2} \right) \right\}^{-1/2} \right].$$

It can be shown that $\frac{M_2}{M_1^2}$ is independent of F. It means $\sigma^* \sqrt{T-t}$ is also independent of F. Consequently, $\frac{\partial d_1}{\partial F} = \frac{\partial d_2}{\partial F} = \frac{1}{F.\sigma^*.\sqrt{T-t}}$.

By analogous way with Vorst method, it can be shown that $F.n(d_1).\frac{\partial d_1}{\partial F} = X.n(d_2).\frac{\partial d_2}{\partial F}$.

$$\Delta = \frac{\partial c}{\partial F}$$

$$= e^{-r(T-t)}\left\{N(d_1) + F.n(d_1).\frac{\partial d_1}{\partial F} - X.n(d_2).\frac{\partial d_2}{\partial F}\right\}$$

$$= \frac{\partial c}{\partial F} = e^{-r(T-t)}N(d_1).$$

$$\Gamma = \frac{\partial^2 c}{\partial F^2} = \frac{\partial \Delta}{\partial F} = e^{-r(T-t)}.n(d_1).\frac{\partial d_1}{\partial F} = e^{-r(T-t)}.n(d_1).\frac{1}{F.\sigma^*\sqrt{T-t}}.$$

$$\Theta = \frac{\partial c}{dt}$$

$$= r.e^{-r(T-t)}\{M_1.N(d_1) - X.N(d_2)\} + e^{-r(T-t)}\frac{\partial}{\partial t}\{M_1.N(d_1) - X.N(d_2)\}$$

$$= r.c + e^{-r(T-t)}\left\{M_1.n(d_1).\frac{\partial d_1}{\partial t} - X.n(d_2).\frac{\partial d_2}{\partial t}\right\}.$$

$$\rho = \frac{\partial c}{dr} = -(T-t).e^{-r(T-t)}\left(e^{M+\frac{1}{2}V}.N(d_1) - X.N(d_2)\right) = -(T-t).\hat{c}$$

To simplify calculation of vega, we define: $U = e^{\sigma^2.\Delta t}$.

$$\frac{\partial U}{\partial \sigma} = 2.\sigma.\Delta t.e^{\sigma^2.\Delta t} = 2.\sigma.\Delta t.U$$

Since we want to calculate the greeks at time $t = t_i$, we can substitute $t_m - t$ with $(m-i).\Delta t$.

$$M_{21} = F^2.e^{\sigma^2(m-i)\Delta t}.\left(\frac{e^{N.\sigma^2.\Delta t}-1}{e^{\sigma^2.\Delta t}-1}\right) = F^2.\frac{U^{(m-i)}}{U-1}.\left(U^N-1\right).$$

$$M_{22} = \frac{F^2.e^{\sigma^2(m-i)\Delta t}}{e^{\sigma^2.\Delta t}-1}\left(e^{\sigma^2.\Delta t}\frac{\left(e^{(N-1)\sigma^2.\Delta t}-1\right)}{\left(e^{\sigma^2.\Delta t}-1\right)}-(N-1)\right)$$

$$= \frac{F^2.U^{(m-i)}}{U-1}\left(U\frac{\left(U^{(N-1)}-1\right)}{(U-1)}-(N-1)\right)$$

$$= \frac{F^2.U^{(m-i+1)}}{U-1}\left(U^{(N-1)}-1\right)-F^2.(N-1)\frac{U^{(m-i)}}{U-1}.$$

Define

$$U^* = \frac{\partial}{\partial U}\left(\frac{U^{(m-i)}}{U-1}\right) = \frac{(m-i)U^{(m-i-1}(U-1)-U^{(m-i)}}{(U-1)^2} =$$

$$\frac{U^{m-i-1}}{(U-1)^2}\left[(m-i-1)U-m+i\right].$$

$$U^{**} = \frac{\partial}{\partial U}\left(\frac{U^{(m-i+1)}}{(U-1)^2}\right)$$

$$= \frac{(m-i+1)U^{(m-i)}(U-1)^2-U^{(m-i+1)}.2.(U-1)}{(U-1)^4}$$

$$= \frac{U^{m-i}}{(U-1)^3}\left[(m-i-1)U-m+i-1\right].$$

$$\frac{\partial M_{21}}{\partial U} = F^2.U^*.\left(U^N-1\right)+F^2.\frac{U^{m-i}}{U-1}.N.U^{(N-1)}.$$

$$\frac{\partial M_{22}}{\partial U} = F^2.U^{**}.\left(U^{(N-1)}-1\right)+F^2.\frac{U^{m+1-i}}{(U-1)^2}.(N-1).U^{(N-2)}-F^2.(N-1).U^*.$$

$$\frac{\partial M_2}{\partial U} = \frac{1}{N^2}\left(\frac{\partial M_{21}}{\partial U}+2.\frac{\partial M_{22}}{\partial U}\right).$$

$$\frac{\partial d_1}{\partial M_2} = -\sigma^2\left[\frac{-1}{2M_2}\log\left(\frac{F(t)}{X^*}\right)\left\{\log\left(\frac{M_2}{M_1^2}\right)\right\}^{-3/2}-\frac{1}{4M_2}\left\{\log\left(\frac{M_2}{M_1^2}\right)\right\}^{-1/2}\right].$$

$$\frac{\partial d_2}{\partial M_2} = -\sigma^2\left[\frac{-1}{2M_2}\log\left(\frac{F(t)}{X^*}\right)\left\{\log\left(\frac{M_2}{(M_1)^2}\right)\right\}^{-3/2} + \frac{1}{4M_2}\left\{\log\left(\frac{M_2}{(M_1)^2}\right)\right\}^{-1/2}\right]$$

$$\frac{\partial d_1}{\partial \sigma} = \frac{\partial d_1}{\partial M_2}.\frac{\partial M_2}{\partial U}.\frac{\partial U}{\partial \sigma}.$$

$$\frac{\partial d_2}{\partial \sigma} = \frac{\partial d_2}{\partial M_2}.\frac{\partial M_2}{\partial U}.\frac{\partial U}{\partial \sigma}.$$

$$\upsilon = \frac{\partial c}{\partial \sigma} = e^{-r(T-t)}\left\{F.n(d_1).\frac{\partial d_1}{\partial \sigma} - X.n(d_2).\frac{\partial d_2}{\partial \sigma}\right\}.$$

E Call price formula using Curran method

$$Cov\left(\log\left(F(t_j)\right), G(T)\right) = Cov\left(\log\left(F(t_i)\right), \frac{1}{N}\sum_{j=m}^{n}\log\left(F(t_j)\right)\right),$$

where $N = n - m + 1$.

$$= \frac{1}{N}.\sum_{j=m}^{i} Cov\left(\log\left(F(t_i)\right), \log\left(F(t_j)\right)\right) +$$

$$\frac{1}{N}.\sum_{j=i+1}^{n} Cov\left(\log\left(F(t_i)\right), \log\left(F(t_j)\right)\right)$$

$$= \frac{\sigma^2}{N}\sum_{j=m}^{i}(t_j - t_0) + \frac{\sigma^2}{N}.\sum_{j=i+1}^{n}(t_i - t_0)$$

$$= \frac{\sigma^2}{N}.\{m.\Delta t + (m+1).\Delta t + \ldots + i.\Delta t + (n-i).i.\Delta t\}$$

$$= \sigma^2.\Delta t\left\{\tfrac{1}{2}.(i-m+1).(m+i) + 2.i.(n-i)\right\}$$

$$= \frac{\sigma^2.\Delta t}{2.N}\left\{-i^2 + (2n+1)i + m(1-m)\right\}$$

Denote coefficient correlation between $\log\left(F(t_i)\right)$ and $G(T)$ as ρ_i. It is defined as

$$\rho_i = \frac{Cov\left(\log\left(F(t_i)\right), \log\left(G(T)\right)\right)}{\sigma_i.\sigma_G}$$

The standard property of the joint normal distribution is *if random variable* X *is normally distributed with mean =* μ_X *and variance =* $\sigma_X{}^2$ *and random variable* Y

is normally distributed with mean $= \mu_Y$ *and variance* $= \sigma_Y{}^2$ *and coefficient correlation between* X *and* Y *is* ρ*, then* X *conditioned on* $Y = y$ *is normally distributed with*

$$\textit{mean} = \mu_X + (y - \mu_y).\rho.\frac{\sigma_X}{\sigma_Y} \textit{ and variance} = (1 - \rho^2).\sigma_X{}^2 .$$

Based on this property , $\log(F(t_i))$ conditional on $\log(G(T)) = x$ will be normally distributed with mean and variance:

$$E\left[(\log(F(t_i))|\log(G(T)) = x)\right] = \log(F_0) - \frac{1}{2}.\sigma^2(t_i - t_0) + (x - \mu_G).\rho_i.\frac{\sigma_i}{\sigma_G}$$

$$Var\left[(\log(F(t_i))|\log(G(T)) = x)\right] = (1 - \rho_i{}^2).\sigma_i{}^2 .$$

Using log-normal distribution property, we obtain

$$E\left[F(t_i)|\log(G(T) = x)\right] =$$

$$F_0.\exp\left[-\frac{1}{2}.\sigma^2(t_i - t_0) + (x - \mu_G).\rho_i.\frac{\sigma_i}{\sigma_G} + \frac{1}{2}.(1 - \rho_i{}^2).\sigma_i{}^2\right].$$

Consequently,

$$E\left[A(T)|G(T) = e^x\right] = E\left[A(T)|\log(G(T) = x)\right]$$

$$= \frac{1}{N}\sum_{i=m}^{n} F_0.\exp\left[-\frac{1}{2}.\sigma^2(t_i - t_0) + (x - \mu_G).\rho_i.\frac{\sigma_i}{\sigma_G} + \frac{1}{2}.(1 - \rho_i{}^2).\sigma_i{}^2\right]$$

$$E\left[A(T)|G(T) = x\right]$$

$$= \frac{1}{N}\sum_{i=m}^{n} F_0.\exp\left[-\frac{1}{2}.\sigma^2(t_i - t_0) + (\log(x) - \mu_G).\rho_i.\frac{\sigma_i}{\sigma_G} + \frac{1}{2}.(1 - \rho_i{}^2).\sigma_i{}^2\right]$$

$$= \frac{1}{N}\sum_{i=m}^{n} F_0.\exp\left[(\log(x) - \mu_G).\rho_i.\frac{\sigma_i}{\sigma_G} - \rho_i{}^2.\sigma_i{}^2\right]$$

The call price is given by

$$c = e^{-r.n.\Delta t}\hat{E}\left[\max(A(T) - X, 0)\right]$$

$$= e^{-r.n.\Delta t}\hat{E}\left(\hat{E}\left[\max(A(T) - X, 0)|G(T) = x\right]\right)$$

$$= e^{-r.n.\Delta t}\int_0^X \hat{E}\left[\max(A(T) - X, 0)|G(T) = x\right].f(x).dx +$$

$$e^{-r.n.\Delta t}\int_X^\infty \hat{E}\left[\max(A(T) - X, 0)|G(T) = x\right].f(x).dx$$

where $f(x)$ is probability density function of $\log(G(T))$.

$$\int_0^X \hat{E}\left[\max\left(A(T)-X,0\right)\middle|G(T)=x\right].f(x).dx \geq$$

$$\int_0^X \max\left[\hat{E}\left(A(T)-X\middle|G(T)=x\right),0\right].f(x).dx$$

$$= \int_{LB}^X E\left(A(T)-X\middle|G(T)=x\right).f(x).dx\,,$$

where $LB = \arg\min\left[x\middle|E\left(A(T)\middle|G(T)=x\right)=X\right]$.

It is clear that $LB < G(T) < X$.

$$\int_X^\infty \hat{E}\left[\max\left(A(T)-X,0\right)\middle|G(T)=x\right].f(x).dx = \int_X^\infty \hat{E}\left(A(T)-X\middle|G(T)=x\right).f(x).dx\,,$$

since $K < G(T) < A(T)$.

Approximation value of call is given as

$$\hat{c} = e^{-r.n.dt}.\left\{\int_{LB}^X \hat{E}\left(A(T)-X\middle|G(T)=x\right).f(x).dx + \int_X^\infty \hat{E}\left(A(T)-X\middle|G(T)=x\right).f(x).dx\right\}$$

$$= e^{-r.n.dt}.\left\{\int_{LB}^\infty \hat{E}\left(A(T)\middle|G(T)=x\right).f(x).dx + X\int_{LB}^\infty f(x).dx\right\}$$

$$= e^{-r.n.dt}.\left\{\frac{1}{N}.\sum_{i=m}^{n} F_0.N\left(\frac{\mu_G+\sigma_i.\sigma_G.\rho_i-\log(LB)}{\sigma_G}\right)+X.N\left(\frac{\mu_G-\log(LB)}{\sigma_G}\right)\right\}$$

since

$$\int_{LB}^\infty \hat{E}\left(A(T)\middle|G(T)=x\right).f(x).dx =$$

$$\int_{LB}^\infty \frac{1}{N}\sum_{i=m}^{n} F_0 \exp\left(\left(\log x-\mu_G\right)\rho_i\frac{\sigma_i}{\sigma_G}-\frac{1}{2}\rho_i^{\,2}\sigma_i^{\,2}\right)\frac{1}{x\sqrt{2\pi}}\exp\left(\frac{-\left(\log x-\mu_G\right)^2}{2\sigma_G^{\,2}}\right)dx$$

$$= \frac{1}{N}\sum_{i=m}^{n} F_0 \int_{LB}^{\infty} \frac{1}{x\sqrt{2\pi}} \exp\left(\frac{\left(\log(x) - \left(\mu_G + \sigma_i.\sigma_G.\rho_i\right)\right)^2}{2.{\sigma_G}^2} \right).dx$$

$$= \frac{1}{N}\sum_{i=m}^{n} F_0 \int_{\log(LB)}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{\left(z - \left(\mu_G + \sigma_i.\sigma_G.\rho_i\right)\right)^2}{2.{\sigma_G}^2} \right).dz$$

$$= \frac{1}{N}\sum_{i=m}^{n} F_0.N\left(\frac{\mu_G + \sigma_i.\sigma_G.\rho_i - \log(LB)}{\sigma_G} \right),$$

and $\int_X^{\infty} f(x).dx = N\left(\frac{\mu_G - \log(LB)}{\sigma_G} \right)$.

LB can be calculated using numerical method (e.g. Newton-Raphson method).

F The Greeks of basket option using log-normal distribution approach

To simplify vega calculation, we define: $U = e^{\sigma^2.\Delta t}$.

$\frac{\partial U}{\partial \sigma} = 2.\sigma.\Delta t.e^{\sigma^2.\Delta t} = 2.\sigma.\Delta t.U$.

Since we want to calculate the greeks at time $t = t_i$, we can substitute $t_n - t$ with $(n-i).\Delta t$.

$$M_{21} = F^2.e^{\sigma^2.\Delta t}.\left(\frac{e^{(n-i).\sigma^2.\Delta t} - 1}{e^{\sigma^2.\Delta t} - 1} \right) = F^2.\frac{U}{U-1}.\left(U^{(n-i)} - 1\right)$$

$$M_{22} = \frac{F^2.e^{\sigma^2.\Delta t}}{e^{\sigma^2.\Delta t} - 1}\left(e^{\sigma^2.\Delta t}\frac{\left(e^{(n-i-1)\sigma^2.\Delta t} - 1\right)}{\left(e^{\sigma^2.\Delta t} - 1\right)} - \left(N^* - 1\right) \right)$$

$$= \frac{F^2.U}{U-1}\left(U\frac{\left(U^{(n-i-1)} - 1\right)}{(U-1)} - \left(N^* - 1\right) \right)$$

$$= \frac{F^2.U^2}{(U-1)^2}\left(U^{(n-i-1)} - 1\right) - F^2.\left(N^* - 1\right)\frac{U}{U-1}$$

Define $U^* = \frac{\partial}{\partial U}\left(\frac{U}{U-1} \right) = \frac{-1}{(U-1)^2}$.

In term of U^*, we obtain

$$U^{**} = \frac{\partial}{\partial U}\left(\frac{U^2}{(U-1)^2}\right) = \frac{-2U}{(U-1)^3}$$

$$\frac{\partial M_{21}}{\partial U} = F^2.U^*.\left(U^{n-i}-1\right) + F^2.\frac{U}{U-1}.(n-i).U^{(n-i-1)}$$

$$\frac{\partial M_{22}}{\partial U} = F^2.U^{**}.\left(U^{(n-i-1)}-1\right) + F^2.\frac{U^2}{(U-1)^2}.(n-i-1).U^{(n-i-2)} - F^2.\left(N^*-1\right).U^*$$

$$\frac{\partial M_2}{\partial U} = \frac{1}{\left(N^*\right)^2}\left(\frac{\partial M_{21}}{\partial U} + 2.\frac{\partial M_{22}}{\partial U}\right)$$

$$\frac{\partial d_1}{\partial \sigma} = \frac{\partial d_1}{\partial M_2}\cdot\frac{\partial M_2}{\partial U}\cdot\frac{\partial U}{\partial \sigma}$$

$$\frac{\partial d_2}{\partial \sigma} = \frac{\partial d_2}{\partial M_2}\cdot\frac{\partial M_2}{\partial U}\cdot\frac{\partial U}{\partial \sigma}$$

$$\upsilon = \frac{\partial c}{\partial \sigma} = \frac{N^*}{N}.e^{-r(T-t)}\left\{F.n(d_1).\frac{\partial d_1}{\partial \sigma} - X^*.n(d_2).\frac{\partial d_2}{\partial \sigma}\right\}.$$

G The Greeks of basket option using log-normal approximation approach

Call option price formula is $c = \exp(-r(T-t))[M_1.N(d_1) - X.N(d_2)]$,

where $M_1 = \sum_{i=1}^{N} a_i.F_i(t)$,

$$M_2 = \sum_{j=1}^{N}\sum_{i=1}^{N} a_i.a_j.F_i(t).F_j(t).\exp\left(\sigma_i.\sigma_j.\rho_{i,j}(T-t)\right),$$

$$V = \log\left(\frac{M_2}{M_1^{\,2}}\right),$$

$$d_1 = \frac{\log(M_1) - \log(X) + \frac{1}{2}V}{\sqrt{V}} \text{ and } d_2 = d_1 - \frac{1}{2}V.$$

$$\Delta_i = \frac{\partial c}{\partial F_i(t)}$$

$$= e^{-r(T-t)}\left[\frac{\partial M_1}{\partial F_i(t)}.N(d_1) + M_1.n(d_1).\frac{\partial d_1}{\partial F_i(t)} - X.n(d_2).\frac{\partial d_2}{\partial F_i(t)}\right]$$

From definitions of M_1, M_2 and V we obtain

$$\frac{\partial M_1}{\partial F_i(t)} = a_i \qquad \text{.....(C.1)}$$

$$\frac{\partial M_2}{\partial F_i(t)} = 2a_i \sum_{j=1}^{N} a_j.F_i(t).F_j(t).\exp\left(\sigma_i.\sigma_j.\rho_{i,j}(T-t)\right) \qquad \text{.....(C.2)}$$

$$\frac{\partial \sqrt{V}}{\partial F_i(t)} = \frac{1}{2\sqrt{V}}.\frac{{M_1}^2}{M_2}.\left[\frac{\frac{\partial M_2}{\partial F_i(t)}.{M_1}^2 - M_2.2.M_1\frac{\partial M_1}{\partial F_i(t)}.}{{M_1}^4}\right]$$

$$= \frac{1}{2M_1.M_2\sqrt{V}}.\left[\frac{\partial M_2}{\partial F_i(t)}.M_1 - 2.\frac{\partial M_1}{\partial F_i(t)}.M_2\right] \qquad \text{.....(C.3)}$$

$$\frac{\partial d_2}{\partial F_i(t)} = \frac{\partial d_1}{\partial F_i(t)} - \frac{\partial \sqrt{V}}{\partial F_i(t)} \text{ since } d_2 = d_1 - \sqrt{V}\,. \qquad \text{.....(C.4)}$$

Using the definitions of d_1, d_2 and the probability density function of standard normal distribution, we can show $M_1.n(d_1) = X.n(d_2)$. Consequently,

$$\Delta_i = e^{-r(T-t)}\left[\frac{\partial M_1}{\partial F_i(t)}.N(d_1) + M_1.n(d_1).\frac{\partial d_1}{\partial F_i(t)} - X.n(d_2).\frac{\partial d_2}{\partial F_i(t)}\right]$$

$$= e^{-r(T-t)}\left[\frac{\partial M_1}{\partial F_i(t)}.N(d_1) + M_1.n(d_1).\frac{\partial d_1}{\partial F_i(t)} - X.n(d_2)\left(\frac{\partial d_1}{\partial F_i(t)} - \frac{\partial \sqrt{V}}{\partial F_i(t)}\right)\right]$$

$$= e^{-r(T-t)}\left[\frac{\partial M_1}{\partial F_i(t)}.N(d_1) + X.n(d_2).\frac{\partial \sqrt{V}}{\partial F_i(t)}\right].$$

$$\Gamma_{i,j} = \frac{\partial^2 c}{\partial F_j(t).\partial F_i(t)}$$

$$= \frac{\partial}{\partial F_j(t)}\left(\frac{\partial c}{\partial F_i(t)}\right)$$

$$= e^{-r(T-t)}\left[a_i.n(d_1)\frac{\partial d_1}{\partial F_j(t)} + X.n(d_2)\frac{\partial^2\sqrt{V}}{\partial F_j(t).\partial F_i(t)} + X\frac{\partial n(d_2)}{\partial F_j(t)}\frac{\partial\sqrt{V}}{\partial F_i(t)}\right].$$

$\frac{\partial\sqrt{V}}{\partial F_i(t)}$ is obtained from eq. (C.3) and in the following we derive $\frac{\partial d_1}{\partial F_j(t)}$, $\frac{\partial^2\sqrt{V}}{\partial F_j(t).\partial F_i(t)}$ and $\frac{\partial n(d_2)}{\partial F_i(t)}$

It is defined that

$$d_1 = \frac{\log(M_1) - \log(X) + \frac{1}{2}V}{\sqrt{V}} = \frac{\log(M_1/X)}{\sqrt{V}} + \frac{1}{2}\sqrt{V}\ .$$

Then $\frac{\partial d_1}{\partial F_i(t)} = \frac{\frac{1}{M_1}.\frac{\partial M_1}{\partial F_i(t)}.\sqrt{V} - \log(M_1/X).\frac{\partial\sqrt{V}}{\partial F_i(t)}}{V} + \frac{1}{2\sqrt{V}}\frac{\partial\sqrt{V}}{\partial F_i(t)}$,

where $\frac{\partial M_1}{\partial F_i(t)}$ and $\frac{\partial\sqrt{V}}{\partial F_i(t)}$ is obtained from equations (C.1) and (C.3).

$$\frac{\partial^2 M_2}{\partial F_j(t).\partial F_i(t)} = \frac{\partial}{\partial F_j(t)}\left(\frac{\partial M_2}{\partial F_i(t)}\right) = 2.a_i.a_j.\exp\left(\sigma_i.\sigma_j.\rho_{i,j}(T-t)\right).$$

Suppose that

$$w_{1,i} = .\frac{\partial M_2}{\partial F_i(t)}.M_1 - 2.\frac{\partial M_1}{\partial F_i(t)}.M_2 \text{ and}$$

$$w_2 = 2.M_1.M_2.\sqrt{V}\ .$$

In terms of $w_{1,i}$ and w_2, $\frac{\partial\sqrt{V}}{\partial F_i(t)}$ in eq. (C.3) can be expressed as $\frac{w_{1,i}}{w_2}$. Then

$$\frac{\partial^2\sqrt{V}}{\partial F_j(t).\partial F_i(t)} = \frac{\frac{\partial w_{1,i}}{\partial F_j(t)}.w_2 - w_{1,i}.\frac{\partial w_2}{\partial F_j(t)}}{{w_2}^2},$$

where $\frac{\partial w_{1,i}}{\partial F_j(t)} = a_j . \frac{\partial M_2}{\partial F_i(t)} + M_1 . \frac{\partial^2 M_2}{\partial F_j(t) . \partial F_i(t)} - 2.a_i . \frac{\partial M_2}{\partial F_j(t)}$ and

$$\frac{\partial w_2}{\partial F_j(t)} = 2.M_2 . \sqrt{V} . \frac{\partial M_1}{\partial F_j(t)} + 2.M_1 . \sqrt{V} . \frac{\partial M_2}{\partial F_j(t)} + 2.M_1 . M_2 . \frac{\partial \sqrt{V}}{\partial F_j(t)} .$$

The probability density function of standard normal distribution defines

$$n(d_2) = \frac{1}{2\sqrt{\pi}} . \exp\left(-\frac{1}{2} . d_2^{\,2} \right).$$

Then $\frac{\partial n(d_2)}{\partial F_i(t)} = -d_2 . n(d_2) . \frac{\partial d_2}{\partial F_i(t)}$,

where $\frac{\partial d_2}{\partial F_i(t)}$ is obtained from eq.(C.4).

$$\rho = \frac{\partial c}{\partial r} = -\left(r\left(T-t\right) \right) . e^{-r(T-t)} \left[M_1 . N\left(d_1\right) - X . N\left(d_2\right) \right] = -\left(r(T-t) \right) . c .$$

$$\Theta = \frac{\partial c}{\partial t}$$

$$= r.c + e^{-r(T-t)} \left[M_1 . n(d_1) . \frac{\partial d_1}{\partial t} - X . n(d_2) . \frac{\partial d_2}{\partial t} \right]$$

$$= r.c + e^{-r(T-t)} \left[M_1 . n(d_1) . \frac{\partial d_1}{\partial t} - X . n(d_2) \left(\frac{\partial d_1}{\partial t} - \frac{\partial \sqrt{V}}{\partial t} \right) \right],$$

since $\frac{\partial d_2}{\partial t} = \frac{\partial d_1}{\partial t} - \frac{\partial \sqrt{V}}{\partial t}$

$$= r.c + e^{-r(T-t)} . X . n(d_2) . \frac{\partial \sqrt{V}}{\partial t} \text{, since } M_1 . n(d_1) = X . n(d_2) .$$

, where $\frac{\partial \sqrt{V}}{\partial t} = \frac{\partial \sqrt{V}}{\partial M_2} . \frac{\partial M_2}{\partial t}$

$$\frac{\partial \sqrt{V}}{\partial M_2} = \frac{1}{2.M_2 . \sqrt{V}} .$$

$$\frac{\partial M_2}{\partial t} = \sum_{j=1}^{N} \sum_{i=1}^{N} - \rho_{i,j} . \sigma_i . \sigma_j . a_i . a_j . F_i(t) . F_j(t) . \exp\left(\sigma_i . \sigma_j . \rho_{i,j} (T-t) \right).$$

Vega (ν) is defined as a matrix (nxn) where $\nu_{i,j} = \begin{cases} \dfrac{\partial c}{\partial \sigma_i}, & i = j \\ \dfrac{\partial c}{\partial \rho_{i,j}}, & i \neq j \end{cases}$.

For $i \neq j$:

$$\nu_{i,j} = \frac{\partial c}{\partial \rho_{i,j}}$$

$$= e^{-r(T-t)}\left[M_1.n(d_1).\frac{\partial d_1}{\partial \rho_{i,j}} - X.n(d_2).\frac{\partial d_2}{\partial \rho_{i,j}} \right]$$

$$= e^{-r(T-t)}\left[M_1 n(d_1)\frac{\partial d_1}{\partial \rho_{i,j}} - Xn(d_2)\left(\frac{\partial d_1}{\partial \rho_{i,j}} - \frac{\partial \sqrt{V}}{\partial \rho_{i,j}} \right) \right], \text{since} = \frac{\partial d_1}{\partial \rho_{i,j}} - \frac{\partial \sqrt{V}}{\partial \rho_{i,j}}$$

$$= e^{-r(T-t)}.X.n(d_2).\frac{\partial \sqrt{V}}{\partial \rho_{i,j}} \text{ , since } M_1.n(d_1) = X.n(d_2).$$

, where $\dfrac{\partial \sqrt{V}}{\partial \rho_{i,j}} = \dfrac{1}{2.M_2.\sqrt{V}}.\dfrac{\partial M_2}{\partial \rho_{i,j}}$.

$$\frac{\partial M_2}{\partial \rho_{i,j}} = 2.a_i.a_j.F_i(t).F_j(t).\sigma_i.\sigma_j.(T-t).\exp(\sigma_i.\sigma_j.\rho_{i,j}(T-t)).$$

For $i = j$:

$$\nu_{i,i} = \frac{\partial c}{\partial \sigma_i}$$

$$= e^{-r(T-t)}\left[M_1.n(d_1)\frac{\partial d_1}{\partial \sigma_i} - X.n(d_2)\frac{\partial d_2}{\partial \sigma_i} \right]$$

$$= e^{-r(T-t)}\left[M_1.n(d_1)\frac{\partial d_1}{\partial \sigma_i} - X.n(d_2)\left(\frac{\partial d_1}{\partial \sigma_i} - \frac{\partial \sqrt{V}}{\partial \sigma_i} \right) \right],$$

$$\text{since } \frac{\partial d_2}{\partial \sigma_i} = \frac{\partial d_1}{\partial \sigma_i} - \frac{\partial \sqrt{V}}{\partial \sigma_i}$$

$$= e^{-r(T-t)}.X.n(d_2)\frac{\partial \sqrt{V}}{\partial \sigma_i}, \text{ since } M_1.n(d_1) = X.n(d_2)$$

, where $\frac{\partial M_2}{\partial \sigma_i} = 2\sum_{j=1}^{N} \rho_{i,j}.\sigma_j.(T-t).a_i.a_j.F_i(t).F_j(t).\exp(\sigma_i.\sigma_j.\rho_{i,j}(T-t))$.

$$\frac{\partial \sqrt{V}}{\partial \sigma_i} = \frac{1}{2.M_2.\sqrt{V}} . \frac{\partial M_2}{\partial \sigma_i} .$$

H Call option price of basket option using reciprocal-gamma distribution approach

$$c = \exp(-r(T-t_0)).\hat{E}(\max(B(T)-X,0))$$

$$= \exp(-r(T-t_0)).H.\hat{E}\left(\max\left(B^*(T)-\frac{X}{H},0\right)\right)$$

$$= \exp(-r(T-t_0)).H.\int_{t_0}^{\infty} \max(B^*(T)-X/H,0).d\psi(B^*(T))$$

$$= \exp(-r(T-t_0)).H.\int_{X/H}^{\infty} (B^*(T)-X/H).g_r(B^*(T),\alpha,\beta).d(B^*(T))$$

$$= \exp(-r(T-t_0)).H\left[\int_{X/H}^{\infty} B^*(T).g_r(B^*(T),\alpha,\beta).d(B^*(T)) - \int_{X/H}^{\infty} (X/H).g_r(B^*(T),\alpha,\beta).d(B^*(T))\right]$$

$$= \exp(-r(T-t_0)).H\left[\int_{X/H}^{\infty} y..g_r(y,\alpha,\beta).dy - \int_{X/H}^{\infty} (X/H).g_r(y,\alpha,\beta).dy\right]$$

$$= \exp(-r(T-t_0)).H\left[\int_{X/H}^{\infty} y.\frac{g\left(\frac{1}{y},\alpha,\beta\right)}{y^2}dy - \frac{X}{H}\int_{X/H}^{\infty} .\frac{g\left(\frac{1}{y},\alpha,\beta\right)}{y^2}.dy\right]$$

$$= \exp(-r(T-t_0)).H\left[\int_{X/H}^{\infty} \frac{g\left(\frac{1}{y},\alpha,\beta\right)}{y}dy - \frac{X}{H}\int_{X/H}^{\infty} .\frac{g\left(\frac{1}{y},\alpha,\beta\right)}{y^2}.dy\right]$$

By substituting $u = \frac{1}{y}$, the call price formula becomes

$$c = \exp(-r(T-t_0)\left[H.\int_{t_0}^{H/X} \frac{g(u,\alpha,\beta).}{u}.du - X.\int_{t_0}^{F/X} g(u,\alpha,\beta).du\right]$$

$$= \exp\left(-r(T-t_0)\right)\left[H.G\left(u,\alpha-1,\beta\right)-X.G\left(u,\alpha,\beta\right)\right], \text{since}$$

$$\frac{g\left(u,\alpha,\beta\right)}{u} = \frac{\exp\left(-u/\beta\right).\left(u/\beta\right)^{\alpha-1}}{u.\beta.\Gamma\left(\alpha\right)} = \frac{\exp\left(-u/\beta\right).\left(u/\beta\right)^{(\alpha-1)-1}}{\beta.\Gamma\left(\alpha-1\right)} = g\left(u,\alpha-1,\beta\right).$$

BOROVKOVA,S.A.: Delft Institute for Applied Mathematics (DIAM), Delft University of Technology, Mekelweg 4, 2628 BT, Delft, The Netherlands.
Phone: +31 (015)2784517
E-mail: S.A.Borovkova@ewi.tudelft.nl

PERMANA, F.J.: Delft Institute for Applied Mathematics (DIAM), Delft University of Technology, Mekelweg 4, 2628 BT, Delft, The Netherlands.
Phone: +31 (015)2784563
E-mail: F.J.Permana@ewi.tudelft.nl

HOW REWARDING IS THE MOVING AVERAGE AND RELATIVE STRENGTH INDEX? EVIDENCE FROM VARIOUS COMPANIES AND THE KUALA LUMPUR COMPOSITE INDEX IN BURSA MALAYSIA

Seow Chiao Ju, Tai Lee Meng, Zainudin Arsad

Universiti Sains Malaysia, Penang, Malaysia

Abstract. Investment in the stock and commodity market has become very crucial to many people in this era. Many investors and traders have been searching for the ultimate method for high profit investment. Technical analysis has make boundless contribution to the forecast of future price trends in the financial and commodity market. This would definitely help in the investment of investors and traders if the forecast were to be significant. Basic t-statistical test has been applied in this research to examine the significance of the signals generated by the respective technical indicators. The main objective of this project is to examine the significance of technical indicators like Moving Average and Relative Strength Index as a forecasting tool focusing on numerical evidence, using data from various companies and the Kuala Lumpur Composite Index with a period of 2 decades. From the research, it is found that the technical indicators are reliable as a forecasting tool, but subjected to certain situation. Signals generated by the Moving Average are found to be reliable only for short term period. Relative Strength Index is found to be better for the forecasting of slightly longer periods. However, the signals generated by both technical indicators are found to be not significant when used in the forecast of long term period. The signals are also found to be less significant during the economic crisis period; Asian Economic Crisis and the World Economic Crisis.

Key-words: Technical analysis, Technical indicators, Moving Average, Relative Strength Index, t-statistical test

1 Introduction

Technical analysis is known as a study of financial and commodity market actions. It is said to be one of the oldest form of financial analysis in the world with its history dating back to the Japanese rice traders trading on the Dojima Rice Exchange in Osaka as early as the 1600s. Charles H. Dow, founder of the Wall Street Journal, can be seen as the grandfather of most technical analysis. He wanted to visualize the development of the economy and thus, he developed the Dow Jones Index. He strongly believed that the stock market characterizes itself by persistent price movements.

Analyst uses technical analysis to evaluate the price variations that occurs everyday. Graphical charts are used vastly in technical analysis. By using the method of assessing the past prices and market behaviors, technical analysis has become one of the most popular method of short term forecasting in the financial

and commodity markets. Technical analysis does not attempt to measure a security's intrinsic value but instead uses charts to identify patterns that suggest future activity. In addition, technical analysis also includes the examination of the current emotional state of the relevant market in its continuous battle between fear and greed.

Technical analysis is not a homogenous body of knowledge. There are many different technical indicators used in the assessment of the current financial and commodity market. In fact, different security companies use different trading rules formed by various technical indicators. Technical analysis is not a scientific way of analyzing the stock and commodity market. It has very little theory and little evidence as to how these indicators are generated.

The role of technical analysis as a forecasting mechanism has increased tremendously over the years. Investors, traders and consumers have been using technical analysis in their daily investment task to ensure that the best investment is being made. Nevertheless, despite the rising popularity of technical analysis, many controversial discussions and debates has been raised over the years about the efficiency and reliability of technical analysis as a forecasting tool in the financial and commodity market.

The ability of technical analysis in producing good forecasts has been investigated by [4]. The data used is the daily close of the Singapore Strait Times Index (STI) from 1 January 1974 to 31 December 1994 (21 years). Their paper has focused on the role of technical analysis in signaling the timing of stock market entry and exit. The full sample is divided into 3 sub-periods of 7 years each. Test statistics were used to test whether the buy and sell signals yield significantly positive return. To investigate on the significance of these indicators, the daily returns of the data are computed.

Using this data, the results showed that the indicators can be used to generate significantly positive return. In other words, technical indicators can play a useful role in the timing of stock market entry and exits and thus enjoy substantial profit. From the results that technical indicators can play a useful role in the timing of stock market entry and exits. By applying the technical indicators, member firms of the Stock Exchange of Singapore (SES) may enjoy substantial profits. Therefore, it is not surprising that most member firms do have their own trading teams that rely heavily on technical analysis.

In another research by [3], the results have indicated that making trading decisions based on moving average rules leads to significantly higher returns than the buy-and-hold policy, even in the presence of transaction costs. In fact, shorter period moving averages give better returns than longer periods. Other studies on technical analysis include [1] and [2]. In contrast to [4] and [3], [2] found out that the Moving Average does not track well the buy-and –sell activities for both the Kuala Lumpur Composite Index and New York Dow Jones Industrial Average Index, giving rise to doubts about the efficacy of technical indicators.

The capability and strength of the technical analysis in forecasting the trend of future price has been questioned by many people. Technical indicators are stimulated using the rising and falling from past prices only. Some technical

indicators are generated by very experienced analysts of the stock market. This has caused technical analysis to be very highly judgmental and very subjective to the interpretation of individual analyst. However, technical analysis is still one of the most well known tools used for short-term forecasting as there is strong evidence that simple forms of technical analysis contain significant forecasting power.

In this paper, the main objective is to examine and testify if technical analysis is really rewarding in computing forecast of the future price trends. Specifically, appropriate statistical test shall be introduced to examine if the buy and sell signals generated by the respective technical indicators yield significantly positive return. Many economists have concluded through various analyses that different economic period that have occurred over the years affects various economic variables significantly. Hence, in addition to the investigation of the reliability of technical analysis, this working paper would also discuss if the respective technical indicators are applicable to different economic periods that have occurred over the years.

This paper is organized as follows. The 2nd Section discusses 2 of the most popular technical indicators that have been widely used in the financial and commodity market. In the 3rd Section, the data and research procedures are discussed. The 4th Section discusses on some of the findings from the research, followed by some conclusion and comments in the final section.

2 Technical Indicators

There are numerous technical indicators used in technical analysis. Generally, technical indicators can be categorized into 2 groups, trend following indicators and counter-trend indicators. In this paper, only 2 technical indicators shall be discussed; Moving Average (MA) which is a trend following indicator and Relative Strength Index (RSI) which is a counter-trend indicator.

Moving Average (MA)

Moving average is one of the most well known methods used in technical analysis. All moving averages are lagging indicators, which is why they fit the category of trend following indicators. When prices are trending, moving averages work well. However, when prices are not trending, moving averages can give misleading signals. A buy signal is generated when the closing price rises above the MA and sell signal is generated when the closing price falls below the MA.

There are a few types of moving averages. One of the simplest moving averages is known as the Simple Moving Average (SMA). SMA is formed by computing the average closing price of a security over N number of days. The mathematical form of SMA can be written as follow:

$$\begin{aligned} \mathrm{SMA}_{t,N} &= \frac{1}{N} \sum_{i=t-N+1}^{t} C_i \\ &= C_t + C_{t-1} + \ldots + C_{t-N+2} + C_{t-N+1} \end{aligned}$$

where $\mathrm{SMA}_{t,N}$ is the simple N-day moving average for day t and C_i is the closing price for period i.

Relative Strength Index (RSI)

The relative strength index (RSI) is a momentum indicator that uses the net difference of closing prices for up days and down days. The value of RSI is expressed as an oscillator with a range from 0 to 100. The index set can be defined as $I_{t,p} = \{i : t - p \le i \le t\}$. The up-closes and the down-closes are defined such that,

$$U_i = \begin{cases} C_i - C_{i-1} & \text{if } C_i > C_{i-1} \\ 0 & \text{otherwise} \end{cases}$$

$$D_i = \begin{cases} C_{i-1} - C_i & \text{if } C_{i-1} > C_i \\ 0 & \text{otherwise} \end{cases}$$

for any $i \in I_{t,p}$ and C_i is the closing price for period i. The N-day SMA for the up-closes and the down-closes are then computed and abbreviated as below:

$$\overline{U}_{t,p} = \text{Average of } U_i \text{ over } I_{t,p}$$
$$\overline{D}_{t,p} = \text{Average of } D_i \text{ over } I_{t,p}$$

With that, the relative strength is calculated by using the formula as shown,

$$RS_{t,p} = \frac{\overline{U}_{t,p}}{\overline{D}_{t,p}}$$

Finally, the RSI at time t for period p is then defined as:

$$RSI_{t,p} = 100 - \frac{100}{1 + RS_{t,p}}$$

In this research, the 'touch' method is being used when generating the respective buy and sell signals. Values above 70 indicate a stock is overbought and may be a sell candidate, while values below 30 indicate the stock is oversold and the price may rise. Values of about 50 often indicate areas of resistance in the corresponding stock. Readings of 100 imply that there are pure upward price movements while readings of 0 imply that there are pure downward price movements. Generally, the longer the time period used, the less frequent and more stable are the trading signals generated. Shorter time periods tend to generate more noise than longer periods.

3 Data and Research Procedures

A total of 10 sets of company data of the daily closing price have been selected from the Main Board. These companies' data are of different ranking, sectors and financial background. In addition, the data of KLCI shall be applied in this research as well. The time period of the data selected is from 1984 until 2004.

In this research, the student-t test is used to test whether the buy and sell signals yield significantly positive return. Since the objective of this research is to testify the significance of technical analysis as a forecasting tool, different forecasting period has been tested for both MA and RSI; a forecast for 2 days ahead, a forecast for 21 days ahead and a forecast for 100 days ahead.

The closing prices of these companies have been used to compute the daily returns, r_t. For the purpose of the analysis on technical indicators, a slightly different method is used to calculate the return from the investment. For example, to calculate a buy return for an *n* day forecast using the *m*-day RSI, the natural logarithm of the difference of the first higher closing price after a buy signal is generated within the *n* days is calculated. If none of the following *n* days show a higher price than the price before, then it is assumed that the trader would sell on the *n*th day. A sell return is then generated using the same concept, only that it would be to calculate the natural logarithm of the difference of the first lower closing price after a sell signal is generated within the *n* days. Again, if none of the following *n* days show a lower price than the price before, then it is assumed that the trader would buy on the *n*th day.

Before applying the student-t test, a basic percentage analysis is used to get a general idea of the outcome that might be obtained. After the respective buy and sell signals have been generated, the buy and sell signals are then compared with the daily returns of the data. From a logical point of view, when a buy signal is generated, the return would be expected to be positive whereas when a sell signal is generated, the return would be expected to be negative. The signals generated are compared and the percentage of "true" and "false" is calculated, whereby "true" is denoted as the day with an expected return, and "false" is denoted as the day without an expected return. The percentage of the respective "true" and "false" are calculated using the formulae as shown:

$$\%\ \text{of "true"} = \frac{\text{Number of true signals}}{\text{Number of signals generated}} \times 100$$

$$\%\ \text{of "false"} = \frac{\text{Number of false signals}}{\text{Number of signals generated}} \times 100$$

In this research, the student t-test has been used to calculate the significance of the signals generated by the technical indicators. From a traders' point of view, it would be logic to say that if one buys a share, a positive return is expected and if one sells its share then a negative return is expected. From the statement above, the hypotheses that would be tested is as follows:

$$H_0 : \bar{r}_{t,\text{buy}} = 0 \qquad H_0 : \bar{r}_{t,\text{sell}} = 0$$
$$H_1 : \bar{r}_{t,\text{buy}} > 0 \qquad H_1 : \bar{r}_{t,\text{sell}} < 0$$

where $\bar{r}_{t,\text{buy}}$ is the return for buy signals and $\bar{r}_{t,\text{sell}}$ is the return for sell signals.

The t-statistic value can be calculated with the formula as shown below:

$$t = \frac{\bar{r}_t - \mu}{s/\sqrt{n}}$$

where: t is the statistic value
$\bar{r}_t$ is the average return
s is the sample standard deviation
n is the number of observations

The hypotheses are tested and the conclusions from the test can be summarized as follows:

Table 1: Summary of tests

Alternative hypothesis	Reject the null hypothesis	Accept the null hypothesis
$\bar{r}_t > 0$	$t \geq t_\alpha$	$t < t_\alpha$
$\bar{r}_t < 0$	$t \leq -t_\alpha$	$t > -t_\alpha$
$\bar{r}_t \neq 0$	$t \leq -t_{\alpha/2}$ or $t \geq t_{\alpha/2}$	$-t_{\alpha/2} < t < t_{\alpha/2}$

From the results above, relevant conclusions can be drawn about the test based on the hypothesis. The t-statistics value has been tested at the level significance of 5%. Note that the data of the companies have been divided into different economic periods. This is with an interest of investigating if different economic periods show any significant differences in the results obtained. The 7 periods are shown as follow:

Table 2: Economic Periods from 1973 to Present

Economic Period	Period Time			Code Name
Developing Period	1 Oct 1973	-	31 Dec 1990	Period 1
Market Advancement	2 Jan 1991	-	5 Jan 1994	Period 2
Stability Period	6 Jan 1994	-	24 Feb 1997	Period 3
Asian Economic Crisis	25 Feb 1997	-	1 Sept 1998	Period 4
1st Recovery Period	2 Sept 1998	-	26 Dec 2000	Period 5
World Economic Crisis	2 Jan 2001	-	21 May 2002	Period 6
2nd Recovery Period	22 May 2002	-	Present	Period 7

4 Empirical Results

The first analysis on both MA and RSI is conducted based on the return of the share for 2 days after a certain buy or sell signal is generated. In general, when a

certain share is bought, a positive return would be expected. Conversely, if a certain share is sold, then a negative return is to be expected. Based on the statement made above, the return calculated from the closing price is compared to the signals generated by the technical indicators.

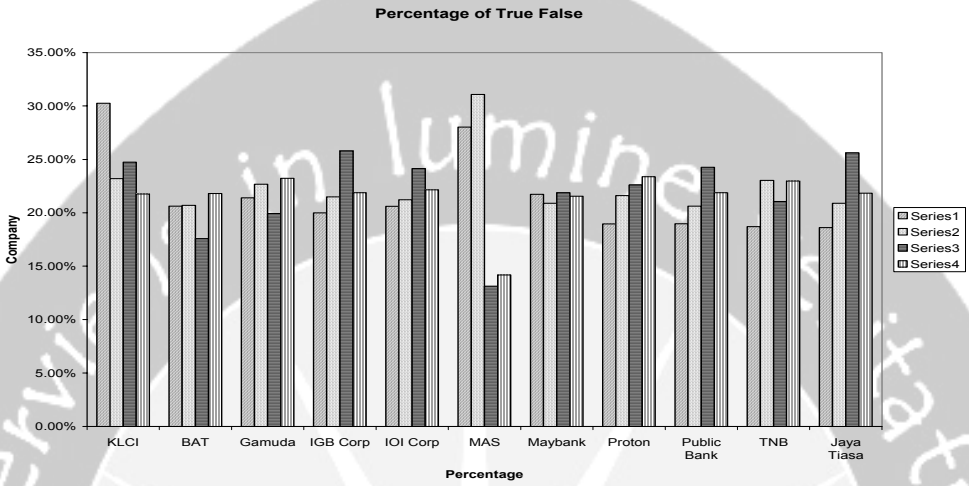


Figure 1: Percentage of True False for 2 Days Ahead for MA

From Figure 1, the percentages of the day 2 days after with the expected return for both buy and sell signals generated by the MA is roughly the same as the percentage without the expected returns when looking at both buy and sell signals.

KLCI shows that the percentage of the second day with the expected return for both buy and sell signals are much higher compared to the rest of the companies. All the other 10 companies are showing, to a certain extent almost the same percentage for both true and false. However, for KLCI, the difference can be distinguished clearly. This could very much be due to the fact that KLCI is a combination of all these 10 companies, and also the remaining 90 companies in the main board, taking into account all the different sectors and different economic backgrounds of each company.

The sell signals again show a higher percentage with the expected return comparably to the buy signals. One of the reasons could be that in the short run, investors would tend to sell immediately after the sell signal has been generated as prolonging the process would increase the loss of an investor. The signals generated are for the prediction of the short run price trend.

Thus, when a sell signal is generated, the prices are liable to fall within the 2 days. An investor would be very much concerned that if the signal is true, the investor would then generate a loss if the share is not sold immediately. Buy signals generated are not that accurate because the investors would prefer to take safer steps when making a choice as to which share should be bought. Hence, investors would opt to wait and analyze the market until the situation is more convincing before making the move. This has once again proved that the majority of the

investors in Malaysia is somewhat less aggressive or is not prone to making high risk investments.

Considering the fact that the percentage of the day with or without an expected return is calculated based on the return and the signals generated, hence, when all the investors have the same psychology as mentioned above, meaning selling the share immediately in the short run, then this would be reflected by the percentage of true in the plot above. This same analysis is conducted by using RSI as the technical indicator to generate buy and sell signals. Figure 2 shows the summary of the analysis.

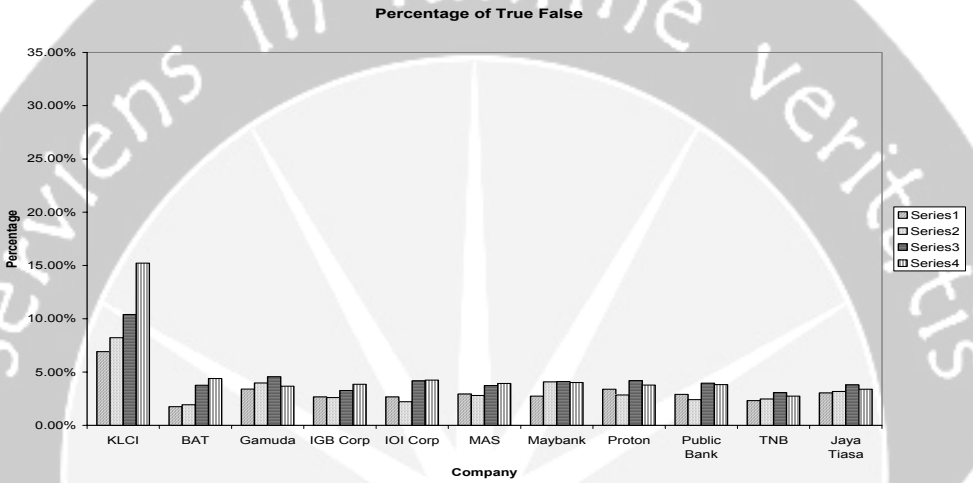


Figure 2: Percentage of True False for 2 Days Ahead for RSI

From Figure 2, it is definitely obvious that the percentage of both days without an expected return is higher than the percentage of both days with an expected return, including KLCI. This has again been a statement to infringe the trust on technical analysis as an accurate forecasting tool. Investors who followed the signals generated by the RSI would also eventually loss out.

The percentage of a day with an expected return and another without an expected return is also fairly high. This is a sign of inconsistency in the signals, resulting in losses in the investments made. With both analyses carried out, it is only fair to say that both technical indicators are not accurate showing inconsistency in this signals generated. This would lead to losses in investments made by investors.

The results have been alarmingly shocking as these two technical indicators have been widely used in the stock and commodity market. Many investors have been dependable on these technical indicators that it has been an essential tool in the investment industry. If it is known that technical indicators are not a good and accurate forecasting tool, will it still be appropriate to treat technical indicators, in this case, MA and RSI as an essential tool in making forecast

Figure 3 shows that the percentage of the day, for 21 days ahead, with an expected return is much higher than the percentage without an expected return. This is applicable to buy signals generated by the MA. As for sell signals, there is still a 50-50 border if MA can be reliable in making the necessary forecast.

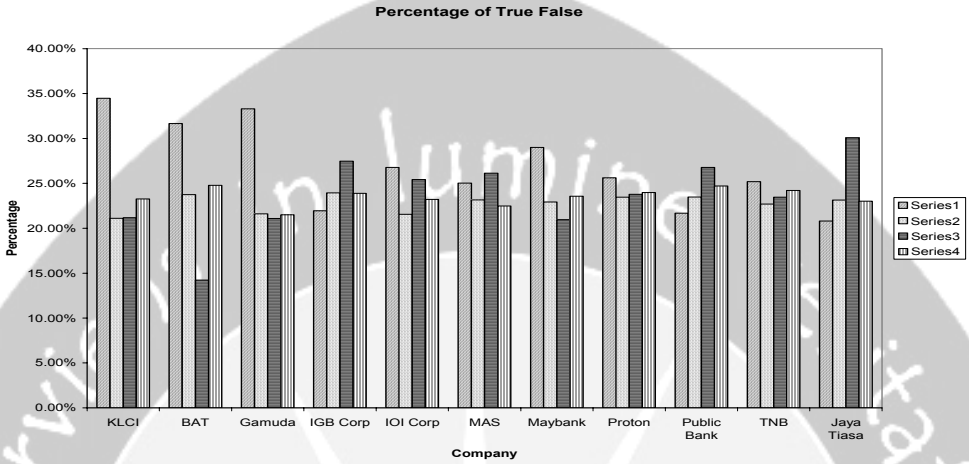


Figure 3: Percentage of True False for 21 Days Ahead for MA

When referring to only the buy signals generated, the profit gained from the investment would be good as the percentage to get the expected return is much higher. Nevertheless, this is only seen for bigger companies but cannot be observed for small companies, such as those ranked 50 and above. In other words, bigger companies are more eligible in using MA as a forecasting tool and the profit gained is promising.

Conceivably, when the signal generated by the MA is "buy", thus an investor would follow the signal and buy the share. With that, a profit is gain from this action. Nevertheless, when the signal generated is "sell", the investors would usually wait, using the "buy and hold" strategy, which does not always apply, and causing losses in the investment. When such a case happens, the percentage of the day without an expected return would generally be higher because the necessary action, which is selling the share, is not taken.

Figure 4 displays the same analysis conducted using RSI as the technical indicator. From the plot, it is rather obvious that the percentage of the day without an expected return is much higher when compared relatively to the percentage of the day with an expected return for both buy and sell signals generated. This would one way or another cause loss to the investors' investment if they were to follow the signals generated by the RSI.

One of the reasons for RSI to show less reliability in forecasting could be due to the fact that RSI has generated lots of "no signal". This is because the principal of RSI is to buy when it is below 30 and sell when the RSI index is above 70. Thus, values of RSI that falls in between 30 to 70 are considered as an area of no movements. There would then be no signals generated, causing investors to not know what to

do. Nevertheless, the forecasting performance of MA is definitely better than the RSI in all of the above 3 cases.

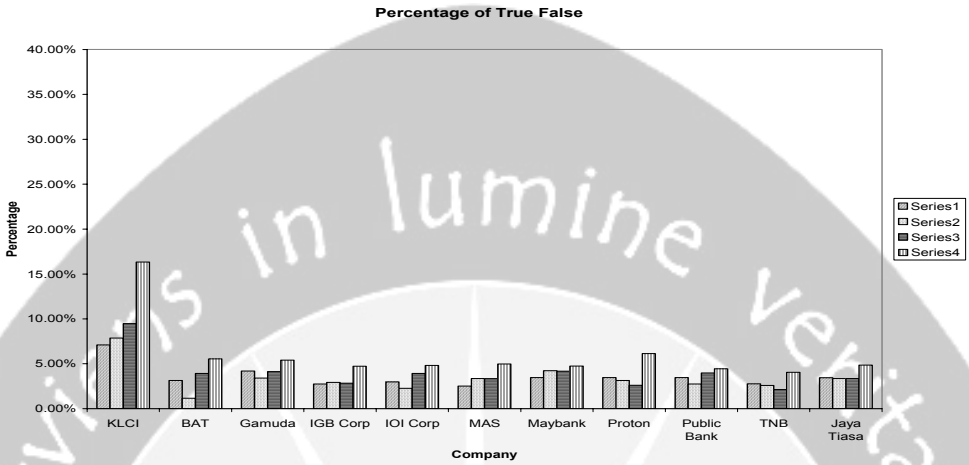


Figure 4: Percentage of True False for 21 Days Ahead for RSI

Finally, the analysis for 100 days ahead is conducted to see if the results shown by the MA and RSI would have improved from the analysis of the 21 days ahead, considering the fact that now, the MA works somewhat better as the days ahead that is predicted increases, whereas the performance of the RSI deteriorates.

From Figure 5, the buy signals generated are more accurate than the sell signals generated. This is almost alike to the results for the analysis for 21 days ahead. Nevertheless, the buy signals generated have a whole different pattern when compared to the analysis for 21 days ahead, in the sense that not only the smaller companies do not see the accuracy of the signals generated by the MA but TNB, being a big company does not see it as well.

With an inconsistency in the results, perhaps it would be wise to take a step further in the analysis of the MA to ensure its reliability and accuracy in the prediction of the price trends in the future. However, it is found that the MA works well in predicting price trends which are further ahead, compared relatively to the prediction of price trends two days later.

From Figure 6, RSI continues to show deterioration in its performance in the prediction of future price trends. This would definitely be a concerned as RSI is very widely used in the investments of the stock and commodity market. As seen in the results of the previous analysis, it is found that the percentage of the day with an expected return is still lower than the percentage of the day without an expected return.

From this basic analysis, a more sophisticated and statistically related method shall be the main approach for the future analysis that will be discussed later. The statistical approach used would ensure that the results and signals generated by different technical indicators are tested statistically and a more concrete answer as

to whether technical analysis is rewarding to the stock and commodity market can be clarified.

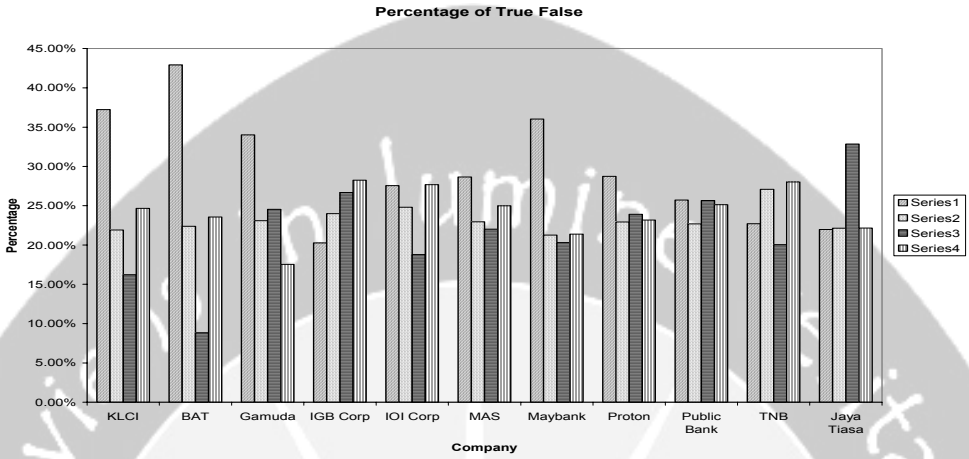


Figure 5: Percentage of True False for 100 Days Ahead for MA

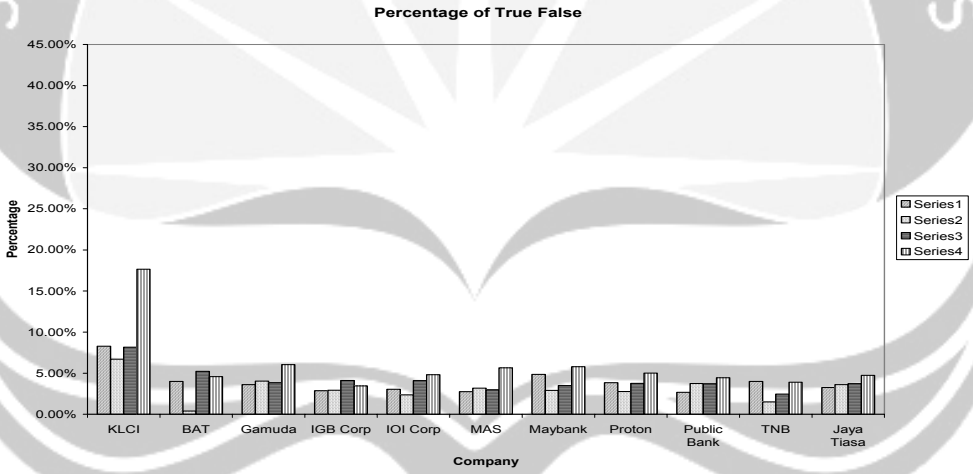


Figure 6: Percentage of True False for 100 Days Ahead for RSI

Student t-test has been conducted on the 2-day Moving Average (MA) and 21-day MA to test the significance of this technical indicator. From the results, it is found that MA would be a good forecasting tool for 2 days MA, which in this case would be a shorter period of forecasting. Nevertheless, MA becomes less dependable when the period is lengthened, for instance, more than 21 days MA. The results of this analysis can be illustrated in the graphs and table shown on the following page.

Table 3 shows the results for the t-test analysis for the 2-day MA. From the results obtained, it can be clarified that the signs for both signals are accurate. In other words, the returns for buy signals are positive, and the returns for sell signals are negative. This strengthens the principle mentioned earlier.

Table 3: t-statistic values for both buy and sell signals (2-day MA)

		Period 1	Period 2	Period 3	Period 4	Period 5	Period 6	Period 7
KLCI	buy	8.055	8.905	7.082	4.492	6.346	5.259	6.670
	sell	-7.820	-6.466	-7.421	-5.790	-4.678	-5.200	-6.390
BAT	buy	6.313	6.500	7.515	4.610	6.843	5.401	7.715
	sell	-5.819	-5.428	-6.763	-4.984	-5.979	-4.530	-6.439
Gamuda	buy		5.728	4.031	4.137	4.804	5.321	6.767
	sell		-3.326	-4.805	-5.071	-4.181	-4.195	-6.597
IGB	buy	9.244	7.682	7.905	4.858	5.566	3.575	7.163
	sell	-8.642	-6.204	-7.528	-5.408	-4.779	-3.572	-5.934
IOI	buy	11.175	8.294	7.236	5.257	6.641	5.219	6.737
	sell	-9.697	-5.687	-6.679	-5.598	-5.671	-4.425	-6.308
MAS	buy	7.370	5.062	5.592	3.216	4.155	2.944	4.341
	sell	-10.137	-6.412	-8.801	-7.067	-4.825	-4.757	-7.609
Maybank	buy	5.004	7.239	7.670	3.671	7.001	3.379	7.163
	sell	-4.839	-6.051	-7.352	-4.937	-5.277	-4.057	-6.563
Proton	buy		5.923	7.763	4.147	6.488	4.540	7.364
	sell		-4.436	-6.305	-5.083	-5.376	-3.535	-6.242
Public Bank	buy	9.362	8.078	5.973	4.118	6.785	3.398	3.785
	sell	-8.719	-4.963	-5.698	-4.852	-5.285	-3.386	-2.865
TNB	buy		6.275	7.749	4.601	6.784	4.838	6.844
	sell		-3.954	-8.102	-5.561	-4.646	-4.841	-6.578
Jaya Tiasa	buy	8.368	6.920	7.736	4.574	6.496	4.190	6.413
	sell	-8.228	-4.662	-6.428	-5.364	-5.255	-4.204	-5.319

From the t-statistic values in Table 3, it is found that the values that have been calculated are within the critical area, indicating that the values are significant at the significant level of 5%. Considering the fact that the values are significant, the null hypotheses for both signals are rejected. The implication behind this statement would be that the MA is able to generate a positive return for both buy and sell signals. Traders would not lose out if they follow the signals generated from the MA.

From Table 3, it can be seen that the t-statistic values are lower during the economic crisis period; Asian Economic Crisis and World Economic Crisis. This would be a sign indicating that traders have to be extra careful during economic crisis periods as it might be misleading, causing losses to the traders themselves. The less significance of the t-statistic values during the economic crisis period could be due to the unpredictable of the situation that might happen next. The economic crisis period can be denoted as the weakest period of all the economic periods. Hence, the prediction might not be accurate anymore as there could be many economic situations that might happen abruptly. Nevertheless, for all other periods, the t-statistic values are fairly high, showing a high significance in the t-statistic values.

The period of the MA is then lengthened to 21 days to test if the MA still works well as a forecasting tool. The results for this analysis are illustrated in Table 4. The

analysis for 21-day MA also shows that the signs for both buy and sell signals are reasonably accurate. This indicates that the returns for the buy signals are positive and the returns for the sell signals are negative. With that, this further strengthens the statement made earlier in this paper, indicating that when a buy signal is generated, a positive return would be expected, whereas when a sell signal is generated, a negative return would be expected.

Table 4: t-statistic values for both buy and sell signals (21-day MA)

		Period 1	Period 2	Period 3	Period 4	Period 5	Period 6	Period 7
KLCI	buy	1.827*	2.219	1.392*	1.050*	1.237*	1.081*	1.682*
	sell	-1.779*	-1.365*	-1.647*	-1.201*	-0.645*	-1.176*	-1.703*
BAT	buy	1.520*	1.752*	2.038	1.161*	1.854*	2.320	2.464
	sell	-1.921*	-1.697*	-2.133	-1.793*	-1.584*	-1.782*	-2.618
Gamuda	buy		1.183*	0.971*	1.162*	1.034*	1.100*	2.128
	sell		-0.421*	-1.785*	-1.187*	-0.820*	-1.191*	-2.050
IGB	buy	2.484	2.160	1.833*	1.122*	1.450*	0.782*	1.947*
	sell	-2.246	-1.624*	-1.905*	-1.270*	-1.144*	-1.021*	-1.680*
IOI	buy	2.732	2.466	1.523*	1.393*	1.611*	1.016*	2.031
	sell	-2.343	-1.344*	-1.591*	-1.398*	-1.565*	-0.916*	-2.025
MAS	buy	1.921*	2.404	2.059	2.008	1.883*	1.225*	1.624*
	sell	-2.017	-1.544*	-1.891*	-1.803*	-1.101*	-1.091*	-1.831*
Maybank	buy	0.723*	1.614*	2.531	1.177*	1.807*	0.701*	2.068
	sell	-0.886*	-1.508*	-2.695	-1.239*	-1.221*	-1.140*	-2.318
Proton	buy		1.959*	2.020	1.252*	1.869*	1.188*	1.385*
	sell		-1.327*	-1.977	-1.268*	-1.331*	-1.038*	-1.651*
Public Bank	buy	2.305	2.429	1.555*	1.213*	1.720*	0.838*	1.277*
	sell	-2.051	-1.161*	-1.624	-1.092*	-1.155*	-1.252*	-0.923*
TNB	buy		1.782*	1.845*	1.117*	1.782*	1.184*	1.523*
	sell		-1.019*	-1.942*	-1.412*	-1.293*	-1.523*	-1.524*
Jaya Tiasa	buy	2.199	1.884*	1.707*	0.944*	2.163	0.984*	2.125
	sell	-2.400	-1.027*	-1.152*	-1.326*	-1.540*	-0.880*	-1.932*

Numbers denoted with a (*) shows values which are not significant at 5% level

In Table 4 most of the values are less than the critical value, signifying insignificance in the t-statistic values. From the table, it is found that MA is a fairly good forecasting tool when used during the developing period for all companies. The significance of the indicator decreases as it moves to a more recent period. Notice that the t-statistic values are not significant during the economic crisis period; Asian Economic Crisis and World Economic Crisis. This result has shown some similarity with the results for the 2-day MA. Traders and investors have to take extra safety steps when making their investment during the economic crisis period, as it is a very weak period and many economic situations might happen abruptly.

The 21-day MA still shows some significant t-statistic values at the significant level of 5%. Hence, it is predicted that the forecasting of the price trends would only get worse as the number of days increases. The 21-day MA is further lengthened to 100-day MA, equivalent to almost 3 months of working days and the result of the t-test are as predicted.

From the table above, the forecasting accuracy done by the MA has deteriorated, with only one significant value when tested at the significant level of 5% when compared to the 2-day MA. Furthermore, there are positive and negative signs which are accurate in the first two analyses are now found to be inaccurate in the 100-day MA. Hence, it can be predicted that as the prediction period prolongs, the role of MA as a forecasting tool deteriorates.

Table 5: t-statistic values for both buy and sell signals (100-day MA)

		Period 1	Period 2	Period 3	Period 4	Period 5	Period 6	Period 7
KLCI	buy	0.372*	0.592*	0.116*	-0.142*	0.218*	0.231*	0.250*
	sell	-0.194*	0.246*	-0.102*	-0.257*	0.077*	-0.047*	-0.130*
BAT	buy	0.451*	0.470*	0.696*	0.365*	0.355*	1.913*	0.923*
	sell	-0.457*	-0.318*	-0.915*	-0.734*	-0.010*	-1.049*	-2.174
Gamuda	buy		0.312*	0.203*	0.073*	0.343*	0.511*	0.505*
	sell		-0.359*	-0.486*	-0.454*	-0.171*	-0.331*	-0.441*
IGB	buy	1.130*	0.695*	0.640*	0.269*	0.737*	0.254*	0.495*
	sell	-0.903*	-0.521*	-0.740*	-0.377*	-0.466*	-0.288*	-0.283*
IOI	buy	0.761*	0.770*	0.481*	0.132*	0.212*	0.596*	0.480*
	sell	-0.679*	-0.275*	-0.761*	-0.388*	0.000*	-0.390*	-0.523*
MAS	buy	0.451*	0.434*	0.464*	0.156*	0.589*	0.240*	0.307*
	sell	-0.535*	-0.211*	-0.482*	-0.608*	-0.351*	-0.181*	-0.397*
Maybank	buy	0.134*	0.486*	0.437*	0.051*	0.380*	0.156*	0.585*
	sell	-0.205*	-0.408*	-0.438*	-0.397*	0.096*	-0.297*	-0.822*
Proton	buy		0.436*	0.658*	0.051*	0.700*	0.590*	0.456*
	sell		-0.980*	-0.597*	-0.358*	-0.259*	-0.610*	-0.522*
Public Bank	buy	0.598*	0.920*	0.171*	0.145*	0.646*	0.564*	0.438*
	sell	-0.578*	-0.337*	-0.278*	-0.545*	-0.622*	-0.493*	-0.523*
TNB	buy		0.688*	0.394*	0.451*	0.723*	0.200*	0.887*
	sell		-0.298*	-0.540*	-0.692*	-0.056*	-0.339*	-1.030*
Jaya Tiasa	buy	0.718*	0.854*	0.497*	0.712*	0.315*	0.400*	0.314*
	sell	-0.750*	-0.399*	-0.375*	-0.449*	-0.165*	-0.181*	-0.438*

From the result shown in the table, the t-statistic values are less significant during the economic crisis period. This is yet again another proof signifying that the significance of the t-statistic values becomes less during this period. In the case of the 100-day MA, there is no particular pattern as to which company would have a higher insignificant value over another company.

With similarity to the research conducted on the MA, simple analyses have been conducted on the Relative Strength Index (RSI) and the results have shown the RSI is not a significant price trend predictor when tested at the significant level of 5%. The result has been a shocked as many traders and investors have been using RSI to forecast future price trends. Therefore, further clarification should be made on the significance of this indicator. The t-test is again applied on the RSI and the results shall be discussed in the next sub sections.

Three different types of trading rules are used for the forecasting of 2 days forward, 5 days forward and 21 days forward. They are 14-Day RSI, 21-Day RSI and 100-Day RSI. This is with the hope to examine the nature of RSI in different periods of forecasting, mainly short term forecasting and long term forecasting. 2 days forward can be denoted as a very short duration, using it to examine the sensitivity

of RSI. 5 days forward is chosen to test the forecasting ability of RSI in a trading week. Finally, 21 days forward is considered as a long term forecasting as it is a one month trading period. The trading rules are then tested to investigate which trading rules give better forecast.

Notice that the significance for the 100 days forecast will not be tested. This is because, as the number of forecast days increases, there would bound to be a day where the price is higher than the price for that day for the generation of the buy return and there would bound to be another day where the price is lower than the price for that day for the generation of a sell return. Thus, it would not be beneficial to test the significance of 100 days forecast as the signals would have a high probability of being significant at the significant level of 5%.

The first analysis is conducted on the forecast for 2 days ahead using all the possible trading rules; 14-day RSI, 21-day RSI and 100-day RSI. A 14-day RSI would indicate 14 days look back on the data. Similarly, the 21-day RSI would indicate 21 days look back and 100-day RSI would mean 100 days look back. The result of the 2 days forecast using the 14-day RSI trading rule is shown in Table 6.

From the results in Table 6, the numbers written in red are the values which are not significant tested at the significant level of 5%. It is fairly obvious that only a handful of the values are significant at the significant level of 5%. This simply signifies that RSI is just not good in making very short period forecasting.

Table 6: t-statistic values for 2 days forecast using 14-day RSI

		Period 1	Period 2	Period 3	Period 4	Period 5	Period 6	Period 7
KLCI	buy	-2.542*	0.350*	-2.568*	-1.023*	1.089*	-2.437*	-2.566*
	sell	1.873*	1.962	0.243*	-1.031*	3.462*	0.183*	1.906*
BAT	buy	-0.151*	0.085*	0.329*	1.508*	1.339*	0.268*	1.754*
	sell	3.399	0.819*	-1.061*	-1.513*	0.000*	-1.208*	-1.164*
Gamuda	buy	-	2.028	-0.453*	-0.703*	2.420	0.006*	-0.710*
	sell	-	-0.055*	-0.898*	0.335*	0.713*	-1.178*	-0.666*
IGBS	buy	-0.473*	1.738	1.521*	-0.543*	-0.544*	-0.360*	0.536*
	sell	0.888*	-0.205*	-0.705*	-2.396	0.873*	-1.554*	-0.095*
IOI	buy	0.128*	0.267*	0.855*	-0.473*	0.540*	0.100*	0.604*
	sell	0.951*	2.055*	0.786*	0.297*	-0.849*	0.485*	1.132*
MAS	buy	-0.616*	-0.296*	2.867	-0.949*	0.562*	0.401*	-1.394*
	sell	0.891*	1.040*	-0.820*	-0.365*	0.156*	1.652*	-0.228*
Maybank	buy	-0.234*	0.135*	-0.303*	-0.227*	-	-1.850*	1.210*
	sell	0.572*	-0.544*	-0.257*	-1.348*	2.910*	-1.056*	-1.118*
Proton	buy	-	-0.110*	0.326*	-1.226*	-0.691*	1.843	-0.827*
	sell	-	-0.716*	-0.083*	-0.905*	1.037*	0.633*	-0.305*
Public	buy	0.373*	0.276*	0.195*	1.048*	0.529*	1.575*	0.734*
Bank	sell	0.814*	1.899*	-0.075*	0.022*	0.855*	-0.559*	0.596*
TNB	buy	-	0.680*	0.991*	-0.913*	0.943*	-0.515*	-0.874*
	sell	-	0.494*	-0.318*	-	0.900*	0.239*	-0.034*
Jaya	buy	0.446*	0.190*	2.156	2.902	-1.335*	0.895*	-2.079*
Tiasa	sell	-0.281*	1.330*	4.214*	-3.087	2.891*	-2.451	-1.723

Traders and investors would definitely lose out if they were to follow the signals generated by the RSI. In addition to the insignificance of the *z*-statistic values, the

signs of the values are also inaccurate. Thus, this has violated the statement made early denoting that if a buy signal is generated a positive return is expected and vice versa for the sell signal. The percentage of significant values for this analysis is only 8.84%, which is extremely low. Therefore, the RSI may not be used to forecast short term price movements.

The analysis is then expanded using the 21-day RSI to forecast 2 days ahead. The result is found in Table 7. When the forecast of 2 days ahead is carried out using the 21-day RSI trading rule, the result is slightly worse than the previous analysis, where the forecast is carried out using the 14-day RSI. Only a handful of the value is found to be significant, denoting that the signs generated are not significant at the significant level of 5%. With comparison to the previous analysis, there are a number of conditions where signals are not generated.

In Period 4, which is during the Asian Economic Crisis Period, only 5 sell signals are generated out of the grand total of 11. This is an illogical situation as during periods of recession, the prices tend to fall abruptly. However, the 21-day RSI did not give an indication to the traders and investors of this sudden fall, leading to future loses. With a longer look back period, the number with significant values decreases. The percentage of significant value is only 6.2%.

Table 7: t-statistic values for 2 days forecast using 21-day RSI

		Period 1	Period 2	Period 3	Period 4	Period 5	Period 6	Period 7
KLCI	buy	-0.683*	0.706*	-0.187*	0.332*	1.126*	-0.529*	-
	sell	0.847*	1.657*	-0.480*	-	1.365*	-0.493*	0.392*
BAT	buy	-0.170*	0.251*	1.785*	0.165*	1.675*	-	0.097*
	sell	2.553	0.108*	-0.019*	-2.332	-0.104*	-	-0.889*
Gamuda	buy	-	-	-0.124*	-0.206*	0.539*	-0.067*	0.800*
	sell	-	0.428*	-1.019*	-0.977*	0.436*	-1.236*	-0.675*
IGBS	buy	-0.151*	-	3.856	0.000*	0.473*	0.000*	0.230*
	sell	0.419*	-0.802*	0.367*	-	1.133*	-2.692	-0.032*
IOI	buy	-0.197*	0.317*	0.071*	1.438*	1.941	0.065*	1.981*
	sell	0.579*	1.117*	-0.380*	-1.545*	-2.421	0.245*	0.000*
MAS	buy	-0.248*	0.683*	-	-0.724*	-	1.667*	-0.978*
	sell	0.289*	0.410*	-	-1.666*	-0.129*	0.917*	-0.166*
Maybank	buy	0.078*	0.355*	2.140	-0.348*	-	-0.662*	0.629*
	sell	0.325*	-0.643*	-1.542*	-	0.986*	-	-0.822*
Proton	buy	-	0.168*	-	-0.861*	0.000*	-	-0.149*
	sell	-	-0.077*	-0.835*	-	1.001*	0.271*	-0.338*
Public Bank	buy	0.396*	-	0.287*	2.353	1.345*	0.385*	0.497*
	sell	-0.078*	1.579*	-0.066*	-0.643*	0.566*	-0.916*	0.349*
TNB	buy	-	0.228*	1.132*	-0.087*	0.861*	-0.376*	-0.506*
	sell	-	0.245*	-2.311	-	0.711*	0.100*	0.469*
Jaya Tiasa	buy	-0.478*	0.033*	0.859*	0.319*	-0.768*	0.595*	-0.221*
	sell	0.352*	1.016*	1.534*	-	0.207*	-0.299*	0.099*

As the analysis is further carried out, the 100-day RSI is used to make forecast for 2 days ahead. The result of the analysis is shown in Table 8. For this analysis, only a handful of signals are generated. This is very much substandard compared to the previous 2 analyses. No signals generated would indicate that there are just no tips or clue as to what a trader or investor should do. From the table, no signals were

generated during the economic crisis period and also the 1st Recovery Period. In other words, traders would not be able to do anything if they were to base their investments on the prediction from the RSI.

Table 8: t-statistic values for 2 days forecast using 100-day RSI

		Period 1	Period 2	Period 3	Period 4	Period 5	Period 6	Period 7
KLCI	buy	-	-	-	-	-	-	-
	sell	-	0.488*	-	-	-	-	-
BAT	buy	-	-	-	-	-	-	-
	sell	1.169*	-	-	-	-	-	-
Gamuda	buy	-	-	-	-	-	-	-
	sell	-	-	-	-	-	-	-
IGBS	buy	-	-	-	-	-	-	-
	sell	0.709*	-	-	-	-	-	-
IOI	buy	-	-	-	-	-	-	-
	sell	-	-	-	-	-	-	-
MAS	buy	-	-	-	-	-	-	-
	sell	-	-	-	-	-	-	-
Maybank	buy	-	-	-	-	-	-	-
	sell	-	-	-	-	-	-	-
Proton	buy	-	-	-	-	-	-	-
	sell	-	-	-	-	-	-	-
Public Bank	buy	-	-	-	-	-	-	-
	sell	-	0.259*	0.154*	-	-	-	0.547*
TNB	buy	-	-	-	-	-	-	-
	sell	-	-1.033*	-	-	-	-	-
Jaya Tiasa	buy	-	-	-	-	-	-	-
	sell	-	0.230*	-2.121	-	-	-	-

.

With reference to these 3 analyses, it would be logic and accurate to conclude that the RSI is not suitable in making very short period forecasting as the signals generated are not significant. The number of signals generated decreases as the number of look back days increases. It does not matter what trading rule is used. What matters is that the results for 2 days forecast are not significant at the significant level of 5%.

It would be wise to check if the performance of the RSI would improve when a longer forecast period is used. Therefore, the analysis is carried out on the forecast for 21 days in advance, using again all the different trading rules available; 14-day RSI, 21-day RSI and 100-day RSI. Table 12 shows the result for the forecast of 21 days ahead using the 14-day RSI. In this case, the results have been very promising compared to the previous 2 forecasts.

From Table 9, it is found that more than more than half of the values are significant at the significant level of 5%, indicating that the signals generated are very much significant. Despite the significance in other periods, the RSI is still found to be less significant during the Asian Economic Crisis. Traders and investors will still face the same dilemma earlier. The only difference would be that there are signals generated. However, the dilemma would not be as big as before as traders and investors could take wiser moves and make smart and wise decisions

as to whether the investment should be made or not. From this analysis, the role of RSI as a forecasting tool seems promising.

Table 9: t-statistic values for 21 days forecast using 14-day RSI

		Period 1	Period 2	Period 3	Period 4	Period 5	Period 6	Period 7
KLCI	buy	-1.354*	2.051	1.732*	-0.943*	3.559	-2.726*	-
	sell	-3.458	-4.124	-3.668	-0.759*	-3.762	-1.160*	-1.783*
BAT	buy	0.075*	2.209	3.447	2.542	2.953	1.357*	2.579
	sell	-5.480	-4.433	-4.206	-2.162	-2.713	-1.570*	-3.550
Gamuda	buy	-	-	0.126*	0.087*	2.106	0.212*	3.117
	sell	-	-5.441	-2.160	-0.790*	-7.104	-2.825	-3.680
IGBS	buy	0.612*	3.536	4.239	-0.028*	1.441*	0.326*	2.926
	sell	-2.804	-4.536	-2.936	-1.912*	-4.430	-3.210	-4.203
IOI	buy	1.094*	1.611*	2.613	2.142	1.632*	0.444*	2.323
	sell	-3.484	-6.589	-3.237	-1.495*	-5.017	-4.650	-3.799
MAS	buy	1.201*	1.325*	4.108	-1.127*	0.992*	2.307	-2.629
	sell	-3.783	-3.691	-2.392	-1.697*	-3.710	-2.551	-0.354*
Maybank	buy	-0.989*	0.293*	2.260	0.853*	-	-1.222*	3.856
	sell	-1.635*	-2.596	-4.519	-1.435*	-5.767	-0.800*	-3.809
Proton	buy	-	-0.933*	2.785	1.211*	1.255*	5.110	-0.259*
	sell	-	-6.044	-4.080	-1.292*	-3.445	-5.871	-1.328*
Public Bank	buy	1.973	1.667*	2.244	2.645	4.229	3.390	0.969*
	sell	-4.054	-7.837	-3.806	-1.536*	-5.150	-3.640	-2.122
TNB	buy	-	1.568*	1.475*	-0.249*	1.978	0.389*	0.171*
	sell	-	-3.481	-3.235	-	-6.031	-2.557	-2.489
Jaya Tiasa	buy	1.356*	1.203*	-0.494*	3.924	-1.340*	-0.324*	0.813*
	sell	-3.649	-6.043	-2.843	-0.649*	-1.916*	-2.202	-3.613

The analysis is persisted using the 21-day MA. The result is displayed in Table 10. Comparing relatively to the analysis using the 14-day RSI, less signals are generated for the 21-day RSI. However, the number of significant values at the significant level of 5% is still roughly the same. With similarity to the previous analysis, period 4, which is the Asian Economic Crisis also, has less significant values compared to the other periods. Again, this indicates that traders have to be extra careful when making investment during this period. Unwanted losses can be generated during this period.

In the earlier analyses, it is found that the 100-day RSI does not generate many signals. This has been true for both 2 days forecast and 5 days forecast. Signals that are not generated can cause uncertainties in the traders and investors decision making. Table 14 shows the result of the 21 days forecast using the 100-day RSI.

From the results shown in Table 11, it is again found that there are no signals generated for the 3 periods; Period 4, Period 5 and Period 6. Although only one value is not significant at the significant level of 5% out all the signals generated, but not being able to generate a signal would simply mean not able to tell a trader or investor what is the action they should take. This would mean that the traders would not be able to make any investment during these periods.

Table 10: t-statistic values for 21 days forecast using 21-day RSI

		Period 1	Period 2	Period 3	Period 4	Period 5	Period 6	Period 7
KLCI	buy	-0.529*	1.963	2.896	2.091	3.631	2.079	-
	sell	-3.026*	-3.127	-2.163	-	-5.600	-2.621	-2.794
BAT	buy	-0.243*	0.811*	1.992	0.165*	2.187	-	0.780*
	sell	-4.195	-4.223	-2.825	-2.332	-1.701*	-	-4.429
Gamuda	buy	-	-	0.119*	1.785*	1.347	0.107*	2.520
	sell	-	-5.106	-2.000	-1.286*	-6.354	-2.556	-2.581
IGBS	buy	0.899*	-	3.250	-0.082*	1.262*	0.305*	2.001
	sell	-2.454	-4.481	-0.770*	-	-3.493	-2.507	-3.228
IOI	buy	1.033*	1.175*	1.645*	3.516	2.334	0.390*	3.550
	sell	-2.634	-5.388	-3.529	-1.739*	-3.127	-4.043	-5.273
MAS	buy	3.038	1.992	-	-0.804	-	2.225	-1.681*
	sell	-4.717	-3.974	-	-1.022*	-3.472	-2.590	-0.211*
Maybank	buy	-0.209*	0.255*	2.596	0.144*	-	0.679*	1.020*
	sell	-1.392*	-2.128	-4.015	-	-5.795	-	-4.457
Proton	buy	-	1.340*	-	1.171*	1.706*	-	3.869
	sell	-	-4.136	-3.974	-	-3.043	-6.190	-2.028
Public Bank	buy	2.932	-	1.199*	2.708	2.809	1.405*	0.542*
	sell	-4.472	-7.366	-2.110	-1.319*	-4.738	-2.830	-1.717*
TNB	buy	-	0.449*	2.279	-0.219*	2.134	0.178*	1.235*
	sell	-	-3.303	-3.431	-	-4.925	-1.055*	-3.100
Jaya Tiasa	buy	-0.080*	0.224*	0.500*	2.016	-1.454*	0.567*	1.468*
	sell	-1.830*	-6.164	-3.047	-	-1.457*	-1.913*	-3.054

Table 11: t-statistic values for 21 days forecast using 100-day RSI

		Period 1	Period 2	Period 3	Period 4	Period 5	Period 6	Period 7
KLCI	buy	-	-	-	-	-	-	-
	sell	-	-2.415	-	-	-	-	-
BAT	buy	-	-	-	-	-	-	-
	sell	-4.225	-	-	-	-	-	-
Gamuda	buy	-	-	-	-	-	-	-
	sell	-	-	-	-	-	-	-
IGBS	buy	-	-	-	-	-	-	-
	sell	-3.532	-	-	-	-	-	-
IOI	buy	-	-	-	-	-	-	-
	sell	-	-	-	-	-	-	-
MAS	buy	-	-	-	-	-	-	-
	sell	-	-	-	-	-	-	-
Maybank	buy	-	-	-	-	-	-	-
	sell	-	-	-	-	-	-	-
Proton	buy	-	-	-	-	-	-	-
	sell	-	-	-	-	-	-	-
Public Bank	buy	-	-	-	-	-	-	-
	sell	-	-2.261	-1.701*	-	-	-	4.685
TNB	buy	-	-	-	-	-	-	-
	sell	-	-25.820	-	-	-	-	-
Jaya Tiasa	buy	-	-	-	-	-	-	-
	sell	-	-2.601	-2.598	-	-	-	-

5 Conclusion

Using the simple it is found that both indicators have shown a very low percentage of the forecast being accurate. This result has come as a shock as MA and RSI are the 2 most widely used technical indicators in the stock and commodity market. However, further statistical analysis has to be carried out in order to ensure the significance of this statement.

Statistical analyses are then brought into the picture of this research. Statistical test has been used to analysis the significance of the technical indicators. The signals generated by the MA have now been found to be significant, but only for the forecast of short term periods; 2 days. The forecast for 21 days are less significant compared to the shorter period but the signals generated can still be used with special cautions. However, the MA is found to be not significant for the forecast of 100 days ahead. The MA is also found to be less significant during the economic crisis periods; Asian Economic Crisis and World Economic Crisis. Thus, it can be concluded that the MA is suitable to forecast for short term periods but not for long term periods.

As for the RSI, different trading rules are being applied; 14-day RSI, 21-day RSI and 100-day RSI. The main difference between the 3 trading rules is that 14-day RSI would have more signals generated whereas 100-day RSI would have the least signals generated. The signals generated by the RSI are found to be significant for the forecast of longer periods, which is 21 days ahead for all the 3 trading rules. For shorter periods, such as 2 days ahead, the signals are found to be not significant. From the analysis, the trading rule is not the main factor affecting the significance of the signals generated. It is the number of days of forecast that makes the difference. Thus, it is concluded that RSI is suitable for the forecast of longer periods.

Therefore, base on this research, investors should be knowledgeable enough to understand the concept behind the MA and RSI. Investors and traders should be caution when making investments based on the signals generated. Different forecasting length and different economic periods have been proved to be a crucial point when making forecast of the future price trends. High profit investments can be made if investors were to analyze the market carefully before applying the MA and RSI.

References

[1] Atmeh, M. A. Dobbs, I. M (2001). Technical analysis and the stochastic properties of the Jordanian Stock Market Index Return. Working Paper, Hashemite University.

[2] Oh, E. S. Oh, E. F. Chai, M. [2004] KLSE and DJIA at the threshold of the Iraq war: a comparison and contrast. Proceedings of the Malaysian Finance Association 6th Annual Symposium, Langkawi, 274-281.

[3] Pampana, C. Sahu, R. (2005) Profitability of Technical Trading Rules in Indian Stock Market: Empirical Evidence. *Proceedings of the International Conference in Economics and Finance (ICEF).* 231 - 235

[4] Wong, W. K. Manzur, M. and Chew, B. K. (2002). How Rewarding Is Technical Analysis? Evidence Form Singapore Stock Market. *Working Paper* No. 0216, Department of Economics, National University of Singapore.

Appendix A: Details of Authors

Seow Chiao Ju: BSc. student, School of Mathematical Sciences, Universiti Sains Malaysia, 11800 Minden, Penang, Malaysia. (Graduating Aug. 2005)
E-mail: chiaoju82@yahoo.co.uk

Tai Lee Meng: BSc. student, School of Mathematical Sciences, Universiti Sains Malaysia, 11800 Minden, Penang, Malaysia. (Graduating Aug. 2005)
E-mail: taileemeng@yahoo.com

Zainudin Arsad: School of Mathematical Sciences, Universiti Sains Malaysia, 11800 Minden, Penang, Malaysia. Fax: 00-60-4-657-0910
E-mail: zainudin@cs.usm.my

On Martingale Valuation of Surplus Process with Safety Function

Adhitya Ronnie Effendie

Dept. Mathematics FMIPA UGM Yogyakarta

Abstract: Using martingale valuation method, we developed a new insurance process model. Furthermore we also add safety function to keep companies' asset for being loss. In this method we used properties of numeraire invariance and no-arbitrage principle and these will be applied into classical Lundberg claim process with special appendix, a safety function.

Keywords: Compound Poisson Process, numeraire, Risk Theory, Martingale valuation

REGRESSION TREES FOR COMPETING RISKS SURVIVAL DATA

Abdul Kudus[a], Noor Akma Ibrahim[b], Mohd. Rizam Abu Bakar[b], Isa Daud[c]

[a] PhD Student, Institute for Mathematical Research, Universiti Putra Malaysia
[b] Institute for Mathematical Research and Dept. of Math., Universiti Putra Malaysia
[c] Dept. of Math., Universiti Putra Malaysia

Abstract. Regression tree method is extended to competing risks survival data. This extension inherits most of the optimal aspects of the classification and regression trees (CART) proposed by Breiman et al. (1]. Rather than fitting single smooth model into data, the method employs to fit a piecewise model that accounts for local variations. Piecewise model can be viewed in the fashion of tree diagram, where the threshold points are considered as cutpoints. It is shown that trees are useful not only in summarizing information in covariates, but also in detecting treatment-covariates interactions. On the basis of the survival model for competing risks, we formulate the approach of tree methodology. The application of this method to contraceptive discontinuation data is presented.
Key-words: Competing risks, Regression Trees, Survival analysis.

1 Introduction

The problem of the determination of functional relations between dependent variable(s) and independent variable(s) is commonly addressed by the regression analysis. Beginning with the simple linear regression up to generalized linear regression, basically we fit one smooth equation for data. This method had the shortcoming, namely one smooth equation that we fit possibly was not so fit for sub-data distributed on predictor space. In other words, the one-equation-method doesn't take into account the local variations on predictor space.

Regression trees as in the case of piecewise regression fit the model by taking into account the local variations in predictor space through recursive partitioning. But, those methods are different in the spirit of threshold determination. Regression trees determined the threshold by outcome-oriented method. By this means we hoped that model will be better in explaining the relationship between dependent and independent variable which is reflected with the small error.

The regression trees procedure works on sub space of predictor variable, and it is really computer intensive. The procedure works to find sub predictor space which the relation between independent and dependent variable was different with another sub predictor space. This was an extra work that must be carried out in order to get the model that as well as possible in explaining the relations between independent and dependent variable. Fortunately, the advancement of computer technology was very supportive this requirement.

Extension of tree techniques to competing risks data is motivated by classification issue associated with multiple endpoint survival data. For example, clinical investigators design study to form prognostic rules. Credit risk analysts collect account information to build up credit scoring criteria. Frequently, in such studies

the outcomes of ultimate interest are competing causes of event, such as relapse or death, late payments or default. Since tree based methods recursively partition the predictor space, they provide a descriptive method for exploratory data analysis.

Recursive partitioning seems to provide a promising means for developing meaningful classification rules for survival data. Considerable efforts have been dedicated to the single endpoint case in the literature. The direct use of the CART procedure by defining appropriate error terms (see, for example [2], [3], [4]), was the major modifications made to CART by many researchers in an effort to overcome the difficulties naturally associated with failure times. The properties of survival data also emerge the idea to construct the tree based on maximizing the between-node separation [5], [6]. While the extension of regression tree to multivariate data problem, such as longitudinal [7], multiple binary [8], [9] and multivariate normal [10], still use classical CART directly, the extension for multivariate survival data adopted the between node separation approach. Multivariate survival regression trees based on frailty model used splitting rule of maximum function of integrated log likelihood [11] and Wald statistic [12] are the recent research on this area. A well-designed pruning algorithm analogous to the CART pruning algorithm was developed for trees grown by between node separation by LeBlanc and Crowley [13]. Growing trees by between node separation brought more flexibility and enlarged the scope of CART from an applied perspective. Splitting data via a measure of difference between two groups remains an accessible tool for allowing researchers to enjoy all the favorable properties of tree techniques, especially when error terms are hard to define.

This paper is concerned with the generalization of tree-based models to competing risks survival data. In attempts to facilitate an extension of tree-based methods to competing risks survival data, more difficulties need to be circumvented either empirically or theoretically.

The remainder of the paper is organized as follows. Section 2 describes the competing risks regression model and its associated likelihood function. Section 3 presents the proposed tree method in detail. Section 3.1 discusses the splitting statistic. In Section 3.2 a likelihood-based pruning procedure is developed followed by the strategy for tree size selection. Section 4 provides an illustration using the discontinuation contraceptive method data. Other related issues are discussed in Section 5.

2 The Competing Risks Regression Model

Consider a typical setting for competing risks survival data, where the data under study consist of n independent observations $\{ (T_i, \delta_i, \boldsymbol{X}_i), i=1,2,\ldots,n\}$. We assume that there are K potential failure times, Y_1, Y_2, …, Y_K, one for each risk. We observes $T = \min(Y_1, Y_2, \ldots, Y_K)$ and a variable $\delta = j$ if $T=Y_j$ for $j = 1, 2, \ldots, K$, that tells the investigator which of the risks caused the event to occur. In this paper, we assumed all K risks are independent. Suppose that we have p covariates X_1, X_2, …, X_p, which could be mixture of continuous, ordinal and categorical variables.

The competing risks probabilities can be summarized by one of two parameters, the crude (or cause-specific) hazard rate or the cumulative incidence function. The crude hazard rate for cause j is defined by

$$\lambda_j(t) = \lim_{\Delta \to 0} \frac{P(t \le T < t + \Delta, \delta = j \mid T \ge t)}{\Delta} \tag{1}$$

This is the rate of occurrence of the *j*th failure cause in the presence of all causes of failure. In the contraceptive discontinuation example, the crude-abandoning hazard rate is the rate at which women with contraceptive use are abandoning the method.

The second parameter is the cumulative incidence function for the *j*th competing risks (see [14]). This function is defined as the probability of the subject failing from cause *j* in the presence of all the competing risks. It is a function of all *K* of the crude hazard rates. The cumulative incidence function is defined by

$$F_j(t) = P(T \le t, \delta = j) = \int_0^t h_j(u) \exp\left\{ -\int_0^u \sum_{i=1}^{K} h_i(v) dv \right\} du \tag{2}$$

The cumulative incidence function is a subdistribution function, that is a nondecreasing function of time with $F_j(\infty) = P(\delta=j)$. In the contraceptive discontinuation example, the abandoning cumulative incidence function is the probability of a woman (who could abandon, switch or failure) abandoning prior to time *t*.

The competing risks regression can be set up on crude hazard rates or cumulative incidence function. In the crude hazard modeling, we regress the crude hazard in (1) through proportional hazards model [15], [16], [17]. This model assumes that the Y_js are independent across event types. We use this model for further analysis in this paper.

Consider now inference on the relationship between crude hazard rate and regression vectors or function $\boldsymbol{X}$. Proportional hazards modeling in which the cause-specific hazard function at time *t* depends on $\boldsymbol{X}$ gives

$$\lambda_j(t, \boldsymbol{X}) = \lambda_{0j}(t) \exp(\boldsymbol{X}\boldsymbol{\beta}_j), \quad j = 1,2,..., K \tag{3}$$

Note that both λ_{0j} and the regression coefficient $\boldsymbol{\beta}_j$ have been permitted to vary arbitrarily over the *K* failure types.

Let we observe $t_1, t_2, \ldots, t_n$ with type of risks $\delta_1, \delta_2, \ldots, \delta_n \in \{0, 1, 2, \ldots, K\}$. If $I_l(t) = I(t_l \ge t)$, where *I* is indicator function, then the method of partial likelihood gives

$$L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \mathrm{K}\ , \boldsymbol{\beta}_K) = \prod_{i=1}^{n} \prod_{j=1}^{K} \left[\frac{\exp(\boldsymbol{X}_i \boldsymbol{\beta}_j)}{\sum_{l=1}^{n} I_l(t_i) \exp(\boldsymbol{X}_i \boldsymbol{\beta}_j)} \right]^{I(\delta_i = j)} \tag{4}$$

The estimation of the $\boldsymbol{\beta}_j$'s can be conducted by applying maximum likelihood method of (4), which reveals *K* independent estimating equations. This result shows that competing risks regression of (3) can be regarded as *K* independent regressions, one for each risk, with failure of type other than *j* is regarded as censored for *j*th estimating equation.

The competing risks regression based on cumulative incidence function is motivated by the fact that sometimes the estimates of crude hazard regression do not agree with impression drawn from plots of cumulative incidence function for each level of a risk factor (18]. Hence, it is reasonable to construct a regression model based on cumulative incidence function directly (18], [19], [20].

3 Regression Trees for Competing Risks Data

We look for a model that fits the data well in the form of *classification*, i.e. such its crude hazard function is represented by

$$\lambda_j(t, s, \boldsymbol{X}) = \lambda_{0sj}(t)\exp\left(\beta_1 I_1(x) + \beta_2 I_2(x) + \mathrm{K} + \beta_q I_q(x)\right), \quad j = 1,2,\ldots,K \qquad (5)$$

where $\lambda_{0sj}(t)$ represents the baseline hazard function of cause j for the individuals in stratum s; the Is denote 'dummy' variables, function of the predictor x, indicating membership to one of the classes other than the reference class [$I_k(z)$ = 1 if the individual with predictor vector z is in class k and =0 otherwise]. The Is will be henceforth referred to as *class indicator variables*. The βs are the log-relative hazards in comparing the specific class of interest to the reference class.

Thus equation (5) represents q + 1 distinct classes: the reference class and q additional classes, the kth class being characterized by a hazard ratio, w.r.t. the reference class, constant in time and given by $\exp(\beta_k)$. Clearly this is simply a stratified Cox regression model with regressor given by class indicator variables.

The idea of CART is adopted in the proposed tree procedure. Its consist of three steps: the splitting rule for growing large tree method by partitioning the data recursively, a method to prune the large tree into a subtree sequence, and the right-sized tree determination.

3.1 The Splitting Rule

For each node the following crude hazard model with single dummy covariate is fitted:

$$\lambda_{ij}(t_i, s, \boldsymbol{X}_i) = \lambda_{0isj}(t)\exp\left(\beta \cdot I(x_i \le c)\right), \quad i = 1,2,\ldots,n; \;\; j = 1,2,\ldots,K \qquad (6)$$

Indicator $I(x_i \le c)$ corresponds to split point c on continuous covariate x. If the covariate is categorical, the indicator function for splitting should be $I(x_i \in A)$ for any subset A of its categories need to be considered. The β is the log-relative hazards between two sibling nodes.

Let $LRS(c)$ denotes the likelihood ratio statistic of model (6) for risk j which compares this model to the "null" model β=0 corresponding to cutpoint c. The best cutpoint c^* is defined as the cutpoint corresponding to the largest likelihood ratio statistic, namely $LRS(c^*) = \max_{c \in C} LRS(c)$, where C denote the set of all permissible cutpoints.

The data are then partitioned into two daughter nodes according to the best cutpoint c^* and the same splitting procedure is applied to either daughter node to split further. This procedure is repeated till a minimum node size restriction or a

minimum event number of type j is reached. The splitting procedure also naturally stops when the node become pure in the sense that all the responses share exactly the same covariate values. This procedure leads to a large initial tree T_0.

3.2 The Pruning of Large Tree and the Right-size Tree

The *information content* of a tree is defined as the sum of partial *LRS* of its terminal nodes. To any tree of size q + 1 we associate a model given by equation (5) where the classes are its q + 1 terminal nodes.

Given any two nested trees, the partial *LRS* comparing the largest to the smallest is a reasonable measure of the information addition corresponding to the pruning of the branch issuing from its internal node. Therefore, we define the *information content* of a tree as the partial *LRS* comparing the tree to the trivial tree consisting of the root node alone. The *information loss* of a tree with respect to the largest tree is simply the difference of the information contents.

Model choice by minimum AIC will yield the right-size tree. The AIC is defined by

$$AIC^{(T)} = -2\left\{\hat{l}^{(T)} - \#\ parameter\right\} \tag{7}$$

where $\hat{l}^{(T)}$ is the log-likelihood of model (5) which is for tree T with q+1 terminal nodes.

4 Example: Contraceptive discontinuation data

To illustrate our approach we consider a sample of 6492 women drawn from the database of the Indonesian Demography and Health Survey (IDHS) 2002. This is the national retrospective database consisting of data on time to contraceptive discontinuation. All subjects are investigated on the history of last episode of contraceptive discontinuation. We observed the length of time of the last contraception use, and we focus on three types of discontinuation in a "competing risks" framework. The outcomes we consider are failure, contraceptive abandonment while in need of family planning, and switching to another contraceptive method. A discontinuation is defined as a contraceptive failure if the woman reported that she became pregnant while using the method. Thus, this definition includes both failures of the method itself and failure owing to incorrect or inconsistent use of the method. Adoption of different method within one month of discontinuation is classified as a method switch, whereas continuation of nonuse for one month or more is classified as contraceptive abandonment. Clearly, contraceptive failure is of interest because it leads directly to an unintended pregnancy. Contraceptive abandonment is also important outcome to study because it leads to immediate risk of unintended pregnancy. Method switching also may lead to an increased risk of unintended pregnancy if use of a modern method is discontinued in favor of a less effective, traditional method. Contraceptive failure is somewhat different from the other two outcomes in that it presumably is an unintentional event, whereas contraceptive abandonment and switching suggest some decision-making and choice on the part of the woman.

We consider some covariates which suppose to be able to explain the rate of discontinuation. The important one is the contraceptive method. For this

analysis, contraceptive methods were grouped into three categories: pills and injectables, IUDs and implants, and other modern methods (mainly condoms). Traditional methods and sterilization were excluded from this study. Pills and injectables were grouped together because they are both short-term hormonal methods. IUDs and implants are longer-term reversible methods that require a health worker to remove them. As such, they are fundamentally different from other reversible methods in that they require the user to be proactive to discontinue use and to have contact with the health system at the time of discontinuation.

The other covariates are woman's education (primary or lower, secondary, university), household economic status (1 – 7 scores), area of residence (urban, rural), age of the women at the start of the episode of use (years), and religion (Islam, non-Islam).

Table 1. Descriptive statistics for variables considered in the model for contraceptive discontinuation

	Percentage of Women	Number of Women	Means	Stdev
Time to discontinuation (month)				
- failure	3.28	213	27.98	21.71
- abandon	34.10	2214	33.05	21.89
- switch	39.20	2545	23.81	20.44
- censored	23.42	1520		
Age at start (year)			27.93	6.93
Socioeconomic scores			2.29	1.73
Residence				
- Rural (0)	55.3	3588		
- Urban (1)	44.7	2904		
Religion				
- Islam (0)	87.8	5701		
- Non-Islam (1)	12.2	791		
Education				
- ≤ Primary (0)	18.3	1187		
- Secondary (1)	54.5	3537		
- University (2)	27.2	1768		
Method				
- Pill/Injectables (0)	81.8	5309		
- IUD/implants (1)	16.3	1058		
- Others (2)	1.9	125		

Result of Kalbfleisch and Prentice [15] is presented for comparison (table 2). Analysis for discontinuation due to failure shows that *Age*, *university education level* and *IUD/implants methods* are statistically significant at level less than 1%.

The regression coefficient shows that the older have a lower risk to experience contraceptive failure, the university-educated-women have the higher risk of contraceptive failure and women with IUD/implants have lower risk than the other methods.

Table 2. Analysis of time to failure of contraceptive methods with crude proportional hazards model

Variable	Coefficients	Standard Error	Wald Chi-Square	Pr > Chi-Square
Age	-0.056953	0.01136	25.15011	0.0001
Socioeco	-0.116885	0.04943	5.59254	0.0180
Residence	0.218572	0.15492	1.99052	0.1583
Religion	0.295574	0.20539	2.07107	0.1501
Secondary	0.094530	0.20494	0.21277	0.6446
University	0.768049	0.23313	10.85352	0.0010
IUD/implant	-1.544351	0.28897	28.56099	0.0001
Other meth's	0.532029	0.51397	1.07150	0.3006

Next, we apply the proposed method to the data set. Unlike the crude ordinary hazard proportional model which concern the quantity of covariate effects on time to discontinuation, our motivation question is "who is more likely to develop contraceptive discontinuation?". We are especially interested in identifying groups of women that have similar characteristics.

We set the minimum node size to 400 observations or 20 failures. Which one is reached first became the stopping rule. Originally we obtained a tree with 7 terminal nodes. It is not surprising to see the first cut on *Age* followed by The *University* covariate as they are also the significant covariate on Table 2.

Through the *LRS* strategy, a sequence of nested trees is obtained. The right-size tree obtained by AIC minimum criteria consists of 3 terminal nodes refers to 3 groups of women. The first group is women whose age younger than 31.4 years with university education level. The second group is younger-than-31.4 years-women with secondary or lower education level. The third group is the older-than-31.4-years –women.

By comparing result of table 2 and figure 1, we see that *Age* and *University* are the significant covariates on both results. Even though, the *IUD/implant*s is not the significant covariate on tree diagram, but the tree diagram shows the easier and more interpretable result. The three terminal groups are sorted in descending order of hazard of failure as we go from left to right. So, the younger women with university education level are the poorest group in terms of time to contraceptive failure rate, followed by younger and secondary or lower education level women, and the best group is the older-than-31.4-years-old-women. This grouping is supported by the comparison of cumulative incidence function which show the probability of contraceptive failure among the three groups (see figure 2).

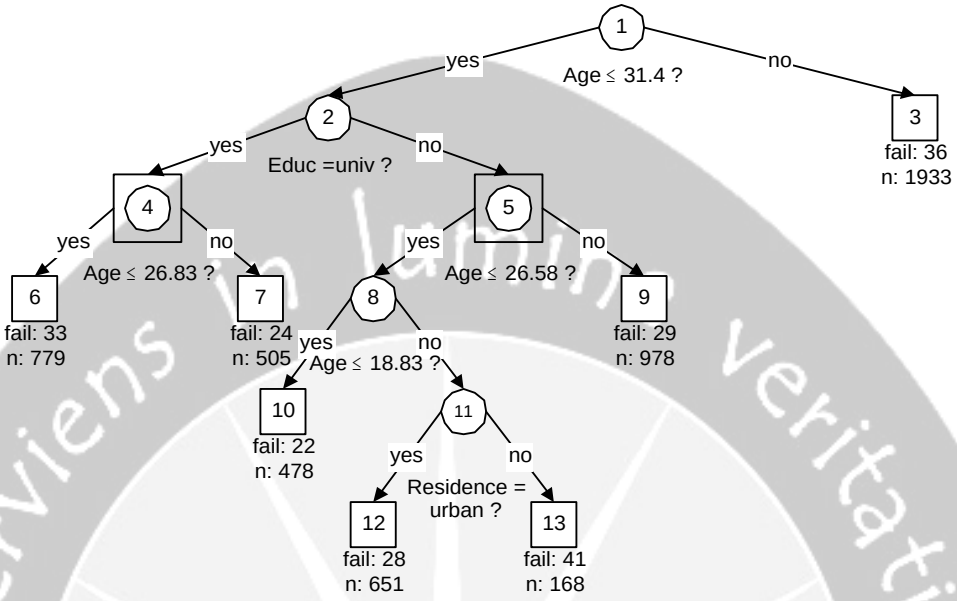


Figure 1. Tree diagram for time to contraceptive failure (the circle-inside-square node is node with pruned branch, *fail* and *n* denotes the number of failure and observations in a terminal node, respectively).

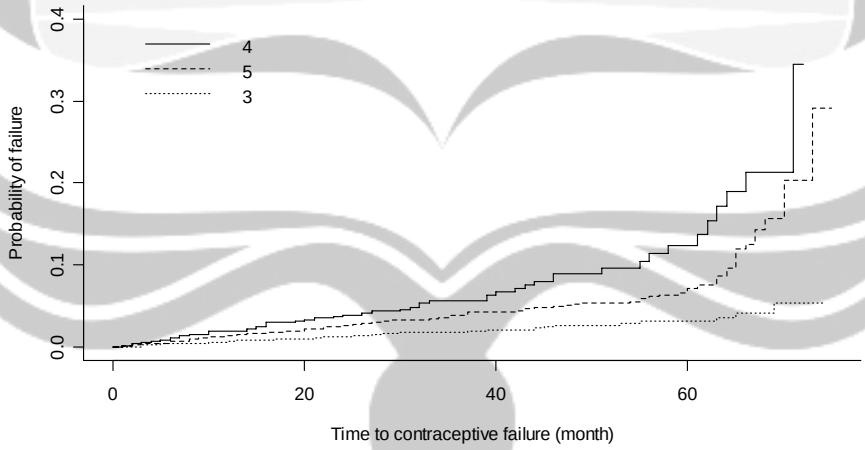


Figure 2. Comparison of Cumulative Incidence Function for 3 groups of women resulted by tree diagram.

We also carried out the same analysis for the two others risks (abandoning and switching). The tree diagram for these two risks enclosed in Appendix 1 and 2.

5 Conclusion

In brief, the crude hazard modeling through proportional hazards model and regression trees for competing risks survival data both appear useful in data exploration, and are somewhat complementary. Usually the ordinary regression approach focuses on the examination of large numbers of model with main (linear) covariate effects and some two factor interactions. Regression tree, on the other hand, is a clustering procedure which builds a tree from covariates. The ordinary regression model seems better suited than regression trees to treat covariate effects which act broadly across the whole data.

Regression tree is capable to uncover the unusual interaction effects that might be missed by the classical regression approach. In terms of classification, regression trees is more directly and simply oriented toward this.

The main emphasis of this paper has been on the flexibility of the tree-growing approach in the analysis competing risks data through crude hazard proportional model. In the example show that regression tree is very useful to discriminate the group of women in terms of contraceptive failure rate.

Acknowledgment

We are grateful to the conference committee for giving the opportunity to present this paper. This research was supported by the Ministry of Science, Technology and Innovation (MOSTI) of Malaysia grant.

References

[1] Breiman, L., Friedman, J., Olshen, R. & C. Stone (1993), *Classification and regression trees*, Chapman and Hall, New York.

[2] Davis, R. & J. Anderson (1989), Exponential survival trees, *Statistics in Medicine*, **8**, 947-962.

[3] Therneau, T., Grambsch, P. & T. Fleming (1990), Martingale based residuals for survival models, *Biometrika*, **77**, 147-160

[4] LeBlanc, M. & J. Crowley (1992), Relative risk trees for censored survival data, *Biometrics,* **48**, 411-425.

[5] Ciampi, A., Thiffault, J., Nakache, J-P. & B. Asselain (1986), Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates, *Computational Statistics & Data Analysis*, **4**, 185-204.

[6] Segal, M. R. (1988), Regression trees for censored data, *Biometrics*, **44**, 35-47.

[7] Segal, M. R. (1992), Tree-structured methods for longitudinal data, *Journal of the American Statistical Association*, **87**, 407-418.

[8] Zhang, H. P. (1998), Classification tree for multiple binary responses, *Journal of the American Statistical Association*, **93**, 180-193.

[9] Lee, S. K. (2003), A study on decision tree for multiple binary responses, *The Korean Communications in Statistics*, **10**, 971-980.

[10] Larsen, D. R. & P. L. Speckman (2004), Multivariate regression trees for analysis of abundance data, *Biometrics*, **60**, 543-549.

[11] Su, X. G. & J. J. Fan (2004), Multivariate survival trees: a maximum likelihood approach based on frailty models, *Biometrics*, **60**, 93-99.

[12] Gao, F., Manatunga, A. K. & S. Chen (2004), Identification of prognostic factors with multivariate survival data, *Computational Statistics & Data Analysis*, **45**, 813-824.

[13] LeBlanc, M. & J. Crowley (1993), Relative risk trees by goodness of split, *Journal of the American Statistical Association*, **88**, 457-467.

[14] Pepe, M. S. & M. Mori (1993), Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data?, *Statistics in Medicine*, **12**, 737-751.

[15] Kalbfleisch, J. D. & R. L. Prentice (1980), *The statistical analysis of failure time data*, John Wiley & Sons, New York.

[16] Lunn, M. & D. McNeil (1995), Applying Cox regression to competing risks, *Biometrics*, **51**, 524-532.

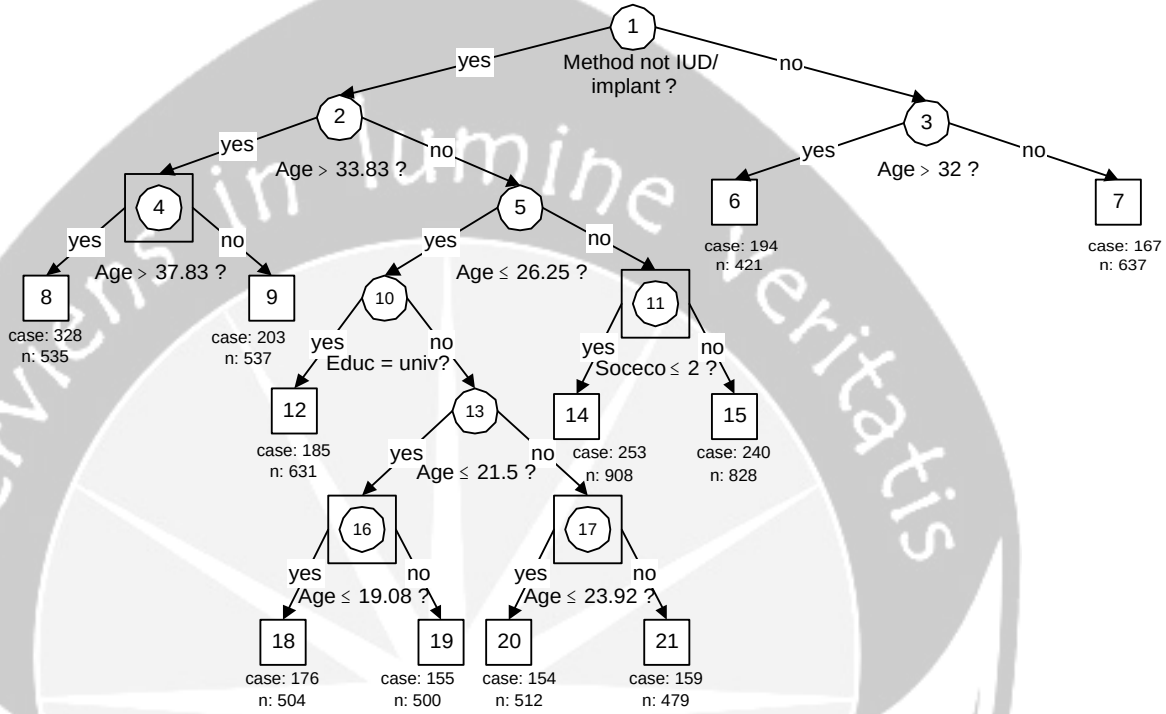
[17] Allison, P. D. (2001), *Survival analysis using the SAS® system: a practical guide*, SAS Publishing, Cary North Carolina.

[18] Klein, J. P. & P. K. Andersen (2005), Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function, *Biometrics*, **61**, 223-229.

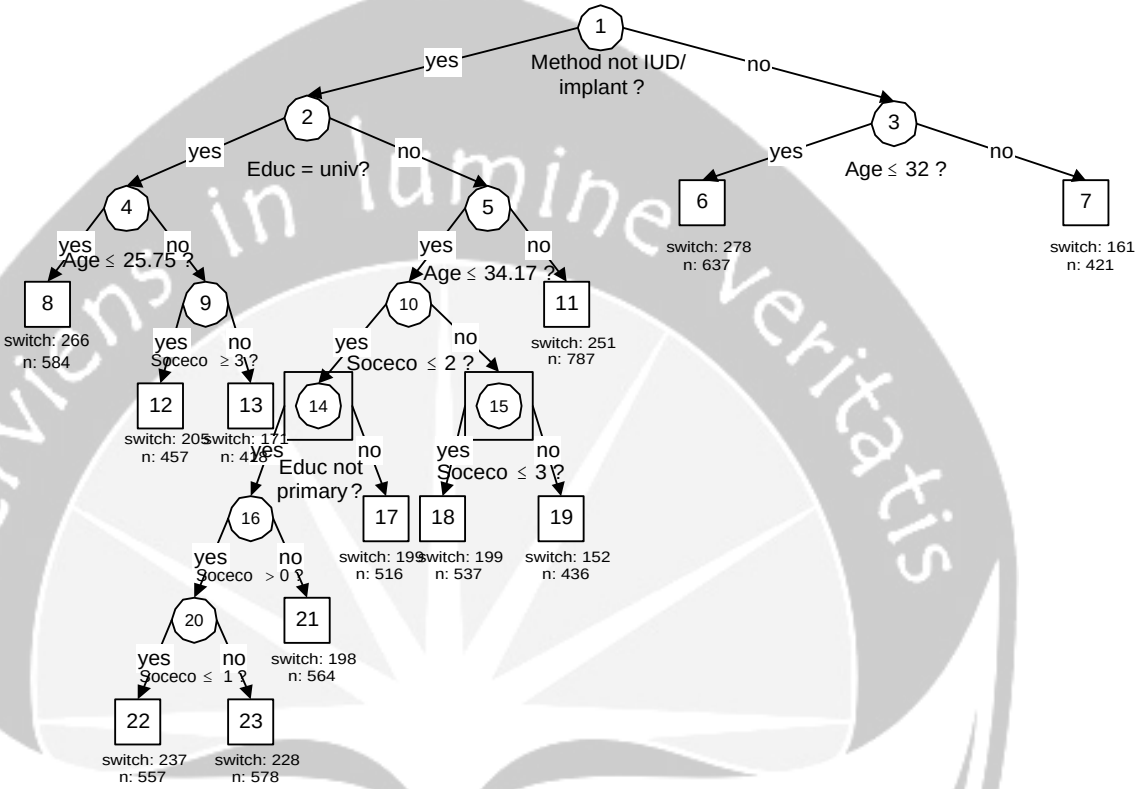
[19] Fine, J. P. & R. J. Gray (1999), A proportional hazards model for the distribution of a competing risk, *Journal of the American Statistical Association*, **94**, 496-509.

[20] Scheike, T. H. & M-J. Zhang (2005), Predicting cumulative incidence probability by direct binomial regression, Department of Biostatistics, University of Copenhagen, Denmark. Available from URL: http://pubhealth.ku.dk/bs/publikationer/rr-05-2.pdf/ [accessed June 15, 2005]

A Tree diagram for time to Abandonment



B Tree diagram for time to Switching



ABDUL KUDUS: PhD student at INSPEM University Putra Malaysia and Lecturer at Department of Statistics, Universitas Islam Bandung, Jl. Taman Sari 1 Bandung 40116, Indonesia. Phone: +62 +22 4203368
E-mail: akudus69@yahoo.com

NOOR AKMA IBRAHIM: Institute for Mathematical Research and Dept. of Mathematics, University Putra Malaysia, UPM 43400, Serdang Selangor DE, Malaysia
E-mail: nakma@putra.upm.edu.my

MOHD. RIZAM ABU BAKAR: Institute for Mathematical Research and Dept. of Mathematics, University Putra Malaysia, UPM 43400, Serdang Selangor DE, Malaysia

ISA DAUD: Dept. of Mathematics, University Putra Malaysia, UPM 43400, Serdang Selangor DE, Malaysia

BATCH PROCESSES, HOW TO MEASURE A PROCESS CAPABILITY WITH A BETTER WAY: A CASE STUDY AT SOFT DRINK FACTORY

I Nyoman Arcana

Widya Mandala Catholic University,
Surabaya, Indonesia.

Abstract. Many products are manufactured in batches such as paint, soft drinks, adhesives, etc. The composition of a product such as soft drinks might be quite uniform throughout the batch. Thus only one observed value of a particular quality characteristic can be obtained. In this case, the sample size used for process control is n = 1; that is, the sample consists of an individual unit. In such situation, the control chart for individual units (IX & MR charts) is useful. The individual measurements on the IX chart are assumed to be uncorrelated. However, dependent or correlated measurements are fairly common in continuous operations such as brewing soft drinks. This study intends to search for ways how both potential and performance capabilities are measured in such situation. The data consists of a brix degree of a soft drink. The data were analyzed using the following steps: calculating the autocorrelation coefficient, reducing the data in such a way so that they become alternate-data, doing the EWMA transformation to get its residue, making an individual control chart, and measuring both potential and performance capabilities. The result of the research shows that a combination of the methods of alternate points and EWMA prediction was very efficient in measuring both potential and performance capabilities. The specialty of the research finding is the ability of the methods used in reducing the false alarm signals.

Key-words: autocorrelation, potential capability, performance capability, alternate-points, EWMA

1 Introduction

Many products are manufactured in batches such as paint, soft drinks, adhesives, etc. The composition of a product such as soft drinks might be quite uniform throughout the batch. Thus only one observed value of a particular quality characteristic can be obtained. In this case, the sample size used for process control is n = 1; that is, the sample consists of an individual unit. In such situation, the control chart for individual units (IX & MR charts) is useful [4, 11]. The control procedure uses the moving range of to successive observation to estimate the process variability. The moving range is defined as $MR_i \mid X_i - X_{i-1} \mid$. The chart for individuals can be interpreted much like an ordinary $\overline{X}$ control chart. A shift in the process average will result in either a point (or points) outside the control limits, or a pattern consisting of a run on one side of the center line.

The individual measurements on the IX chart are assumed to be uncorrelated [1, 5, 8]. However, dependent or correlated measurements are fairly common in continuous operations such as brewing soft drinks. Even there are times when an assignable cause is known, but cannot be easily eliminated because it stems from

an integral part of the process. [10] recommended the use of alternate-points. In addition, [9] found the usefulness of exponential weighted moving average (EWMA) for forecasting when there is positive autocorrelation and the process average changes slowly.

When a systematic time-related drift is present in a process, [5] recommended modeling the time series with one of the auto regressive, integrated, moving average (ARIMA) models. Once a reasonable model is established, future process measurements may be reliably predicted from the current and previous observations due to the auto correlated nature of data. When such a model fits the data well, differences between forecasted and actual measurements, called forecast errors or "residuals" should be relative small. In addition, a correctly specified model removes autocorrelation from the forecast error, leaving only random residuals, which may then be correctly monitored on an IX & MR chart.

When the IX & MR chart for residuals indicates a good state of control, measuring process capability can be done. This research focuses on measuring potential capability and performance capability. Furthermore, the measuring process is based on the process recommended by [5].

2 Materials and Methods

The data used in this paper is secondary data from a soft drink company. This data can be found in [3]. The data were taken from company "M" that produced carbonated soft drink. The observed characteristic was the brix level of the soft drink. Every measurement was the result of an examination of one bottle which was randomly taken every 30 minute production. So the sub-sample (n) =1.

Data was statistically analyzed through the following steps: to make lag one plot measurements, to find out autocorrelation coefficient, to reduce the data to become alternate data, to do EWMA transformation and get the residual, to make individual control chart, and to measure potential and performance capability.

3 Finding and discussion

The data of this paper is brix levels of soft drink from the factory. The data consists of 258 sub-samples, while the size of each sub-sample (n) is 1. Every sub-sample is taken with interval 30 minutes, the data is provided on appendix A (the data have been transformed to protect the secret of data). The factory determines LSL, USL and Target: 0.85, 1.15 and 1.00 respectively (they have been transformed, too).

Because n = 1, statistical quality control uses Individual (IX) & Moving Range (MR) chart. The control chart is shown in Figure 1. Notice that the process is apparently out-of-control but operating at an acceptable level because the whole points lay in USL and LSL limits. It would be of interest to know whether the out-of-control signals are indicating real problems or whether they are false alarm signals.

As shown in Figure 1, if the brix measurement taken at time t minus 1 is above the centerline, there is a strong likelihood the next measurement at time t will also be above the centerline. Likewise, when the measurement at t minus 1 is below the centerline, it is very probably the brix measurement for time period t is also below.

Because there is little change from one brix measurement to the next, the moving ranges are relatively small, but display good control on the moving range chart. However, these small ranges cause $\overline{MR}$ to also be small, resulting in very "tight" control limits for the IX chart. As the process average for brix gradually drifts up and down due to subtle changes in the blending process, a number of brix readings are push outside of these narrow control limits. In this situation, any inherent drifts in the process average will dramatically increases the false alarm rate of Shewhart charts to the point of rendering them ineffective, and possibly even misleading. In addition, the standard deviation estimated from this $\overline{R}$ value is virtually worthless for estimating process capability [6].

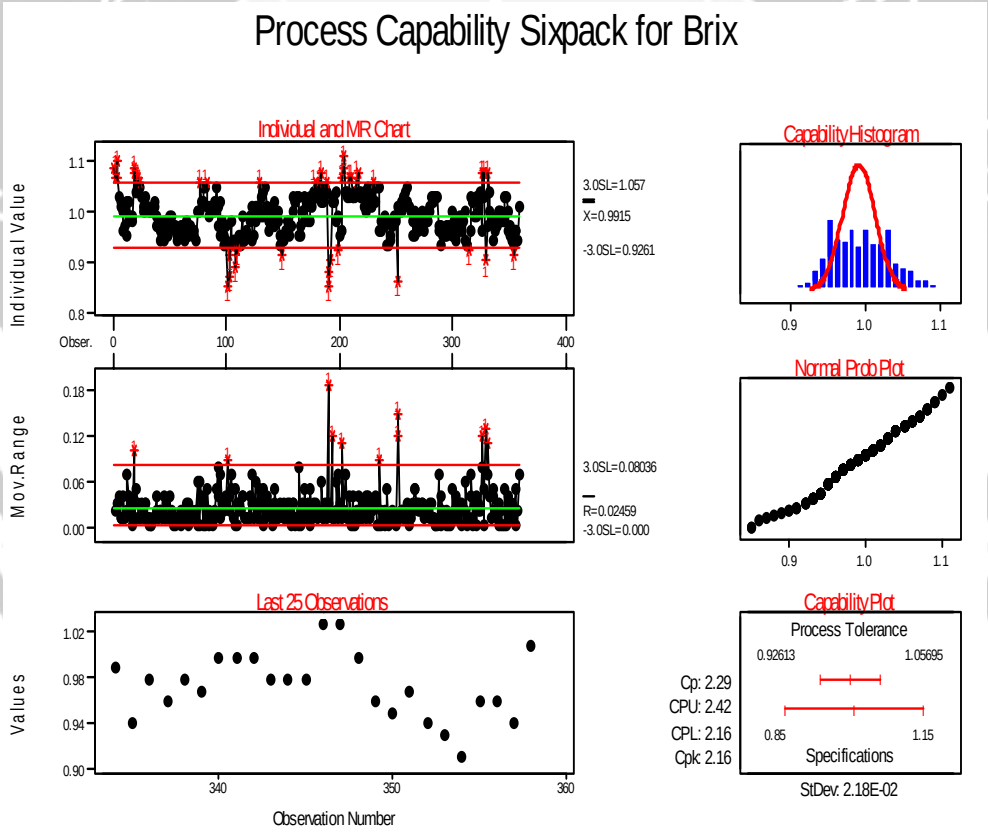


Fig. 1 IX & MR chart for brix level

If the cause of this drift is part of the process (as it often is in continuous process), or cannot be readily adjusted, the operator has an out-of-control signal he cannot correct, leading to either the chart being ignored or the process being over adjusted. The cyclical drift of the process average also negates the usefulness of any Western Electric run rules, thus contributing even more interpretation difficulties.

This charting problem is direct result of high autocorrelation, which is a measure of the degree to which individual measurement from a single process are related to each other [5]. Autocorrelation is occasionally referred to as serial correlation, as it occur at time series of measurements, like those produce in the processing industries. Lack of independence in consecutive observations is typically detected in one of two ways: plotting the data on a scatter diagram and looking for a pattern, or calculating the simple autocorrelation and testing to determine if it is significant.

With the first approach, the measurement at time t (labeled X_{t-1}) is considered the x coordinate, while the measurement at time t (labeled X_t) becomes the y coordinate for each point plotted on the scatter diagram. X_{t-1} is often referred to in the statistical literature as the measurement for the "lag one" time period. If analysis of the scatter diagram reveals a pattern then a reasonable prediction of X_t can be made from X_{t-1}, implying that consecutive measurement are not independent. If so, standard Shewhart control charts should not be chosen to monitor this process.

In this research, the writer considered that the use of one cycle for measuring process capability is enough. This data is the first 73 measurements from appendix A presented in Table 1. The first column lists the measurement number in chronological order from the time period of 1 to 73, while the second column contains the actual brix measurements at time t, labeled as X_t. In the third column, the measurements of column two have been shifted down one row to become X_{t-1}, the lag one time period. Thus, for time period two, X_2 equals 1.07 while the brix measurement for the lag one time period X_{2-1}, or X_1, is 1.09.

Table 1 73 brix measurements displaying autocorrelation

Time Period, t	Y = X_t	x = X_{t-1}	$X_t - \overline{X}$	$X_{t-1} - \overline{X}$	$(X_t - \overline{X})x(X_{t-1} - \overline{X})$	$(X_t - \overline{X})^2$
1	1.09	_	0.098356	_	_	0.009674
2	1.07	1.09	0.078356	0.098356	0.007707	0.00614
3	1.10	1.07	0.108356	0.078356	0.00849	0.011741
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
73	0.94	0.96	-0.05164	-0.03164	0.001634	0.002667
Total	72.39				0.100082	0.129203
Average	0.9916					

$$\hat{\rho}_1 = \frac{\sum_{t=2}^{73}\left(X_t - \overline{X}\right)\left(X_{t-1} - \overline{X}\right)}{\sum_{t=1}^{73}\left(X_t - \overline{X}\right)^2} = 0.775 \; > \; \frac{Z_{\alpha/2}}{\sqrt{m}} = \frac{1.96}{\sqrt{73}} = 0.229$$

Each brix measurement, along with corresponding lag one time period measurement, is plotted as a point on the scatter diagram exhibited in Figure 2.

Note that the scales on both axes are identical for an autocorrelation scatter diagram. Although data are collected over 73 time periods, there are only 72 points because the first measurement (at t = 1) has no corresponding lag one time period.

Because the plot points are grouped in a fairly tight ellipsoid around the 45-degree line, the existence of positive autocorrelation is quite likely. Positive autocorrelation means there is a high probability that the current brix reading will be higher than average when the immediately previous reading is higher than average. Negative autocorrelation is the reverse: there is a high probability X_t will be low given that X_{t-1} was high. Positive autocorrelation may be due to the "inertia" of a process. The brix of soft drink cannot be instantaneously shifted higher or lower.

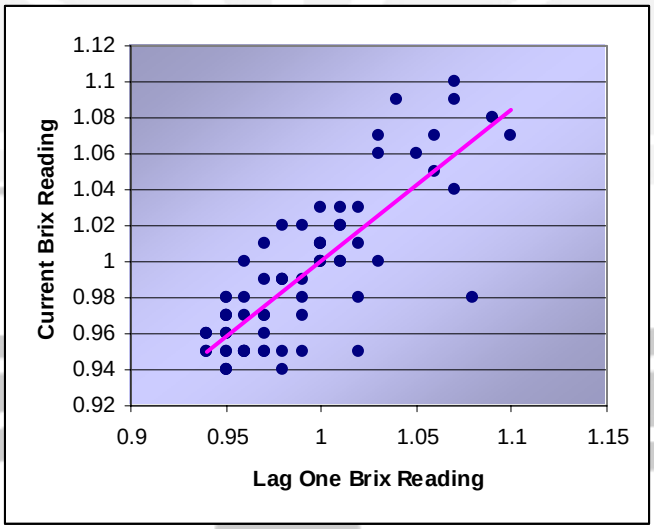


Fig. 2 Scatter diagram revealing presence of autocorrelation.

The presence of either positive or negative autocorrelation implies the process average is not stable, but moving. In many situations, this movement is inherent to the process, especially in the chemical and processing industries. If so, these drifts in the process average cannot be easily or quickly eliminated and must definitely be taken into consideration when conducting a capability study on this process.

Autocorrelation may be more precisely quantified by computing an estimate of ρ, the coefficient of autocorrelation. Its formula is given below, where t represents the lag period, while m is the number of time periods [2]. Thus, ρ_1 measures the amount of correlation between the current measurement and the lag one time period measurement.

$$\hat{\rho}_1 = \frac{\sum_{t=2}^{73}\left(X_t - \overline{X}\right)\left(X_{t-1} - \overline{X}\right)}{\sum_{t=1}^{73}\left(X_t - \overline{X}\right)^2}$$

The estimate of ρ will always be between -1 and +1. A negative ρ value implies negative autocorrelation, whereas a positive value indicates the presence of positive autocorrelation. P values close to zero mean little autocorrelation is present. Autocorrelations begins creating problems for the calculation of meaningful control limits when ρ is greater than 0.6 [5]. For the data in Table 1, autocorrelation of the current measurement with the lag one measurement is estimated as 0.775.

For an α of level 0.05, autocorrelation is significant if the absolute value of the estimate for ρ is greater than 1.96 divided by the square root of m. In the brix in table 1, the ρ_1 is estimated as 0.775, while the number of time period is 73.

$$\left|\hat{\rho}_1\right| = \left|0.775\right| > \frac{Z_{\alpha/2}}{\sqrt{m}} = \frac{1.96}{\sqrt{73}} = 0.229$$

As 0.775 is substantially greater than 0.223, autocorrelation with the lag one time period is certainly significant at the α equal 0.05 level. Thus traditional control charts should not be chosen to monitor the brix level of this brewing soft drink because the brix measurements cannot be considered independent. For this reason, [10] recommended to make a control chart by plotting only alternate points, that is, every other point, so that the plotting points are independent. This research took alternate data, that is 1, 3, 5, ... 71, so that it obtained 36 measurements (points), as showed in table 2.

Table 2 Alternate points from 73 consecutive measurement available are given in table 1.

Sub-sample Number	Brix	Sub-sample Number	Brix	Sub-sample Number	Brix	Sub-sample Number	Brix
1	1.09	19	1.09	37	1.01	55	0.96
3	1.10	21	1.07	39	0.99	57	0.94
5	1.03	23	1.05	41	0.96	59	0.95
7	1.01	25	1.03	43	0.95	61	0.95
9	0.96	27	1.00	45	0.98	63	0.94
11	0.95	29	1.00	47	0.96	65	0.98

13	0.98	31	1.00	49	0.99	67	0.95
15	0.99	33	1.00	51	0.99	69	0.95
17	0.98	35	1.01	53	0.97	71	0.95

Autocorrelation of these 36 measurements with the lag one measurement is estimated as $\hat{\rho} = 0.689$. It is lower than the previous $\hat{\rho}$; however, it is still significant because of $|\hat{\rho}| = 0.689 > \frac{z_{\alpha/2}}{\sqrt{m}} = \frac{1.96}{\sqrt{36}} = 0.327$.

Therefore, the processing should be continued with EWMA (exponentially weighted moving average) transformation.

[9] have found exponentially weighted moving average (EWMA) useful for forecasting when there is positive autocorrelation and the process average changes slowly. The EWMA models predicts the measurement for time period t, $\hat{X}_t$, by taking a weighted average of the previous measurement, X_{t-1}, and the forecast for X_{t-1}, labeled $\hat{X}_{t-1}$. The formula is:

$$\hat{X}_t = \lambda X_{t-1} + (1-\lambda)\hat{X}_{t-1}$$

The symbol λ is the weighting factor, which is a positive number less than 1. λ values close to 1 give more weight to the most recent observation and less to previous ones, whereas values close to 0 weigh older observations more than current ones. The precise λ for a particular process is frequently selected as the one which minimizes the sum of the squared residuals (the forecast error) or which has insignificant autocorrelation residual.

The application of the EWMA model to 36 alternate measurements with λ equal to 0.9 is presented in Table 3.

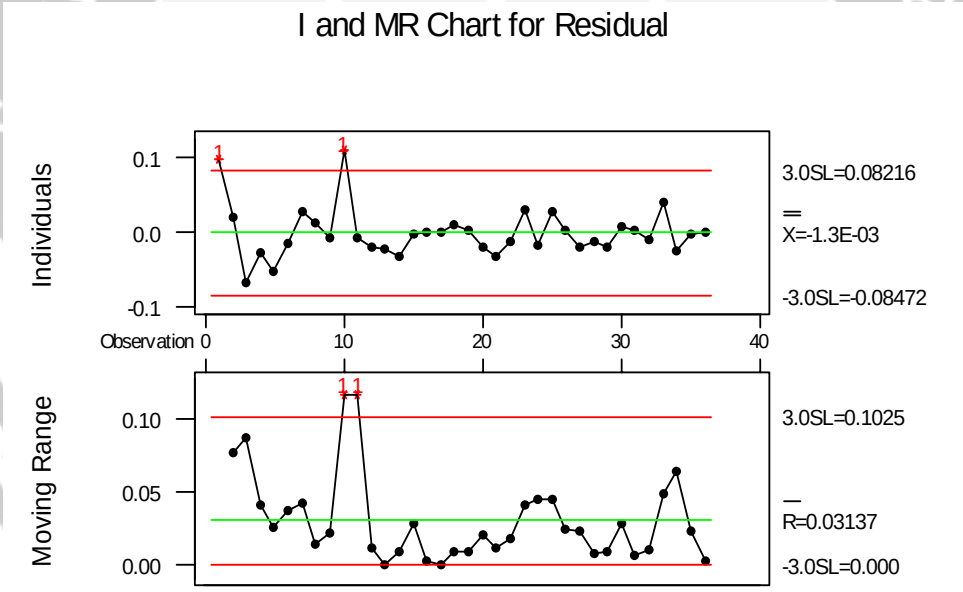
Table 3 Application of the EWMA model to brix measurements.

Actual X_t	Forecast $\hat{X}_t$	Residual (Res_t) $(X-\hat{X}_t)$	Res_{t-1}	$Res_t - \overline{Res}$	$Res_{t-1} - \overline{Res}$	(5)x(6)	$(5)^2$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1.09	0.99194	0.0980	_	0.099332	_	_	0.00986
1.1	1.08	0.02	0.0980	0.021276	0.099332	0.00211	0.00045
1.03	1.098	-0.068	0.02	-0.06672	0.021276	-0.0014	0.00445
1.01	1.037	-0.027	-0.068	-0.02572	-0.06672	0.00171	0.00066
0.96	1.013	-0.053	-0.027	-0.05172	-0.02572	0.00133	0.00267
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

0.95	0.95	0	-0.003	0.001276	-0.00172	-2.2E-06	1.63E-06
	Total	-0.0459				0.002514	0.04027
	Average	-0.0012					

$$\hat{\rho}_1 = \frac{\sum_{t=2}^{36}\left(Res_t - \overline{Res}\right)\left(Res_{t-1} - \overline{Res}\right)}{\sum_{t=1}^{36}\left(Res_t - \overline{Res}\right)^2} = 0.0624 < \frac{Z_{\alpha/2}}{\sqrt{m}} = \frac{1.96}{\sqrt{36}} = 0.327$$

As 0.0624 is substantially less than 0.327, autocorrelation with the lag one time period is certainly not significant at the α equal 0.05 level. Because no significant autocorrelation exists between the residuals, these residuals can be used to make IX & MR chart. The IX & MR chart is shown in Figure 3.



An analysis of Figure 3 shows that there are out-of-control points on the IX chart at point 1 and 10 and out-of-control point on MR chart at point 9 and 10. These points (residuals) are discarded from the data (residuals) and the new IX & MR chart is computed with remaining residuals. After being revised twice, the result is shown in Figure 4.

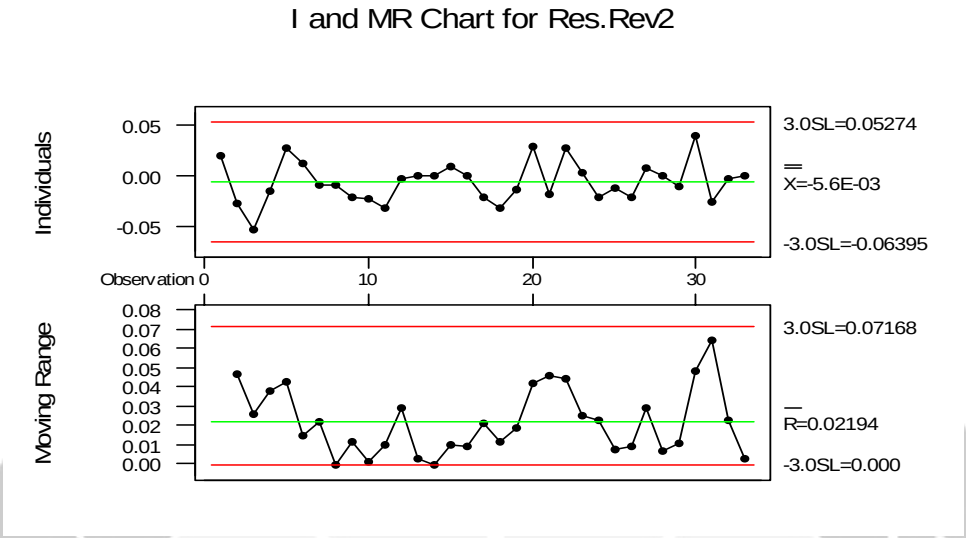


Fig. 4 IX & MR chart after being revised twice.

Figure 4 shows that there is no out-of-control point on the IX & MR chart. This means that the process is stable. The result obtained from this chart is shown as follows:

CL = -0.0056, UCL = 0.05274 and LCL = -0.06395 for IX chart; and $\overline{MR}$ = 0.02194 for MR chart.

Because the IX & MR chart of residual exhibits statistically controlled, the $\hat{\sigma}_{ST.AUTO}$ is estimated from the average moving range of 0.02194,

$$\hat{\sigma}_{ST.AUTO} = \frac{\overline{MR}}{1.128} = 0.01945$$

With a LSL of 1.15 and an USL of 0.85 for brix, potential capability, $C_{P.AUTO}$ is estimated as:

$$C_{P.AUTO} = \frac{Tolerance}{6\hat{\sigma}_{ST.Auto}} = \frac{0.3}{0.116702} = 2.570647.$$

This reveals the process capability ($6\sigma_{P.AUTO}$) is less then the tolerance (USL-LSL), is the most desirable case. A capability index of 1.33 is considered by most companies to be a de facto standard with even large values desired [4].

From the forecasted brix levels exhibited in Table 3, $\hat{\mu}_L$ and $\hat{\mu}_H$ are found to be 0.941 and 1.098, respectively. With the standard deviation estimated as 0.01945, the performance capability, $\hat{C}_{PK.AUTO}$, is estimated as:

$$\hat{C}_{PK.AUTO} = Minimum\left[\frac{\hat{\mu}_L - LSL}{3\hat{\sigma}_{ST.AUTO}}, \frac{USL - \hat{\mu}_H}{3\hat{\sigma}_{ST.AUTO}}\right]$$

$$= Minimum\left[\frac{0.941 - 0.85}{3(0.01945)}, \frac{1.15 - 1.098}{3(0.01945)}\right]$$

$$= Minimum[1.559526,\ 0.891158] = 0.891.$$

As $\hat{C}_{PK.AUTO}$ is less than 1.0, work must begin to either shrink $\hat{\sigma}_{ST.AUTO}$ or diminish the amount of drifting in the process average.

Figure 5 presents the IX and MR chart, normal probability plot, and capability plot for the residual.

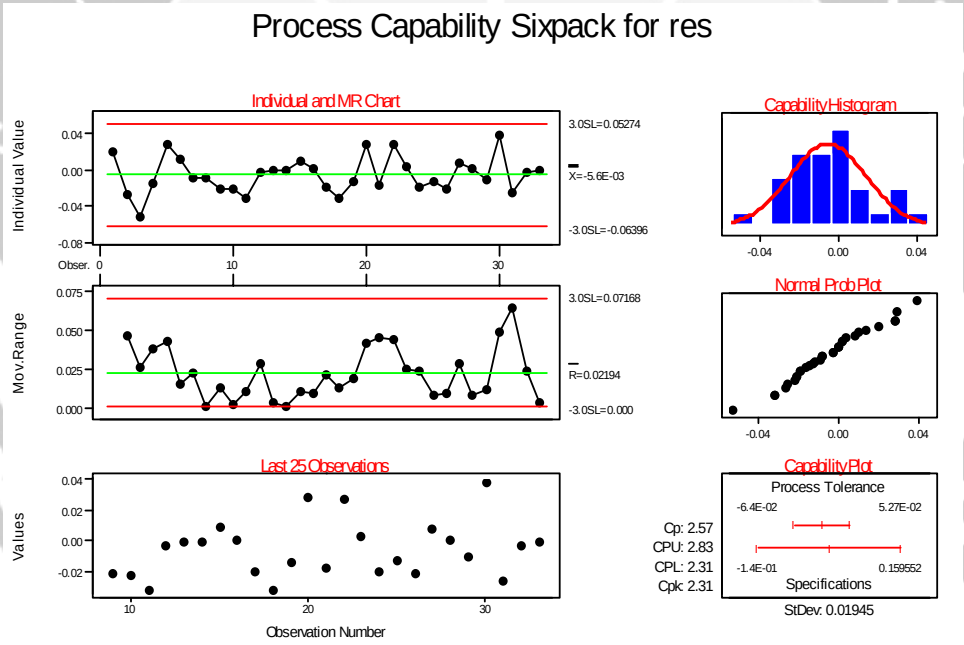


Fig. 5 Process capability sixpack for Residual.

The $\hat{\sigma}_{ST.AUTO}$ = 0.01945 obtained is used to set up IX & MR for the 73 brix levels in Table 1. The control chart is shown in Figure 6.

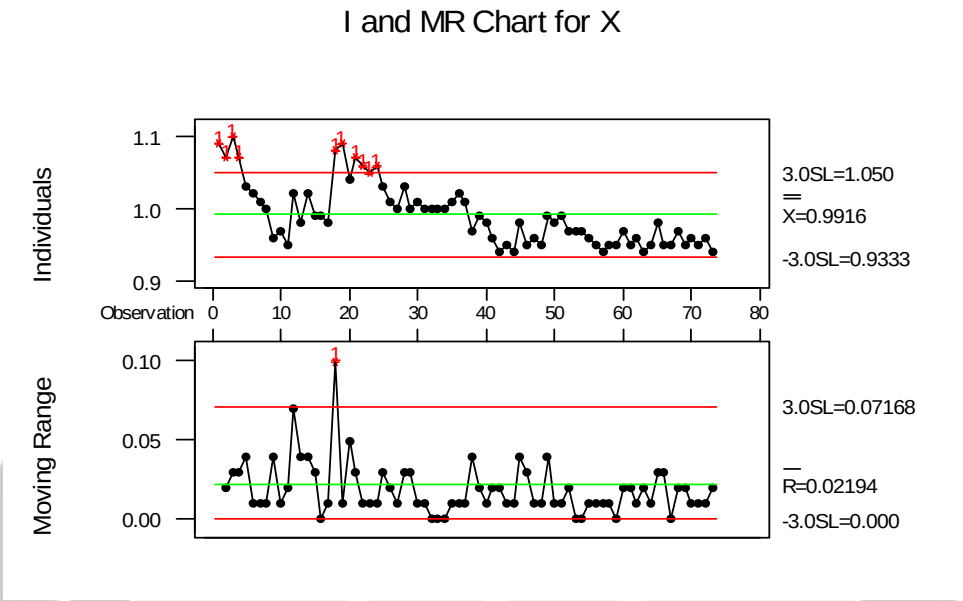


Fig. 6 IX & MR for the 73 brix levels with $\hat{\sigma}_{ST.AUTO}$ = 0.01945

If Figure 6 is compared to Figure B.3 in appendix B, there are differences in the number of points which are outside of the control limits. In figure 6 there are 10 points, while in Figure B.3 there are 32 points. In other words, there are 22 false alarm signals reduced.

4 Conclusion

The result of the research shows that a combination of the methods of alternate-points and EWMA prediction was very efficient in measuring both potential and performance capabilities. The potential capability index obtained in this research is 2.57, is the most desirable case. Moreover, the performance capability index obtained is less than 1.0 (i.e. 0.89) which means work must begin to either shrink $\hat{\sigma}_{ST.AUTO}$ or diminish the amount of drifting in the process average. Finally, this research finding reveals the ability of the methods used in reducing the false alarm signals.

References

[1] Albin, L. S., L. Kang, & G. Shea (1997), An X and EWMA Chart for Individual Observations, *Journal of Quality Technology*, **29**, 41-48.

[2] Alwan, L. C. (1992), Autocorrelation: Fixed versus Variable Control Limits, *Quality Engineering*, **4**, 167-188.

[3] Arcana, N. (2002), Telaah Pengendalian Mutu di Perusahaan "M" dan Upaya Perbaikannya dengan Penerapan Statistika, *Theses,* IPB, Bogor.
(Quality Control at "M" Company and its Improvement Efforts Using Statistics)**.**

[4] Besterfield, D.H. (1994), *Quality Control, 4th*. Prentice-Hall, Inc., New Jersey

[5] Bothe, D.R. (1997), *Measuring Process Capability,* McGraw-Hill Book Company, New York.

[6] Crayer, J.D. & T.P. Ryan (1990), The Estimation of Sigma for an X Chart: $\overline{MR}/d_2$ or S/c4?, *Journal of Quality Technology*, **22**, 187-192.

[7] Harris, T. J. & W. H. Ross (1991), Statistical Process Control Procedures for Correlated Observation, *Canadian Journal of Chemical Engineering***, 69**, 48-57.

[8] Montgomery, D.C. (1991), *Introduction to Statistical Quality Control, 2nd*. John Wiley & Sons, Inc., New York.

[9] Montgomery, D. C. & C.M. Mastrangelo (1991), Some Statistical Process Control Methods for Autocorrelated Data, *Journal Of Quality Technology***, 23**, 179-193.

[10] Quesenberry, C.P. (1997), *SPC Methods for Quality Improvement*, John Wiley & Sons, Inc., New York.

[11] Roes, K. C. B., R. J. M. M. Does, & Y. Schurink (1993), Shewhart-Type Control Charts for Individual Observations, *Journal of Quality Technology***, 25**, 188-198.

Appendix

A Brix levels of soft drink from the factory (the data have been transformed to protect the secret of data).

Sub-sample Number	Brix	Sub-sample Number	Brix	Sub-sample Number	Brix	Sub-sample Number	Brix	Sub-sample Number	Brix
1	1.09	73	0.94	145	1	217	1.04	289	1
2	1.07	74	1	146	0.98	218	1.03	290	0.9
3	1.1	75	1	147	0.95	219	1.04	291	0.9
4	1.07	76	1	148	0.93	220	1.03	292	1
5	1.03	77	1.06	149	0.91	221	1.01	293	0.9

6	1.02	78	1.02	150	0.94	222	1.04	294	0.9
7	1.01	79	1.02	151	0.95	223	1	295	1
8	1	80	1.05	152	0.96	224	1.01	296	1
9	0.96	81	1.05	153	0.95	225	1.05	297	1
10	0.97	82	1.06	154	0.98	226	1.02	298	1
11	0.95	83	1.01	155	0.95	227	1.03	299	0.9
12	1.02	84	1.01	156	0.97	228	1.01	300	1
13	0.98	85	0.99	157	1	229	1.03	301	1
14	1.02	86	1	158	1	230	1.06	302	0.9
15	0.99	87	1.01	159	1	231	1.02	303	0.9
16	0.99	88	1.01	160	1	232	1.05	304	1
17	0.98	89	1.02	161	1	233	1.05	305	1
18	1.08	90	1	162	1	234	1.05	306	1
19	1.09	91	1.05	163	1	235	0.96	307	1
20	1.04	92	0.97	164	0.9	236	0.96	308	1
21	1.07	93	1.01	165	0.9	237	0.98	309	1
22	1.06	94	1.02	166	1	238	0.97	310	1
23	1.05	95	0.95	167	1	239	0.97	311	0.9
24	1.06	96	0.98	168	1	240	0.99	312	0.9
25	1.03	97	0.95	169	1	241	0.98	313	0.93
26	1.01	98	0.93	170	1	242	0.93	314	0.92
27	1	99	0.97	171	1	243	0.93	315	0.96
28	1.03	100	0.94	172	1	244	0.94	316	1.03
29	1	101	0.85	173	1	245	0.95	317	1.02
30	1.01	102	0.87	174	1	246	0.97	318	1.02
31	1	103	0.91	175	1	247	0.98	319	0.99
32	1	104	0.98	176	1	248	0.99	320	0.99
33	1	105	0.93	177	1	249	0.99	321	1.03
34	1	106	0.94	178	1.1	250	0.98	322	1
35	1.01	107	0.94	179	1	251	0.86	323	1.02
36	1.02	108	0.89	180	1.1	252	1.01	324	1.04
37	1.01	109	0.92	181	1	253	1	325	0.96
38	0.97	110	0.92	182	1.1	254	1.02	326	1.08
39	0.99	111	0.95	183	1.1	255	1	327	1.08
40	0.98	112	0.93	184	1	256	1.01	328	1.03
41	0.96	113	0.95	185	1	257	1.03	329	0.9
42	0.94	114	0.99	186	1	258	1.04	330	0.97
43	0.95	115	0.95	187	1.1	259	1.05	331	1.08
44	0.94	116	0.94	188	1.1	260	1.02	332	1.04
45	0.98	117	0.95	189	1	261	1	333	1
46	0.95	118	0.96	190	0.9	262	1	334	0.99

47	0.96	119	0.95	191	0.9	263	1	335	0.94
48	0.95	120	0.99	192	0.9	264	1	336	0.98
49	0.99	121	0.98	193	1	265	0.9	337	0.96
50	0.98	122	1.01	194	1	266	1	338	0.98
51	0.99	123	0.94	195	1	267	0.9	339	0.97
52	0.97	124	0.95	196	1	268	1	340	1
53	0.97	125	1.01	197	1	269	1	341	1
54	0.97	126	0.98	198	1	270	1	342	1
55	0.96	127	1	199	0.9	271	1	343	0.98
56	0.95	128	1.01	200	0.9	272	1	344	0.98
57	0.94	129	1.06	201	1	273	1	345	0.98
58	0.95	130	1.03	202	1.1	274	1	346	1.03
59	0.95	131	0.99	203	1.1	275	1	347	1.03
60	0.97	132	0.99	204	1.1	276	1	348	1
61	0.95	133	1.04	205	1	277	1	349	0.96
62	0.96	134	1.05	206	1	278	1	350	0.95
63	0.94	135	1.01	207	1	279	1	351	0.97
64	0.95	136	0.99	208	1	280	1	352	0.94
65	0.98	137	0.96	209	1.07	281	1	353	0.93
66	0.95	138	0.98	210	1.07	282	1	354	0.91
67	0.95	139	1	211	1.03	283	1	355	0.96
68	0.97	140	0.99	212	1.03	284	1	356	0.96
69	0.95	141	0.99	213	1.03	285	1	357	0.94
70	0.96	142	1.02	214	1.04	286	1	358	1.01
71	0.95	143	0.97	215	1.06	287	1		
72	0.96	144	0.99	216	1.08	288	1		

B IX & MR chart for the 73 brix levels in table 1 without alternate points and EWMA transformation.

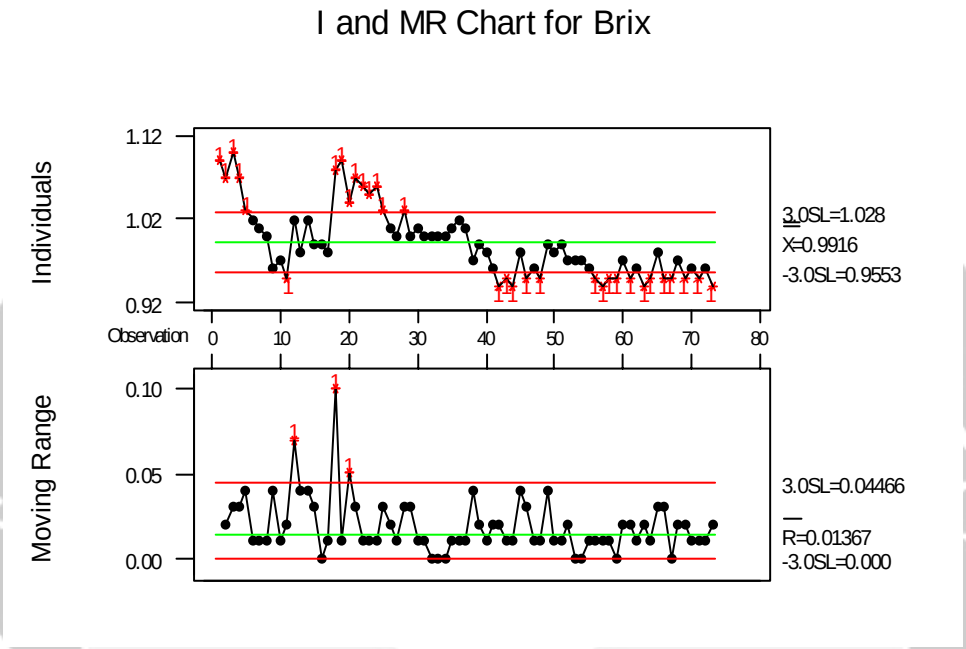


Fig. B.1 IX & MR chart for the 73 brix levels in table 1

An analysis of Figure A shows that there are out-of-control points on the IX chart at point 1, 2, 3, 4, 5, 11, 18, 19, 20, 21, 22, 23, 24, 25, 28, 42, 43, 44, 46, 48, 56, 57, 58, 59, 61, 63, 64, 66, 67, 69, 71, and 73 These points are discarded from the data and the new IX & MR chart is computed with remaining brix levels. After being revised fourth, the result is shown in Figure B.

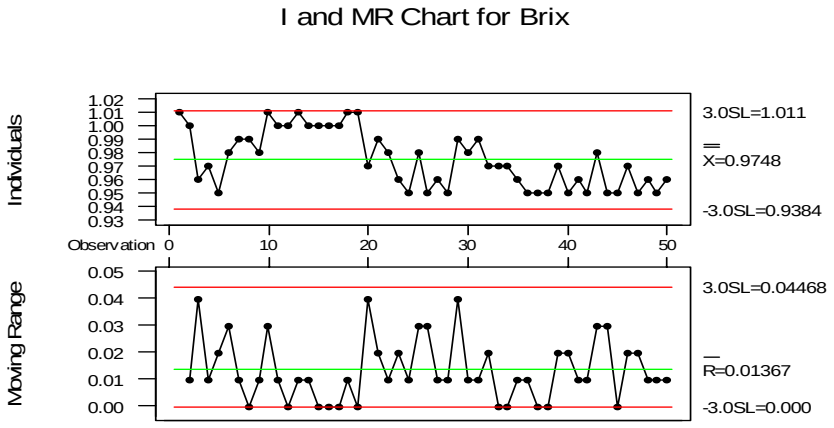


Fig. B.2 IX & MR chart after being revised fourth.

Because the IX & MR chart after being revised fourth exhibits statistically controlled, the $\hat{\sigma}_{ST.AUTO}$ is estimated from the average moving range of 0.01367, $\hat{\sigma}_{ST.AUTO} = \dfrac{\overline{MR}}{1.128}$ = 0.012119 This $\hat{\sigma}_{ST.AUTO}$ is used to set up IX & MR for the 73 brix levels in Table 1. The control chart is shown in Figure B.3.

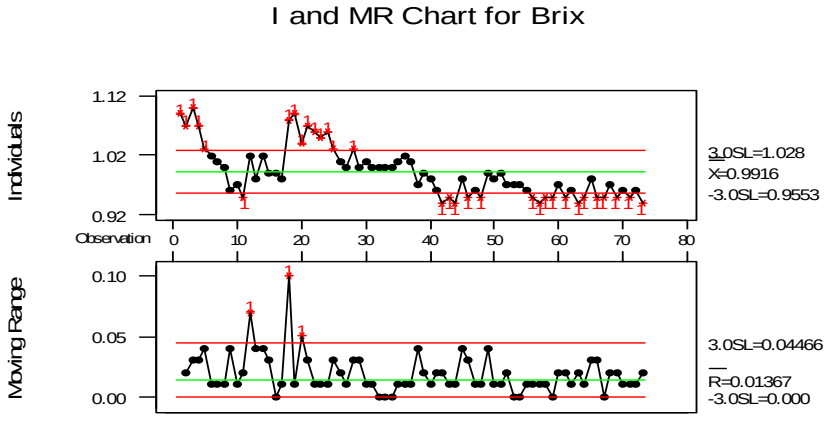


Fig. B.3 IX & MR for the 73 brix levels with $\hat{\sigma}_{ST.AUTO}$ = 0.012119

An analysis of Figure C shows that there are 32 points are out side of the control limits.

I NYOMAN ARCANA: Widya Mandala Catholic University, Surabaya, Indonesia.
E-mail: arcana@mail.wima.ac.id

ON THE STABILITY OF NEURAL NETWORKS IN PERIODIC ENVIRONMENT

Sariyasa

IKIP Negeri Singaraja, Bali, Indonesia

Abstract. The dynamics of Hopfield-type neural networks subjected to periodic external stimuli are investigated. Delays are incorporated in the networks and the networks parameters are assumed to be periodic. The investigation is focused on the dynamics of the networks in encoding external stimuli which vary periodically with time and recalling the encoded patterns associated with the external stimuli. In particular sufficient conditions for the exponential stability of the networks toward encoded patterns associated with the external stimuli are established.
Key-words: Hopfield-type neural networks, time delays, exponential stability, periodic environment

1 Introduction

Most of the theoretical studies on neural networks is predominantly concerned with autonomous systems containing temporally uniform network parameters and external input stimuli. A study dealing with time-varying stimuli or network parameters appears to be scarce; such studies are however important to understand the dynamical characteristics of neuron behaviour in time-varying environments.

In studying neural networks, it is generally assumed that the environment within which the network operates does not change with time or stationary. Frequently, however, the environment of interest is non-stationary which means that the parameters of the information-bearing signals generated by the environment vary with time [7]. Thus it is necessary to develop a model than can track the temporal variations of the environment in which it operates.

It has been reported (see [3], [6], [4]) that assemblies of cells in the visual cortex oscillate synchronously in response to external stimuli. Such a synchrony is a manifestation of the encoding process of temporally varying external stimuli. In [9] it is shown that a chaotic attractor can be converted to any one of a large number of possible time periodic motions by the introduction of time dependent perturbations to the network parameters. Thus it is worthwhile to consider time-dependent parameters in the equations governing the dynamics of neurons.

Since delays are naturally present in biological neurons through synaptic transmissions, finite conduction velocities along the axon, and neural processing of input stimuli, we incorporate delays in the processing parts of the network's architectures.

Applications of neural networks depend on the dynamics of the governing system of neural networks. For instance in solving problems in parallel computation and signal processing involving optimization one has to design models of neural networks possessing a unique equilibrium of temporally static or dynamic type so as to avoid spurious behaviour related to local minima (see for instance [1], [5], [8]). In applications, it is important to improve the rate of convergence to equilibrium state of the network and hence to reduce the computation time. If the equilibrium of the network is exponentially asymptotically stable, then the convergence is fast for real time computation. Thus it is required to design a network that possesses exponential stability.

We are interested in studying the effect of a temporally varying external stimuli on the dynamics of the Hopfield-type neural networks; in particular we investigate the dynamics of a neural network in encoding external stimuli which vary periodically with time and recalling the encoded patterns associated with the external stimuli. We will derive sufficient conditions for the existence (or encoding) of a globally attractive (associative recall) periodic solution (or pattern) associated with a given periodic external stimuli.

2 Existence and Exponential Stability of Periodic Solutions

We consider the dynamics of the neural network in the following form:

$$\frac{dx_i(t)}{dt} = -a_i(t)x_i(t) + \sum_{j=1}^{n} b_{ij}(t)\tanh\big(x_j(t-\tau_{ij})\big) + f_i(t), \quad t > 0 \tag{1}$$

in which $\tau_{ij} \geq 0$ denotes delays, $i \in \mathfrak{I} = \{1, 2, \ldots, n\}$ and $a_i(\cdot)$, $b_{ij}(\cdot)$, $f_i(\cdot)$ denote continuous real-valued functions defined on $(-\infty, \infty)$ and are periodic with period $\omega > 0$ so that

$$a_i(t+\omega) = a_i(t), \quad b_{ij}(t+\omega) = b_{ij}(t), \quad f_i(t+\omega) = f_i(t), \quad t \in \mathbb{R}.$$

In (1), $a_i(\cdot)$ denote quantitative measures of the neuronal dissipation or negative feedback terms; $b_{ij}(\cdot)$ denotes the neuron gains, and $f_i(\cdot)$ denote the periodic external stimuli.

The equation (1) is supplemented with an initial condition of the form

$$x_i(s) = \varphi_i(s), \quad s \in [-\tau, 0], \quad \varphi_i \in C[-\tau, 0], \quad \tau = \max_{i,j\in\mathfrak{I}}\{\tau_{ij}\}$$

where $C[-\tau, 0]$ denotes the space of all continuous real-valued functions defined on $[-\tau, 0]$ endowed with the supremum norm $\|\cdot\|$ defined by

$$\|x_i(t)\| = \sup_{s\in[-\tau,0]} |x_i(t+s)|.$$

The presence of the delays in the networks may affect the convergence rate. Accordingly, it is necessary to eliminate the effects of delay on the convergence rate. Thus, in the following we derive sufficient conditions independent of the delay for the exponential stability of solutions of (1).

Let $x_i(t) = x_i(t, \varphi_i)$ and $y_i(t) = y_i(t, \psi_i)$, $t > 0$ denote arbitrary solutions of (1) corresponding to initial conditions $\varphi_i = \varphi_i(s)$ and $\psi_i = \psi_i(s)$ defined for $s \in [-\tau, 0]$ respectively.

The following result establishes the exponential stability of the solutions of (1).

Theorem 2.1. Assume that the coefficients $a_i(\cdot)$ and $b_{ij}(\cdot)$ are bounded and continuous on $(-\infty, \infty)$ such that

$$0 < \inf_{t\in\mathbb{R}} a_i(t) = \underline{a}_i > \sum_{j=1}^{n} \bar{b}_{ji}, \quad i \in \mathfrak{I} \tag{2}$$

where $\bar{b}_{ji} = \sup_{t\in\mathbb{R}} |b_{ij}(t)|$; suppose the time delay τ_{ij} is a constant, $\tau_{ij} \geq 0$. Then solutions of the system (1) is exponentially stable in the sense that there exist positive numbers M and r such that

$$\sum_{i=1}^{n} \|x_i(t) - y_i(t)\| \leq Me^{-rt} \sum_{i=1}^{n} \|x_i(0) - y_i(0)\|, \quad t > 0.$$

Proof. Consider the function $g_i : [0, \infty) \to \mathbb{R}$ defined by

$$g_i(w_i) = \underline{a}_i - w_i - \sum_{j=1}^{n} \bar{b}_{ji} e^{w_i \tau_{ji}}, \quad w_i \geq 0, \quad i \in \mathfrak{I}. \tag{3}$$

We have from (2) and (3) that

$$g_i(0) = \underline{a}_i - \sum_{j=1}^{n} \bar{b}_{ji} > 0 \quad \text{for all } i \in \mathfrak{I}.$$

Since $g_i(\cdot)$ is continuous on $[0, \infty)$ and $g_i(w_i) \to -\infty$ as $w_i \to \infty$, there exists an $\epsilon_i^* \in [0, \infty)$ such that

$$g_i(\epsilon_i^*) = \underline{a}_i - \epsilon_i^* - \sum_{j=1}^{n} \bar{b}_{ji} e^{\epsilon_i^* \tau_{ji}} = 0 \quad \text{for all } i \in \mathfrak{I}.$$

By choosing $\epsilon = \min_{i\in\mathfrak{I}} \{\epsilon_i^*\}$ we have

$$g_i(\epsilon) = \underline{a}_i - \epsilon - \sum_{j=1}^{n} \bar{b}_{ji} e^{\epsilon_i^* \tau_{ji}} \geq 0 \quad \text{for all } i \in \mathfrak{I}. \tag{4}$$

Let $x_i(t)$ and $y_i(t)$ denote any two solutions of (1). Then we have from (1) that

$$\frac{d^+}{dt}|x_i(t)-y_i(t)| \le -\underline{a}_i|x_i(t)-y_i(t)| + \sum_{j=1}^{n}\bar{b}_{ij}|x_j(t-\tau_{ij})-y_j(t-\tau_{ij})|, \;\; t>0.$$

Now we define

$$z_i(t) = e^{\epsilon t}|x_i(t)-y_i(t)| \tag{5}$$

and construct Lyapunov functional $V(t)$ as follows

$$V(t) = \sum_{i=1}^{n}\left(z_i(t) + \sum_{j=1}^{n}\bar{b}_{ij}e^{\epsilon\tau_{ij}}\int_{t-\tau_{ij}}^{t} z_j(s)\,ds\right), \quad t>0.$$

Note that $V(t) > 0$ for $t > 0$ and $V(0)$ is finite. Calculating the upper right derivative of V along the solutions of (1) we obtain

$$\begin{aligned}\frac{d^+V}{dt} &\le \sum_{i=1}^{n}\left(-\big(\underline{a}_i-\epsilon\big)z_i(t) + \sum_{j=1}^{n}\bar{b}_{ij}e^{\epsilon\tau_{ij}}z_j(t)\right)\\ &= -\sum_{i=1}^{n}\left(\underline{a}_i - \epsilon - \sum_{j=1}^{n}\bar{b}_{ji}e^{\epsilon\tau_{ji}}\right)z_i(t), \quad t>0. \end{aligned}\tag{6}$$

By using (4) in (6) we have $\frac{d^+V(t)}{dt} \le 0$ for $t > 0$ and so $V(t) \le V(0)$ for $t > 0$. This implies that

$$\begin{aligned}\sum_{i=1}^{n} z_i(t) &\le \sum_{i=1}^{n}\left(z_i(0) + \sum_{j=1}^{n}\bar{b}_{ij}e^{\epsilon\tau_{ij}}\int_{-\tau_{ij}}^{0} z_j(s)\,ds\right)\\ &= \sum_{i=1}^{n}\left(z_i(0) + \sum_{j=1}^{n}\bar{b}_{ji}e^{\epsilon\tau_{ji}}\int_{-\tau_{ji}}^{0} z_i(s)\,ds\right), \quad t>0 \end{aligned}\tag{7}$$

Using (5) in (7) and noting that $\tau = \max_{i,j\in\mathfrak{I}}\{\tau_{ij}\}$, we derive

$$\begin{aligned}\sum_{i=1}^{n}|x_i(t)-y_i(t)|e^{\epsilon t} &\le \sum_{i=1}^{n}\left(|x_i(0)-y_i(0)| + \sum_{j=1}^{n}\bar{b}_{ji}e^{\epsilon\tau_{ji}}\int_{-\tau_{ji}}^{0}|x_j(s)-y_j(s)|e^{\epsilon s}\,ds\right)\\ &\le \sum_{i=1}^{n}\left(\sup_{s\in[-\tau,0]}|x_i(s)-y_i(s)| + \sum_{j=1}^{n}\tau\bar{b}_{ji}e^{\epsilon\tau}\sup_{s\in[-\tau,0]}|x_j(s)-y_j(s)|\right)\\ &= \sum_{i=1}^{n}\left(1+\sum_{j=1}^{n}\tau\bar{b}_{ji}e^{\epsilon\tau}\right)\|x_i(0)-y_i(0)\|. \end{aligned}\tag{8}$$

Thus we obtain from (8),

$$\sum_{i=1}^{n} |x_i(t) - y_i(t)| \le e^{-\epsilon t} \sum_{i=1}^{n} \left(1 + \sum_{j=1}^{n} \tau \bar{b}_{ji} e^{\epsilon \tau} \right) \|x_i(0) - y_i(0)\|, \quad t > 0$$

from which we derive

$$\begin{aligned} \sum_{i=1}^{n} \|x_i(t) - y_i(t)\| = \sup_{s \in [t-\tau, t]} |x_i(s) - y_i(s)| &\le e^{-\epsilon t} \sum_{i=1}^{n} \left(1 + \sum_{j=1}^{n} \tau \bar{b}_{ji} e^{\epsilon \tau} \right) \|x_i(0) - y_i(0)\| \\ &= M e^{-\epsilon t} \sum_{i=1}^{n} \|x_i(0) - y_i(0)\|, \quad t > 0 \end{aligned} \tag{9}$$

where $M = 1 + \sum_{j=1}^{n} \tau \bar{b}_{ji} e^{\epsilon \tau}$. Since $M \ge 1$, $\epsilon > 0$, $x_i(t)$ and $y_i(t)$ for $i \in \mathfrak{I}$ denote arbitrary solutions of (1), the exponential stability of (1) follows from (9). This completes the proof.

The following result establishes the existence of exponentially stable periodic solutions. These periodic solutions represent the encoded neural pattern corresponding to the periodic input stimuli denoted by f_i.

Theorem 2.2. Assume that the hypotheses of Theorem 2.1 hold. Suppose that $a_i(\cdot)$, $b_{ij}(\cdot)$, and $f_i(\cdot)$ are periodic with period $\omega > 0$. Let $\tau_{ij} \ge 0$ be a constant. Then the system (1) has a periodic solution $x_i^*(t)$ with period ω which is exponentially stable.

Proof. We only need to show the existence of the periodic solution $x_i^*(t)$. The exponential stability will follow from Theorem 2.1. Let $\underline{a}_i$ and $\bar{b}_{ji}$ be as in Theorem 2.1. Then we have from (9) that

$$\sum_{i=1}^{n} \|x_i(t) - y_i(t)\| \le M e^{-\epsilon t} \sum_{i=1}^{n} \|x_i(0) - y_i(0)\|, \quad t > 0. \tag{10}$$

Choose a positive integer m large enough such that

$$M e^{-\epsilon m \omega} = \rho < 1.$$

Then we have from (10) with $t = m\omega$ that

$$\sum_{i=1}^{n} \|x_i(m\omega) - y_i(m\omega)\| \le \rho \|x_i(0) - y_i(0)\|.$$

Define a map $T : C[-\tau, 0] \longmapsto C[-\tau, 0]$ by the following

$$T(\varphi_i) = x_i(\omega, \varphi_i)$$

where $x_i(\omega, \varphi_i)$ denotes the solution in $C[-\tau, 0]$ of (1) corresponding to the initial value $\varphi_i \in C[-\tau, 0]$. Then we have from the uniqueness of solutions of (1) that

$$T^2(\varphi_i) = T(x_i(\omega, \varphi_i)) = x_i(2\omega, \varphi_i)$$

and in general,

$$T^m(\varphi_i) = x_i(m\omega, \varphi_i).$$

It then follows from

$$\begin{aligned} \|T^m(\varphi_i) - T^m(\psi_i)\| &= \|x_i(m\omega, \varphi_i) - x(m\omega, \psi_i)\| \\ &\leq \rho\|\varphi_i - \psi_i\| < \|\varphi_i - \psi_i\| \end{aligned}$$

that the mapping T^m is a contraction on the Banach space $C[-\tau, 0]$ of continuous functions since $C[-\tau, 0]$ is the space of continuous functions defined on a bounded closed interval. By the contraction mapping principle (see for instance Coppel [1965], p. 11], the mapping T^m has a fixed point say φ_i^* such that

$$T^m(\varphi_i^*) = \varphi_i^*.$$

Since

$$T^m(T(\varphi_i^*)) = T(T^m(\varphi_i^*)) = T\varphi_i^*,$$

it follows that

$$T(\varphi_i^*) = \varphi_i^* \qquad \text{or} \quad x_i(\omega, \varphi_i^*) = \varphi_i^*.$$

Again from the uniqueness of solutions of (1), we obtain from $x_i(\omega, \varphi_i^*) = \varphi_i^*$ that

$$\begin{aligned} x_i(t + \omega, \varphi_i^*) &= x_i(t, x_i(\omega, \varphi_i^*)) \\ &= x_i(t, \varphi_i^*) \quad \text{for all } t. \end{aligned}$$

Thus $x_i(t, \varphi_i^*)$ is a periodic solution of (1) and the existence of a periodic solution of (1) is established. The exponential stability follows from Theorem 2.1 and the proof is complete.

3 Summary

Exponential stability of delayed Hopfield-type neural networks subjected to periodic external stimuli has been studied. The network parameters are assumed to be periodic. We have obtain sufficient conditions for the existence (or encoding) of a globally attractive (associative recall) periodic solution (or pattern) associated with a given periodic external stimuli.

References

[1] Cohen, M. A. and Grossberg, S. (1983), Absolute stability of global pattern formation and parallel memory storage by competitive neural networks, *IEEE Trans. Systems Man Cybernet.*, **13**, 815–826.

[2] Coppel, W. A. (1965), *Stability and Asymptotic Behavior of Differential Equations*, D. C. Heath and Company, Boston.

[3] Eckhorn, R., Bauer, R., Jordan, W., Brosch, M., Kruse, W., Munk, M., and Reitboeck, H.J. (1988), Coherent oscillations : A mechanism of feature linking in the visual cortex ?, *Biol. Cybern.*, **60**, 121–130.

[4] Engel, A.K., Kreiter, A.K., König, P., and Singer, W. (1991), Synchronization of oscillatory neuronal responses between striate and extrastriate visual cortical areas of the cat, *Proc. Natl. Acad. Sci.*, **88**, 6048–6052.

[5] Fang, Y. and Kincaid, T. G. (1996), Stability analysis of dynamical neural networks, *IEEE Trans. Neural Networks*, **7**, 996–1006.

[6] Gray, C.M., König, P., Engel, A.K., and Singer, W. (1989), Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties, *Nature*, **338**, 334–337.

[7] Haykin, S. (1999), *Neural Networks: A Comprehensive Foundation*, 2nd Edition, Prentice Hall, New Jersey.

[8] Jin, L. and Gupta, M. M. (1996), Globally asymptotical stability of discrete-time analog neural networks, *IEEE Trans. Neural Networks*, **7**, 1024–1031.

[9] Ott, E., Grebogi, C., and Yorke, J.A. (1990), Controlling chaos, *Phys. Rev. Lett.*, **64**, 1196–1199.

Sariyasa: Jurusan Pendidikan Matematika, IKIP Negeri Singaraja, Jl. A. Yani 67 Singaraja 81116, Bali – Indonesia.
E-mail: sariyasa64@yahoo.com

BARENBLATT'S SOLUTION: AN APPLICATION FOR DIFFUSION PROCESS OF AN IMPULSIVE SOURCE AND THE LIMITING CASE FOR NON-POROUS MEDIA

H. Budiman[a], E. Cahyono[a]

[a] Universitas Haluoleo, Kendari, Indonesia

Abstract. In this paper we discuss Barenblatt's solution of a diffusion equation in porous media and fundamental solution of the heat equation. We observe that in the limiting case Barenblatt's solution approaches fundamental solution. We exploit the sum of Barenblatt's solution (SBS) to model the diffusion process of the liquid that is produced by an impulsive source of pollution.
Key-words: heat equation, diffusion equation, Barenblatt's solution, fundamental solution

1 Introduction

The paper is motivated by a problem appearing in chemical industry. Often, a network of pipes installed in the ground under the factory. Chemical liquid is transported from one industrial process to the other via the pipes. There is a possibility that the pipes leak the liquid to the surrounding area, and may pollute the ground water underneath the factory as illustrated in Figure 1. This can be indicated by the chemical concentration of the ground water, which is continuously monitored at several observation wells inside the complex of the factory.

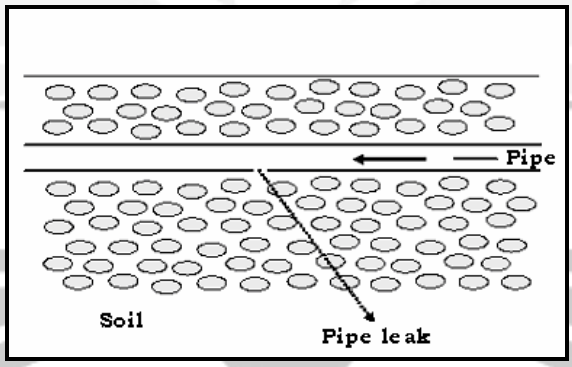


Figure 1. Illustrative plot of a pipe beneath the surface.

It is desirable to develop a method to find the source of pollution based on the historical data collected from the observation wells. To do so we assume the chemical pollution diffuses in a two-dimensional uniform medium. Hence, the diffusion of the pollution is governed by the diffusion equation in porous media. This model uses the so-called Barenblatt's solution. It exhibits singularity at the origin, interpreted as the source of mass (of pollution). The sum of Barenblatt's solutions (SBS) to the diffusion equation in porous will be a candidate to develop such method. Its singular point, interpreted as the position where the pipe,

continually leaks some amount of chemical 'mass'. This solution may provide a way to trace the source by formulating the issue into the so-called signaling problem.

The focus of this paper is the sum of Barenblatt's solution (SBS). We will exploit SBS to model the diffusion process of the liquid that is produced by an impulsive source of pollution. Moreover, we compare Barenblatt's solution to the fundamental solution of diffusion equation for nonporous media (heat equation).

2 Diffusion equation, Barenblatt's and fundamental solutions

We consider mass transfer in one-dimensional space ($\Re$). We write the spatial variable in the form $x \in \Re$, temporal variable $t > 0.$ The state variable $u = u(x,t)$ represents mass density at the point x and time *t*. The mass transfer in a one-dimensional non-porous medium satisfies the diffusion equation

$$u_t = \partial_x \left(K(u) \cdot u_x \right), \tag{1}$$

where $K(u)$ is the diffusion rate of the medium. Equation (1) was first derived by Fourier, recent derivation is given in [2] and for the case of constant diffusion rate can be found in some standard textbook such as [5:510-512].

Consider a constant diffusion rate and all variables in the normalized form without writing the straightforward. Based on this restriction, equation (1) becomes

$$u_t = u_{xx}. \tag{2}$$

The fundamental solution of (2) given by

$$u(x,t) = \frac{1}{\sqrt{t}} \exp\left(-\frac{x^2}{4t}\right), \text{ for } t > 0. \tag{3}$$

Note that the fundamental solution exhibits singularity at the origin at $t = 0$. Interpreting the origin as the source of mass, it releases an amount of mass to the surrounding area. At $t = 0$ the released mass occupies a single point (the origin), causing the mass density infinite.

On the other hand, the diffusion process in a porous medium in normalized form is given by (see [4])

$$u_t = \partial_x^2 \left(u^r \right), \tag{4}$$

where $r > 1$ is a parameter related to the pore size. Writing (4) in the form of (2) we have

$$u_t = \partial_x \left(r u^{r-1} \cdot u_x \right). \tag{5}$$

Hence, (4) is a special case of (1) for $K(u) = r u^{r-1}$. Observe that for r approaches 1, (4) tends to (2). It is natural to ask whether (4) has a solution that looks like (3), and approaches (3) as r approaches 1.

Equation (4) give a solution (see [4]) the so-called Barenblatt's solution in the form

$$u(x,t) = \frac{1}{t^{\left(\frac{1}{r+1}\right)}}\left[b - \left(\frac{r-1}{2r(r+1)}\right)\frac{x^2}{t^{\left(\frac{2}{r+1}\right)}}\right]^{\frac{1}{r-1}}.$$

where b is a constant, and the restriction on

$$b - \left(\frac{r-1}{2r(r+1)}\right)\frac{x^2}{t^{\left(\frac{2}{r+1}\right)}} > 0. \qquad (6)$$

The restriction is to guarantee the non-negative value of the state variable that represent the mass density. Solving (6) with respect to x gives

$$-x_1 < x < x_1.$$

where

$$x_1 = \frac{\sqrt{2br\left(r^2-1\right)t^{\frac{2}{r+1}}}}{r-1}.$$

Barenblatt's solution also has singularity at the origin at $t = 0$. Hence we may apply the same interpretation as fundamental solution by considering the restriction.

We now consider a modified Barenblatt's solution in the form

$$u(x,t) = \begin{cases} \dfrac{1}{t^{\left(\frac{1}{r+1}\right)}}\left[b - \left(\dfrac{r-1}{2r(r+1)}\right)\dfrac{x^2}{t^{\left(\frac{2}{r+1}\right)}}\right]^{\frac{1}{r-1}} & \text{for } -x_1 < x < x_1 \\ 0 & \text{for } x \le -x_1 \text{ or } x \ge x_1 \end{cases}.$$

Figure 2 shows plots of fundamental solution and modified Barenblatt's solutions for various value of r. The plots are taken for $b = 1$ and at time $t = 1$. Observe that for $r \to 1$ the Barenblatt's solution close to the fundamental solution.

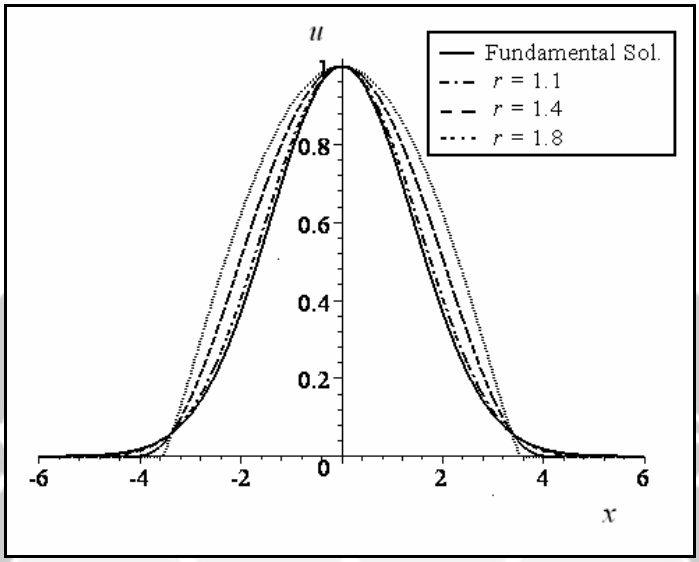


Figure 2. Fundamental and Barenblatt's solutions for various value of r.

On the other hand, the plot of Barenblatt's solution for various time is given in Figure 3. Interpreting u as the mass density, and the origin is a source that releases an amount of mass in a single pulse, the mass diffusion is as follows. At the time close to zero, the mass density at the origin is very large but it quickly spread out to the surrounding area.

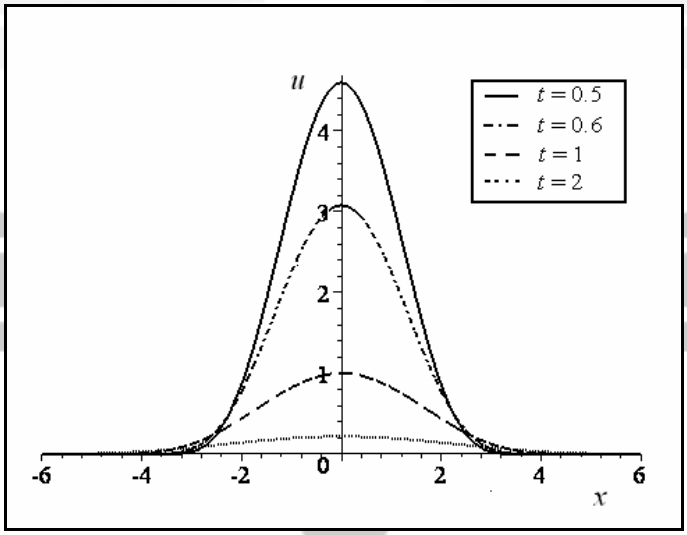


Figure 3. Barenblatt's solution for various time r=1.2.

3 The sum of Barenblatt's solution (SBS)

The discussion in this section will follow [1], but for more general case. We consider n-dimensional diffusion equation in porous. In normalized form it has the form

$$u_t - \Delta(u^r) = 0 \text{ in } \Re^n x(0,\infty) \tag{7}$$

where $u \geq 0$ and $r > 1$ constant. Barenblatt's solution of (6) is

$$u(x,t) = \frac{1}{t^{\alpha}}\left(b - \frac{r-1}{2r}\beta\frac{|x|^2}{t^{2\beta}}\right)^{\frac{1}{r-1}} \tag{8}$$

where

$$\alpha = \frac{n}{n(r-1)+2} \text{ and } \beta = \frac{1}{n(r-1)+2}.$$

Note that the restriction on $u \geq 0$ resulting in

$$|x|^2 \leq \frac{2brt^{2\beta}}{(r-1)\beta}.$$

Writing

$$R = \frac{2brt^{2\beta}}{(r-1)\beta},$$

we now define a modified Barenblatt's solution in $\Re^n$ of the form

$$u_n(x,t) = \begin{cases} \dfrac{a_n}{(t-t_n)^{\left(\frac{1}{r+1}\right)}}\left[1b - \left(\dfrac{r-1}{2r(r+1)}\right)\dfrac{|x|^2}{(t-t_n)^{\left(\frac{2}{r+1}\right)}}\right]^{\frac{1}{r-1}} & \text{for } |x|^2 < R \\ 0 & \text{for } |x|^2 \geq R \end{cases}$$

Observe that we have multiplied by a_n and translate the time by t_n.

Suppose at $t = t_1$ the source (the origin) starts to release an amount of mass in a single pulse. Before it release another amount of mass at $t = t_2$, i.e. $t_1 < t \leq t_2$, the mass transport is governed by

$$u_1(x,t).$$

At $t = t_2$ the source releases an amount of mass again, and in the interval $t_2 < t \le t_3$ the mass transport satisfies

$$u(x,t) = u_1(x,t) + u_2(x,t).$$

The first term of the right hand side represents the mass transport released by the source at $t = t_1$, and second term represents the mass transport released at $t = t_2$.

We now consider the condition that the source continually releases an amount of mass at $t = t_n$ for $n = 1,2,\cdots,N$. The associated mathematical expression of this condition is the sum of Barenblatt's solutions given by

$$u(x,t) = \sum_{n=1}^{N} u_n(x,t) \quad \text{for } t_N < t \le t_{N+1}.$$

4 Conclusion and further research

We have discussed Barenblatt's solution of a diffusion equation in porous media. We have also investigated that in the limiting case Barenblatt's solution approaches fundamental solution of heat equation. We propose the SBS to model the diffusion of mass impulsively released by a source in porous media. The SBS, however, continually exhibits singularity at the origin, i.e. when the source is releasing an amount of mass.

The future research will be focused on the application of the SBS to trace the source of pollution underneath the surface. This will be dealt with signaling problem.

Acknowledgment

The first author is partly supported by SP4-project, Jurusan Matematika FMIPA, Universitas Haluoleo.

References

[1] Cahyono, E. (2004), Modeling of mass transfer produced by impulsive and continuous sources, accepted in *Proceedings of Konferensi Nasional Matematika XII*, Denpasar, Indonesia.

[2] Cahyono, E. (2005), Modeling of wood drying: a step function approach to the diffusivity, *Proceedings of the 2nd International Conference on Research and Education in Mathematics*, Kuala Lumpur, Malaysia, Editors: Bekbaev, U.Dj., A. Kilicman, Z.K. Eshkuvatov, I.S.Rakhimov, H. Midi, M.R.Md. Said, W.Z.W. Ali, & Z. Abbas, 358 – 364.

[3] Evans, L.C. (1998), *Partial differential equations*, *Graduate Studies in Mathematics Vol. 19*, American Mathematical Society, Providence, Rhode Island.

[4] Hornung, U. (1997), Introduction, in *Homogenization and porous media*, Editor: Hornung, U., Springer-Verlag, Berlin, 1 – 23.

[5] Kreyszig, E. (1999), *Advanced engineering mathematics*, John Wiley & Son, Singapore.

H. BUDIMAN: Department of Mathematics, FMIPA, Universitas Haluoleo, Kampus Bumi Tridharma Anduonohu Kendari 93232, Indonesia. Phone
E-mail: herdi_budiman@unhalu.ac.id

E. CAHYONO: Department of Mathematics, FMIPA, Universitas Haluoleo, Kampus Bumi Tridharma Anduonohu Kendari 93232, Indonesia. Phone
E-mail: edi_cahyono@unhalu.ac.id

A QUASI-LINEAR DIFFUSIVITY APPROACH FOR DIFFUSION PROCESS OF LUMBER DRYING

La Gubu[a], E. Cahyono[a]

[a] Universitas Haluoleo, Kendari, Indonesia

Abstract. A good drying process is very much of interest of lumber and timber industries. The good process prevents the timbers from developing surface cracks and several other defects. Moisture Content (MC) of wood is an important aspects on wood drying. The drying process is described as an initial and boundary value problem. The MC is a function of spatial and temporal variables, and it is governed by a non-linear diffusion equation. In this paper the diffusion rate is approximated with a quasi-linear function. The solution is computed numerically by using a finite difference method. It is also shown that the solution is remarkably match with real data from industry.

Key-words: diffusion equation, wood, wood and lumber drying, finite difference method.

1 Introduction

A good drying process is very much of the interest of lumber and timber industries. This process may prevent the lumber from developing surface cracks and several other defects. It may reduce lumber weight by a factor two or more, which means a reduced transportation cost. It increases the lumber strength; nails, screws and glue hold better, paint and finishes adhere well. Dry lumber is a better thermal insulator than the wet one (see [1]).

The moisture content (MC) of lumber is an important aspect on lumber drying. MC of lumber is defined as the ratio of the mass of water contained in the lumber to the mass of the lumber without water. MC of some fresh log cut from a tree may be above 100%. Industries dry lumbers to have MC around 6% to 20%.

To have good control on lumber drying, middle and large sizes timber industries dry lumber in ovens. The moisture content of the lumber before entering the oven varies around 50% to 70%. In the drying process, the MC needs to be brought down to about 10%-15%. The drying process in the oven is done by controlling the Equilibrium Moisture Content (EMC), i.e. the air humidity in the oven. To make the process faster, the EMC should be lower, and vice versa.

Drying process decreases the cross-sectional dimension of the lumber up to ten percent. Lumber which is dried too quickly, leaving the surface much drier than the inside, may develop cracks on the surface. If one surface is drier than the other, the lumber may bend. A good drying process should not develop these mal-forms, except reducing dimension. Therefore a good process should dry the lumber evenly. Understanding this mechanism is extremely important to find an optimal drying time.

While drying process of the surface of lumber is directly controlled by setting the EMC, drying the inside part very much depends on the surface and also the type of lumber, hence is not easily controlled. This paper discusses modeling of the drying process of the inside part of the lumber due to the given EMC. It is intended to understand the process better.

2 Mathematical model

We consider a spatial variable $x \in \Re^n$, and the temporal variable $t > 0$. A state variable representing the mass density (of water) at the point x inside a bounded domain (lumber) and time t is denoted by $\rho = \rho(x,t)$.

$$\frac{\partial \rho}{\partial t} = \nabla \bullet (K \cdot \nabla \rho). \tag{1}$$

A derivation of equation (1) for the case K constant can be found in [6].

In the industrial application, what we mean lumber is actually a board, in which the length and the width are much larger than the thickness. In this case we may consider one-dimensional diffusion model with respect to the thickness. Industrial data suggests that the diffusive factor K is a function of ρ as reported by Cahyono [2]. In this paper we attempt to develop a simple, non-linear model. We consider a diffusion equation in one dimension of the form

$$\frac{\partial \rho}{\partial t} = \partial_x (K(\rho) \cdot \partial_x \rho). \tag{2}$$

The initial condition is given by

$$\rho = \rho_0 \qquad \text{in } 0 < x < d \text{ at } t = 0, \tag{3}$$

where d is the thickness of the lumber, and at the boundary we have a Dirichlet-type condition

$$\rho = f(t) \qquad \text{on } x = 0 \text{ and } x = d, \tag{4}$$

which represent the conditioning of the EMC inside the oven. Note that in this paper we neglect the shrinkage of the lumber during the process.

A simple example

We discuss a simple example in normalized form, where the domain is the interval $[0,1]$. Consider the problem with a constant diffusivity and boundary condition $f(t) \equiv 1$. Hence we have an initial and boundary value problem

$$\begin{aligned} \rho_t &= \rho_{xx} && \text{for } 0 < x < 1 \\ \rho &= 0 && \text{for } 0 < x < 1 \text{ and } t = 0 \\ \rho &= 1 && \text{for } x = 0 \text{ and } x = 1. \end{aligned}$$

An analytical solution can be computed using separation of variables (see [3]), yielding

$$\rho(x,t) = \sum_{n=1}^{\infty} \frac{2}{(2n-1)\pi} e^{-(2n-1)^2 \pi^2 t} \sin\big((2n-1)\pi x\big).$$

3 Quasi-linear approach to the diffusivity

We are interested in approximating the model by using a simple non-linear model with continuous diffusion rate, $K(u)$. This leads to a quasi-linear diffusivity, i.e. piece-wise linear $K(u)$, see Figure 1. For this approximation we try to find the value of $K(u)$ at a finite points and applying linear interpolation for intervals between those points. The choice of those points will be based on the comparison of the solution with the real data.

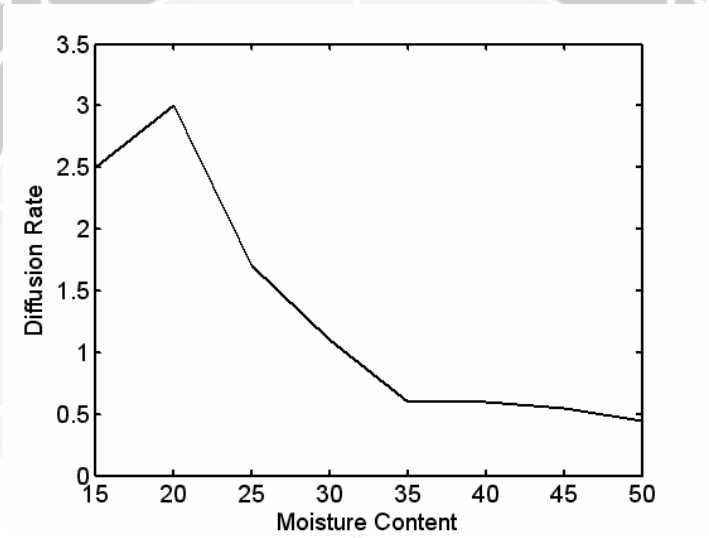


Figure 1. Piece-wise approximation to $K(u)$.

4 Numerical solution and comparison with real data

The blocks of lumber considered in industries have typical length and width that are much larger than the thickness. Hence, we may only consider one dimensional case, i.e. the thickness. The data is gathered from observation in lumber industry for durian wood (*Durio zibethinus*). The thickness of the lumber is d = 5 cm. The data consists of time, EMC and moisture content at the center of the lumber. Note that the measurement of MC at center of the lumber, however, includes its surrounding area, but small, about hundreds or thousands of the pore size of wood. Hence, the MC does not refer only at a single point, rather an average quantity in its neighborhood. This technique is known as Representative Elementary Volume (REV), see [4].

EMC can be automatically controlled. It depends on the temperature and the humidity inside the oven. A software is applied to set the temperature and control fans, so to have the desired EMC. The data was recorded for the period of 4 days during the drying process.

The numerical solution of the model (2) – (4) is computed using finite difference method. This method have been widely discussed in standard books such as [5]. In our finite difference scheme, the interval [0,4] is divided into partitions. For the numerical computation we use Δx =0.1, and the time step $\Delta t = 0.01$. Let the approximate solution be denoted by $\rho(x_i, t_j) = U_i^j$. The finite difference method gives

$$U_i^{j+1} = U_i^j + K(U_i^j) \cdot \frac{\Delta t}{(\Delta x)^2} \left(U_{i-1}^j - 2 \cdot U_i^j + U_{i+1}^j \right)$$

where the diffusivity $K(U_i^j)$ is a quasi-linear (or piecewise linear) function. The value at several points are given in Table 1, and the value between in the interval of two adjacent points of ρ is approximated linearly through those points.

Water Content (ρ)	**Diffusivity** ($K(\rho)$)
15	2.50
20	3.00
25	1.70
30	1.10
35	0.60
40	0.60
45	0.50
50	0.45

Table 1. The diffusion rate of the lumber approximated by a piecewise linear function.

The EMC which acts as the boundary condition for the numerical scheme is computed as follows. Let τ_k is *k*-th data of time which corresponds to the *k*-th data of EMC, say EMC(*k*). The boundary condition at $t_0 = \tau_0 = 0$ is $U_0^0 = U_N^0 = \text{EMC}(0)$. At $t = t_j$ and $j > 0$ there are two cases. The first, there is τ_k such that $\tau_k = t_j$. Hence, we apply $U_0^j = U_N^j = \text{EMC}(k)$. The second, $\tau_{k-1} < t_j < \tau_k$ for some $k \geq 0$. The boundary condition is approximated by

$$U_0^j = U_N^j = \frac{(\tau_k - t_j) \cdot \text{EMC}(k-1) + (t_j - \tau_{k-1}) \cdot \text{EMC}(k)}{\tau_k - \tau_{k-1}}.$$

Note that this approximation is just a linear interpolation of the EMC data.

Figure 2 shows MC of the numerical results and the real data and EMC from industry. The numerical solution gives a remarkable match with the real data. It shows a smoother process of drying, which is more desirable by the industries to improve their current methods of drying.

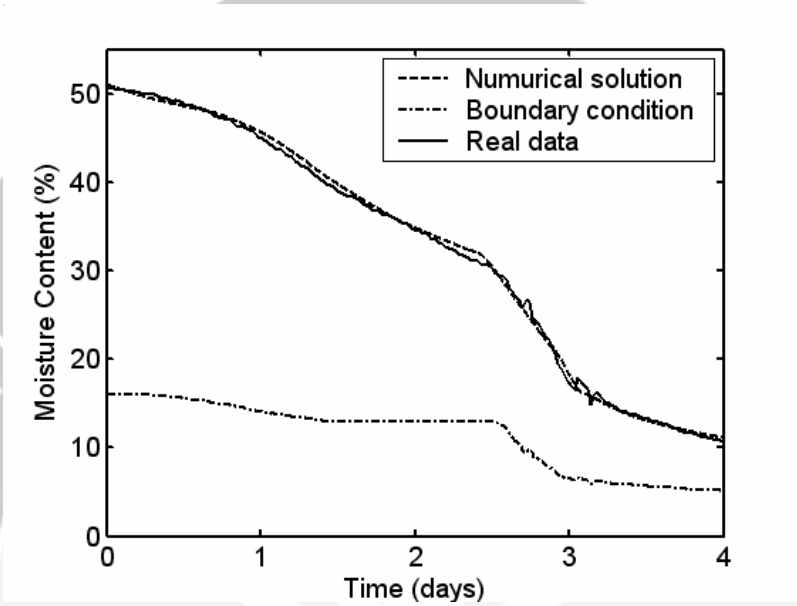


Figure 2. Numerical result (dashed line), MC (solid line) and EMC (dashed-dotted line).

5 Conclusion and further research

We have developed a simple non-linear model for lumber drying. This leads to a piecewise quadratic diffusion equation; the diffusion rate is a quasi-linear function of the state variable. While we do not solve the model explicitly using analytical tools, we can solve it numerically. The numerical solution gives a remarkable match with industrial data. It shows a smoother process of drying, which is desirable by the industries to improve their current methods of drying.

The model yields a diffusion rate function which is not smooth at finite points. The future research will be focused on looking for a smooth functions to approximate the diffusion rate.

Acknowledgment

The first author is partly supported by SP4-project, Jurusan Matematika FMIPA Universitas Haluoleo, 2005-2006.

References

[1] ____ (1988), *Dry kiln operator's manual*, Hardwood Research Council, U.S. Department of Agriculture, Memphis.

[2] Cahyono, E. (2005), Modeling of wood drying: a step function approach to the diffusivity, *Proceedings of the 2nd International Conference on Research and Education in Mathematics*, Kuala Lumpur, Malaysia, Editors: Bekbaev, U.Dj., A. Kilicman, Z.K. Eshkuvatov, I.S.Rakhimov, H. Midi, M.R.Md. Said, W.Z.W. Ali, & Z. Abbas, 358 – 364.

[3] Cahyono, E., D. C. Tjang & La Gubu (2003), Modeling of wood drying, *Proceedings of International Conference. on Mathematics. And Its Applications*, Yogyakarta, Indonesia, Editors: Aryati, L., Supama, B. Surodjo & Ch.R. Indrati, 227-233.

[4] Hornung, U. (1997), Introduction, in *Homogenization and porous media*, Editor: Hornung, U., Springer-Verlag, Berlin, 1 – 23.

[5] Morton, K. W. & D. F. Mayers (1996), *Numerical Solution of Partial Differential Equations*, Cambridge University Press, Cambridge.

[6] Kreyszig, E. (1999), *Advanced engineering mathematics*, John Wiley & Son, Singapore.

LA GUBU: Department of Mathematics, FMIPA, Universitas Haluoleo, Kampus Bumi Tridharma Anduonohu Kendari 93232, Indonesia. Phone: +62-401-391929
E-mail: yahoo2001@yahoo.com

E. CAHYONO: Department of Mathematics, FMIPA, Universitas Haluoleo, Kampus Bumi Tridharma Anduonohu Kendari 93232, Indonesia. Phone: +62-401-391929
E-mail: edi_cahyono@unhalu.ac.id

MODELING OF DIFFUSION PROCESS IN NON-LINEAR MEDIA

Mukhsar[a], E. Cahyono[a]

[a] Universitas Haluoleo, Kendari, Indonesia

Abstract. In this paper we discuss a non-linear diffusion problem appearing in a wood or lumber drying. The moisture content (MC) of the lumber, which is a function of spatial and temporal variables, is governed by a diffusion equation. Observation in industries shows that the diffusion rate depends on the MC. Its mathematical expression, however, is still unknown. We propose a method to approximate the diffusion rate numerically based on a Steepest Decent Method (SDM). This method exploits measurements of MC in the center of the lumber, and the Equilibrium Moisture Content (EMC) which acts as the boundary condition of the drying process.

Key-words: diffusion equation, wood, wood or lumber drying, steepest decent method

1 Introduction

Lumber drying is one of the most time- and energy-consuming in processing wood products. The anatomical structure of woods limits how rapidly water can move through and out of wood. In addition, the sensitivity of the structure to stresses set up in drying limit the drying rate; rapid drying causes defects such as surface and internal checks, collapse, splits, and warp, see [1].

Drying time and susceptibility to many drying defects increase as a rate that is more than proportional to wood thickness. The variability of wood properties farther complicates drying. Its species has different properties and even within species, variability in drying rate and sensitivity to drying defects impose limitations on the development of standard drying procedures. The interactions of wood, water, heat and stress during drying are complex.

The moisture content (MC) of lumber is an important aspect on lumber drying. MC of lumber is defined as the ratio of the mass of water contained in the lumber to the mass of the lumber without water. MC of some fresh log cut from a tree may be above 100%.

Drying process is governed by a diffusion equation. Observation in industries shows that the diffusion rate is a function of the MC, in [2] it approximated by a step function. The mathematical or numerical expression of the diffusion rate, however, is still unknown. The purpose of this paper is to develop a method to determine the diffusion rate of the wood. The diffusion rate is responsible for the drying time of wood. Therefore it is important for finding optimal drying time.

2 Problem formulation

We consider the drying process of a block of lumber. In industries it has typical length and width that are much larger than the thickness. Hence, we may assume the process occurs in one-dimensional medium with respect to the thickness. Without lost of generality, we assume the thickness is 2. Writing the MC as $u = u(x,t)$, inside the lumber u satisfies (see [2])

$$u_t = \partial_x (K(u)\partial_x u). \tag{1}$$

Measurements in industries usually provide EMC which acts as the boundary condition during the drying process

$$u(-1,t) = u(1,t) = f(t), \tag{2}$$

MC (in the center of the lumber)

$$u(0,t) = h(t) \tag{3}$$

and the MC when the drying starts

$$u(x,0) = C \tag{4}$$

which is the initial condition for process. The initial condition is desired to be constant. Based on (2) – (4) we would like to find $K(u)$, such that (1) is satisfied.

Solving this problem is very important for industries. This will help the industries to improve current methods of drying to obtain an efficient method which is fast but it does not create defects on the lumber. Having the analytical or numerical expressions of $K(u)$ will help to set EMC in order to obtain the desired MC before the drying is implemented.

3 Formulation of drying process

The drying process is somewhat different from the discussion in the previous section. The boundary condition (2) acts as a control, so the solution to the equation (1) behaves in a desirable way. In this section we consider special case for constant K, namely D. Hence, (1) becomes

$$u_t = D u_{xx}. \tag{5}$$

The boundary condition and the initial condition are given by (2) and (4), respectively. Note that $f(t)$ in (2) is assumed to be piecewise smooth. This case has been discussed in many standard books, e.g. [4,5].

To solve (5) for $u(x, t)$ with the conditions given by (2) – (4), we will follow [4,5]. First we consider the case of homogeneous boundary conditions, i.e. $f(t) \equiv 0$. The solution is sought in the form of a series of eigen functions of the homogeneous problem. Thus

$$u(x,t) = \sum_{n=1}^{\infty} u_n(t)\phi_n(x). \tag{6}$$

where $\phi_n(x)$ is normalized eigen functions of the Sturm-Liouville eigen value problem with the associated boundary conditions. Applying separation of variables, we have

$$u_n(t) = e^{-\lambda_n D t}.$$

Note that (6) and the derivatives u_x, u_{xx}, u_t converge uniformly in the interval $-1 < x < 1$.

To solve the original problem for nonzero $f(t)$, we transform the problem to the homogeneous boundary condition by defining a new function.

$$v(x,t) = u(x,t) - U(x,t) = u(x,t) - [A(t) + xB(t)],$$

where $U(x,t)$ is obtained from the steady state solution.

In order to solve the stated equation, we must have

$$v_t - Dv_{xx} = -[A'(t) + xB'(t)],$$
$$v(x,0) = C - [A(0) + xB(0)].$$

The function $A(t)$, $B(t)$ are chosen such that the linear function $A(t)+x\ B(t)$ satisfies the non homogeneous boundary conditions (2). This requires that

$$A(t) - B(t) = A(t) + B(t) = f(t).$$

This results in

$$A(t) \equiv 0, \text{ and } B(t) = \frac{f(t)}{2}.$$

As for the homogeneous boundary conditions, we consider the trial solution in the form

$$v(x,t) = \sum_{n=1}^{\infty} v_n(t)\,\phi_n(x).$$

Hence, the solution of the given problem is

$$u(x,t) = U(x,t) + \sum_{n=1}^{\infty} v_n(t)\,\phi_n(x),$$

or

$$u(x,t) = \frac{1}{2} x f(t) + \sum_{n=1}^{\infty} v_n(t)\,\phi_n(x).$$

4 Numerical scheme for computing the diffusion rate

In this part, we recall again the main problem discussed is section 2. To solve the problem we define a function $F(K(u))$ in the form

$$F(K(u)) = u_t - \partial_x\left(K(u)\cdot\partial_x u\right).$$

We are looking for the diffusion rate $K(u)$, such that the function $F(K(u))$ minimum.

To solve this we will apply a steepest descent method as discussed in [3,6]. Suppose we are given a point x^{k}. To find the next point $\mathrm{x}^{\mathrm{k+1}}$, we start at x^{k} and move by an amount and the direction of $-\alpha\nabla f(\mathrm{x}^{\mathrm{k}})$, where α_k is a positive scalar called the step size. The above procedure leads to the following iterative algorithm

$$\mathrm{x}^{\mathrm{k+1}}=\mathrm{x}^{\mathrm{k}}-\alpha_k\nabla F(\mathrm{x}^{\mathrm{k}}).$$

Observe that the method of steepest descent moves in orthogonal steps, as stated in the following theorem.

Theorem 3.1. *If* $\left\{\mathrm{x}^{k}\right\}_{k=0}^{\infty}$ *is a steepest descent sequence for a given function* $F:\Re^n\to\Re$, *then for each k the vector* $\mathrm{x}^{k+1}-\mathrm{x}^{k}$ *is orthogonal to the vector* $\mathrm{x}^{k+2}-\mathrm{x}^{k+1}$.

Proof of theorem 3.1 detailed in [3].

The above theorem 3.1 implies that $\nabla F(\mathrm{x}^{k})$ is parallel to the tangent plane to the level set $\left\{F(\mathrm{x})=F(\mathrm{x}^{\mathrm{k+1}})\right\}$ at $\mathrm{x}^{\mathrm{k+1}}$. Note that as each new point is generated by the steepest descent algorithm, the corresponding value of the function F decreases in value.

Theorem 3.2. *If* $\left\{\mathrm{x}^{k}\right\}_{k=0}^{\infty}$ *is the steepest descent sequence for* $F:\Re^n\to\Re$ *and if* $\nabla F(\mathrm{x}^{k})\neq 0$ *then* $F(\mathrm{x}^{\mathrm{k+1}})<F(\mathrm{x}^{\mathrm{k}})$.

Proof of theorem 3.2 detailed in [3].

In the theorem 3.2, we used the assumption that $\nabla F(\mathrm{x}^{k})\neq 0$ to prove that the algorithm possesses a descent property, that is $F(\mathrm{x}^{\mathrm{k+1}})<F(\mathrm{x}^{\mathrm{k}})$ if $\nabla F(\mathrm{x}^{k})\neq 0$, if for some k, we have $\nabla F(\mathrm{x}^{k})=0$, then the point x^{k} satisfies the First-Order Necessary Condition (FONC). In this case $\mathrm{x}^{\mathrm{k+1}}=\mathrm{x}^{\mathrm{k}}$, we can use the theorem 3.2 as the basis for a stopping (termination) criterion for the algorithm.

The condition $\nabla F(\mathrm{x}^{k+1})=0$, however is not directly suitable as a practical stopping criterion, since the numerical computation of the gradient will rarely be identically equal to zero. A practical stopping criterion is to check if the norm

$\left\|\nabla F(\mathrm{x}^k)\right\|$ of the gradient is less than a prescribed value, in which case we stop. Alternatively, we may compute the absolute difference $\left|F(\mathrm{x}^{k+1})-F(\mathrm{x}^k)\right|$ between objective function values for every two successive iterate and if the difference is less than a prescribed value then we stop.

5 Conclusion and future research

We have discussed a problem of determining the diffusion rate of lumber drying which depends on the MC. This leads to a non-linear diffusion equation. We propose a method to compute the diffusion rate by exploiting data that is available from industries. The data consists of MC (the value of the state variable) in the center of the lumber, the initial MC (the initial condition) and the EMC (the boundary condition). The future research will be focused on the implementation of the proposed method.

Acknowledgment

The first author is partly supported by the IAEUP (FK8PT) project. He also would like to thank Ir. F.S. Rembon, M.Sc. for the arrangement of the sponsorship.

References

[1] ____ (1988), *Dry kiln operator's manual*, Hardwood Research Council, U.S. Department of Agriculture, Memphis.

[2] Cahyono, E. (2005), Modeling of wood drying: a step function approach to the diffusivity, *Proceedings of the 2nd International Conference on Research and Education in Mathematics*, Kuala Lumpur, Malaysia, Editors: Bekbaev, U.Dj., A. Kilicman, Z.K. Eshkuvatov, I.S.Rakhimov, H. Midi, M.R.Md. Said, W.Z.W. Ali, & Z. Abbas, 358 – 364.

[3] Chong, E.K.P. & H.Z. Stanislaw (1996), *An introduction to optimization*, John Wiley & Sons, Inc.

[4] Erick, Z. (1989), *Partial differential equation of applied mathematics*, Second edition, John Wiley & Sons.

[5] Mark, A.P. (1998), *Partial differential equations and boundary-value problems with applications*, Third edition, McGraw-Hill, Singapore.

[6] Rao, S.S. (1995), *Optimization, theory and application*, Second edition, New age international(P) limited, India.

MUKHSAR: Department of Mathematics, FMIPA, Universitas Haluoleo, Kampus Bumi Tridharma Anduonohu Kendari 93232, Indonesia. Phone: +62-401-391929
E-mail: mukhsar@unhalu.ac.id

E. CAHYONO: Department of Mathematics, FMIPA, Universitas Haluoleo, Kampus Bumi Tridharma Anduonohu Kendari 93232, Indonesia. Phone: +62-401-391929
E-mail: edi_cahyono@unhalu.ac.id

Unsteady Viscous Incompressible Flow In A Porous Annulus Subject To Slip At The Inner Surface

Nirmal C. Sacheti, E. A. Hamza, S. C. Rajvanshi

Dept. of Math. & Statistics, College of Science, Sultan Qaboos University, Sultanate of Oman

Abstract: A fully developed, axisymmetric, pulsatile motion of an incompressible Newtonian fluid, under the action of an oscillatory pressure gradient, has been considered in this work. The flow is assumed to take place in the annular space between two coaxial circular cylinders, the outer one a porous cylinder of uniform permeability and the inner one a naturally permeable tube. There arises a coupled flow, which has been analysed by solving the Navier-Stokes equations in the free space region and the Darcy's equation in the porous region, together with the Beavers-Joseph slip condition at the free fluid-porous medium interface. Using an appropriate set of similarity variables, the governing partial differential equations have been transformed to a system of nonlinear ordinary differential equations. The solution of the resulting system, together with appropriate boundary conditions, has been obtained, for a special case, by a perturbation approach. It has been assumed that both the frequency of pulsation and the suction parameter are small (<< 1). The variation of velocity profiles and pressure drop has been illustrated in a number of cases of interest. Some preliminary results using a numerical method to solve the system will also be presented.

Keywords: Mathematical Physics; Porous Media

Boundary Layer Flow over a Stretching Sheet in a Porous Medium Filled with a Micropolar Fluid

Sharidan Shafie[1], Norsarahaida Amin[1], Ioan Pop[2]

[1)] Department of Mathematics, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia
[2)] Faculty of Mathematics, University of Cluj, Romania

Abstract: The forced convection boundary layer flow over a stretching sheet in a porous medium filled with an incompressible micropolar fluid is investigated. The governing boundary layer equations are transformed into a numerically equivalent system of nonlinear ordinary differential equations. The resulting equations are solved numerically using the implicit finite-difference scheme known as Keller-box method. The flow pattern depends on two non-dimensional parameters and a material parameter Δ. A comparison between the numerical solution and the analytical solution when $\Delta = 0$ (Newtonian fluid) is presented with the results being shown in a table.

Keyword: boundary layer, stretching sheet, porous medium, micropolar fluid, numerical results

THE CHANGE OF WOOD DIMENSION DEPENDING ON THE MOISTURE CONTENT: A CRITICAL STEP ON MODELING OF LUMBER DRYING

E. Cahyono[a]

[a] Universitas Haluoleo, Kendari, Indonesia

Abstract. In this paper a model for wood or lumber drying is proposed, based on the diffusion process of the moisture content (MC) of the lumber. The length and width of the lumber are considered to be much larger than its thickness, hence we consider only diffusive process in the direction of the thickness, and therefore we apply a one-dimensional diffusion equation. The improvement to the current model is on taking into account the change of wood dimension during the process. A numerical scheme based on a finite difference method is proposed. It is also proved that for fixed media and a constant diffusion rate, the model is the standard diffusion equation.
Key-words: diffusion equation, wood, wood and lumber drying, finite difference method.

1 Introduction

For various reasons, a good drying process is very much of the interest of lumber and timber industries. A good drying process may prevent the lumber from developing surface cracks and several other defects. It may reduce lumber weight by a factor two or more, which means a reduced transportation cost. It increases the lumber strength; nails, screws and glue hold better, paint and finishes adhere well. Dry lumber is a better thermal insulator than the wet one [1].

The moisture content (MC) of lumber is an important aspect on lumber drying. MC of lumber is defined as the ratio of the mass of water contained in the lumber to the mass of the lumber without water. MC of some fresh log cut from a tree may be above 100%.

Middle and large sizes timber industries dry lumber in ovens. The moisture content of the lumber before entering the oven varies around 50% to 70%. In the drying process, the MC needs to be brought down to about 10%-15%. The drying process in the oven is done by controlling the Equilibrium Moisture Content (EMC), i.e. the air humidity in the oven. To make the process faster, the EMC should be lower, and vice versa.

The drying process starts from the surface of the lumber and goes to the inside. It decreases the cross-sectional dimension of the lumber up to ten percent. Lumber which is dried too quickly, leaving the surface much drier than the inside, may develop cracks on the surface. If one surface is drier than the other, the lumber may bend. A good drying process should not develop these mal-forms, except reducing dimension. This is illustrated in Figure 1. Therefore a good process should dry the lumber evenly. Understanding this mechanism is extremely important to find an optimal drying time. This paper focused on the developing a

model that taking into account the change of lumber dimension during the drying process which is responsible for the mal-forms of the lumbers.

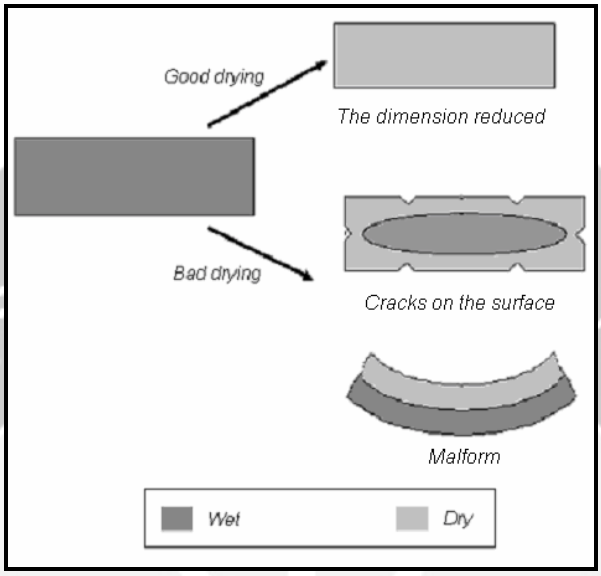


Figure 1. The change of wood dimension during the drying process.

2 Mathematical model

We consider $(q(t),t) \in \Re^n \times \Re$ representing spatial and temporal variables, respectively. The state variable $\rho = \rho(q(t),t)$ stands for the moisture content at time t and at the point $q(t)$. At $q(t)$ the flux of the moisture content to the surrounding area, denoted by v, is assumed to be proportional to the gradient of ρ with respect to $q(t)$ plus ρ brought by the rate of motion of $q(t)$, see Figure 2. Hence, we have

$$v = K \cdot \nabla\rho + \rho \frac{dq}{dt} \tag{1}$$

Note that in general K is not constant, but it may be a function of $q(t)$ and ρ.

We now consider a bounded set in $\Re^n$, at time t it is denoted by $\Omega(t)$ and the boundary is by $\partial\Omega(t)$. The total mass leaving the region is therefore the integral of the normal component (1) along the boundary

$$\int_{\Omega(t)} \left(K \cdot \nabla\rho + \rho \frac{dq}{dt} \right) \bullet \mathbf{n}\, dA .$$

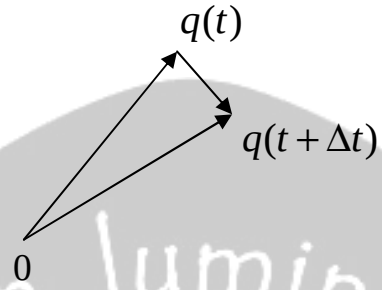


Figure 2. Point displacement as a function of time.

On the other hand, the total mass in $\Omega(t)$ is the integral of the density over the region

$$m = \int_{\Omega(t)} \rho \, dV \, .$$

The decrease of mass in $\Omega(t)$ is given by

$$\frac{\partial m}{\partial t} = \int_{\Omega(t)} \frac{\partial \rho}{\partial t} dV + \int_{\Omega'(t)} \rho \, dV \, .$$

Assuming the mass conserved, the decrease of mass in $\Omega(t)$ is equal to the mass leaving the boundary. Hence, we have

$$\int_{\partial\Omega(t)} \left(K \cdot \nabla\rho + \rho \frac{dq}{dt} \right) \bullet \mathbf{n} \, dA = \int_{\Omega(t)} \frac{\partial \rho}{\partial t} dV + \int_{\Omega'(t)} \rho \, dV \, . \tag{2}$$

Applying Gauss Divergence Theorem, (2) can be written in the form

$$\int_{\Omega(t)} \left(\frac{\partial \rho}{\partial t} - \nabla\left(K \cdot \nabla\rho\right) \right) dV = \int_{\partial\Omega(t)} \rho \frac{dq}{dt} \bullet \mathbf{n} \, dA - \int_{\Omega'(t)} \rho \, dV \, . \tag{3}$$

Observe that the presence of the right hand side of (3) is due to the change in the dimension during the diffusion process. If there is no change of the dimension during the process, $\Omega(t) = \Omega_0$ and $q(t) = x$ are hold. Therefore, we have $dq/dt = 0$ and $\Omega'(t) = 0$. In this case the right hand side of (3) is vanished. This results in

$$\int_{\Omega_0} \left(\frac{\partial \rho}{\partial t} - \nabla\left(K \cdot \nabla\rho\right) \right) dV = 0 \, . \tag{4}$$

Assuming the integrand is sufficiently smooth and (4) are satisfied for any region Ω_0 in $\Re^n$, it gives the well-known diffusion equation

$$\frac{\partial \rho}{\partial t} = \nabla(K \cdot \nabla \rho). \tag{5}$$

A simpler derivation of (5) for the case K constant can easily be found in standard textbooks, e.g. [6]. Observation in industries, however, shows that K is a function of ρ, see [2]. In this case, equation (5) is more general than diffusion equation for porous media discussed in [4].

Note:
In industrial applications, the measurement of MC at the point $q(t)$ includes its surrounding area, but small, about hundreds or thousands of the pore size of wood. Hence, the state variable does not refer only at a single point, rather an average quantity in the neighborhood of $q(t)$. This technique is known as Representative Elementary Volume (REV), see [5].

3 One-dimensional case

Lumbers considered in industries have typical length and width that are much larger than the thickness. Hence, the diffusion may be assumed as a one-dimensional process with respect to the thickness. Equation (3), therefore, becomes

$$\int_{I(t)} \left(\frac{\partial \rho}{\partial t} - \partial_q \left(K \cdot \partial_q \rho \right) \right) dq = \begin{pmatrix} \rho(b_I(t), t) \cdot b_I'(t) \\ -\rho(a_I(t), t) \cdot a_I'(t) \end{pmatrix} - \int_{I'(t)} \rho \, dq \,. \tag{5}$$

We have used notation $I(t) = [a_I(t), b_I(t)]$ for any interval along the thickness of the lumber. We will also use $q_1(t)$ and $q_2(t)$ to denote the end points of the lumber thickness at time t. Without lost of generality, at $t = 0$ we consider $q_1(0) = -1$ and $q_2(0) = 1$. The function $q_1(t)$ and $q_2(t)$ may be recorded from the drying process in industry.

During the drying process, the boundary condition is given by the EMC which is a function of time. Hence, we have

$$\rho(q_1(t), t) = \rho(q_2(t), t) = f(t). \tag{6}$$

On the other hand, $t = 0$ the MC of the lumber should be homogeneous which results in the initial condition

$$\rho(q, 0) = 1, \text{ for } -1 < q < 1. \tag{7}$$

A simple example
We discuss a simple example in normalized form, where the domain is the interval $[0,1]$. Consider the problem with a constant diffusivity and boundary condition $f(t) \equiv 1.$ Hence we have an initial and boundary value problem

$$\begin{aligned} \rho_t &= \rho_{xx} && \text{for } 0 < x < 1 \\ \rho &= 1 && \text{for } 0 < x < 1 \text{ and } t = 0 \\ \rho &= 0 && \text{for } x = 0 \text{ and } x = 1. \end{aligned}$$

An analytical solution can be computed using separation of variables, yielding

$$\rho(x,t) = \sum_{n=1}^{\infty} \frac{2}{(2n-1)\pi} e^{-(2n-1)^2 \pi^2 t} \sin\left((2n-1)\pi x\right).$$

This equation can be interpreted as the dynamics of the MC inside the lumber, see [3].

4 Proposed numerical scheme

To solve (1) for the given boundary and initial conditions (6) and (7) we propose a method as follows. First, we rewrite (5) in the form

$$\frac{\partial \rho}{\partial t}\left(\int_{I(t)} \rho \, dq \right) = \int_{I(t)} \left(\partial_q \left(K \cdot \partial_q \rho\right)\right) dq + \begin{pmatrix} \rho\left(b_I(t),t\right) \cdot b_I'(t) \\ -\rho\left(a_I(t),t\right) \cdot a_I'(t) \end{pmatrix}. \quad (8)$$

Divide interval $[q_1(t), q_2(t)]$ into M partitions which are evenly spaced with nodes $q_1(t) = a_0 < a_1 < a_2 < \cdots < a_M = q_2(t)$. The width of the partitions is $h(t)$. Note that $a_m = a_m(t)$ a function of time for $m = 1,2,\cdots,M$. Instead of working on any interval in $[q_1(t), q_2(t)]$, rather we work on the interval $I_m = [a_m, a_{m+1}]$. We also write $a_m(t_n) = a_m^n$, $h(t_n) = h^n$, $\rho\left(a_m^n, t_n\right) = \rho_m^n$ and $K\left(\rho_m^n\right) = K_m^n$. The time step is written in the form $t_n = t_0 + n \cdot \Delta t$.

Assuming that the value of constant in each interval, we have the following approximation;

$$\frac{\partial}{\partial t}\left(\int_{I_m} \rho \, dq \right) = \frac{\partial}{\partial t}\left(\rho_m^n\right) = \frac{\rho_m^{n+1} - \rho_m^n}{\Delta t},$$

$$\int_{I_m} \left(\partial_q \left(K \cdot \partial_q \rho\right)\right) dq = K^n_{m+1}\left(\frac{\rho^n_{m+1} - \rho^n_m}{h^n}\right) - K^n_m\left(\frac{\rho^n_m - \rho^n_{m-1}}{h^n}\right)$$

and

$$\begin{pmatrix} \rho(a_{m+1}(t_n), t_n) \cdot a_{m+1}'(t_n) \\ -\rho(a_m(t_n), t_n) \cdot a_m'(t_n) \end{pmatrix} = \rho^n_{m+1}\, q'^n_{m+1} - \rho^n_m\, q'^n_m .$$

Applying this approximation leads to the numerical scheme to solve (8) in the form

$$\rho^{n+1}_m = \rho^n_m \frac{h^n}{h^{n+1}} + \frac{\Delta t}{h^{n+1}}\left[\begin{array}{l} K^n_{m+1}\left(\dfrac{\rho^n_{m+1} - \rho^n_m}{h^n}\right) - K^n_m\left(\dfrac{\rho^n_m - \rho^n_{m-1}}{h^n}\right) \\ + \rho^n_{m+1}\, q'^n_{m+1} - \rho^n_m\, q'^n_m \end{array}\right]. \quad (9)$$

For the case of a constant diffusion rate and the diffusion process occurs in fixed media, we have $h^{n+1} = h^n \equiv h$ and $q'^n_{m+1} = q'^n_m \equiv 0$. This leads to the well-known finite difference method for a standard diffusion equation

$$\rho^{n+1}_m = \rho^n_m + \frac{\Delta t}{h^2} K\left[\rho^n_{m+1} - 2\rho^n_m + \rho^n_{m-1}\right],$$

which is widely discussed in many standard textbooks, e.g. [7].

5 Conclusion and further research

We have developed a model for diffusion process of lumber drying by taking into account the shrinkage of the lumber thickness. This model leads to an integral equation that is more general than the (integral) equation which is the 'primitive' of the well-known diffusion equation.

We have proposed a numerical scheme to solve the model based on a finite difference method. For the case of a constant diffusion rate and the diffusion process occurs in fixed media, the scheme is nothing else but the standard finite difference method for the standard diffusion equation.

The future research will be focus on several topics. The first is on the comparisons the numerical results with real data recorded from industry. The second will be considering the point displacement directly depends on the state variable, as observed in industry.

References

[1] ____ (1988), *Dry kiln operator's manual*, Hardwood Research Council, U.S. Department of Agriculture, Memphis.

[2] Cahyono, E. (2005), Modeling of wood drying: a step function approach to the diffusivity, *Proceedings of the 2nd International Conference on Research and Education in Mathematics*, Kuala Lumpur, Malaysia, Editors: Bekbaev, U.Dj., A. Kilicman, Z.K. Eshkuvatov, I.S.Rakhimov, H. Midi, M.R.Md. Said, W.Z.W. Ali, & Z. Abbas, 358 – 364.

[3] Cahyono, E., D. C. Tjang & La Gubu (2003), Modeling of wood drying, *Proceedings of International Conference. on Mathematics. And Its Applications*, Yogyakarta, Indonesia, Editors: Aryati, L., Supama, B. Surodjo & Ch.R. Indrati, 227-233.

[4] Evans, L.C. (1998), *Partial differential equations*, *Graduate Studies in Mathematics Vol. 19*, American Mathematical Society, Providence, Rhode Island.

[5] Hornung, U. (1997), Introduction, in *Homogenization and porous media*, Editor: Hornung, U., Springer-Verlag, Berlin, 1 – 23.

[6] Kreyszig, E. (1999), *Advanced engineering mathematics*, John Wiley & Son, Singapore.

[7] Morton K. W. & D. F. Mayers (1996), *Numerical Solution of Partial Differential Equations*, Cambridge University Press, Cambridge.

E. CAHYONO: Department of Mathematics, FMIPA, Universitas Haluoleo, Kampus Bumi Tridharma Anduonohu, Kendari 93232, Indonesia. Phone: +62-401-391929
E-mail: edi_cahyono@unhalu.ac.id

Numerical Reconstruction of Electrical Impedance Tomography using Levenberg-Marquardt Algorithms with a-posteriori parameter choice rule

A.D. Garnadi

Dept. of Mathematics, FMIPA-IPB. Jl Meranti, Kampus IPB Darmaga, Bogor

Abstract: We report a numerical study of image reconstructions from Electrical Impedance Tomography. The reconstructions based on Levenberg-Marquardt Algorithms with a parameter choice rule.

The algorithms which already shown to be monotonic for a class of non-linear inverse problems; is applied to EIT which happens to be falls into the same classes of the nonlinear inverse problems. We adopt an a-posteriori strategy in choosing the parameter within the algorithms which not yet investigated.

Keywords : Inverse Problems, nonlinear ill-posed, Reconstruction Algorithms

INTEGER WAVELET TRANSFORM FOR DISTRIBUTED LOSSLESS MEDICAL IMAGE COMPRESSION

P. Rahmiati[a], A.B. Suksmono[a], T.L.R. Mengko[b], A.Handayani[b], A. Fajri[b]

[a] Laboratorium Telekomunikasi Radio dan Gelombanbg Mikro ITB, Indonesia
[b] Imaging and Image Processing Research Group ITB, Indonesia

Abstract. This paper presents a distributed lossless medical image compression using integer wavelet transform (IWT). We will investigate the capability of IWT in JPEG2000 coding standard to provide a scalable image coding and also we will compare the time required to conclude the process between distributed and non-distributed system.
A medical image usually has a large amount of data in order to get a high image quality which is crucial to support a right diagnosis; therefore we need to compress the data in order to lessen storage memory and bandwidth transmission.

IWT is a reversible operation; the image can be fully reconstructed from the integer transform coefficients. In JPEG 2000 coding standard, IWT is used as the standard transformer, it can compress the image losslessly.

IWT codes the image into several resolution layers. When the image is synthesized from its transform coefficients, each coefficient contributes only to a specific region in the reconstruction so hat we can choose only a specific significant area (Region of Interest, RoI).

In this paper the system will be implemented in distributed and non-distributed client-image server architecture. On the first laboratory experiment, evaluation on the quality refinement on 32×32 to 256×256 pixels Rio on a non-distributed system shows that this scheme has considerably reduced the number of bytes transferred from server to client during one single image access period, while maintaining good image quality on user-defined medically-significant areas (Rio). And on the second laboratory experiment, evaluation on the required time to process a 1024x1024 pixels image shows that the distributed system requires less processing time than the non-distributed one.
Key-words: Lossless, medical image, scalable coding, integer wavelet transform, JPEG 2000, RoI.

1 Introduction

Wavelet transform is a tool for the analysis of transient, or time-varying, non-stationary phenomena. It specifies the frequency content of *f(t)* as a function of *t*. The main idea is to select a mother wavelet, "a small" wave which has its energy concentrated in time, and use it to explore the properties of *f(t)* in an interval. The mother wavelet is then translated to another interval of *t* and used in the same way. Different resolutions of *f(t)* are explored by scaling the mother wavelet.

IWT transforms integer inputs sequences into integer output sequences of the same length, and it is perfect reconstruction in that the IWT is reversible, i.e. the original image can be recovered exactly.

Focus of a medical analysis is commonly limited on a specific area of an image which is significant to the diagnosis, therefore a perfect reconstruction of digital medical image is actually crucial only on this part. Quality-layered image representation provided by wavelet techniques in JPEG2000 image coding standard makes it possible to encode and decode a specific part of the image (RoI) in a better quality compared to the rest.

Distributed system proposed in this paper is set up in clients-image server architecture which image data bank is put on the server-side and the encoding applications are put on several clients which work in parallel way in the middle of the system. End client-side application request images from server while providing supports for interactive RoI selection and carrying out determination of resolution layers which correspond to the RoI. The difference between the distributed and the non-distributed system is that there is no several clients in the middle of the non-distributed system so that image data bank and encoding applications are put on the server side.

2. Methods

2.1 Wavelets

A wavelet is a "small wave", which has its energy concentrated in time to give a tool for the analysis of transient, non-stationary, or time-varying phenomena. It is a mathematical concept to decompose a function $f(t)$ into sets of other functions known as wavelet bases $\psi_{a,b}(t)$.

$$f(t) = \sum_{t} a_{a,b} \psi_{a,b}(t)$$

We need to use a suitable family of functions $\psi_{a,b}(t)$ in order to get an efficient representation of the signal *f* using only a few coefficients $a_{a,b}$. The wavelet bases $\psi_{a,b}(t)$ should match the features of the data we want to represent.

In order to get the variable time-frequency resolution, we have to define a mother wavelet or prototype function ψ (*t*). Basis functions , $\psi_{a,b}(t)$ are the scaled and translated version of the prototype.

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi \frac{(t-a)}{b}$$

Here, a is the translating coefficient and b is the scaling coefficient.

Resembles with the Fourier Transform Theorem, we have the *classic wavelet transform*:

$$f(t) = \sum_{a,b} a_{a,b} \psi_{a,b}(t)$$

$$\text{or} \quad a_{a,b} = \int_{-\infty}^{\infty} f(t) \psi_{a,b}(t)$$

2.2 Multi Resolution Analysis

MRA analyzes the signal at different frequencies with different resolutions.

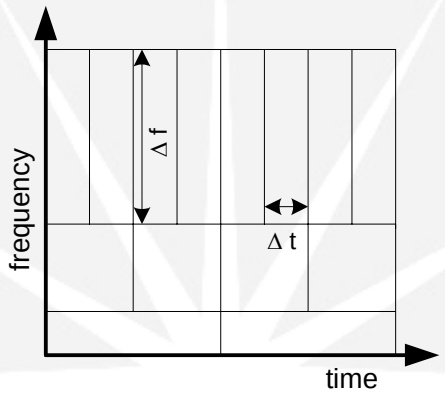


Figure 1. Multiresolution analysis on wavelet transform

The time and frequency resolution can not be arbitrary small, but according to the Heisenberg uncertainty principle:

$$\Delta(t)\Delta(f) \geq \frac{1}{4\pi}$$

we can see that there is a tradeoff between time resolution and frequency resolution. At low frequencies, the frequency resolution is better but the time resolution is poorer and adversely at higher frequencies the time resolution gets better but the frequency resolution gets poorer.

Wavelet transform of a signal is a multiresolution representation of that signal where the wavelets are the basis functions, which at each resolution level de-correlate the signal.

2.3 Generating Wavelets Using Two-Channel Filterbanks

Using a 2 channel filterbanks called synthesis, we can generate wavelet transform coefficients.

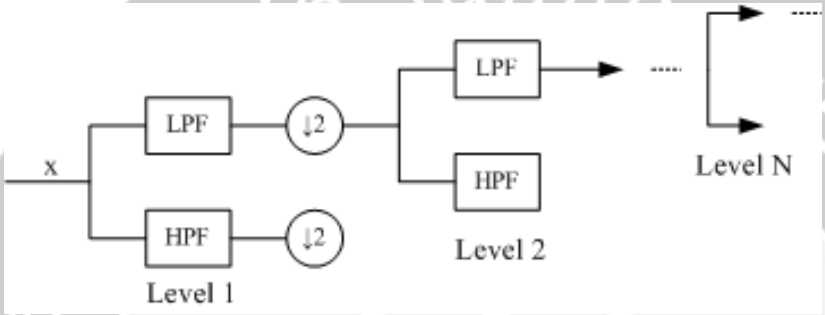


Figure 2. N-level wavelet transform of one-dimensional signal

In each decomposition level, we pass the signal through a lowpass filter (LPF) and a highpass filter (HPF), and then down sampling the output of each filter. Figure 2 depicts N-level wavelet decomposition of one-dimensional signal and Figure 3 depicts N-level wavelet reconstruction of a one-dimensional signal.

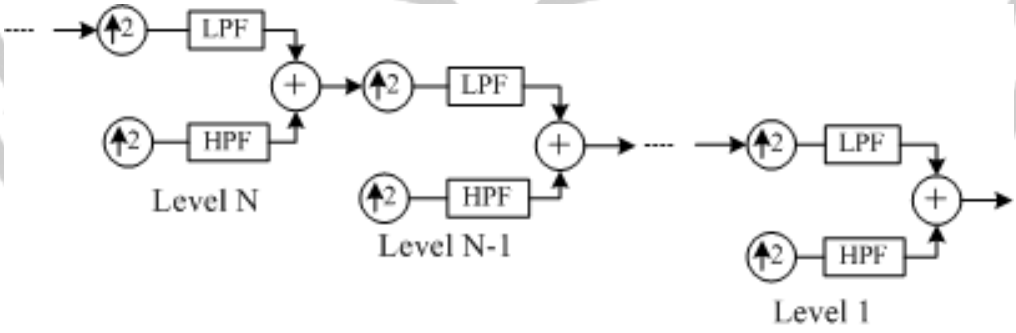


Figure 3. N-level wavelet reconstruction of a one-dimensional signal

For a two dimensional image, one level decomposition procedure consists of one dimensional image data filtering in the direction of row followed by one dimensional image data filtering in the direction of columns utilizing wavelet low pass and high pass analysis filter. Detailed wavelet decomposition diagram is shown in Figure 4.

Main information of each level is contained in its lowest level frequency subband image (LL); further addressed as approximation component. The three consecutive higher frequency subband images build up detail components; each contains horizontal detail (LH), vertical detail (HL), and diagonal detail (HH). Subsequent level subband images are generated by performing similar decomposition on approximation component of the previous decomposition level, such that approximation component of decomposition level d is a reduced resolution version of the original image, having width and height reduced by a factor of 2^d. Complete image representation for every level is synthesized from its approximation component altogether with the three corresponding detail components. Varying the subband components taken into account in image synthesis process with regards to either their decomposition level or their spatial entities consequently will produce various representations of image quality and resolution.

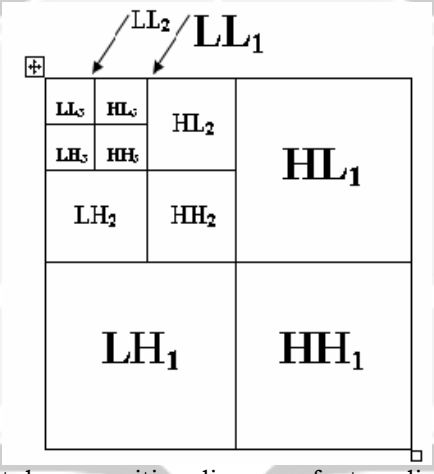


Figure 4. Wavelet decomposition diagram of a two-dimensional image.

2.4 Perfect Reconstruction

For signal analysis, a wavelet $\psi(t)$ is considered to be a bandpass windows function that stops at least the zero frequency.

$$\int_{-\infty}^{\infty} \psi(t)dt = \hat{\psi}(0) = 0$$

This property enables the transform to annihilate the flat segments of analog signal to provide a better understanding of its details.

If $\psi(t)$ has vanishing moments of higher than order, meaning that

$$\int_{-\infty}^{\infty} t^k \psi_t dt = 0 \quad k = 0,...m-1,$$

for some integer $m \geq 2$, then even the "smooth polynomial" segments are annihilate and the wavelet transform can better reveal the details of the signal on each octave band by considering different scale a.

In the construction of wavelet, we will use the multiscale structure of Multi Resolution Analysis (MRA). Suppose that that an MRA $\{V_{(n)}\}$ of L^2 is generated by some scaling function $\phi(t)$. Then if the integer translates of $\phi(t)$ locally reproduce all polynomials of degree $\leq m-1$ and if $\psi(t)$ is constructed to be orthogonal of to all integer translates of $\phi(t)$, then in each octave band, the wavelets $\psi(2^n t-k)$ are orthogonal to V_n

Any orthogonal wavelet $\psi(t)$ provide an orthogonal basis

$$\{\psi_{j,k}(t): j,k \in Z\}$$

of the finite energy space L^2. Any signal $f(t) \in L^2$ has a Fourier representation

$$f(t) = \sum_{j,k} \hat{d}_{j,k}\, \psi_{j,k}(t)$$

The general Fourier coefficients $\hat{d}_{j,k}$ of $f(t)$ possess very significant time-frequency information of the signal $f(t)$. For each $j,k \in Z$, the coefficient

$$d_{j,k} = \int_{-\infty}^{\infty} \hat{f}(t)\overline{\psi_{j,k}(t)}dt = (W_\psi f\left(\frac{k}{2^j}, \frac{1}{2^j}\right)$$

is the value of the wavelet transform of $f(t)$, with the orthogonal wavelet $\psi(t)$ itself as the analyzing wavelet, at the time-scale location

$$(b,a) = \left(\frac{k}{2^j}, \frac{1}{2^j}\right) \quad .$$

A plot of time-frequency location called Dyadic sampling grid is shown in Figure 5.

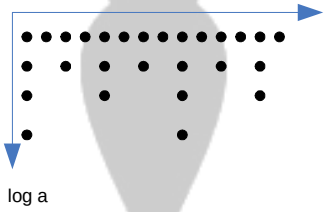


Figure 5. Dyadic sampling grid

Observe that although the wavelet transformation information of the signal $f(t)$ contained in the coefficient sequence $\left\{\hat{d}_{j,k}\right\}$ is only available on a very sparse set in the time-scale domain, this information is sufficient to determine the signal uniquely or give perfect reconstruction. It is to be noted that the analysis filter and synthesis filter are transposes as well as inverses of each other; the whole filter bank is orthogonal. When they are inverses, but not necessarily transposes, they are bi-orthogonal [1],[4]. In this experiment, bi-orthogonal reversible 5-3 filter is used.

The coefficients of analysis and synthesis of 5/3 Daubechies filter are shown in Table 1 and Table 2.

Table 1. Analysis Filter Coefficients

SYNTHESIS FILTER COEFFICIENTS		
I	**LPF**	HPF
0	1	$\frac{6}{8}$
±1	$\frac{1}{2}$	$\frac{-2}{8}$
±2		$\frac{-1}{8}$

Table 2. Synthesis Filter Coefficients

ANALYSIS FILTER COEFFICIENTS		
i	**LPF**	HPF
0	$\frac{6}{8}$	1
±1	$\frac{2}{8}$	$\frac{-1}{2}$
±2	$\frac{-1}{8}$	

2.5 Integer Wavelet Transform

In many applications, especially in image processing, the input data consist of integer samples. But all of the wavelet coefficients are floating point values even though the input samples are integers because filter coefficients used in transform filters are mostly rational or real numbers. Integer data is very important and

useful for fast implementation of the discrete wavelet transform, particularly in hardware, due to its efficient storage and encoding.

The principle of this transform is simple and illustrated here for one dimensional case. Given a data vector of N integers x_i, where $i = 0,1,...,N-1$, we define $k = N/2$ and we compute the transform vector y_i by calculating the odd and even components of y separately. The N/2 odd components y_{2i+1} (where $i = 0,1,...,k-1$) are calculated as differences of the x_i's. The become the detail (high frequency) transform coefficients. Each of the even components y_{2i} is calculated as weighted average and becomes as low-frequency transform coefficients.

Given an integer input sequence x_i, its forward IWT will also be an integer sequence y_i, which will be computed depending on whether the length N of the integer sequence is even or odd.

The main feature of the particular IWT described here is the use of truncation which is denoted by "floor" symbols. It is used to produce integer transform coefficients y_i and also integer reconstructed data items z_i.

If the signal length N is even (i.e. $N = 2k$), then integer transform coefficients y_i and integer reconstructed data items z_i are computed in the following steps.

$$y_{2i+1} = \begin{cases} x_{2i+1} - \lfloor (x_{2i} + x_{2i+2})/2 \rfloor, & i = 0,1,...,k-2 \\ x_{2i+1} - x_{2i}, & i = k-1 \end{cases}$$

$$y_{2i} = \begin{cases} x_{2i} + \lfloor y_{2i+1}/2 \rfloor, & i = 0 \\ x_{2i} - \lfloor (y_{2i-1} + y_{2i+1})/4 \rfloor, & i = 0,1,...,k-1 \end{cases}$$

$$z_{2i} = \begin{cases} y_{2i} - \lfloor y_{2i+1}/2 \rfloor, & i = 0 \\ y_{2i} - \lfloor (y_{2i-1} + y_{2i+1})/4 \rfloor, & i = 0,1,2,...,k-1 \end{cases}$$

$$z_{2i+1} = \begin{cases} y_{2i+1} + \lfloor (z_{2i} + z_{2i+2})/2 \rfloor, & i = 0,1,...,k-2 \\ y_{2i+1} - z_{2i}, & i = k-1 \end{cases}$$

And if the signal length N is odd (i.e. $N = 2k-1$), then integer transform coefficients y_i and integer reconstructed data items z_i are computed in the following steps.

$$y_{2i+1} = x_{2i+1} - \lfloor (x_{2i} + x_{2i+2})/2 \rfloor, \quad i = 0,1,...,k-1$$

$$y_{2i} = \begin{cases} x_{2i} + \lfloor y_{2i+1}/2 \rfloor, & i = 0 \\ x_{2i} - \lfloor (y_{2i-1} + y_{2i+1})/4 \rfloor, & i = 0,1,...,k-1 \\ x_{2i} + \lfloor y_{2i-1}/2 \rfloor, & i = k \end{cases}$$

$$z_{2i} = \begin{cases} y_{2i} - \lfloor y_{2i+1}/2 \rfloor, & i = 0 \\ y_{2i} - \lfloor (y_{2i-1} + y_{2i+1})/4 \rfloor, & i = 0,1,2,...,k-1 \\ y_{2i} - \lfloor y_{2i+1}/2 \rfloor & i = k \end{cases}$$

$$z_{2i+1} = y_{2i+1} + \lfloor (z_{2i} + z_{2i+2})/2 \rfloor, \quad i = 0,1,...,k-1$$

Because of truncation, some information is lost when y_i are calculated. However truncation is also used in the calculation of z_i, which restores the lost information. Thus, the equations above are true forward and inverse IWT that reconstruct the original data items exactly.

2.6 Overview of the JPEG2000 Coding Standard

In this paper we will use the IWT on JPEG2000 , which is an image coding standard designed to incorporate compression of different type of images (bi-level, gray level, color, and multi component) with various imaging models (real time transmission, image library archival, limited buffer and bandwidth resources, etc). Some JPEG2000 features correspond to the application presented in this paper are as follow [6], [7]:

- Lossless and lossy compression.

  JPEG2000 utilizes two types of wavelet filter. Daubechies 9/7 floating point wavelet filter provides lossy compression due to floating point quantization errors, yet yields better compression ratio. Biorthogonal 5/3 integer wavelet filter supports lossless compression at the cost of higher compression bit rate. Reversible transform with integer filter is used to produce a scalable bit stream which builds up scalable quality image representation.

- Good performance on low compression bit rate.

  JPEG2000 demonstrates a significantly better low bit rate performance compared to existing compression method.

- Random codestream access and processing.

JPEG2000 coding is conducted independently in the level of spatially non-overlapping image parts called codeblocks. Hence, every codeblock could be considered as autonomous smaller image which builds up the original image. Since spatial location information is self-contained in the codeblock, spatial random access and processing of JPEG2000 codestream is possible.

- Progressive pixel accuracy and resolution

Multiresolution feature of discrete wavelet transform brings forth the possibility to reconstruct image in various level of resolution and in consequence various level of pixel accuracy.

2.7 Distributed System

Considering that a medical image has a huge amount of data we will make a parallel distributed process client-server network as shown on Figure--------.

This scheme has a system manager on the server and a number of clients which may acts either as a distributed process client or an end-client. Distributed processor client receives encoding task from system manager and performs it in parallel way with the other similar positioned clients. End-client requests for an image, retrieves it in the compressed form, and reconstructs it for viewing.

When an end-client request an image from the server, system manager in the server will divide the image into N tiles and distribute them to the N distributed processor clients. Each client will perform transformation for each given tile, and then send the outputs to the end client. The end client will inverse transform and reconstruct the image. This scheme is expected to be able to shorten the computation time.

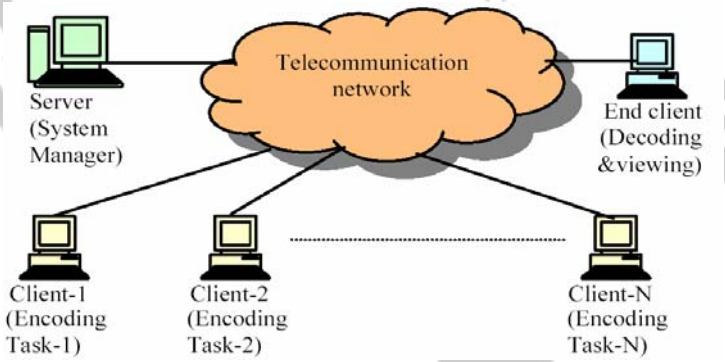
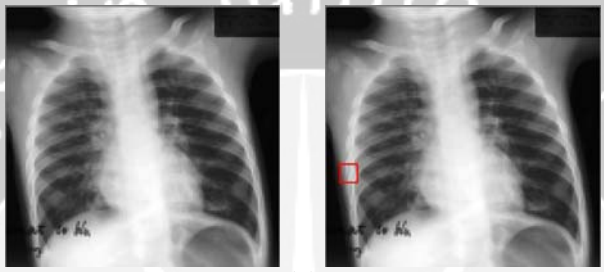


Figure 6. Distributed system architecture.

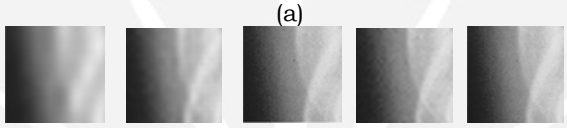
3 EXPERIMENTS & DISCUSSION

In the first experiment will evaluate the performance of IWT in compressing and decompressing a medical image progressively from lossy to lossless quality in a non-distributed system.

The test image incorporated in the first experiment was a digitized x-ray thorax image, scanned at 96dpi vertical and horizontal resolution. The digitized image was represented as gray level (8 bit depth) bitmap image xrayA.bmp with the dimension of 2002x1915 pixels, shown in Figure 7.



(a)



(b)

Figure 7. Result of the non distributed system (a).Original test image (left) and first resolution level reconstruction image with 128x128 RoI within red box(right), (b).Detailed view of RoI refinement from 1st to 5th level

Table 3. Performance of RoI Quality Refinement: RoI PSNR Value

RoI PSNR value of xrayA.jp2					
RoI Size (pixels)	**RoI PSNR(dB) value per Resolution Level**				
	1*	**2**	**3**	**4**	**5**
32x32	30.343	32.917	33.85	35.939	Inf
64x64	30.325	32.394	33.087	35.038	Inf
128x128	30.427	32.998	33.645	35.423	Inf
256x256	29.194	32.51	33.083	34.806	Inf
2002x 1915 (full size)	27.211	32.339	32.888	34.242	Inf

*Image general picture first retrieved by client

Table 4: Performance of RoI Quality Refinement : Retrieved Bytes Statistic

Retrieved Bytes Statistics of xrayA.jp2				
RoI Size (pixels)	**Retrieved Bytes per Resolution Level (kb)**			
	2	**3**	**4**	**5**
32x32	8.36	8.02	8.35	16.94
64x64	8.36	8.02	8.35	16.94
128x128	8.36	8.02	16.70	33.88
256x256	8.36	8.02	33.39	76.22
2002x 1915 (full size)	23.81	94.81	405.98	1633.88

Table 3 remarks the general increase of PSNR value from the lowest resolution level (lossy) to the highest resolution level (lossless). PSNR level reaches infinity at the retrieval of the highest resolution level which involves all resolution layers of compressed image.

Number of bytes retrieved for every level of quality refinement is provided in Table 4. It shows that the first until the fourth resolution level could be obtained by retrieving a small number of bytes. This scheme will less burden the limited bandwidth networks but still has the ability to provide good image quality on the end client-side.

The objective of the second experiment is to compare the performance between the distributed and non-distributed system for distributed lossless medical image compression. The test image incorporated in the experiment was a gray level (8 bit depth) bitmap image with the dimension of 1024 x1024 pixels.

In this evaluation scheme, image is encoded using lossless wavelet compression with decomposition level of 5. The fifth decomposition level subband images coded-data is set as the first data transferred to the client in answer to client and end-client side request.

General picture in the client and end-client side is generated by reconstructing the given data and resizing the resulted image into original image resolution. Table 5. shows that the encoding time of distributed system is shorter than non-distributed system on the same level. The same with the decoding time, distributed system has shorter decoding time than non-distributed system. The encoding time of the distributed system is 88.30% on average, shorter than the distributed system, while the decoding time of the distributed system is 81.33% on average, shorter than the non-distributed system.

Table 5. Encoding Time

Level	ENCODING TIME (s)		DECODING TIME (s)	
	Distributed	Non Distributed	Distributed	Non Distributed
1	0.239	0.289	0.284	0.312
2	0.237	0.278	0.242	0.319
3	0.261	0.284	0.252	0.31
4	0.262	0.291	0.254	0.328
5	0.263	0.287	0.256	0.316

Table 6. Achieved Bitrates

Level	ACHIEVED BITRATE (bpp)			
	ENCODING		DECODING	
	Distributed	Non Distributed	Distributed	Non Distributed
1	3.22882845	3.4186478	3.2288285	3.4186478
2	3.0109711	3.2060547	3.0109711	3.2060547
3	2.9778137	3.1582642	2.965538	3.1582642
4	2.9594345	3.1475983	2.9577179	3.1475983
5	2.95895385	3.1467743	2.9589539	3.1467743

Table 7. Downloaded Bytes

Level	DOWNLOADED BYTES			
	ENCODING		DECODING	
	Distributed	Non Distributed	Distributed	Non Distributed
1	423209	448089	423209	448089
2	394654	420224	394654	420224
3	390308	413960	390308	413960
4	387899	412562	387899	412562
5	387836	412454	387836	412454

Table 6 shows that achieved bitrate of the distributed system is 94.14 % lower on average than non-distributed system and Table 7 shows that downloaded bytes of the distributed system is consequently has the same average number 94.14% shorter than the non-distributed system.

4 Conclusions

The proposed scheme, IWT for distributed lossless medical image compression, will considerably reduce the number of transferred bytes during one access period, thus less-burdening the network while maintaining good image quality on medically-significant area. However, performance effectiveness of the scheme is very much dependent on RoI intensity characteristics and its codeblocks affiliation.

Performing the JPEG2000 encoding tasks on the distributed system being developed may significantly be able to share the computation loads and to shorten the processing time while maintaining good image quality.

References

[1] C.K. Chui (1997), Wavelets: A Mathematical Tool for Signal Processing, Siam, Philadelphia, 59-90.

[2] C.S. Burrus, R.A. Gopinath, H. Guo (1998), Introduction to Wavelets and Wavelet Transforms: A Primer, Prentice-Hall International Inc., New Jersey, 41 – 49.

[3] D. Salomon (2000), Data Compression: The Complete Reference, Springer, New York, 535 – 534, 567 – 579.

[4] G. Strang & T. Nguyen (1997), Wavelets and Filter Banks, Wellesley-Cambridge Press, Wellesley, 27 – 35.

[5] ISO/IEC 1544-1,"Information Technology-JPEG2000 Image Coding System-Part 1: Core Coding System",2000.

[6] Skodras, A.N. Christopoulos, C.A. Ebrahimi, T. (2001), JPEG2000 Still Image Compression Standard, *IEEE Signal Processing Magazine*, 36-58.

[7] Skodras, A.N. Christopoulos, C.A. Ebrahimi, T. (2000), JPEG2000: The Upcoming Still Image Compression Standard, *Proceedings of the 11th Portuguese Conference on Pattern Recognition (REPCA00D20)*, 359-366.

P. RAHMIATI: Department of Electrical Engineering, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia. Phone: +62 +22 250 1661
E-mail: pauline332@students.itb.co.id

A.B. SUKSMONO: Department of Electrical Engineering, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia. Phone: +62 +22 250 1661
E-mail: suksmono@ltrgm.ee.itb.ac.id

T.L.R. MENGKO: Department of Electrical Engineering, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia. Phone: +62 +22 250 2260
E-mail: tmengko@itb.ac.id

A. HANDAYANI: Department of Electrical Engineering, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia. Phone: +62 +22 250 2260
E-mail: astri@biomed.ee.itb.ac.id

A. FAJRI: Department of Electrical Engineering, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia. Phone: +62 +22 250 2260
E-mail: fajri@freebsd.or.id

DISTRIBUTED AND PROGRESSIVE MULTIGRID V-CYCLE PHASE UNWRAPPING FOR MRI APPLICATION

Dyah Ekashanti O. Dewi, Andriyan B. Suksmono, Tati Latifah R. Mengko

Institut Teknologi Bandung (ITB), Indonesia

Abstract. We describe the development of Multigrid approach that can be used for solving Partial Differential Equation (PDE) problem of wrapped phase image for Magnetic Resonance Imaging (MRI) application. This paper proposes a modified Multigrid V-Cycle Phase Unwrapping (PU) which generates the solution in gradual manner by decomposing a complete Multigrid V-Cycle structure into small pieces of V-Cycle scheme and performing the solution based on the level of the structure. The wrapped phase is first transferred to the coarser grids iteratively generating approximation solution on each grid until it reaches the coarsest grid. Then, the approximation solutions are distributed to be solved partially producing the intermediate solutions in the finest grid. Progressive refinement is accomplished by superposing the unwrapped intermediate solutions from the subsequent stages immediately after each process is done. By utilizing this scheme, the absolute phase image is built progressively, allowing the users to quickly become conversant with the image before the PU process is fully completed. A preliminary evaluation on simulated as well as MRI complex raw data shows promising progressive results and feasibility to apply in a variety of MRI applications.
Key-words: distributed, progressive, multigrid V-Cycle, phase image, phase unwrapping, MRI application

1 Introduction

In coherent imaging modalities, phase processing is essential. For instance, the phase part of MRI image is able to provide some physical properties of interests which are of use for MRI applications. However, the acquired phase is naturally bounded on (-π,π] producing ambiguity problem, called wrapped phase. Any values beyond this range will be wrapped back on itself to cause the sudden artificial phase jumps near boundaries [1]. Two-Dimensional PU is a computational method which estimates the unwrapped phase from its wrapped form in the principal value of modulo 2π. The principal value has to be unwrapped by eliminating, or at least, reducing the discontinuities. It can be stated mathematically in (1) as finding the estimated absolute phase from the given wrapped phase.

$$\phi^u(x,y) = \phi^w(x,y) + k(x,y).2\pi \quad (1)$$

where ϕ^u is the estimated absolute phase, ϕ^w is the principal value from the wrapped phase, and $k(x,y)$ is the integer function [2].

Nowadays, such technique has blossomed into many algorithms and been applied in a number of applications. Principally, the existing PU algorithms can be grouped into two main categories, Local PU and Global PU methods. The Local PU (Path Following) unwraps the phase based on the integration path which connects the residues identified in the wrapped phase field and cuts, whereas the Global PU

(Minimum Norm) impresses on the solution by modeling the wrapped phase into PDE and solving it using numerical methods [2].

Multigrid is one of the Global PU which is expressed in term of recursion. It unwraps the phase by solving the PDE of the wrapped phase over a hierarchy of grids. The first practical Multigrid technique was pioneered by Brandt in 1973. Multigrid in PU is first proposed by Pritt [5] for InSAR. The implementation of such method to MRI application has been presented in [6] for water-fat separation. Furthermore, applying weighting value from the scaled magnitude data to the algorithm improves the PU performance [6]. The basic idea of progressive method in PU has also been initiated by [7] utilizing FFT in recursive manner as well as [8] which improves the Multigrid PU by reprocessing the residual error of the phase images. Additionally, [9] has introduced the preliminary evaluation on progressive Multigrid V-Cycle PU. Following the previous works, we can develop the distributed and progressive Multigrid V-Cycle PU.

2 Multigrid Phase Unwrapping

The use of Multigrid in PU is derived from least-squares PU problem. The least-squares error notion is known formally as a minimum norm in the L^2 sense. That is, the sum of the squared differences between the gradients of the solution and those of the measurement is minimized. It can be formulated as unweighted or weighted version [2].

2.1 Unweighted Least-Squares Phase Unwrapping

As motivation for Global PU problem, the general unweighted Minimum L^p-Norm approach is presented. Let the wrapped phase $\phi^w_{i,j}$ and the adored unwrapped phase values $\phi^u_{i,j}$ are sampled on a rectangular grid [2][5]. The row and column partial derivatives of the wrapped phase are defined by

$$\Delta^x_{i,j} = \phi^w_{i+1,j} - \phi^w_{i,j}\ ,\ i = 0,...,M-1\ ,\ j = 0,...,N \tag{2}$$

$$\Delta^y_{i,j} = \phi^w_{i,j+1} - \phi^w_{i,j}\ ,\ i = 0,...,M\ ,\ j = 0,...,N-1 \tag{3}$$

The value 2π is added or subtracted as necessary to ensure that they lie in the interval $(-\pi,\pi]$. The derivatives must be corrected at the grid boundaries using (4) to assure that the Gauss-Seidel relaxation converges to the correct solution.

$$\Delta^x_{0,j} = -\Delta^x_{1,j}\ ,\ \ \Delta^x_{M+1,j} = -\Delta^x_{M,j}\ ,\ \ \Delta^y_{i,0} = -\Delta^y_{i,1}\ ,\ \ \Delta^y_{i,N+1} = -\Delta^y_{i,N} \tag{4}$$

The intermediate unwrapped function is utilized to minimize the discrete function

$$J = \in^p = \sum_{i=0}^{M-1}\sum_{j=0}^{N}\left|\phi^u_{i+1,j} - \phi^u_{i,j} - \Delta^x_{i,j}\right|^p + \sum_{i=0}^{M}\sum_{j=0}^{N-1}\left|\phi^u_{i,j+1} - \phi^u_{i,j} - \Delta^y_{i,j}\right|^p \tag{5}$$

The least-squares PU solves the discretized PDE problem with efficient mathematical methods. Additional constraints, such as smoothness, can be imposed through regularization methods. This technique is generally favored as they lead to linear equation where the resulting mathematics is tractable and amenable to efficient methods of solution [2]. For p = 2, we get the Least-Squares equation like so

$$\epsilon^2 = \sum_{i=0}^{M-2}\sum_{j=0}^{N-1}\left|\phi^u_{i+1,j} - \phi^u_{i,j} - \Delta^x_{i,j}\right|^2 + \sum_{i=0}^{M-1}\sum_{j=0}^{N-2}\left|\phi^u_{i,j+1} - \phi^u_{i,j} - \Delta^y_{i,j}\right|^2 \quad (6)$$

Through differentiating (6) on $\phi^u_{i,j}$ and setting the result equals to zero yield the linear equation, regarded as a discretized PDE $\nabla^2\phi = \rho_{i,j}$ known as Poisson's equation, where ∇^2 is the phase Laplacian operator $\partial/\partial x^2 + \partial/\partial y^y$ [5].

$$\left(\phi^u_{i+1,j} - 2\phi^u_{i,j} + \phi^u_{i-1,j}\right) + \left(\phi^u_{i,j+1} - 2\phi^u_{i,j} + \phi^u_{i,j-1}\right) = \rho_{i,j} \quad (7)$$

$$\rho_{i,j} = \Delta^x_{i,j} - \Delta^x_{i-1,j} + \Delta^y_{i,j} - \Delta^y_{i,j-1} \quad (8)$$

The classical method for solving such equation is called Gauss-Seidel relaxation. The unwrapped phase solution array $\phi^u_{i,j}$ is set to be zero (or a predefined value) and performs the following updates iteratively until convergence.

$$\phi^u_{i,j}(n+1) = [(\phi^u_{i+1,j}(n) + \phi^u_{i-1,j}(n) + \phi^u_{i,j+1}(n) + \phi^u_{i,j-1}(n)) - \rho_{i,j}]/4 \quad (9)$$

In this regard, *n* is the number of iteration. Yet, it is not practical due to its extremely slow convergence. Alternatively, Red-Black Gauss-Seidel yields better convergence result.

2.2 Weighted Least-Squares Phase Unwrapping

The unweighted version provides adverse results due to the disability of solving the residue problem. Therefore, weighted least-squares PU enhances the previous one. Such technique uses a set of weights (e.g, quality maps or masks) to avoid integrating through the residues, accommodates the residues in some fashion, isolates the regions of low signal-to-noise, or imposes other properties or preferences on the expected solution. When certain phase values are corrupted by the noise, aliasing, or other degradations, the phase values are zero-weighted to avoid affecting the PU process [2][5]. In practice, an array of weights $0 \le w_{i,j} \le 1$ is given to the phase data. The minimized weighted Least-Squares function becomes

$$\epsilon^2 = \sum_{i,j} U(i,j)\left(\phi^w_{i+1,j} - \phi^w_{i,j} - \Delta^x_{i,j}\right)^2 + \sum_{i,j} V(i,j)\left(\phi^w_{i,j+1} - \phi^w_{i,j} - \Delta^y_{i,j}\right)^2 \quad (10)$$

where

$$U(i,j) = \min\left(w^2_{i+1,j}, w^2_{i,j}\right), \qquad V(i,j) = \min\left(w^2_{i,j+1}, w^2_{i,j}\right) \quad (11)$$

The weighted Least-Squares solution to this problem yields the weighted Phase Laplacian, which is defined by

$$U(i,j)\Delta^x_{i,j} - U(i-1,j)\Delta^x_{i-1,j} + V(i,j)\Delta^y_{i,j} - V(i,j)\Delta^y_{i,j-1} = c_{i,j} \quad (12)$$

The classical Gauss-Seidel relaxation method for weighted PU is defined as follows

$$\phi^u_{i,j} = \frac{U(i,j)\phi^u_{i+1,j} + U(i-1,j)\phi^u_{i-1,j} + V(i,j)\phi^u_{i,j+1} + V(i,j-1)\phi^u_{i,j-1} - c_{i,j}}{U(i,j) + U(i-1,j) + V(i,j) + V(i,j-1)} \quad (13)$$

Undesirably, the Gauss-Seidel relaxation converges too slowly in practical use.

2.3 Multigrid V-Cycle Phase Unwrapping

Multigrid is a prime source of important advances in algorithmic efficiency, finding a rapidly increasing number of users. Multigrid offers the possibility of solving large classes of problems, especially the elliptic equation with non-constant coefficients with hardly any loss in efficiency, even nonlinear equations with comparable speed. It basically accelerates the convergence of classical Global PU by approximating the smooth part of the error on coarser grid and refining the non-smooth part with a small number of iterations on the fine grid. At this point, the classical error-smoothing scheme, Gauss-Seidel relaxation, is applied iteratively to approximate the solution until some convergence criterion is satisfied [3][4]. The defect on coarse grid is defined by a fine-to-coarse transfer operator called Restriction operator. In this regard, we use Full Weighting operator as defined in (14)

$$c_{i,j} = \frac{1}{16}(f_{2i-1,2j-1} + f_{2i+1,2j-1} + f_{2i-1,2j+1} + f_{2i+1,2j+1}) + \frac{1}{8}(f_{2i,2j-1} + f_{2i,2j+1} + f_{2i-1,2j} + f_{2i+1,2j}) + \frac{1}{4} f_{2i,2j} \quad (14)$$

The coarse-to-fine operator, known as Prolongation operator, transfers the intermediate solution into new finer grid by interpolating and adding the coarse-grid correction. Bilinear Interpolation is one example of such operator. It is given by

$$\begin{aligned} f_{2i,2j} &= c_{i,j} \\ f_{2i-1,2j} &= \frac{1}{2}(c_{i,j} + c_{i-1,j}), \qquad f_{2i+1,2j} = \frac{1}{2}(c_{i,j} + c_{i+1,j}) \\ f_{2i,2j-1} &= \frac{1}{2}(c_{i,j} + c_{i,j-1}), \qquad f_{2i,2j+1} = \frac{1}{2}(c_{i,j} + c_{i,j+1}) \\ f_{2i+1,2j+1} &= \frac{1}{4}(c_{i,j} + c_{i+1,j} + c_{i,j+1} + c_{i+1,j+1}) \end{aligned} \quad (15)$$

Multigrid can be illustrated as a pyramid which represents the grid building with each grid is one-half the resolution of its predecessor. It is shown as follows

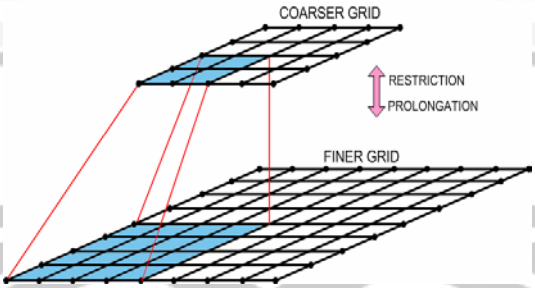


Fig.1. Grid building in Multigrid PU

The order in which the grids are visited is called multigrid cycle. Multigrid V-Cycle, which sweeps the grids based on V-alphabet like structure, is the simplest type of Multigrid schedule [2][3][4]. Generally speaking, the Multigrid V-Cycle PU transfers the wrapped phase to the next coarser grids by performing Gauss Seidel relaxation and restricting the residual error gradients until it reaches the coarsest grid. Then, the intermediate solution is transferred back to the finer grids. This recursive procedure is terminated once the finest grid is accomplished. In this regard, each prolonged grid is relaxed over again. The Multigrid V-Cycle PU schedule diagram is given in Figure 2.

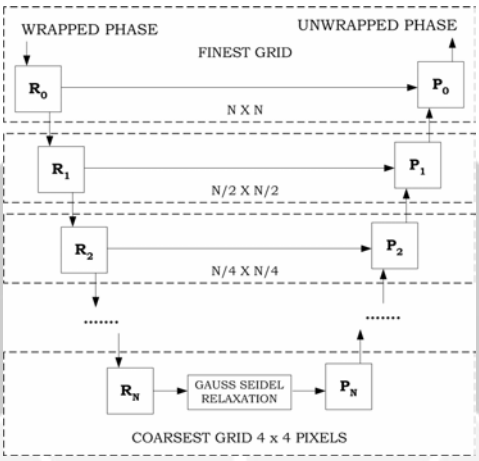


Fig. 2. Multigrid V-Cycle PU schedule diagram

2.4 Distributed and Progressive Multigrid V-Cycle Phase Unwrapping

The proposed distributed and progressive Multigrid V-Cycle implements the Multigrid V-Cycle PU scheme by decomposing a complete V-Cycle order into small pieces of V-Cycle structure and performing the solution based on the level of the structure. Firstly, the wrapped phase is transferred to the coarser grid iteratively by means of restriction operator until the coarsest grid is obtained. The restriction process on each grid generates approximation solution. Then, the approximation solutions are distributed to be solved partially by employing the prolongation operator and Gauss-Seidel relaxation. Progressive refinement is accomplished by superposing the unwrapped intermediate solution from the subsequent distributed scheme immediately after the process is completed. The distributed and progressive Multigrid V-Cycle PU schedule diagram is depicted in Figure 3.

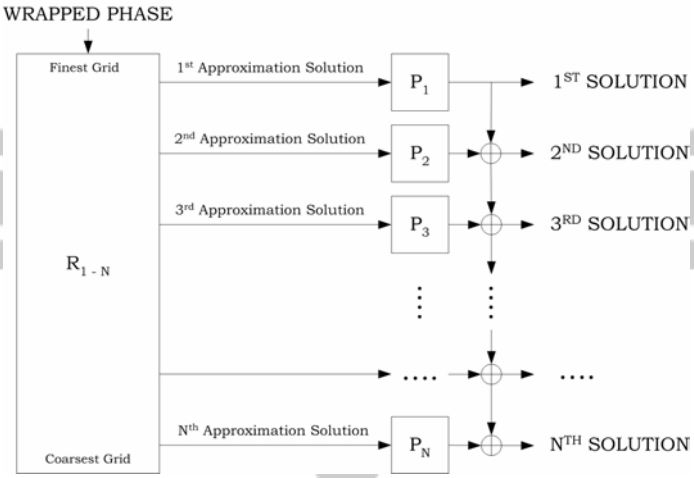


Fig. 3. Distributed and Progressive Multigrid V-Cycle PU schedule diagram

$R_{1\text{-}N}$ block defines the Restriction process from finest to coarsest grid in iterative manner, whereas P_i blocks refer to the Multiple Prolongation processes from the coarser to finer grid in distributed fashion, where i is the level of the grid. The detail of each block can be shown in Figure 4 and Figure 5 respectively.

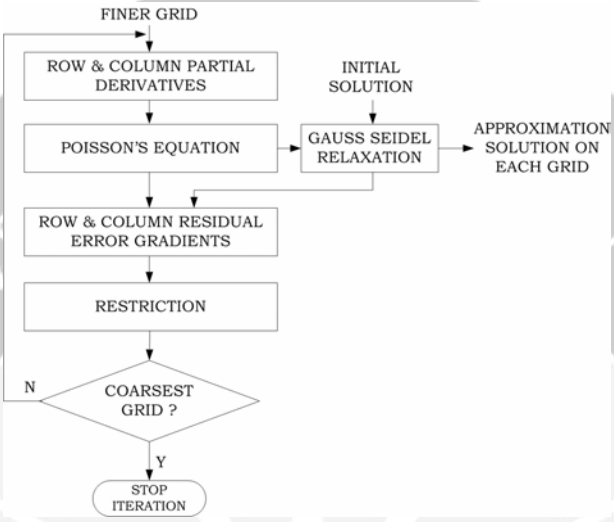


Fig. 4. $R_{1\text{-}N}$ block diagram

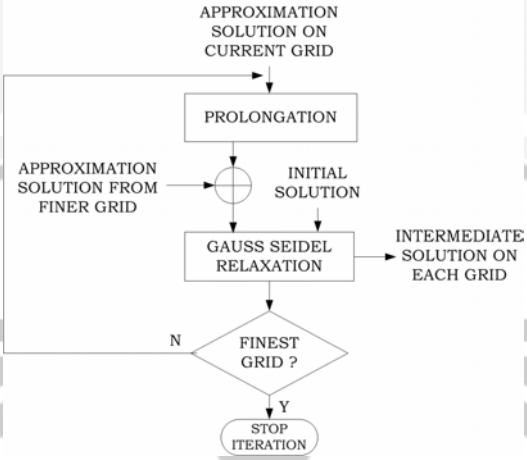


Fig. 5. P_i block diagram

In detail, the distributed and progressive system divides the Multigrid V-Cycle procedure into multiple subsystems. The process is started by transferring the wrapped phase into the coarser grid by way of block $R_{1\text{-}N}$. In the initial stage of Restriction process, the first residual error gradients are set to be the row and column partial derivatives of the wrapped phase, (i.e., $e_0^{wx} = \Delta_0^x$ and $e_0^{wy} = \Delta_0^y$), while the initial unwrapped phase is set to be zero ($\phi_0^u \equiv 0$). Then, the 1st intermediate

solution is performed by applying the P_1 block. The process continues to apply the scheme in the following distributed blocks. Each P_i block immediately employs Gauss Seidel relaxation iteratively until the finest grid is attained. The 2nd solution is obtained by the second P_2 result and added to the 1st solution subsequently. The next solutions are achieved in the same way as the preceding grids by superposing them to the previous solutions. At stage-N, the system yields an Nth solution ϕ_N^u and residual error gradients e_N^{wx} and e_N^{wy}. By utilizing this scheme, the unwrapped phase information is gradually improved, yielding better ones at later time, while the residual error gradients are reduced in stages. Additionally, by incorporating computational strength of the distributed and progressive Multigrid V-Cycle PU, it is potential to realize a fast PU system for future real-time MRI application.

3 Experimental Results and Discussion

In the experiment we utilize simulated as well as actual MRI complex raw data. The simulated data is merely performed in unweighted Multigrid V-Cycle PU manner, while the MRI complex raw data is tested by means of weighted version.

The images in Figure 6 and Figure 7 demonstrate the original, unwrapped, and rewrapped phase images of simulated data by using the multigrid V-Cycle PU and the proposed distributed and progressive Multigrid V-Cycle PU schemes. In this regard, the distributed and progressive test on simulated data decomposes the solution into 5 stages. The measurement of the estimated unwrapped phase images to the original one is defined by Peak Signal to Noise Ratio (PSNR) value.

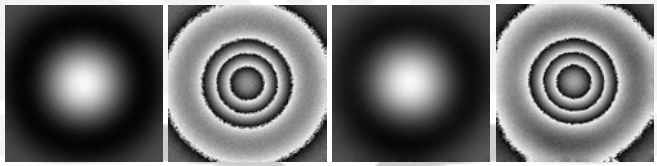


Fig. 6. Original Unwrapped (a) and Wrapped (b) images, Unwrapped (c) and Rewrapped (d) results of simulated data using Multigrid V-Cycle PU

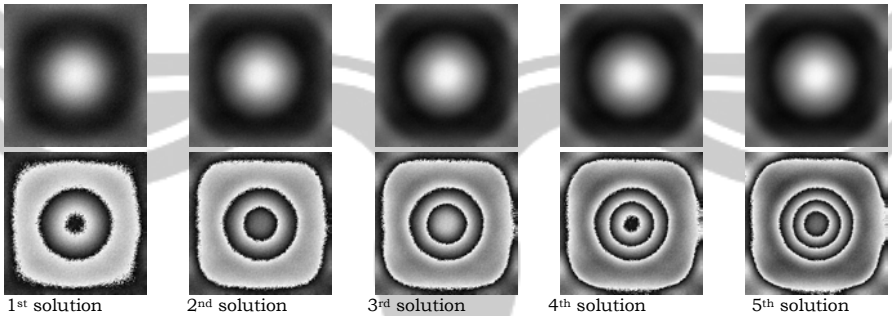


Fig. 7. Unwrapped (Top) and Rewrapped (Bottom) simulated data results of the distributed and progressive Multigrid V-Cycle PU in each stage of solutions

By observing the images in Figure 6 and 7, we can infer that the proposed PU algorithm visually does a comparative job in unwrapping the simulated data in distributed fashion. It can be shown from the improved number of fringe lines on each stage of the rewrapped images. Yet, some occasional failures still exist in the discontinuous areas. From the quantitative measurement, the PSNR value of Multigrid V-Cycle PU test is 28.3660 dB, while the distributed and progressive method yields an increased number of PSNR value in stages, ranging from 17.2087 dB to 21.5789 dB. Although these results are still far from the previous method, the proposed technique is capable of solving the solution in progressive way.

The second experiment is performed using the MRI phase image. Figure 8 presents the original, unwrapped, and rewrapped MRI phase image solutions of Multigrid V-Cycle PU, while Figure 9 displays the unwrapped and rewrapped results of the proposed distributed and progressive Multigrid V-Cycle PU which decomposes the solution into 6 stages.

Fig. 8. Original Wrapped (a), Unwrapped (b) and Rewrapped (c) results of MRI phase image using Multigrid V-Cycle PU

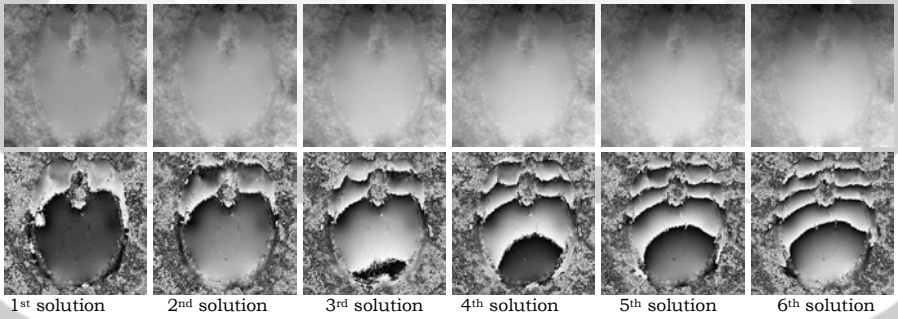


Fig. 9. Unwrapped (Top) and Rewrapped (Bottom) MRI phase results of the distributed and progressive Multigrid V-Cycle PU in each stage of solutions

From the results above, it is shown that the proposed technique is capable of presenting the solution in distributed and progressive way. Nevertheless, the algorithm does not adequately solve the existing discontinuity problem in some noisy and inconsistent areas, such as in nose and eyes part of the object.

4 Summary

We have presented a distributed and progressive Multigrid V-Cycle PU method for simulated data as well as actual MRI phase images. A preliminary evaluation on simulated data as well as MRI complex raw data shows promising progressive improvements and feasibility to apply in MRI application. The increasing PSNR value of the proposed method results explains that by distributing the V-Cycle

scheme on PU, the progressive solution can be generated. The existing failures still calls for restoration of the algorithm. For further direction, the distributed and progressive Multigrid PU method can be developed for real-time MRI applications.

Acknowledgment

The authors would like to acknowledge Prof. Akira Hirose of Graduate School of Frontier Informatics, The University of Tokyo, for donating [2] and Prof John M. Pauly of Radiological Sciences Laboratory, Stanford University of Medicine for the permission to use the actual MRI complex raw data.

References

[1] Z.-P. Liang (1996), A model based method for phase unwrapping, *IEEE Trans. Med. Imag.*, **15**, 893-897.

[2] D.C. Ghiglia and M.D. Pritt (1998), *Two-dimensional phase unwrapping : Theory, algorithms, and aoftware*, John Wiley & Sons, Inc.

[3] P. Wesseling (1992), *An introduction to multigrid methods*, A Volume in Pure and Applied Mathematics, John Wiley & Sons Ltd.

[4] W.H. Press, S.A, Teukolsky, W.T. Vetterling, B.P. Flannery (1997), *Numerical recipes in C, The art of scientific programming*, Cambridge Univ. Press.

[5] M.D. Pritt (1996), Phase unwrapping by means of multigrid techniques for interferometric SAR", *IEEE Trans. Geosci. Remote Sensing*, **34**, 728-738.

[6] D.E.O Dewi, A.B. Suksmono, T.L.R. Mengko (2005), Multigrid phase unwrapping of MRI phase images for water – fat separation application, *Proceedings of the 3rd APT Telemedicine Workshop 2005, Malaysia,* 221-225.

[7] A.B. Suksmono, A. Hirose (2003), Recursive transform based phase unwrapping, *Proceedings of International Conference Image Processing (ICIP) 2003, Barcelona.*

[8] A.B. Suksmono (2005), Improving the multigrid phase unwrapping algorithm by reprocessing the residual error and its application to MRI phase image processing, *Proceedings of the 3rd APT Telemedicine Workshop 2005, Malaysia*, 243-247.

[9] D. E. O Dewi, A. B. Suksmono, T. L. R. Mengko (2005), Progressive multigrid v-cycle phase unwrapping for MRI phase images, *Proceedings of The 7th International Workshop on Enterprise Networking and Computing in Healthcare Industry* (*Healthcom 2005*), Korea, 368-371.

DYAH EKASHANTI O. DEWI: Member of Imaging and Image Processing Research Group (I2PRG), Master student at Biomedical Engineering Program, Department of Electrical Engineering, ITB, Jalan Ganesha 10 Bandung 40132, Indonesia.
E-mail: deo_dewi@biomed.ee.itb.ac.id

ANDRIYAN B. SUKSMONO: Member of I2PRG, Faculty member of Radio Telecommunication and Microwave Laboratory, Department of Electrical Engineering, ITB, Indonesia.
E-mail: suksmono@ltrgm.ee.itb.ac.id

TATI LATIFAH R. MENGKO: Member of I2PRG, Faculty member of Electronics and Component Laboratory & Biomedical Engineering Program, Department of Electrical Engineering, ITB, Indonesia.
E-mail: tmengko@itb.ac.id

Analysis of a Non-Linear System by a New Technique Based On the Continuation Method

S. Kadry

University of Technology of Belfort-Montbeliard, France

Abstract. The Continuation method is illustrated for the solution of non-linear differential system. A proposed technique to get the best possible partition of [0, 1] for this method. This technique searches for a set of $\{t_i\} \in [0,1]$ involving a passage from 0 to 1 with minimum operations and acceptable solution.
Key words: Continuation method, Newton method, numerical analysis and non-linear Partial Differential Equation.

1 Introduction

In this study, our objective is the analysis of the resolution of finite non-linear algebraic systems. These systems are presented by the following shape:

$$F(u) = 0 \tag{1}$$

Where, $F : IR^n \to IR^n$ is a non-linear function.

For $u = (u_1, u_2, ..., u_n)^T \in IR^n$, $F(u) = (F_1(u), F_2(u), ..., F_n(u))^T \in IR^n$ each of the scalar functions $F_i(u)$, is non linear ($F_i(u) : IR^n \to IR^n$).

Our goal is the setting up of efficient and powerful algorithms to solve (1) using the Continuation method with a new technique of interval partition. The origin of the non-linear discreet systems is varied. We are concerned by considered equations that come from the discretization of continuous models of (ordinary or partial) differential equations, met in the mechanical sciences [1] (ex: bending of a beam, membrane distortions, fluids out-flow, problems of potential...) and in several domains of engineering.

Without loss of generality, we consider the non-linear Poisson equation [2, 7], (has one or two dimensional space).

For Ω opened on *IR*, search for $u : \overline{\Omega} \to IR, (\overline{\Omega} = \Omega \cup \partial\Omega)$, which verifies:

$$\begin{cases} -\dfrac{d^2u}{dx^2} + f(u) = g(x), in\,\Omega \\ u = 0 \quad on\,\partial\Omega \end{cases} \tag{2}$$

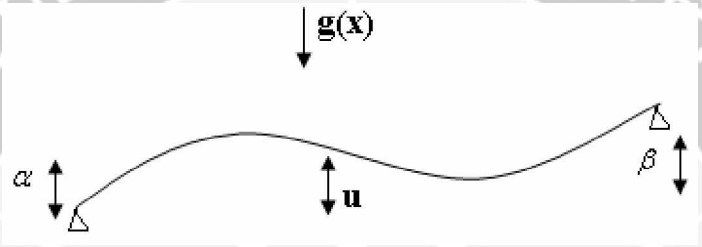
$f : IR \to IR$, is generally a real function that has a real non-linear variable. The

equation (2) models the bending of a beam.

2 Discretization of the problem

Let's consider the following problem: Being given a non-linear function f, with one real variable. Find a U function two times continuously derivable on [0, 1] as:

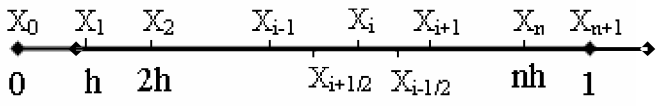
$$\begin{cases} -\dfrac{d^2u}{dx^2} + c(x)u = g(x) \quad 0 < x < 1 \\ u(0) = \alpha \,,\, u(1) = \beta \qquad \alpha, \beta \in IR \end{cases} \tag{3}$$



This mechanical situation problem is the one of bending the beam, stretched according to its axis by a linear load strength g(x) and merely supported its ends. Then the moment of bending non-linear f (u) in the point of x abscissa is solution of the problem (3) with c(x) = f/EI(x), or E the Young module of the material, I(x) is the principal moment of inactivity of the beam section at the x point.
Except for some rare cases, a formula that permits to get u(x) explicitly doesn't exist for all x∈ [0,1]. It therefore requests to find a means to approach the values of the problem solution (3) more accurately. A method to reach this goal consists in finding a number of finite parameters $\{u_i / i = 1,...,n\}$, as either an approximation of $u(x_i), i = 1,...,n.$ We are interested in the method of finite differences.

3 Method of the finite differences

let **n** is positive, put $h = \dfrac{1}{n+1}$ or **h** is the step of discretization (supposed to be uniform here); $x_i = ih$ for $i = 0,...,n+1, \{x_i\}$ are the discretization nodes.



Besides, it's possible to demonstrate that u is a regular function (for example u is

class C^4) that [8]:

$$\frac{du}{dx_i} = \frac{u(x_{i+1/2}) - u(x_{i-1/2})}{h} + O(h^2)$$

then

$$\frac{d^2u}{dx_i^2} = \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2} + O(h^2) \tag{4}$$

To solve (3) numerically, based on (4), and calculate the values u_i (we note $u_i \cong u(x_i)$) with $1 \le i \le n$, cautious to be approximately $u(x_i)$ (after replacing the formula with the differences (4) in (3)):

$$\begin{cases} \dfrac{-u_{i-1} + 2u_i - u_{i+1}}{h^2} + c(x_i)u_i = g(x_i) \\ u_0 = \alpha, \quad u_{n+1} = \beta \end{cases} \tag{5}$$

The problem (5) is called **approach problem** (or discreet problem) gotten, by a method of finite differences, by opposition to the problem (3) declares a continuous problem. The vectorial shape of (5) is presented as follows:

Find $u = (u_1, ..., u_n)^T$ $(with\ u_0 = \alpha\ et\ u_{n+1} = \beta)$, as $F(u) = 0$ with:

$$F(u) = \begin{bmatrix} F_1(u) \\ F_2(u) \\ . \\ . \\ . \\ F_{n-1}(u) \\ F_n(u) \end{bmatrix} = \begin{bmatrix} \dfrac{-\alpha + (2 + c_1)u_1 - u_2}{h^2} - f(u_1) \\ \dfrac{-u_1 + (2 + c_2)u_2 - u_3}{h^2} - f(u_2) \\ . \\ . \\ . \\ \dfrac{-u_{n-2} + (2 + c_{n-1})u_{n-1} - u_n}{h^2} - f(u_{n-1}) \\ \dfrac{-u_{n-1} + (2 + c_n)u_n - \beta}{h^2} - f(u_n) \end{bmatrix} \tag{6}$$

We, therefore, represent by (6) a non-linear system of n equations for the n unknown $(u_1, ..., u_n)$. Hence, to solve this system, it is necessary to linearize while using one of the following iterative methods: method of the successive approximations, Newton's method, Newton-cord and Shamanski's methods [6]. These methods look for a linearization by a highly determined procedure. For

example, the use of Newton's method is based on the following formula:

$$u^{j+1} = u^j - [DF(u^j)]^{-1} F(u^j),\ j = 0,1,2,3,... \tag{7}$$

by means of DF which is the Jacobian matrix (Jacobian). Therefore DF (u) of (6) =

$$\frac{1}{h^2}\begin{bmatrix} 2+c_1-h^2\dfrac{\partial f(u_1)}{\partial u_1} & -1 & 0 & . & . & . & 0 \\ -1 & 2+c_2-h^2\dfrac{\partial f(u_2)}{\partial u_2} & -1 & 0 & . & . & 0 \\ 0 & & & & & & . \\ . & & & & & & . \\ . & & & & & & . \\ . & & & & & & . \\ . & & & & & & 0 \\ & & & & & & -1 \\ 0 & . & & . & . & 0 & -1 \quad 2+c_n-h^2\dfrac{\partial f(u_n)}{\partial u_n} \end{bmatrix}$$

Which is a sparse matrix (three diagonals) dependent on the problem solution (6). Consequently, to calculate u^{j+1} from u^j, it is done by solving the following system:

$$DF(u^n)(u^n - u^{n+1}) = F(u^n)$$

The algorithm of resolution proceeds as follows:

$j = 0$

$s = 1 \quad \{s \text{ is a mesure of convergence}\}$

while $s > \varepsilon$ *and* $j \le n\max\{\max imal\ number\ of\ iteration\}$

Calculate $F(u^j)$

Calculate $DF(u^j)$

Solve $DF(u^j)y = F(u^j)$ $\{$*by a linear solver*$\}$

Calculate $u^{j+1} = u^j - y$

Calculate $s = \dfrac{\left\|F(u^{j+1})\right\|}{\left\|F(u^0)\right\|}$

$j = j + 1$

end

For such an iterative method (Newton's method), it is normal to ask the following questions:

1) **Existence:** Is the method well defined? $DF^{-1}(u^j)$ It exists at every iteration.

2) **Convergence**: The continuation $\{u^j\}$ is its convergent in the way that $\lim_{j\to\infty} u^j = u$ where u verifies $F(u) = 0$?

For example, for the problem (5), we demonstrate the following:

Theorem 3.1. *Let u be a solution of the system (5). Let* $\overline{f} = maz_{1\le i\le n}\left|\dfrac{\partial f(u_i)}{u_i}\right|$. *if c, h and* $\overline{f}$ *verify the condition* $c > h^2\overline{f}, \forall i, 1 \le i \le n,$ *then* $DF^{-1}(u)$ *exists.*

Proof. It is obvious that the matrix $DF(u)$ is symmetrical, $DF(u)$ is also defined positive. Indeed:

$\forall x \in IR^n$ *we have* $x^t DF(u)x =$

$$\frac{1}{h^2}\begin{bmatrix} (2 + c_1 - h^2\dfrac{\partial f(u_1)}{\partial u_1})x_1 - x_2, -x_1 + (2 + c_2 - h^2\dfrac{\partial f(u_2)}{\partial u_2})x_2 - x_3, \ldots, \\ -x_{n-1} + (2 + c_n - h^2\dfrac{\partial f(u_n)}{\partial u_n})x_n \end{bmatrix}\begin{bmatrix} x_1 \\ . \\ . \\ . \\ x_n \end{bmatrix}$$

With

$$2 + c_i - h^2 \frac{\partial f(u_i)}{\partial u_i} > 2 \Rightarrow x^t DF(u)x \geq \frac{1}{h^2}\left[x_1^2 + x_n^2 + \sum_{i=2}^{n}(x_i - x_{i-1})^2\right] \geq 0 \therefore$$

The outcome of this result is that if the initial condition u^0 of the iterative algorithm is chosen close to *u* then $DF^{-1}(u^0)$ on hand. If the sequence of the iterations $\{u^j\}$ exists in the region, then $DF^{-1}(u^j)$ exist.
In this concern, we can find more precise results on the convergence of Newton's method, for example in [2]**:**

Theorem 3.2. *If F is continuously derivable two times in relation to the variables* u_j, $1 \leq j \leq n$, *and if* $\bar{u}$ *is as* $F(\bar{u})$*=0 and if* $DF(\bar{u})$ *is then regular, the continuation* (u^j) *defined by Newton's method converges towards u when* $n \to \infty$ *provided that* u^0 *is chosen sufficiently close to u.*

It appears therefore that the choice of u^0 sufficiently close to *u* is fundamental.

4 Method of Continuation

The above stated iterative methods give the solutions that converge locally toward the solution a of $F(u) = 0$, so when the initial condition is chosen close to the exact solution. The question of choosing the initial condition is asked therefore to find an efficient method permitting to guarantee a perceptive choice. It is then the method of **Continuation** that permits to introduce a precise approach proposes forcing recognition to a parameter $t \in [0,1]$ and hence of $t = 0$, to make $t = 1$ to the solution a of $F(u)$. This method has been introduced by [3]. From that time, it has been used by several authors **Avila** [4], [5, 6]. This last [6] introduces it in the setting of the equations in the non-linear partial differential, while leaning on the method of the topological degree of Leray-Schauder. This technique demonstrated all its power in the analysis of the solutions existence of the non-linear **PDE**.

5 Principle of the Continuation method

The Continuation method consists of proposing the problem $F(u) = 0,$ in the setting of a related problems parameterized by a variable $t \in [0,1].$

Either $F : [0,1] \times IR^n \to IR^n$
$(t,u) \to F(t,u)$

As $F(1,u) = F(u), \forall u \in IR^n$.

We also supposes that for $t = 0$, the problem $F(0,u^0) = 0$ admits a unique solution u^0, capable to be calculated by the use of a simple algorithm. The choice of the initial condition for the answer calculation $t = 1$, depend undoubtedly on u^0. So while following the related solutions $F(t,u^t) = 0$, from $t = 0$, one can make $t = 1$ to the solution of $F(u) = 0$ under precise hypotheses on the related functions $F(.,.)$. It is the principle of the **Continuation** method**.** Numerically, it results by the successive research of the problem solutions, $F(t_j, a_j) = 0, 0 < j < m$ with $0 = t_0 < t_1 < t_2 < ... < t_m = 1$, or $\{t_i\}$ is a discretization of $[0,1]$. The ultimate phase of this process is finite for $t_m = 1$, giving then, $a_m = a$ verifying: $F(t_m, a_m) = F(1, a_m) = F(a_m) = 0.$

Suppose the existence of the solution a_j (or of an approximation $\overline{a_j}$ of a_j) as: $F(t_j, a_j) = 0$ (or $\left\| F(t_j, \overline{a_j}) \right\| \le \varepsilon_j, \varepsilon_j$ small). Search for the pair $\{t_{j+1}, a_{j+1}\}$ with $t_j < t_{j+1} \le 1, a_{j+1} \in IR^n$, as $F(t_{j+1}, a_{j+1}) = 0$. Conduct the resolution of the non- linear system $F(t_{j+1}, a_{j+1}) = 0$ by a method of Newton type, Newton cord or Shamanski getting the continuation therefore: $\{a_{j+1}^k / k = 0,1,2,...,k_{j+1}\}$ with $a_{j+1}^0 = a_j^{k_j}, k_j$ denoting the indication of the last applied iteration in the approximation of a_j.

Example 5.1. The equation of Poisson

Let the following equation of Poisson: $-\Delta u + f(u) = g$ is solved thus by a method of Continuation. To the parameter $t \in [0,1]$, look for the solution u_t of the problem: $-\Delta u_t + tf(u_t) = g$. The discretization of this model by the method of the **finite Elements** gives then $Au_t + tIf(u_t) = G$, or A is the matrix of rigidity, $If(u_t)$ is the diagonal matrix:

$$\begin{pmatrix} f(u_{t,1}) & 0 & . & . & 0 \\ 0 & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & 0 \\ 0 & . & . & 0 & f(u_{t,n}) \end{pmatrix}$$

When $t = 0$, the previous model is linearized to give u_0 the solution of: $Au_0 = G$ resolute by an adequate linear solver. As indicated above, we pass from $t_0 = 0$ to $t_m = 1$, by the resolution of *n* non-linear systems. The method performance is going to be bound therefore to a reduction of the number of step m with minimum iterations to every $t_j, j = 1,...,m.$

6 Method of Newton - Continuation

An application of Newton method to every step $t_j, j = 1,...,m.$ gives:

$$\left.\begin{array}{l} x^{j,k+1} = x^{j,k} - DF_x(x^{j,k}, t_j).F(x^{j,k}, t_j), \\ x^{1,0} = x^0, x^{j+1,0} = x^{j,m_j} \end{array}\right\} k = 0,..., m_j - 1;\ j = 1,..., N-1$$

$$x^{N,k+1} = x^{N,k} - DF_x(x^{N,k},1).F(x^{N,k},1)$$

with DF_x the Jacobean of F in relation to x and $0 = t_0 < t_1 < ... < t_m = 1.$
At this level, it would be necessary to study the **Convergence** and the **choice of the discretization** of $\{t_i\}$.

7 Convergence

For the convergence of this method, we give two results that extract respectively from [4, 3]:

Theorem 7.1. *Let* $F : [0,1] \times IR^n \to IR^n$, *suppose that* DF_x (the *Jacobean of F in relation to* $x \in IR^n$) *exist and continuous on* $[0,1] \times IR^n$. *Besides,* DF_x^{-1} *exists for all* $t \in [0,1]$. *Then a discretization exists* $t_i, 0 = t_0 < t_0 < ... < t_m = 1$, *and N-1 entire,* $m_1, m_2,..., m_{N-1}$ *as* $\lim_{k \to \infty} x^{N,K} = x(1)$ *∴.*

Theorem 7.2. *Let* $F : IR^n \to IR^n$ *a function continuously derivable in relation to the components of the vector in* IR^n, *suppose that DF (x) is regular for all* $x \in IR^n$ *and that* $\left\| DF_x^{-1} \right\| \le \beta \ \forall x \in IR^n$ *then the continuation method converges.*

In [6], we demonstrate that in the case of Poisson non-linear equation, if the

function $f(.)$ is monotony in the sense that $\int_{\Omega}(f(u_1)-f(u_2))d\Omega \geq 0$ the hypotheses of these theorems are then well verified.

The two previous theorems demonstrate the existence of a discretization so $\{t_i\}$ at $[0,1]$ for which we have convergence. **It doesn't give a method to distribute the $\{t_i\}$ leading to the convergence of the Continuation method. In what follows, we intend to approach this question.**

8 Choosing the Discretization of $\{t_i\}$

The choice of the discretization $\{t_i\}$ consists of estimating the step $\tau_i = t_{i+1} - t_i$, for t_i given, in order to achieve two main objectives:

- Minimize the number"n" of the discretization in the interval [0, 1].
- Reduce, to the minimum, the numbers of iterations, "k_i" necessary for the convergence of the Newton method used to every discretization $\{t_i\}$.

Let's note that it is demonstrated in [4], that if $t_{i+1} - t_i$ is sufficiently small, "k_i" to every step would be equal to 1.

These two criteria minimize the function to two variables $\sum_{i=1}^{n} k_i$. The principle is to have $k_i = 1$ and to optimize the value of n. Let's consider then the case of non-linear **Poisson** and continuous problem. Let $F(t,u) = -\Delta u + tf(u)$, we deduced easily that $F_t(t,u) = f(u)$, knowing that ($F_t(t,u)$ the derivative of *F* in relation to t). Let's consider the approached solution u_i to the discretization t_i, c to d $F(t_i,u_i) \cong 0$. This solution will be, according to the principle of the continuation method, the 1st iteration of the problem to the discretization t_{i+1} and let's try to estimate $F(t_{i+1},u_i)$. We verify the following identity easily:

$$F(t_{i+1},u_i) = -\Delta u_i + t_{i+1} f(u_i) = -\Delta u_i + (t_i + \tau_i) f(u_i) \cong (t_{i+1} - t_i) f(u_i)$$

It gives:

$$\|F(t_{i+1},u_i\| = (t_{i+1} - t_i)\| f(u_i)\| \qquad (8)$$

suppose that for t_i, the choice of the initial condition u_0^i, is managed by the condition

$\frac{\|F(t_i,u_i^0)\|}{\|F(1,u_0)\|} \leq tol_0$ (tol_0 =given tolerance). So the equation (8) gives:

$$\frac{\tau_i \left\| f(u_i) \right\|}{\left\| f(u_0) \right\|} \leq tol_0 \text{ , let}$$

$$\tau_i = \frac{tol_0}{\left\| f(u_i) \right\|} \left\| f(u_0) \right\| \quad (9)$$

Then the determination of tol_0 becomes an indispensable step for the implementation of the discretization descended from (9). We propose for it the following algorithm:

9 Algorithm (Matlab)

```
function [u,i,k] = continuation(tol,tol0,F,DF,A,f,g,kmax)
% input:
%- tol is the precision of calculation to every t
%- tol0 is the precision asked in the choice of ti
%- F=F(t,u)=Au + tf - g = 0
%- DF is the Jacobean matrix in relation to u
% output:
%- u the solution of F(1,u) = 0
%- i the number of step
%- k the total number of iterations
%- kmax the maximal number of iterations
%- Resolution of F(0,u) = 0
u0 = A\g ;
t = min(1,t + tau);
i = i + 1;
[u,verif,ki] = newton(F,DF,t,u,tol,kmax);
tau = tol0*norm(f(u0)) / norm(f(u));
k = k + ki;
if (ki > 1)
tol0 = tol0/2;
end
if (i > 1)
i = i-1;
end
end
```

10 Applications

In this paragraph, we will validate the theories exposed to the previous paragraph. We carry out numerical tests dividing in two parts:

1- Resolution of the equation of non-linear Poisson has one dimension. The discrete system obtained by approach-finished differences is solved by the method of Newton and the method of Continuation. The choice of initial condition in this last method illustrates its importance.

2 - Study of the best possible partition of [0, 1] in continuation method. This

study searches a set of $\{t_i\} \in [0,1]$ involving a passage from 0 to 1 with the minimum operations.

Application 10.1. In this application we will solve the following non-linear differential equation:

$$\begin{cases} -\dfrac{d^2u}{dx^2} + e^u = 2 + e^{x(1-x)} \\ u(0) = u(1) = 0 \end{cases}$$

where $u = x(1-x)$ is an exact solution.

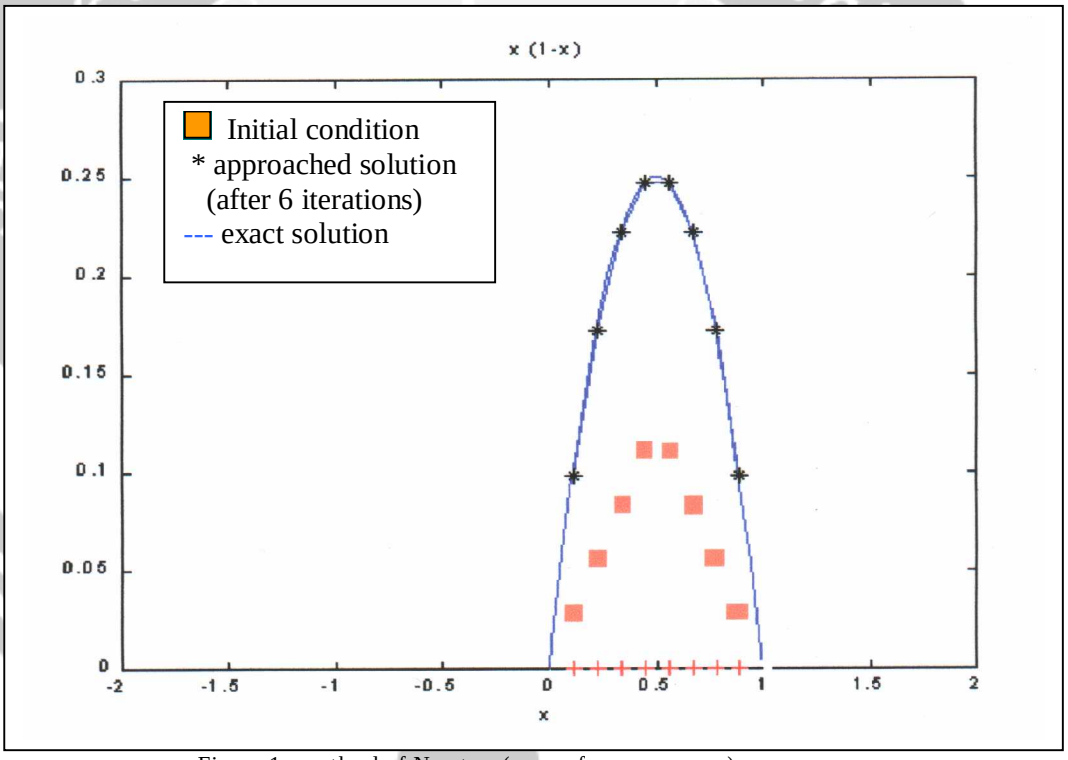


Figure1: method of Newton (case of convergence)

The figure 1 gives the solution of this model with a Discretization of 8 points of the domain [0, 1], by using the method of Newton (represent the initial solution, * the approached solution, the continuous curve represents the exact solution).

we notice that the method of Newton converges in this example, that is because the initial condition is chosen well. On the other hand if this condition is far from the exact solution (u_0 = (-3.4, 9.75, 2.875, -37.5, 0.375, 2.5, -0.1038, -0.0331), the method of Newton diverge (fig. 2).

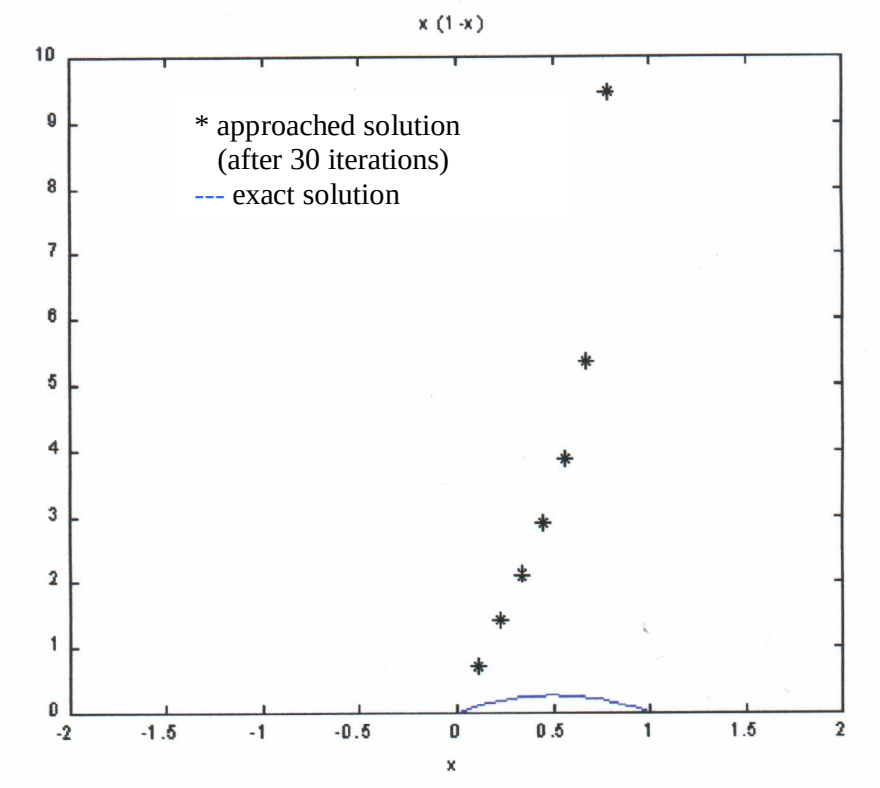


Figure 2: method of Newton (case of divergence after 30 iterations)

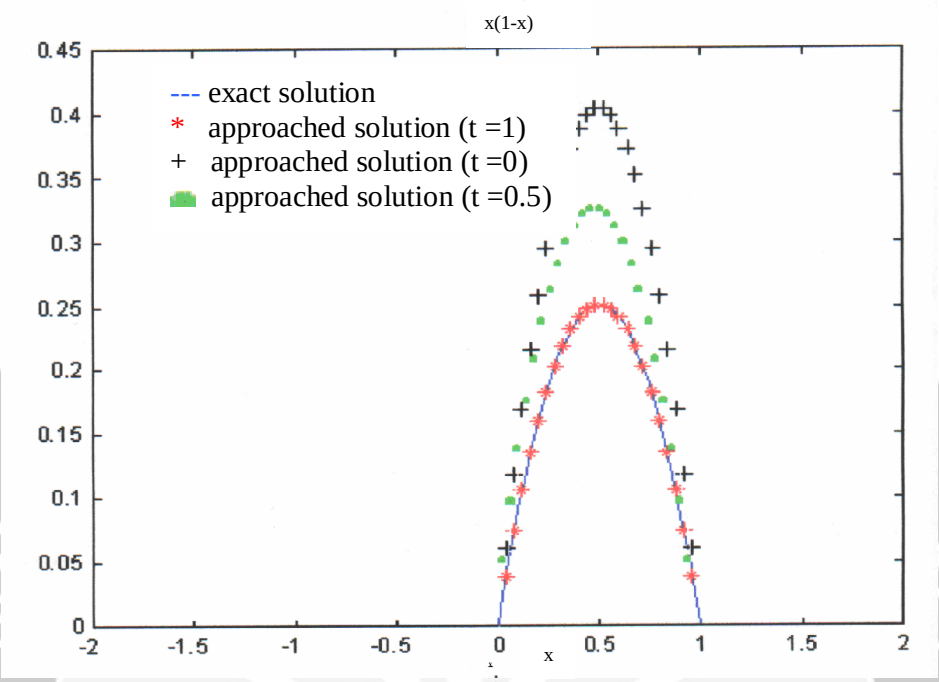


Figure 3: method of continuation

The figure 3 represents the solution while using the method of continuation with a discretization of 24 points of the domain [0, 1], and a step of t=0.1 and k=1(number of iterations for this step), immediately we notice the convergence of this method.

Application 10.2. In this test, we study the choice of the partition of the $\{t_i\}$. On the basis of the previous example, while using the method of continuation, but with a step equals to $\frac{1}{3}$, i.e. we divide the interval [0, 1] into 3 parts, in this case we notice that the final solution in t =1, with K_i=1(i=1, 2, 3) is not close to the exact solution of figure 4, therefore the partition of them $\{t_i\}$ influences directly the solution. In other words of the convergence, it is required to study it or to find the best partition, the minimum of iterations for every step of the partition with, at the same time, an acceptable solution (i.e. when $\|U_{calculated}-U_{exact}\| / \|U_{exact}\| \leq$ tol)

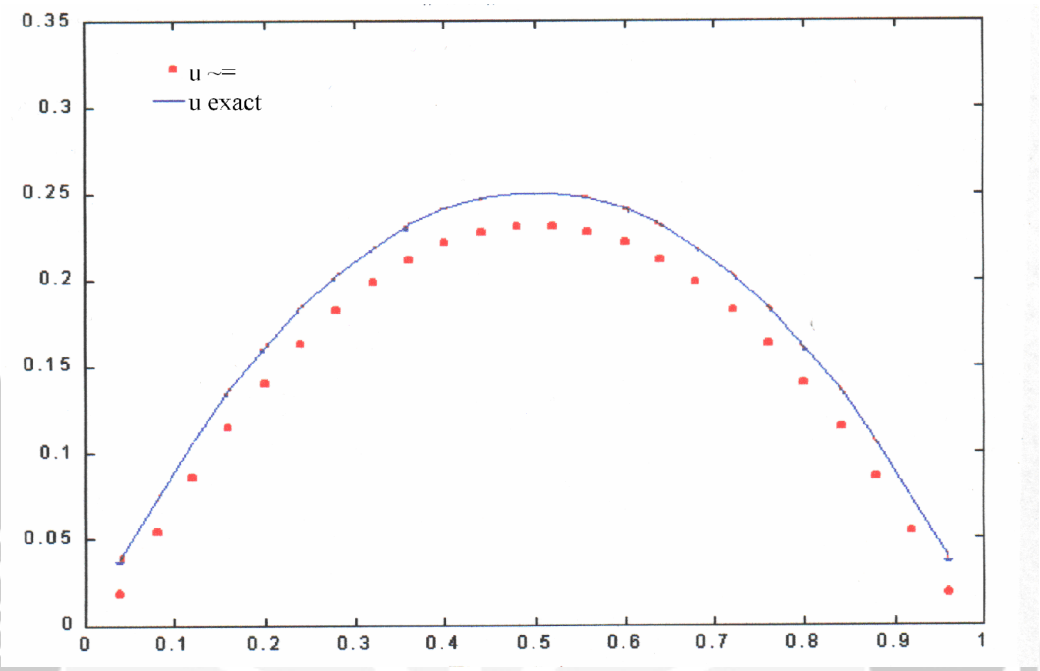


Figure 4

For that, we will make two approaches:

First approach

If we apply the theorem of the choice of partitions stated in previous paragraph, We get the following table:

Choice of tol_0	Number of t_i	$\sum k_i$ (k_i=1)	acceptable (a) or unacceptable (u) solution	flops
0.01	261	261	a	60552
0.1	27	27	a	6264
0.5	6	6	a	1392
0.6	5	5	a	1160
0.7	1	1	u	928
.....	1	1	u	....

Therefore the optimal choice according to that approaches is gotten for tol_0=0.6(fixed for all partitions). Because it gives the best partition (n=5), with k_i=1(fix in this approach), and a minimum of flops with an acceptable solution.

Second approach

In this approach, tol_0 is not anymore constant. We must apply the algorithm

expressed previously, and the formula giving τ_i according to tol_0, $f(u_i)$ and $f(u_0)$. We get the following table:

Choice of ε	Number of t_i	$\sum k_i$($k_i$vary)	acceptable (a) or unacceptable(u) solution	flops
0.01	8	16	a	3879
0.1	7	14	a	3363
0.5	5	10	a	2331
0.6	4	8	a	1815
0.7	1	1	u	…
…	1	1	u	….

Therefore the optimal choice according to that approaches is gotten for tol_0=0.6 because it gives the best partition (n=4*t), with k=8(fixed in this approach), and a minimum of flops with an acceptable solution.

11 Conclusion

In this paper, our goal is to design efficient and powerful algorithms to solve a non-linear differential system. The iterative methods (Newton, Newton-Raphson, Shamanski...) give solutions that converge locally to the solution a of $F(u)=0$, whenever the initial condition is chosen close to the exact solution. The question of choosing the initial condition is asked therefore to find an efficient method permitting to guarantee a perceptive choice. It is then the method of Continuation which enables to introduce a rigorous approach that uses a parameter t in [0,1] to force convergence. We proposed also a new technique for the partitioning of the parameter set [0,1]. This technique has shown a great potential to analyze existence of solutions of non-linear PDEs.

12 References

[1]: T. Gmûr, Dynamique des structures (1997). Presses Polytechnique et Universitaires Romandes.

[2]: J. Rappaz et M. Picasso (2001), Introduction à l'analyse numérique.

[3]: J.M. Ortega and W.C. Rheinboldt(1970), iterative solution of non-linear equations.

[4]: W.C. Rheinboldt (1974), Methods for solving systems of non-linear equations.

[5]: K. Allgower and T. Georg (1995), Finite element method.

[6]: S. Kadry and N. Nassif (2001), memoire DEA modelisation calcul intensif.

[7]: G. Evans (2000), Numerical methods for partial differential equations.

[8]: A. Quarteroni(2000), numerical mathematics.

Seifedine Kadry: Ph D student at M3M laboratory, University of technology of Belfort-Montbeliard(FRANCE)
E-mail: skadry@gmail.com

A CONVERGENCE ANALYSIS OF THE BROYDEN-SD METHOD FOR THE UNCONSTRAINED OPTIMIZATION

[a] Mustafa Mamat, [a] Yosza Dasril, [a] Ismail Mohd and [b] Leong Wah June

[a] Jabatan Matematik, Fakulti Sains dan Teknologi, Kolej Universiti Sains dan Teknologi Malaysia (KUSTEM), 21030 Kuala Terengganu.
[b] Jabatan Matematik, Fakulti Sains dan Alam Sekitar, Universiti Putra Malaysia 43400 UPM Serdang Selangor

Abstract. In this article we discuss the convergence on combination of search direction for the Broyden and the steepest descent methods. In particular, we analyse the relation between superlinear convergence and the summability of $\left\|x^{(k+1)} - x^*\right\|$, and the relation between the two angles θ_k and υ_k where θ_k and υ_k denote respectively, the angle between s_k and $B_k s_k$ and the angle between s_k and $-g_k$.

Keywords: Broyden-SD method, superlinearly convergent, search direction combination, Hessian approximation.

1. Introduction

Consider the unconstrained optimization problem

$$\min f(x) \tag{1.1}$$

where f is twice continuously differentiable function from $\Re^n$ into $\Re$.

The Quasi-Newton methods are known to be among the most popular methods used to solve problem (1.1). The Quasi-Newton methods are an iterative method, whereby at the $(k+1)$th iteration, $x^{(k+1)}$ is obtained using the following equation

$$x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)} \tag{1.2}$$

where $d^{(k)}$ denotes the search direction and α_k its stepsize. The search direction, $d^{(k)}$ is calculated using

$$d^{(k)} = -B_k^{-1} g_k \tag{1.3}$$

The quantity $g_k = \nabla f\left(x^{(k)}\right)$ denotes the gradient of f at $x^{(k)}$ while B_k is the Hessian approximation $\nabla^2 f\left(x^{(k)}\right)$ that fulfills the Quasi-Newton equation

$$B_k s_k = y_k \tag{1.4}$$

Observe that the step length in (1.2) may be obtained by taking $\min_{\alpha \geq 0} f\left(x + \alpha d^{(k)}\right)$.

Within the Quasi-Newton methods, the Broyden family of method forms a very important class. In the Broyden family, the Hessian approximation for the B_k update is generated using Broyden formula [1].

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k} + \phi_k \left(s_k^T B_k s_k\right) w_k w_k^T \tag{1.5}$$

where $s_k = x^{(k+1)} - x^{(k)}$, $y_k = g_{k+1} - g_k$ and $w_k = \frac{y_k}{y_k^T s_k} - \frac{B_k s_k}{s_k^T B_k s_k}$. In formula (1.5) ϕ_k is a parameter.

According to Dennis and More [4], there are exist two update formulas which contained in the Broyden family namely, the BFGS update formula for $\phi_k = 0$ and the DFP update formula for $\phi_k = 1$. Consequently, (1.5) may be written as

$$B_{k+1} = (1-\phi_k) B_{k+1}^{BFGS} + \phi_k B_k^{DFP} \tag{1.6}$$

We are making the assumption that the stepsize, α_k fulfills both Wolfe's conditions, that,

$$f\left(x^{(k)} + \alpha_k d^{(k)}\right) \leq f\left(x^{(k)}\right) + \beta_1 \alpha_k g^{(k)T} d^{(k)} \tag{1.7}$$

$$g\left(x^{(k)} + \alpha_k d^{(k)}\right)^T d_k \geq \beta_2 g_k^T d^{(k)} \tag{1.8}$$

with $0 < \beta_1 < \frac{1}{2}$ and $\beta_1 < \beta_2 < 1$.

If we let $\phi_k \in [0,1]$ then (1.5) is called the Broyden convex family. However, if $\phi_k \in [0,1-\sigma]$ for $\sigma \in (0,1]$ then (1.5) is called the restricted Broyden family (Byrd, Nocedal dan Yuan, 1987).

Suppose that x^* is a minimizer for f and let the Hessian matrix $G\left(x^*\right)$ for f at x^* be positive definite. Dennis and More [4] proved that if the stepsize is always taken to be $\alpha_k = 1$, either for the BFGS update formula or the DFP update formula, and if it satisfies

$$\sum_{k=0}^{\infty} \left\| x^{(k+1)} - x^* \right\| < \infty \tag{1.9}$$

then $\left\{x^{(k)}\right\}$ converges to x^* at a superlinear rate.

2. Combination of QN-SD Search Direction

The QN-SD search direction is characterized by

$$d^{(k)} = -\eta_k H_k g_k - \delta_k g_k \tag{2.1}$$

where parameters $\eta > 0$ and $\delta > 0$ respectively.

The substitution of equation (2.1) into equation (1.2) produces

$$x^{(k+1)} = x^{(k)} + \alpha_k \left[-\eta_k H_k g_k - \delta_k g_k \right]$$

Then, we obtain the stepsize by using

$$\min_{\alpha,\delta,\eta>0} f\left(x^{(k+1)}\right) = \min_{\alpha,\delta,\eta>0} f\left(x^{(k)} - \alpha\left[\eta H_k g_k + \delta g_k\right]\right)$$

with α_k satisfying for (1.7) and (1.8).

Consequently we obtain,

$$x^{(k+1)} = x^{(k)} + \alpha_k \left(-\eta_k^* H_k g_k - \delta_k^* g_k \right).$$

Details can be referred in [5].

3. Convergence Analysis

By using QN-SD search direction combination, Mustafa *et al* [6] had proved that the Quasi-Newton method coverges globally at a superlinear rate. In what follows is a collection of assumptions and theorems that will be useful in our discussion on the Broyden-SD method.

Definition 3.1
The level set defined by $L = \left\{ x : f(x) \le f\left(x^{(0)}\right) \right\}$ is bounded where $x^{(0)}$ is the initial point.

Assumption 3.1
[a] The objective function f is twice continuously differentiable.
[b] The level set L is convex. Moreover, there exist positive constants c_1 and c_2 satisfying

$$c_1 \|z\|^2 \le z^T G(x) z \le c_2 \|z\|^2$$

for all $z \in \Re^n$ and $x \in L$, where $G(x)$ is the Hessian matrix for f.

Assumption 3.2
The Hessian matrix is Lipschitz continuous at the point x^*, which is exists a positive constant c_3 satisfying

$$\left\| G(x) - G\left(x^*\right) \right\| \le c_3 \left\| x - x^* \right\|$$

for all x in a neighborhood of x^*.

Theorem 3.1 (Global Convergence)
Suppose that the Assumption 3.1 is satisfied. Then

$$\lim_{k\to\infty} \| g_k \| = 0 .$$

Proof: Refer [6].

Theorem 3.2 (Superlinear convergence)
Suppose that the Assumption 3.2 is satisfied. Furthermore, suppose that $\{x^{(k)}\} \to x^*$ and sequences $\{\| B_k \|\}$ and $\{\|H_k\|\}$ are bounded. If $x^{(k+1)} = x^{(k)} + d^{(k)}$ is satisfied for all sufficiently large k and if

$$\lim_{k\to\infty} \frac{\left\| \left[B_k - G(x^*) d^{(k)} \right] \right\|}{\left\| d^{(k)} \right\|} = 0 \qquad (3.1)$$

then

$$\lim_{k\to\infty} \frac{\left\| x^{(k+1)} - x^* \right\|}{\left\| x^{(k+1)} - x^{(k)} \right\|} = 0 \qquad (3.2)$$

By defining $e_k = \max \left\{ \left\| x^{(k+1)} - x^* \right\|, \left\| x^{(k)} - x^* \right\| \right\}$, then the Assumption 3.1 together with Theorem 3.1 ensures that $x^{(k)}$ approaches x^*. As a consequence, (1.9) is satisfied. This, in turn, strengthens superlinear convergence.

Lemma 3.1
If the Assumptions 3.1 and 3.2 are fulfilled, then exist a sequence of numbers $\{ \varepsilon_k \}$ with

$$\frac{\left\| y_k - G(x^*) s_k \right\|}{\left\| s_k \right\|} \le \varepsilon_k \text{ and } \sum_{k=1}^{\infty} \varepsilon_k < \infty .$$

Let the angle between s_k and $B_k s_k$ and the angle between d_k and $-g_k$ be denoted by θ_k and υ_k respectively. Then define $\tilde{s}_k = G(x^*)^{1/2} s_k$, $\tilde{y}_k = G(x^*)^{-1/2} y_k$ and $\tilde{B}_k = G(x^*)^{-1/2} B_k G(x^*)^{-1/2}$. Next, let $\tilde{\theta}_k$ denote the angle between $\tilde{s}_k$ and $\tilde{B}_k \tilde{s}_k$. What is the relation between θ_k and $\tilde{\theta}_k$? This question is answered in the lemma as follows.

Lemma 3.2
Suppose that Assumption 3.1 is fulfilled, then $\cos\theta_k \ge c_4 \cos \tilde{\theta}_k$, c_4 being a positive constant.
Proof:
From the definition we obtain

$$\cos\tilde{\theta}_k = \frac{\tilde{s}_k^T \tilde{B}_k \tilde{s}_k}{\|\tilde{s}_k\|\|\tilde{B}_k \tilde{s}_k\|} = \frac{s_k^T B_k s_k}{\left\|G(x^*)^{1/2} s_k\right\|\left\|G(x^*)^{-1/2} B_k s_k\right\|}$$

$$= \frac{s_k^T B_k s_k}{\sqrt{\left(s_k^T G(x^*) s_k\right)\left[(B_k s_k)^T G(x^*)^{-1}(B_k s_k)\right]}} \tag{3.3}$$

With Assumption 3.1 (b), it is clear that (3.3) can be written as

$$\cos\tilde{\theta}_k \le \frac{s_k^T B_k s_k}{c_3 c_3' \|s_k\|\|B_k s_k\|}$$

$$\le \frac{1}{c_4}\cos\theta_k$$

$$\cos\theta_k \ge c_4 \cos\tilde{\theta}_k \tag{3.4}$$

where $c_4 = c_3 c_3'$ is a positive constant.

The following lemma provides a relation between θ_k and υ_k.

Lemma 3.3

Suppose that the conditions in Assumption 3.1 are fulfilled. If $\cos\theta_k \ge c_5$, then $\cos\upsilon_k \ge c_6 \cos\theta_k$ with c_5 and c_6 being positive constants satisfying all sufficiently large k.

Proof:

We know that

$$-g_k^T d^{(k)} = -g_k^T B_k^{-1} B_k d^{(k)} \tag{3.5}$$

Using (2.1), we can obtain

$$d^{(k)T} = -\eta_k g_k^T B_k^{-1} - \delta_k g_k^T$$

so that

$$d^{(k)T} B_k d^{(k)} = \left(-\eta_k g_k^T B_k^{-1} - \delta_k g_k^T\right) B_k d^{(k)} \tag{3.6}$$

The combination of (3.5) and (3.6) yield

$$-g_k^T d^{(k)} = \frac{1}{\eta_k}\left[d^{(k)T} B_k d^{(k)} + \delta_k g_k^T B_k d_k\right]$$

$$\frac{-g_k^T d^{(k)}}{d_k^T B_k d^{(k)}} = \frac{1}{\eta_k}\left[1 + \frac{\delta_k g_k^T B_k d^{(k)}}{d^{(k)T} B_k d^{(k)}}\right]$$

$$\geq \frac{1}{\eta_k}\left[1-\frac{\delta_k\|g_k\|}{c_5\|d^{(k)}\|}\right]$$

$$\geq \frac{c_6}{\eta_k} \tag{3.7}$$

Further from the definition, $-g_k^T d^{(k)} = \|g_k\|\,\|d^{(k)}\|\cos\upsilon_k$ and $d^{(k)T} B_k d^{(k)} = \|d^{(k)}\|\|B_k d^{(k)}\|\ \text{kos}\,\theta_k$ so that (3.7) can be written as

$$\frac{\|g_k\|\cos\upsilon_k}{\|B_k d^{(k)}\|\cos\theta_k} \geq \frac{c_6}{\eta_k} \tag{3.8}$$

Since $\|B_k d^{(k)}\| \geq \eta_k\|g_k\|$, upon combining with (3.8) and simplifying we obtain

$$\frac{kos\,\upsilon_k}{kos\,\theta_k} \geq c_6$$

As a result, Lemma 3.2 is proved.

□

Lemma 3.4

If $\{s_k\}$ and $\{y_k\}$ satisfy the Lemma 3.1 then $\lim_{k\to\infty}\frac{\|[B_k - G(x^*)]s_k\|}{\|s_k\|} = 0$ and sequences $\{\|B_k\|\}$ and $\{\|H_k\|\}$ are bounded.

Thus, we can conclude from Theorem 3.1 and Lemma 3.4 that the resstricted Broyden-SD method converges superlinearly.

4. Conclusion

In this paper we determined the suitable values of η and δ. By combination of QN-SD search direction we proved that the convergent is superlinear and the stepsize have to satisfy the Wolfe conditions [7].

Convergence analysis may also be discussed by means of the trace and the determinant of B_{k+1} as suggested by Byrd, Nocedal dan Yuan [2] and Byrd and Nocedal [3]. By using a similar idea we shall create a modification of the Broyden-SD method in the future.

5. References

[1] Broyden, C.G (1967), *Quasi-Newton methods and application to function minimization,* Math. Comp., 21, pp 368-381.
[2] Byrd, R.H., Nocedal, J., and Yuan, Y.X (1987), *Global Convergence of a class of Quasi-Newton Methods on Convex Problems*, SIAM J. Numerical Analysis, Vol 24 (5), pp 1171-1189.
[3] Byrd, R.H., and Nocedal, J., (1989), *A tool for the analysis of Quasi-Newton Methods with Application to unconstrained optimization*, SIAM J. Numerical Analysis, Vol 26 (3), pp 727-739.
[4] Dennis, J.E., and More, J.J, (1977), *Quasi-Newton methods, motivation and theory*, SIAM Review 19, pp 46-89.
[5] Mustafa Mamat, Yosza Dasril and Ismail Mohd, (2004), *Algoritma Arah Carian Kaedah BFGS-SD Dalam Pengoptimuman*, Prosiding Seminar Kebangsaan Sains Pemutusan, UUM at Holiday Inn Resort Penang, 15 – 17 December 2004.
[6] Mustafa Mamat, Yosza Dasril and Ismail Mohd, (2004), *Kaedah Quasi-Newton untuk pengoptimuman tak berkekangan*, Prosiding Simposium Kebangsaan Sains Matematik ke 12, at UIA, Gombak Selangor, 23 – 25 December 2004.
[7] Mustafa Mamat, Yosza Dasril and Ismail Mohd, (2005), *Penumpuan Superlinear Gabungan Arah Carian Quasi-Newton dan Penurunan Tercuram Untuk Masalah Pengoptimuman Tak Berkekangan*, Jurnal Matematika UTM Malaysia – accepted for publication.

MUSTAFA MAMAT: Jabatan Matematik, Fakulti Sains dan Teknologi, Kolej Universiti Sains dan Teknologi Malaysia (KUSTEM), 21030 Kuala Terengganu.
E-mail: mus@kustem.edu.my

YOSZA DASRIL: Jabatan Matematik, Fakulti Sains dan Teknologi, Kolej Universiti Sains dan Teknologi Malaysia (KUSTEM), 21030 Kuala Terengganu.
E-mail: yosza@kustem.edu.my

ISMAIL MOHD: Jabatan Matematik, Fakulti Sains dan Teknologi, Kolej Universiti Sains dan Teknologi Malaysia (KUSTEM), 21030 Kuala Terengganu.
E-mail: ismailmd@kustem.edu.my

LEONG WAH JUNE: Jabatan Matematik, Fakulti Sains dan Alam Sekitar, Universiti Putra Malaysia 43400 UPM Serdang Selangor
E-mail: leongwj@putra.upm.edu.my

A MODIFIED LEARNING VECTOR QUANTIZATION WITH GENERATING UNIFORM RANDOM VARIATE FOR TRAINING VECTOR ON SIGNATURE RECOGNITION

Mohammad Isa Irawan[a]

ITS, Surabaya, Indonesia

Abstract. In this paper will be developed Learning Vector Quantization (LVQ) neural network on signature recognition. The purpose of this research is to reduce the number of computing done on learning phase and also recognition. The number of features will be proportional ever greater of the size the pattern input, so that time learning become longer. In principle, this method combine between statistical methods and neural network, what divided to become three phase that is preliminary stage, learning stage and recognition stage. We hope with this modified neural network when learning process and signature recognition become accurate and quicker.

Key-words: modified LVQ, signature recognition, uniform random variate

1. Introduction

Pattern recognition has a long history within electrical engineering but has recently become much more widespread as the automated capture of signals and images has become cheaper. Many of the applications of neural network and statistics method are to classification, and so are within the field of pattern recognition. LVQ is well-known prototype based clustering method, which describes a cluster by a center and possibly some size and shape parameters. Closely related approaches are K-Means Clustering [4, 6] and Fuzzy Clustering [2, 8]. Heidemann [7], proposed a new algorithm for vector quantization, the Activity Equalization Vector Quantization (AEV). It is based on the winner takes all rule with an additional supervision of the average node activities. Villman, et al.[10], combine approaches Generalized LVQ (LVQ) with the neighborhood oriented learning in the neural gas network (NG). They obtain a supervised version of the NG (SNG), that more robust than GLVQ. Sabourin [9], use local granulometric size distribution for signature verification, that fundamental problem in the field of off-line signature verification is the lack of a signature representation based on shape description and pertinent features. And so, Chalechale and Mertins [1], use a novel fast method for line segment extraction in Persian signature recognition. Basically, a lot of modified neural network used for classification and also pattern recognition, start from simplest method like Perceptron and LVQ until the more complex method like Backpropagation and Adaptive Resonance Theory [ART).

Generally, neural network for clustering and also recognition, sum of features or long of training vector as much (M X N) according to amount of pixel in pattern to be recognized. This matter will become problem if size of ever greater pattern, so that require sufficient time at learning process. In this paper, we propose modified LVQ, representing combination between statistical method and neural networks. First, sample taken away from some pattern to be recognized.

after that, threshold of image from the pattern into binary form. Length of training vector, (M x N) reduced to become (p x q) dividedly is region from the pattern. Sum of binary data from each region which is in the form of integer data. Here in after, generating Uniform random variate as much K to each, every pattern to be used as training and reference vector. It is clear will be described at section 3.

2. Learning Vector Quantization (LVQ)

Learning Vector Quantization (LVQ) is mainly influenced by the standard algorithm by Kohonen, that is a pattern classification method in which each output unit represents a particular class or category [3, 5]. Reference vector and also training vector can be in the form of data real, binary and also bipolar as weighted vector or features of pattern will be recognized.

After training, an LVQ net classifies an input vector by assigning it to the same class as the output unit that has its weight vector (reference vector) closest to the input vector. The architecture of an LVQ neural net, shown in Figure 1.

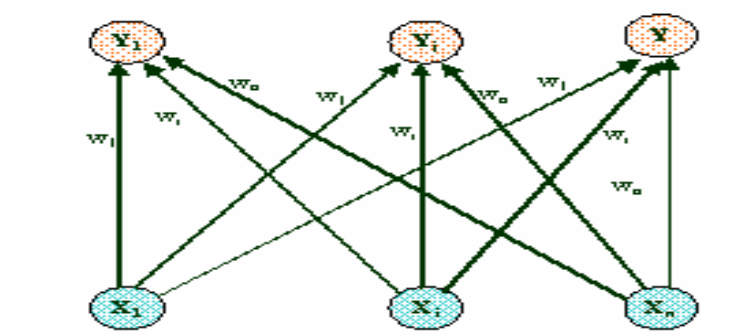


Figure 1. LVQ neural network

Algorithm from this network follow as :
Step 0. Initialize reference vector, learning and rate.
Step 1. While stopping condition is false, do steps 2-6.
Step 2. For each training input vector **x**, do step 3-4.
Step 3. Find J so that $||\mathbf{x} - w_J||$ is a minimum.
Step 4. Update wJ as follows :
If T = C_J, then

$$w_J(new) = w_J(old) + a[\mathbf{x} - w_J(old)]$$

if T $\neq$ C_J, then

$$w_J(new) = w_J(old) - \beta[\mathbf{x} - w_J(old)]$$

Step 5. Reduce learning rate.
Step 6. Test stopping condition:
The condition may specify a fixed number of iterations or learning rate reaching a sufficiently small value.

The nomenclature as follows :

x	training vector (x1,…,xi, …, xn)
T	correct category for the training vector

w_j weight vector for jth output unit (w1j, …, wij,…, wnj).
C_j category represented by jth output unit.
$||x - w_j||$ Euclidean distance between input vector and weight vector for jth output unit.

3. Modified LVQ

Methodology from this modified LVQ divided to become 3 phase. Preliminary Stage more using statistic approach, that is to estimate parameter of Uniform distribution from sample data to generate data to be used as training vector and reference at learning stage by LVQ. At recognition stage used Euclidian measure to look for the winner cluster. In general of this methodology is explained at Figure 2.

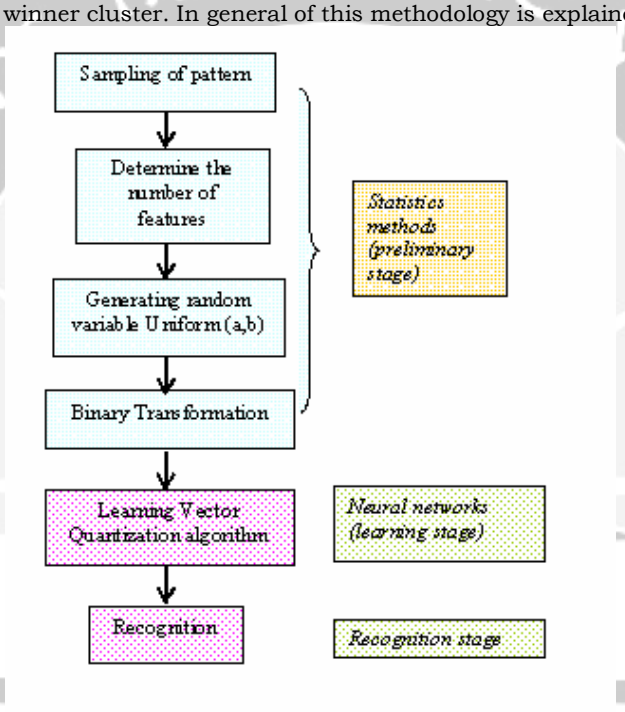


Figure 2. Methodology of Modified LVQ

Preliminary stage
At this phase, first is to take all taken by n sample from each object (pattern). And so from (M X N) the features pattern reduced pattern dividedly become (p x q) features (region). As visible illustration of Figure 3.

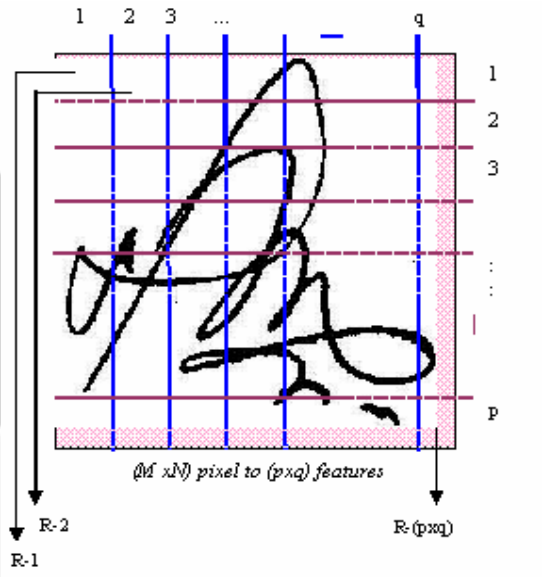


Figure 3. Ilustration : segmentation process

Binary data result of the thresholding pattern at each region ranked among by following:

$$x_k = \sum_{i=1}^{M/p} \sum_{j=1}^{N/q} p_{ij} \,, \qquad (1)$$

k = 1, 2,3 ..., (p x q)

which x_k represent amount of pixel value at one particular region and p_{ij} pixel value from i^{th} line and from j^{th} column of j^{th} at one particular region.

Determination p and q values depend on complexity of pattern to be recognized. For example at recognition of letter, p and q values earn compared to smaller than signature recognition.

x represent vector with length (p x q) as features of a pattern. Assumedly is sample data have Uniform distribution, each n sample of the pattern anticipated by a parameter value α and β, representing value of minimum and maximum value of pixel [of] each region from each pattern sample:

$$\alpha_k = \min \{ x_{ik} \} \qquad (2)$$

and

$$\beta_k = \max \{ x_{ik} \} \qquad (3)$$

i = 1, 2, ...n

j = 1, 2, ..., (pxq)

α_k represent parameter k^{th} feature of α parameter. And so do for parameter β_k

Next step, generating Uniform random variate x features from each, every pattern as much K runs by function

function y=uni(a,b)
y=rand(1)(b-a)+a;*

If there are any as much S of pattern of reference hence there is as much (K X S) training (reference) vector. Advantage from this method that is not require a lot of sample, because pattern awakened by random pursuant to distribution parameter Data of result of evocation which is in the form of data of integer, then transformed to binary form according to the following function :

$$x_k \begin{cases} 1, x_k \geq \gamma \\ 0, x_k < \gamma \end{cases} \quad , \text{k =1, 2, ..., (pxq)} \tag{4}$$

Value γ is grand mean of pixel value from all evocation data :

$$\gamma = \frac{1}{npq} \sum_{i=1}^{pq} \sum_{j=1}^{m} x_{ij} \tag{5}$$

x_{ij} is value of pixel from i^{th} feature and j^{th} data.

Learning Stage
At this phase is data of result of evocation which have transformed to binary data used by as much S as reference vector and as much (K X (S-1)) as training vector. Learning according to algorithm of LVQ to get new weighted (reference) vector.

Recognition Stage
The last phase is recognition stage used Euclidean distance min $||x - w_j||$, which x* representing new features vector to be recognized after transformed to binary data.

4. Experimental Planning

Experiment will be done by off line at PC Intel Pentium IV 1.2 GHz , RAM 128 MB by using MATLAB. Pattern to be recognized in the form of signature. In researching into this will be tried to recognize as much 20 signature. Each pattern will be taken as much 10 sample, 3 pattern for evocation of data and 7 pattern for validation. Figure 4 representing example of pattern signature to be used in researching into this.

Figure 4. Examples of signature pattern

At phase of thresholding pattern, require to be done by detect boundary of each pattern. This is caused by size of signature pattern will tend to vary in each people even if have been provided by a place of the size which remain to. From result of simulation from preliminary research, really this method enough rely on, because from 5 kinds of pattern of each 4 attempt, about 85% recognized truly. There are two performances will be measured, that is validity of recognition and time executed. Others, performance of this method also will be compared by LVQ network, K-Means and Euclidian distance.

We hope this modified LVQ is effective and efficient at signature recognition. From (M X N) features reduced to become (p x q) features so that the very costly computing can be reduced to become $\frac{pq}{MN} x100\ \%$, equally the time execute will be minimized . Others, estimation of parameter of each feature done with statistic approach is so that expected more effective.

5. Summary

A modified learning for LVQ has been presented to signature recognition. The purpose of the modification is to reduce computing cost. It must be tested for several times of training and testing in order to know effectiveness of the method. We hope that we will finishing the computer program in order to test the method.

Acknowledgment

The author would like to thank to the chairman of A2 program who support my accommodation during attending ICAM05 conference in Bandung, Indonesia.

References

[1] Chalechale A. and A. Mertins [2002], Line Segment Distribution of Sketches for Persian Signature Recognition, in *IEEE Transaction On Pattern Analysis and Machine Intelligence,* vol 25 No.2

[2] Bezdek., JC. [1981], *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, USA

[3] Borgelt C., Girimonte D., and Acciani G. [2005], Learning Vector Quantization : Cluster and Cluter Number, in *http://fuzzy.cs.uni-magdeburg.de*

[4] Everitt. B.S [1981], *Cluster Analysis*. Heinemann, London, UK

[5] Fausett L. [1994], *Fundamentals of Neural Network : Architectures, Algorithm, and Applications*, Prentice-Hall Inc.

[6] Hartigan J.A. and M.A. Wong. [1979], A k-means Clustering Algorithm in *Applied Statistics* 28:100–108. Blackwell, Oxford, UK.

[7] Heidemann G. [2001], Efficient Vector Quantization Using the WTA-Rule with Activity Equalization, in *Neural Processing Letters*, Vol 13(1):17-30

[8] Hoppner F., Klawonn F., Kruse R., and T. Runkler [1999], *Fuzzy Cluster Analysis*. J. Wiley & Sons, Chi Chester, UK

[9] Sabourin R. [1997], Off-Line Signature Verification by Local Granulometric Size Distributions, *in IEEE Transaction On Pattern Analysis and Machine Intelligence,* Vol. 19, No. 9.

[10] Villmann T., Hammer B., and Strickert M.[2005], Supervised Neural Gas for Learning Vector Quantization, in *http://www.informatik.uni-osnabrueck.de*

Appendix

MOHAMMAD ISA IRAWAN: Department of Mathematics, Institut Teknologi Sepuluh Nopember, Kampus ITS Keputih – Sukolilo, Surabaya, Indonesia. Phone: +62 +31 5943354 /Fax: +62 +31 5996506 e-mail: m_isa_irawan@yahoo.com

Analysis Relation of Ergodic with Mixing and Markov Shift

Henry Junus Wattimanela

Jurusan Matematika, Universitas Pattimura, Ambon, Indonesia

Abstract: Ergodic theory is the study of the qualitative properties of actions of groups on spaces. The space has some structure (e.g. the space is measure space, or a topological space, or a smooth manifold) and each element of the group acts as a transformation on the space and preserves the given structure (e.g. each element acts as a measure – preserving transformation, or continuous transformation, or a smooth transformation. If T is a measure – preserving transformation of a probability space, we have deduced from the ergodic theorem that T is ergodic if and only if $\forall A, B \in \beta$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} m\left(T^{-1}A \cap B\right) = m(A)\, m(B).$$

In this paper, we will focus of our discussion about mixing in ergodic theory, especially about weak - mixing and strong - mixing. Furthermore, we will analysis relation of ergodic with mixing and Markov shift.

Keywords: transformation, measure, preserving, ergodic, mixing, Markov shift.

THE POTENTIAL DISTRIBUTION MEASUREMENTS IN A LAYERED TRANSVERSELY ISOTROPIC MEDIA WITH LAYERS HAVING EXPONENTIALLY VARYING CONDUCTIVITY

Sri Mardiyati[a], Peg Foo Siew[b]

[a] University of Indonesia

[b] Curtin University of Technology, Western Australia

Abstract. The electrical potential due to a source point of current supplied at the surface of a layered transversely isotropic media is discussed. A recurrence relation is derived to calculate the potential distribution on the surface and on the first layer. This relation is applicable to general cases in which layers have exponentially varying conductivity. A finite element scheme is built to compute the potential distribution on each node in a layered earth. The recurrence relation is used to check the accuracy of the scheme against results obtainable using The Chave's algorithm.

Key-words: transversely isotropic, transverse conductivity, vertical conductivity, direct current, electrical potential, recurrence relation, finite element, infinite element.

1 Introduction

In direct current resistivity studies, the electrical potential, ϕ, satisfies the partial differential equation, $\nabla.(\sigma\nabla\phi) = 0$, where σ denotes the conductivity of the medium. Many authors have investigated the forward problem of determining the potential distribution in a layered isotropic media. Sri Niwas and Uphadhyay [10] investigated a layered earth model with an anisotropic inhomogeneous transition layer. Stoyer and Wait [11] analyzed a two-layer model, where the lower layer has an exponentially varying conductivity. Sato and Sampaio [8] considered a half-space whose resistivity varies as the power of an expression depending linearly on depth. Banerjee *et al* [1] discussed a multilayered isotropic earth with one layer having an exponentially varying conductivity. Kim and Lee [5] derived recurrence relations for calculating the apparent resistivity of a multilayered isotropic earth where the resistivity of each layer is exponentially varying. In many situations, the medium may be transversely-isotropic, that is, the conductivity is the same in all-horizontal directions but has a different value for vertical current flow. In the present study, we consider a transversely isotropic medium, and a recurrence relation is derived for transverse and vertical conductivity which are exponentially varying in depth. The finite element scheme is built to compute a potential distribution in every node on the domain. The recurrence relation is used to evaluate

Hankel transforms which represent the potential distribution on the surface and on the first layer. These Hankel transforms are calculated by using Chave's algorithm at a certain depth. The accuracy of the results obtained using our code is compared with that obtainable using Chave's algorithm [2].

2 Formulation of The Problem

Using the standard suffix notation, the electrical potential in an aeolotropic medium, ϕ, satisfies the equation

$$\sigma_{ik}\,\frac{\partial^2\phi}{\partial x_i \partial x_k} = 0 \quad i,k = 1,2,3$$

where repeated suffices indicate summation, σ_{ik} is the electrical conductivity and x_i denotes the coordinate directions. For a transversely isotropic medium, the above equation takes on a simpler form,

$$\sigma_l(\frac{\partial^2\phi}{\partial r^2} + \frac{1}{r}\frac{\partial\phi}{\partial r}) + \frac{\partial}{\partial z}(\sigma_v\frac{\partial\phi}{\partial z}) = 0 \tag{1}$$

where axial symmetry is assumed. σ_l and σ_v are respectively, the transverse and vertical conductivity, and the radial and vertical coordinates (r, z) are taken with z increasing downwards from the ground surface. We will assume that both the transverse and vertical conductivity are functions of z only. If a source point of current is supplied at the rate of I amp at the origin of the coordinate system, the conditions to be satisfied become:

1. The vertical component of the current density must be zero at ground surface

$$-\sigma_v(0)\frac{\partial\phi}{\partial z}(r,0) = \frac{I}{2\pi}\frac{\delta(r)}{r}$$

2. The electrical potential must approximate zero at infinite distance

$$\phi(r,z) \to 0 \ \texttt{as}\ z \to \infty \ \texttt{or}\ r \to \infty$$

3. The electrical potential must be continuous at each of the boundary planes in the earth

$$\phi_k = \phi_{k+1},\ \texttt{k=1,2,...,N-1}$$

4. The vertical component of the current density must be continuous at each of the boundary planes in the earth

$$\sigma_v^{(k)}(h_k)\frac{\partial\phi_k}{\partial z} = \sigma_v^{(k+1)}(h_k)\frac{\partial\phi_{k+1}}{\partial z},\ \texttt{k=1,2,...,N-1}$$

where a subscript (k) or superscript (k) denote the property in the k^{th} layer.

3 Derivation of Solution for N-Layer Earth

The electrical potential in the transversely isotropic media satisfies the partial differential equation (1) or :

$$[\frac{\partial^2\phi}{\partial r^2}+\frac{1}{r}\frac{\partial\phi}{\partial r}]+\frac{1}{\sigma_l}\frac{\partial\sigma_v}{\partial z}\frac{\partial\phi}{\partial z}+\frac{\sigma_v}{\sigma_l}\frac{\partial^2\phi}{\partial z^2}=0.$$

By using separation of variables, the general solution to (1) is given by

$$\phi(r,z)=\int_0^\infty A(\lambda)H(z,\lambda)J_0(r\lambda)d\lambda \tag{2}$$

where $J_0(r\lambda)$ is the Bessel function of order 0, $H(z,\lambda)$ satisfies the following differential equation

$$H_{zz}+\frac{\partial\sigma_v}{\partial z}\frac{1}{\sigma_v}H_z-\lambda^2\frac{\sigma_l}{\sigma_v}H=0, \tag{3}$$

and $A(\lambda)$ is an arbitrary function of λ. If N-layer of transversely isotropic earth is considered and the conductivity for k^{th} layer is assumed:

$$\sigma_l^{(k)}=b_ke^{(-c_kz)} \quad \texttt{and} \quad \sigma_v^{(k)}=a_ke^{(-c_kz)}$$

where $h_{k-1}<z<h_k$ and a_k,b_k,c_k are real constants, then a solution in k^{th} layer is named by $\phi_k=R_k(r)\ H_k(z)$ where $R_k(r)$ is the Bessel function of order zero and the function of $H_k(z)$ satisfies equation (3). The general solution in each layer is

$$\phi_k(r,z)=\int_0^\infty[A_k(\lambda)e^{(K_k^+z)}+B_k(\lambda)e^{(K_k^-z)}]J_0(\lambda r)d\lambda$$

where

$$K_k^+=\frac{c_ka_k-\sqrt{a_k(a_kc_k^2+4\lambda^2b_k)}}{2a_k} \quad \texttt{and} \quad K_k^-=\frac{c_ka_k+\sqrt{a_k(a_kc_k^2+4\lambda^2b_k)}}{2a_k}$$

and A_k, B_k are arbitrary functions of λ, that will be determined by boundary conditions. Using the boundary condition (3), we arrive at

$$A_k(\lambda)e^{(K_k^+h_k)}+B_k(\lambda)e^{(K_k^-h_k)}=A_{k+1}(\lambda)e^{(K_{k+1}^+h_k)}+B_{k+1}(\lambda)e^{(K_{k+1}^-h_k)} \tag{4}$$

and using the boundary condition (4), we get

$$\begin{aligned}\sigma_v^{(k)}[K_k^+A_k(\lambda)e^{(K_k^+h_k)}+K_k^-B_k(\lambda)e^{(K_k^-h_k)}]=\\ \sigma_v^{(k+1)}[K_{k+1}^+A_{k+1}(\lambda)e^{(K_{k+1}^+h_k)}+K_{k+1}^-B_{k+1}(\lambda)e^{(K_{k+1}^-h_k)}]\end{aligned} \tag{5}$$

Dividing equation (4) by equation (5), we obtain

$$\begin{aligned}\frac{A_k(\lambda)e^{(K_k^+h_k)}+B_k(\lambda)e^{(K_k^-h_k)}}{K_k^+A_k(\lambda)e^{(K_k^+h_k)}+K_k^-B_k(\lambda)e^{(K_k^-h_k)}}=\frac{\sigma_v^{(k)}(h_k)}{\sigma_v^{(k+1)}(h_k)}\times\\ \frac{A_{k+1}(\lambda)e^{(K_{k+1}^+h_k)}+B_{k+1}(\lambda)e^{(K_{k+1}^-h_k)}}{K_{k+1}^+A_{k+1}(\lambda)e^{(K_{k+1}^+h_k)}+K_{k+1}^-B_{k+1}(\lambda)e^{(K_{k+1}^-h_k)}}\end{aligned} \tag{6}$$

Define a new function

$$Y_k(\lambda) = \frac{A_k(\lambda)e^{(K_k^+ h_{k-1})} + B_k(\lambda)e^{(K_k^- h_{k-1})}}{K_k^+ A_k(\lambda)e^{(K_k^+ h_{k-1})} + K_k^- B_k(\lambda)e^{(K_k^- h_{k-1})}} \tag{7}$$

and named $M = \frac{A_k(\lambda)}{B_k(\lambda)}$. If M is substituted into equation (7), then equation (7) becomes:

$$Y_k(\lambda) = \frac{Me^{(K_k^+ h_{k-1})} + e^{(K_k^- h_{k-1})}}{K_k^+ Me^{(K_k^+ h_{k-1})} + K_k^- e^{(K_k^- h_{k-1})}}. \tag{8}$$

By substituting equation (8) into equation (6), the next equation is obtained:

$$\frac{Me^{(K_k^+ h_k)} + e^{(K_k^- h_k)}}{K_k^+ Me^{(K_k^+ h_k)} + K_k^- e^{(K_k^- h_k)}} = \frac{\sigma_v^{(k)}(h_k)}{\sigma_v^{(k+1)}(h_k)} \times Y_{k+1}(\lambda). \tag{9}$$

Using the equation (8) for M, we obtain

$$M = -\frac{e^{(K_k^- h_{k-1})}[1 - Y_k(\lambda)K_k^-]}{e^{(K_k^+ h_{k-1})}[1 - Y_k(\lambda)K_k^+]} \tag{10}$$

By substituting equation (10) into equation (9), we get

$$\frac{e^{(K_k^- h_k)} - e^{(K_k^- h_{k-1})}e^{(K_k^+ t_k)}\frac{[1-Y_k(\lambda)K_k^-]}{[1-Y_k(\lambda)K_k^+]}}{K_k^- e^{(K_k^- h_k)} - K_k^+ e^{(K_k^- h_{k-1})}e^{(K_k^+ t_k)}\frac{[1-Y_k(\lambda)K_k^-]}{[1-Y_k(\lambda)K_k^+]}} = \frac{\sigma_v^{(k)}(h_k)}{\sigma_v^{(k+1)}(h_k)} \times Y_{k+1}(\lambda) \tag{11}$$

where $t_k = h_k - h_{k-1}$. The left hand side of equation (11) is multiplied by $e^{(-K_k^- h_{k-1})}$, and new parameters

$$T^+ = e^{(K_k^+ t_k)}, \;\; T^- = e^{(K_k^- t_k)}, \;\; O_k = \frac{\sigma_v^{(k)}(h_k)}{\sigma_v^{(k+1)}(h_k)}$$

are replaced, so the equation (11)becomes:

$$\frac{T^- - T^+\frac{[1-Y_k(\lambda)K_k^-]}{[1-Y_k(\lambda)K_k^+]}}{K_k^- T^- - K_k^+ T^+\frac{[1-Y_k(\lambda)K_k^-]}{[1-Y_k(\lambda)K_k^+]}} = O_k(Y_{k+1}(\lambda)) \tag{12}$$

The above equation is simplified, and the result becomes:

$$\frac{T^-[1 - Y_k(\lambda)K_k^+] - T^+[1 - Y_k(\lambda)K_k^-]}{K_k^- T^-[1 - Y_k(\lambda)K_k^+] - K_k^+ T^+[1 - Y_k(\lambda)K_k^-]} = O_k Y_{k+1}(\lambda)$$

or

$$Y_k(\lambda)[T^+K_k^- - T^-K_k^+ + O_k Y_{k+1}(\lambda)T^-K_k^-K_k^+ - O_k Y_{k+1}(\lambda)T^+K_k^+K_k^-] =$$
$$T^+ - T^- + O_k Y_{k+1}(\lambda)[T^-K_k^- - T^+K_k^+]$$

or

$$Y_k(\lambda) = \frac{(T^+ - T^-) + O_k Y_{k+1}(\lambda)[T^- K_k^- - T^+ K_k^+]}{[T^+ K_k^- - T^- K_k^+] + O_k Y_{k+1}(\lambda)(2\lambda^2 \frac{b_k}{a_k})[T^+ - T^-]}. \tag{13}$$

The above equation is a recursive relation between Y_k and Y_{k+1}. Using the boundary condition (2), $\phi_k(r)(z) \to 0$ as $z \to \infty$ then $A_k(\lambda)$ must be zero on the lowermost k^{th} layer. From the equation (7), we obtain

$$Y_k(\lambda) = \frac{B_k(\lambda) e^{(K_k^- h_{k-1})}}{K_k^- B_k(\lambda) e^{(K_k^- h_{k-1})}} = \frac{1}{K_k^-}$$

Using the recursive relation (13), we can find $Y_k(\lambda)$ for k = 1,2,...,N - 1. Since our point of interest is to obtain an expression for potential ϕ_1 at the ground surface, we need to assume that the first layer is isotropic. So the transverse and the vertical conductivity are same to a constant, those are $c = 0$, $\sigma_l = b_1$ and $\sigma_v = b_1$. So, on the first layer, the differential equation becomes

$$\frac{\partial^2 H}{\partial z^2} - \lambda^2 \frac{b_1}{a_1} H = 0$$

and the solution of this differential equation is

$$H(z, \lambda) = c_1 e^{(\sqrt{\frac{b_1}{a_1}}\lambda z)} + c_2 e^{(-\sqrt{\frac{b_1}{a_1}}\lambda z)}$$

The potential on the first layer is

$$\phi_1(z, \lambda) = \int_0^\infty [A_1(\lambda) e^{(\sqrt{\frac{b_1}{a_1}}\lambda z)} + B_1(\lambda) e^{(-\sqrt{\frac{b_1}{a_1}}\lambda z)}] J_0(\lambda r) d\lambda \tag{14}$$

Using the boundary condition (1), the following result is obtained:

$$-\sigma_v^1 \int_0^\infty \sqrt{\frac{b_1}{a_1}} [A_1(\lambda) - B_1(\lambda)] \lambda J_0(\lambda r) d\lambda = \frac{I}{2\pi} \frac{\delta(r)}{r} \tag{15}$$

Inverting equation (15) by the Fourier Bessel integral (Watson 1958), we get

$$A_1(\lambda) - B_1(\lambda) = \frac{-I}{2\pi\sqrt{a_1 b_1}}.$$

Substituting $B_1(\lambda)$ into equation (14) the result is as follows:

$$\begin{aligned} \phi_1(z, \lambda) = \int_0^\infty A_1(\lambda) [e^{(\sqrt{\frac{b_1}{a_1}}\lambda z)} + e^{-(\sqrt{\frac{b_1}{a_1}}\lambda z)}] J_0(\lambda r) d\lambda \\ + \frac{I}{2\pi\sqrt{a_1 b_1}} \int_0^\infty e^{-(\sqrt{\frac{b_1}{a_1}}\lambda z)} J_0(\lambda r) d\lambda. \end{aligned} \tag{16}$$

By substituting the Lipshitz integral

$$\int_0^\infty e^{(-\lambda z)} J_0(\lambda r) d\lambda = \frac{1}{\sqrt{r^2 + z^2}}$$

into equation (16) we arrive to the following result:

$$\phi_1(z,\lambda) = \frac{1}{2\pi\sqrt{a_1 b_1}\sqrt{r^2+f^2z^2}} + \int_0^\infty A_1(\lambda)[e^{(\sqrt{\frac{b_1}{a_1}}\lambda z)} + e^{(-\sqrt{\frac{b_1}{a_1}}\lambda z)}]J_0(\lambda r)d\lambda.$$

The first part of the right hand side of this equation is the normal potential created by a source point of current in an anisotropy ground and the second part is the distributing potential caused by subsurface layer. This solution is identical to the solution in an isotropic medium. So the potential on the first layer (16) can be represented as

$$\phi_1(z,\lambda) = \frac{I}{2\pi\sqrt{a_1 b_1}} \int_0^\infty [e^{-(f_1\lambda z)} + \theta_1(\lambda)[e^{(f_1\lambda z)} + e^{(-f_1\lambda z)}]]J_0(\lambda r)d\lambda \tag{17}$$

where

$$\theta_1(\lambda) = \frac{2\pi A_1(\lambda)\sqrt{a_1 b_1}}{I}$$

and the coefficient of anisotropy $f_1 = \sqrt{\frac{b_1}{a_1}}$. To evaluate the electrical potential at the ground surface, we must compute the function $\theta_1(\lambda)$ which can be derived by using the boundary condition (3) and (4) at $z = h_1$. From the boundary (3), we get

$$\frac{I}{2\pi\sqrt{a_1 b_1}}[e^{-(f_1\lambda h_1)} + \theta_1(\lambda)[e^{(f_1\lambda h_1)} + e^{(-f_1\lambda h_1)}]] = A_2(\lambda)e^{(K_2^+ h_1)} + B_2(\lambda)e^{(K_2^- h_1)} \tag{18}$$

and from the boundary (4), we obtain

$$\frac{a_1 I}{2\pi\sqrt{a_1 b_1}}[(-f_1\lambda)e^{-(f_1\lambda h_1)} + \theta_1(\lambda)[(f_1\lambda)e^{(f_1\lambda h_1)} - (f_1\lambda)e^{(-f_1\lambda h_1)}]] = \sigma_v^{(2)}(h_1)[A_2(\lambda)K_2^+ e^{(K_2^+ h_1)} + B_2(\lambda)K_2^- e^{(K_2^- h_1)}] \tag{19}$$

Dividing equation (18) by equation (19) and using equation (7) yields

$$\frac{e^{-(f_1\lambda h_1)} + \theta_1(\lambda)[e^{(f_1\lambda h_1)} + e^{(-f_1\lambda h_1)}]}{a_1[(-f_1\lambda)e^{-(f_1\lambda h_1)} + \theta_1(\lambda)[(f_1\lambda)e^{(f_1\lambda h_1)} - (f_1\lambda)e^{(-f_1\lambda h_1)}]]} = \frac{a_1}{\sigma_v^{(2)}(h_1)}Y_2(\lambda)$$

or

$$\theta_1(\lambda) = \frac{[\frac{a_1}{\sigma_v^{(2)}(h_1)}Y_2(\lambda)(-f_1\lambda) - 1]}{e^{2(f_1\lambda h_1)} + 1 - \frac{a_1}{\sigma_v^{(2)}(h_1)}Y_2(\lambda)f_1\lambda[e^{2(f_1\lambda h_1)} - 1]}.$$

Finally the electrical potential at the ground surface is obtained by the following equation

$$\phi_1(r,0) = \frac{I}{2\pi\sqrt{a_1 b_1}} \int_0^\infty [1 + 2\theta_1(\lambda)]J_0(\lambda r)d\lambda \tag{20}$$

4 The Finite Element Scheme

The boundary value problem, defined by equation (1) and 4 boundary conditions, can only be solved analytically for certain special forms of conductivity profile. We developed in this paper a general numerical technique capable of dealing with any profile of conductivity utilizing finite and infinite elements. The variational statement corresponding to the boundary value problem is to find $\phi(r,z) \in H_0^1(\Omega)$ such that

$$\int_\Omega [\sigma_l \frac{\partial \phi}{\partial r}\frac{\partial w}{\partial r} + \sigma_v \frac{\partial \phi}{\partial z}\frac{\partial w}{\partial z}] r \; dr \; dz \; = \; \frac{-I}{2\pi} \; w(0,0) \;\; \forall \;\; w(r,z) \in H_0^1(\Omega) \qquad (21)$$

where *w(r,z)* is a weight function and $H_0^1(\Omega)$ is given by

$$H_0^1 = \{v \mid v, \frac{\partial v}{\partial r}, \texttt{ are square integrable and } v(r,z) = 0 \texttt{ on } r = \infty \texttt{ and } z = \infty\}.$$

To solve the above variational boundary value problem, we pose the problem in a finite dimensional subspace and use the Galerkin method for the finite element descritization. The computation domain is divided into two regions, $\Omega_f = \{(r,z) | 0 \leq r \leq L \;\; \text{and} \;\; 0 \leq z \leq D\}$ and $\Omega_I = \{(r,z) | r > L \text{ or } z > D\}$. The finite element mesh for Ω_f is generated by mapping a nine-point isoparametric element to the rz-plane using the following coordinate transformation

$$T_e : r = \sum_{i=1}^{9} r_i \psi_i \;\; \texttt{and} \;\; z = \sum_{i=1}^{9} z_i \psi_i$$

where

$$\psi_1 = \frac{1}{4}(\xi^2 - \xi)(\eta^2 - \eta), \;\; \psi_2 = \frac{1}{2}(1 - \xi^2)(\eta^2 - \eta), \;\; \psi_3 = \frac{1}{4}(\xi^2 + \xi)(\eta^2 - \eta),$$

$$\psi_4 = \frac{1}{2}(\xi^2 - \xi)(1 - \eta^2), \;\; \psi_5 = (1 - \xi^2)(1 - \eta^2), \;\; \psi_6 = \frac{1}{2}(\xi^2 + \xi)(1 - \eta^2),$$

$$\psi_7 = \frac{1}{4}(\xi^2 - \xi)(\eta^2 + \eta), \;\; \psi_8 = \frac{1}{2}(1 - \xi^2)(\eta^2 + \eta), \;\; \psi_9 = \frac{1}{4}(\xi^2 + \xi)(\eta^2 + \eta).$$

For the region Ω_I, the method as outlined in Zienkiewicz [12] is used for mapping to the master infinite element with the following element shape functions

$$\tau_1 = (\eta^2 - \eta)(\frac{-\xi}{1 - \xi}), \;\; \tau_2 = \frac{1}{2}(\eta^2 - \eta)(\frac{1 + \xi}{1 - \xi}), \;\; \tau_3 = 2(1 - \eta^2)(\frac{-\xi}{1 - \xi}),$$

$$\tau_4 = (1 - \eta^2)(\frac{1 + \xi}{1 - \xi}), \;\; \tau_5 = (\eta^2 + \eta)(\frac{-\xi}{1 - \xi}), \;\; \tau_6 = \frac{1}{2}(\eta^2 + \eta)(\frac{1 + \xi}{1 - \xi}).$$

Through the use of the Galerkin approximation and assembling process, the following system of equations can be obtained from equation (21)

$$K\alpha = F$$

where $K = \{K_{ij}\}$ is the system stiffness matrix, $F = \{F_i\}$ is the load vector and $\alpha = \{\phi_i\}$ represents the nodal values of $\phi(r,z)$.

r	chave's algorithm	FE method	relative error
1.5	0.025363487860	0.025389510009	0.001025968871
2	0.018517624010	0.018519459338	0.000099112499
2.5	0.014412195910	0.014411707393	0.000033896084
3	0.011677402880	0.011676382690	0.000087364460
3.5	0.009726177353	0.009724956001	0.000125573692
4	0.008264975152	0.008263657720	0.000159399390
4.5	0.007130705679	0.007129333054	0.000192494973
5	0.006225504125	0.006224093581	0.000226575065
5.5	0.005487082569	0.005485641530	0.000262623896
6	0.004873905636	0.004872437022	0.000301321796
6.5	0.004357208257	0.004355712758	0.000343224127
7	0.003916434506	0.003914911607	0.000388848326
7.5	0.003536500936	0.003534949437	0.000438710191
8	0.003206086300	0.003204504567	0.000493353220
8.5	0.002916524943	0.002914911035	0.000553366774
9	0.002661069135	0.002659420870	0.000619399541
9.5	0.002434384411	0.002432699391	0.000692174988
10	0.002232196355	0.002230471967	0.000772507309

Table 1: The potential at $z = 0$ on a 2-layer earth with $\sigma_v^{(1)} = 3$, $\sigma_l^{(1)} = 5$, $h_1 = 10$ and $\sigma_v^{(2)} = 2e^{(0.2z)}$,$\sigma_l^{(2)} = 3e^{(0.2z)}$

5 The Model Test

2-layer and 3-layer media are chosen to validate the finite element scheme. From the above discussion, the calculation of the electrical potential at the ground surface and on the first layer for N-layer media need the calculation of $Y_2(\lambda)$ which can be calculated by using the following equations:

- for 2-layer media:

$$Y_2(\lambda) = \frac{1}{K_2^-} = \frac{2a_2}{c_2 a_2 + \sqrt{a_2(a_2 c_2^2 + 4\lambda^2 b_2)}}$$

- for 3-layer media:

$$Y_3(\lambda) = \frac{1}{K_3^-} = \frac{2a_3}{c_3 a_3 + \sqrt{a_3(a_3 c_3^2 + 4\lambda^2 b_3)}}$$

and

$$Y_2(\lambda) = \frac{(T^+ - T^-) + O_2 Y_3(\lambda)[K_2^- T^- - K_2^+ T^+]}{(K_2^- T^- - K_2^+ T^+) + O_2 Y_3(\lambda)(2\lambda^2 \frac{b_2}{a_2})[T^+ - T^-]}$$

where

$$O_2 = \frac{\sigma_v^{(2)}(h_2)}{\sigma_v^{(3)}(h_2)},\ T^+ = e^{K_2^+ t_2},\ T^- = e^{K_2^- t_2},\ t_2 = h_2 - h_1$$

r	Chave's algorithm	FE Method	relative error
1.5	0.004013168891	0.004012851343	0.000079126498
2	0.003897609657	0.003897322963	0.000073556365
2.5	0.003758759871	0.003758091295	0.000177871432
3	0.003602238209	0.003601524918	0.000198013279
3.5	0.003433527628	0.003432669250	0.000249998862
4	0.003257615340	0.003256751688	0.000265117858
4.5	0.003078780812	0.003077880234	0.000292511242
5	0.002900513310	0.002899628422	0.000305079793
5.5	0.002725524927	0.002724641609	0.000324090964
6	0.002555823056	0.002554954899	0.000339678053
6.5	0.002392812164	0.002391953520	0.000358843044
7	0.002237403452	0.002236553277	0.000379982877
7.5	0.002090118959	0.002089276538	0.000403049308
8	0.001951183447	0.001950343900	0.000430275790
8.5	0.001820601369	0.001819766571	0.000458528712
9	0.001698219058	0.001697384303	0.000491547304
9.5	0.001583773558	0.001582941975	0.000525064329
10	0.001476930101	0.001476098439	0.000563101801

Table 2: The potential at $z = 5$ on a 2-layer earth with $\sigma_v^{(1)} = 3$, $\sigma_l^{(1)} = 5$, $h_1 = 10$ and $\sigma_v^{(2)} = 2e^{(0.2z)}$,$\sigma_l^{(2)} = 3e^{(0.2z)}$

and

$$K_2^+ = \frac{c_2a_2 - \sqrt{a_2(a_2c_2^2 + 4\lambda^2 b_2)}}{2a_2},\ K_2^- = \frac{c_2a_2 + \sqrt{a_2(a_2c_2^2 + 4\lambda^2 b_2)}}{2a_2}$$

2-layer Media

Assume the conductivity is as follows:

- 1^{st}-layer:

$$\sigma_v^1 = 3,\ \ \sigma_l^1 = 5,\ \ f_1 = \sqrt{\frac{5}{3}}\ \texttt{and}\ h_1 = 10$$

- 2^{nd}-layer:

$$\sigma_v^2 = 2\ e^{0.2z}\ \ \texttt{and}\ \ \sigma_l^2 = 3\ e^{0.2z}$$

To calculate the potential on the ground which satisfies equation (20) and potential in the first layer which satisfies equation (17), we need to calculate:

$$Y_2(\lambda) = \frac{4}{0.4 + \sqrt{0.16 + 24\lambda^2}} = \frac{10}{1 + \sqrt{1 + 150\lambda^2}}$$

r	Chave's algorithm	FE Method	relative error
1.5	0.002128299919	0.002127766406	0.000250675666
2	0.002088410560	0.002087870806	0.000258452055
2.5	0.002039082083	0.002038490575	0.000290085429
3	0.001981590985	0.001980990305	0.000303130164
3.5	0.001917315714	0.001916678142	0.000332533654
4	0.001847663968	0.001847021597	0.000347666573
4.5	0.001774009850	0.001773346660	0.000373836707
5	0.001697644327	0.001696981626	0.000390365043
5.5	0.001619740191	0.001619068938	0.000414420167
6	0.001541330851	0.001540663186	0.000433174357
6.5	0.001463301124	0.001462632083	0.000457213481
7	0.001386387462	0.001385722743	0.000479461203
7.5	0.001311185138	0.001310522213	0.000505592216
8	0.001238159952	0.001237500679	0.000532461900
8.5	0.001167662463	0.001167005762	0.000562406535
9	0.001099943240	0.001099288971	0.000594820693
9.5	0.001035168003	0.001034515889	0.000629959580
10	0.000973431920	0.000972780887	0.000668802190

Table 3: The potential at $z = 7$ on a 2-layer earth with $\sigma_v^{(1)} = 3$, $\sigma_l^{(1)} = 5$, $h_1 = 10$ and $\sigma_v^{(2)} = 2e^{(0.2z)}$,$\sigma_l^{(2)} = 3e^{(0.2z)}$

and

$$\theta_1(\lambda) = \frac{[1.5e^{-2}Y_2(\lambda)(-\lambda\sqrt{\frac{5}{3}})] - 1}{e^{(20\sqrt{\frac{5}{3}}\lambda)} + 1 - [1.5e^{(-2)}Y_2(\lambda)(\lambda\sqrt{\frac{5}{3}})][e^{20\sqrt{\frac{5}{3}}\lambda} - 1]}$$

Using the above result, the potential on the ground surface is represented by the following equation:

$$\phi(r,0) = \frac{1}{2\pi\sqrt{15}} \int_0^\infty [1 + 2\theta_1(\lambda)] J_0(\lambda r) d\lambda$$

and the potential on the first layer is represented as follows:

$$\phi_1(z,r) = \frac{1}{2\pi\sqrt{15}} \int_0^\infty \{e^{-(\sqrt{\frac{5}{3}}\lambda z)} + \theta_1(\lambda)[e^{(\sqrt{\frac{5}{3}}\lambda z)} + e^{(-\sqrt{\frac{5}{3}}\lambda z)}]\} J_0(\lambda r) d\lambda$$

Using Chave's algorithm, we calculate the integral

$$\phi(z,r,\lambda) = \int_0^\infty f(\lambda)\ J_0(\lambda r)\ d\lambda$$

where

- For potential on the ground, z=0:

$$f(\lambda) = \frac{1}{2\pi\sqrt{15}}[1 + 2\theta_1(\lambda)]$$

r	Chave's algorithm	FE Method	relative error
2	0.019817109030	0.019958477944	0.007133679983
3	0.012967136090	0.012986659884	0.001505636546
4	0.009541257387	0.009548584984	0.000767990706
5	0.007484829391	0.007490866296	0.000806552118
6	0.006113004067	0.006118622723	0.000919131729
7	0.005132307695	0.005137652368	0.001041378132
8	0.004396034041	0.004401139217	0.001161314028
9	0.003822707608	0.003827578685	0.001274247863
10	0.003363467777	0.003368101422	0.001377639183
11	0.002987242635	0.002991633622	0.001469913073
12	0.002673335245	0.002677479031	0.001550043530
13	0.002407431278	0.002411324927	0.001617345856
14	0.002179317294	0.002182959824	0.001671408753
15	0.001981511695	0.001984904069	0.001712013110
16	0.001808409012	0.001811553999	0.001739090537
17	0.001655726282	0.001658628247	0.001752684022
18	0.001520134080	0.001522798739	0.001752910506
19	0.001399004218	0.001401438375	0.001739921130
20	0.001290233282	0.001292444570	0.001713866811

Table 4: The potential at $z = 0$ on a 3-layer earth with conductivity $\sigma_v^{(1)} = 3$, $\sigma_l^{(1)} = 5$, $h_1 = 10$, $\sigma_v^{(2)} = 2e^{(0.2z)}$, $\sigma_l^{(2)} = 3e^{(0.2z)}$, $h_2 = 20$, $\sigma_v^{(3)} = e^{(0.2z)}$, $\sigma_l^{(3)} = 2e^{(0.2z)}$

- for potential on the first layer:

$$f(\lambda) = \frac{1}{2\pi\sqrt{15}}\{e^{-(\sqrt{\frac{5}{3}}\lambda z)} + \theta_1(\lambda)[e^{(\sqrt{\frac{5}{3}}\lambda z)} + e^{(-\sqrt{\frac{5}{3}}\lambda z)}]\}$$

The table 1,2 and 3 give the corresponding results for a 2-layer earth at three different values of z. The maximum relative error is small with the largest being about 0.001 for the case z=0.

3-layer Media

Assume the conductivity is as follows:

- 1^{st}-layer:

$$\sigma_v^1 = 3, \;\; \sigma_l^1 = 5, \;\; f_1 = \sqrt{\frac{5}{3}} \;\; \texttt{and} \;\; h_1 = 10$$

- 2^{nd}-layer:

$$\sigma_v^2 = 2\, e^{0.2z}, \;\; \sigma_l^2 = 3\, e^{0.2z} \;\; \texttt{and} \;\; h_2 = 20$$

- 3^{rd}-layer:

$$\sigma_v^3 = e^{0.2z} \text{ and } \sigma_l^3 = 2\, e^{0.2z}$$

The calculation of the potential on the ground which satisfies equation (20) and potential in the first layer which satisfies equation (17), requires calculation of

$$\theta_1(\lambda) = \frac{[\frac{a_1}{\sigma_v^{(2)}(h_1)} Y_2(\lambda)(-f_1\lambda) - 1]}{e^{2(f_1\lambda h_1)} + 1 - \frac{a_1}{\sigma_v^{(2)}(h_1)} Y_2(\lambda) f_1\lambda [e^{2(f_1\lambda h_1)} - 1]}$$

where

$$Y_2(\lambda) = \frac{(T^+ - T^-) + O_2 Y_3(\lambda)[K_2^- T^- - K_2^+ T^+]}{(K_2^- T^- - K_2^+ T^+) + O_2 Y_3(\lambda)(2\lambda^2 \frac{b_2}{a_2})[T^+ - T^-]}.$$

r	Chave's algorithm	FE Method	relative error
2	0.016533760850	0.016601283072	0.004083899762
3	0.011852219960	0.011862504577	0.000867737608
4	0.009045275091	0.009051819829	0.000723553229
5	0.007224413149	0.007230335165	0.000819722776
6	0.005960267249	0.005965879258	0.000941570028
7	0.005035382721	0.005040748009	0.001065517419
8	0.004330795932	0.004335928592	0.001185153972
9	0.003776741366	0.003781640309	0.001297134891
10	0.003329869495	0.003334529081	0.001399329916
11	0.002961930709	0.002966344872	0.001490299211
12	0.002653774259	0.002657938196	0.001569062246
13	0.002391980847	0.002395891629	0.001634955399
14	0.002166879401	0.002170536185	0.001687580766
15	0.001971329913	0.001974733873	0.001726732790
16	0.001799949245	0.001803103399	0.001752357189
17	0.001648603295	0.001651512277	0.001764513033
18	0.001514064780	0.001516734585	0.001763336044
19	0.001393777114	0.001396214831	0.001749000594
20	0.001285688030	0.001287901575	0.001721681270

Table 5: The potential at $z = 1$ on a 3-layer earth with conductivity $\sigma_v^{(1)} = 3$, $\sigma_l^{(1)} = 5$, $h_1 = 10$, $\sigma_v^{(2)} = 2e^{(0.2z)}$, $\sigma_l^{(2)} = 3e^{(0.2z)}$, $h_2 = 20$, $\sigma_v^{(3)} = e^{(0.2z)}$, $\sigma_l^{(3)} = 2e^{(0.2z)}$

The value of $Y_2(\lambda)$ needs calculations in the following terms:

1.
$$Y_3(\lambda) = \frac{1}{K_3^-} = \frac{2}{0.2 + \sqrt{0.04 + 8\lambda^2}} = \frac{10}{1 + \sqrt{1 + 200\lambda^2}}$$

2.
$$K_2^+ = 0.1 - \sqrt{0.01 + 1.5\lambda^2}, \;\; K_2^- = 0.1 + \sqrt{0.01 + 1.5\lambda^2}$$

3. If $t_2 = h_2 - h_1 = 10$ then

$$T^+ = e^{10K_2^+} = e^{(1-\sqrt{1+150\lambda^2})}, \quad T^- = e^{(10K_2^-} = e^{1+\sqrt{1+150\lambda^2})}$$

4.
$$K_2^+T^+ = (0.1 - \sqrt{0.01+1.5\lambda^2})e^{(1-\sqrt{1+150\lambda^2})}$$

5.
$$K_2^+T^- = (0.1 - \sqrt{0.01+1.5\lambda^2})e^{(1+\sqrt{1+150\lambda^2})}$$

6.
$$K_2^-T^+ = (0.1 + \sqrt{0.01+1.5\lambda^2})e^{(1-\sqrt{1+150\lambda^2})}$$

7.
$$K_2^-T^- = (0.1 + \sqrt{0.01+1.5\lambda^2})e^{(1+\sqrt{1+150\lambda^2})}$$

8.
$$O_2 = \frac{\sigma_v^{(2)}(h_2)}{\sigma_v^{(3)}(h_2)} = 2$$

r	Chave's algorithm	FE Method	relative error
2	0.005370739851	0.005378540567	0.001452447189
3	0.005060927218	0.005068178336	0.001432764726
4	0.004696535358	0.004703469314	0.001476398126
5	0.004314753319	0.004321459966	0.001554352359
6	0.003940966124	0.003947449443	0.001645109041
7	0.003589579382	0.003595812132	0.001736345498
8	0.003267096819	0.003273048483	0.001821698079
9	0.002975154842	0.002980801079	0.001897796014
10	0.002712740052	0.002718064316	0.001962688609
11	0.002477586760	0.002482579527	0.002015173426
12	0.002266977660	0.002271635266	0.002054544287
13	0.002078167234	0.002082490733	0.002080438441
14	0.001908584869	0.001912579136	0.002092789828
15	0.001755916857	0.001759589794	0.002091748812
16	0.001618125116	0.001621486974	0.002077625498
17	0.001493435166	0.001496497982	0.002050853006
18	0.001380311002	0.001383088114	0.002011946580
19	0.001277425945	0.001279931583	0.001961474174
20	0.001183633976	0.001185882933	0.001900044309

Table 6: The potential at $z = 5$ on a 3-layer earth with conductivity $\sigma_v^{(1)} = 3$, $\sigma_l^{(1)} = 5$, $h_1 = 10$, $\sigma_v^{(2)} = 2e^{(0.2z)}$, $\sigma_l^{(2)} = 3e^{(0.2z)}$, $h_2 = 20$, $\sigma_v^{(3)} = e^{(0.2z)}$, $\sigma_l^{(3)} = 2e^{(0.2z)}$

r	Chave's algorithm	FE Method	relative error
2	0.003237517203	0.003247647720	0.003129100593
3	0.003147889343	0.003157744370	0.003130677710
4	0.003031568659	0.003041070222	0.003134206765
5	0.002895626413	0.002904712562	0.003137887180
6	0.002746984459	0.002755607237	0.003138997737
7	0.002591744913	0.002599868922	0.003134571215
8	0.002434860419	0.002442462075	0.003122008942
9	0.002280076361	0.002287143217	0.003099394442
10	0.002130041479	0.002136571215	0.003065544058
11	0.001986495491	0.001992494544	0.003019917753
12	0.001850471837	0.001855953824	0.002962480644
13	0.001722482048	0.001727466111	0.002893535527
14	0.001602668047	0.001607177370	0.002813635056
15	0.001490920709	0.001494981182	0.002723466765
16	0.001386968222	0.001390607331	0.002623786863
17	0.001290439877	0.001293685817	0.002515374841
18	0.001200910891	0.001203791889	0.002399010636
19	0.001117933150	0.001120476951	0.002275450012
20	0.001041055697	0.001043289205	0.002145426039

Table 7: The potential at $z = 8$ on a 3-layer earth with conductivity $\sigma_v^{(1)} = 3$, $\sigma_l^{(1)} = 5$, $h_1 = 10$, $\sigma_v^{(2)} = 2e^{(0.2z)}$, $\sigma_l^{(2)} = 3e^{(0.2z)}$, $h_2 = 20$, $\sigma_v^{(3)} = e^{(0.2z)}$, $\sigma_l^{(3)} = 2e^{(0.2z)}$

Using the above result, the potential on the ground surface is represented by the following equation:

$$\phi(r,0) = \frac{1}{2\pi\sqrt{15}} \int_0^\infty [1 + 2\theta_1(\lambda)] J_0(\lambda r) d\lambda$$

and the potential on the first layer is:

$$\phi_1(z,r) = \frac{1}{2\pi\sqrt{15}} \int_0^\infty \{e^{-(\sqrt{\frac{5}{3}}\lambda z)} + \theta_1(\lambda)[e^{(\sqrt{\frac{5}{3}}\lambda z)} + e^{(-\sqrt{\frac{5}{3}}\lambda z)}]\} J_0(\lambda r) d\lambda$$

Using Chave's algorithm, we calculate the integral

$$\phi(z,r,\lambda) = \int_0^\infty f(\lambda)\ J_0(\lambda r)\ d\lambda$$

where

- For potential on the ground, z=0:

$$f(\lambda) = \frac{1}{2\pi\sqrt{15}}[1 + 2\theta_1(\lambda)].$$

r	Chave's algorithm	FE Method	relative error
2	0.003237517203	0.003247647720	0.003129100593
3	0.003147889343	0.003157744370	0.003130677710
4	0.003031568659	0.003041070222	0.003134206765
5	0.002895626413	0.002904712562	0.003137887180
6	0.002746984459	0.002755607237	0.003138997737
7	0.002591744913	0.002599868922	0.003134571215
8	0.002434860419	0.002442462075	0.003122008942
9	0.002280076361	0.002287143217	0.003099394442
10	0.002130041479	0.002136571215	0.003065544058
11	0.001986495491	0.001992494544	0.003019917753
12	0.001850471837	0.001855953824	0.002962480644
13	0.001722482048	0.001727466111	0.002893535527
14	0.001602668047	0.001607177370	0.002813635056
15	0.001490920709	0.001494981182	0.002723466765
16	0.001386968222	0.001390607331	0.002623786863
17	0.001290439877	0.001293685817	0.002515374841
18	0.001200910891	0.001203791889	0.002399010636
19	0.001117933150	0.001120476951	0.002275450012
20	0.001041055697	0.001043289205	0.002145426039

Table 8: The potential at $z = 8$ on a 3-layer earth with conductivity $\sigma_v^{(1)} = 3$, $\sigma_l^{(1)} = 5$, $h_1 = 10$, $\sigma_v^{(2)} = 2e^{(0.2z)}$, $\sigma_l^{(2)} = 3e^{(0.2z)}$, $h_2 = 20$, $\sigma_v^{(3)} = e^{(0.2z)}$, $\sigma_l^{(3)} = 2e^{(0.2z)}$

- for potential on the first layer:

$$f(\lambda) = \frac{1}{2\pi\sqrt{15}}\{e^{-(\sqrt{\frac{5}{3}}\lambda z)} + \theta_1(\lambda)[e^{(\sqrt{\frac{5}{3}}\lambda z)} + e^{(-\sqrt{\frac{5}{3}}\lambda z)}]\}.$$

The table 4, 5, 6, 7, 8 and 9 show the comparison of the results from using the finite element code with those obtained by using Chave's algorithm for specific value of z. For this model, the maximum relative error is about 0.007 for the case z=0.

6 CONCLUSION

The model test for 2-layer and 3-layer media confirm the accuracy of the finite element code. Although in most practical problems only the potential on the surface of the earth is measurable, our scheme does give the potential at every node in the computational domain. Further research will investigate the application of the code to the inverse problem of determining various conductivity profiles based on surface measurements and will derive the problem to obtain a general representation of potential distribution on other layers.

r	Chave's algorithm	FE Method	relative error
2	0.002531311711	0.002545344415	0.005543649144
3	0.002478340750	0.002491411962	0.005274178702
4	0.002407975586	0.002420493389	0.005198475879
5	0.002323409003	0.002335010669	0.004993380840
6	0.002228051776	0.002238869469	0.004855225142
7	0.002125233067	0.002135159622	0.004670807712
8	0.002017981960	0.002027085024	0.004510973924
9	0.001908898737	0.001917162948	0.004329308224
10	0.001800104711	0.001807583904	0.004154865522
11	0.001693249969	0.001699969788	0.003968592425
12	0.001589557193	0.001595569008	0.003782069010
13	0.001489883318	0.001495231957	0.003589971735
14	0.001394785902	0.001399521766	0.003395405699
15	0.001304586290	0.001308760639	0.003199749248
16	0.001219425291	0.001223086191	0.003002151937
17	0.001139309847	0.001142507827	0.002806944931
18	0.001064150762	0.001066930163	0.002611848903
19	0.000993792318	0.000996199730	0.002422450000
20	0.000928035016	0.000930110561	0.002236494599

Table 9: The potential at $z = 10$ on a 3-layer earth with conductivity $\sigma_v^{(1)} = 3$, $\sigma_l^{(1)} = 5$, $h_1 = 10$, $\sigma_v^{(2)} = 2e^{(0.2z)}$, $\sigma_l^{(2)} = 3e^{(0.2z)}$, $h_2 = 20$, $\sigma_v^{(3)} = e^{(0.2z)}$, $\sigma_l^{(3)} = 2e^{(0.2z)}$

References

[1] Banerjee B., B. J. Sengupta and B. P. Pal (1980), "Apparent Resistivity of a Multilayered Earth with a Layer Having Exponentially Varying Conductivity", *Geophysical Prospecting*, vol.28, p.435-452.

[2] Chave, Alan D. (1983), "Numerical Integration of Related Hankel Transforms by Quadrature and Ccontinued Fraction Expansion", *Geophysics*, vol.48, p.1671-1686.

[3] Eric B.Becker, Graham F.Carey and J.Tinsley Oden, (1981) *Finite Elements, An introduction, Volume I*, [Prentice-Hall Inc].

[4] Grant, F. S. and West, G. F. (1965), *Interpretation Theory in Applied Geophysics*, [McGraw - Hill Book Company].

[5] Kim, H.S. and Lee, K. (1996), "Response of A Multilayered Earth with Layers Having Exponentially Varying Resistivities", *Geophysics*, vol.61, p.180-191.

[6] Oldenburg D.W. (1978), "The Interpretation of Direct Current Resistivity Measurements", *Geophysics*, vol.43, p.610-625.

[7] Sampaio E.S. (1976), "Electrical Sounding of A Half Space Whose Resistivity or Its Inverse Function Varies Linearly With Depth", *Geophysical Prospecting*, vol.24, p.112-122.

[8] Sato H.K. and Sampaio E.S. (1980), "Electrical Sounding Of A Half-Space With A Monotonic Continuous Variation Of The Resistivity With Depth", *Geophysical Prospecting*, vol.28, p.967 - 976.

[9] Sri Mardiyati, Peg Foo siew, Yong Hong Wu (2003), "Finite Element Approach to Direct Current Resistivity in Aeolotropy Media - The Forward Problem", *Proceeding of the SEAMS-GMU Conference 2003 on Mathematics and its Aplication"*, p 349-357.

[10] Sri Niwas and Upadhayay S. K., "Theoretical Resistivity Sounding Results over A Transition Layer Model", *Geophysical Prospecting*, 22, p 279-296, 1974.

[11] Stoyer. C.H. and Wait. J.R. (1977), "Resistivity Probing of an Exponential Earth with a Homogeneous Overburden", *Geoexploration*, vol.15, p.11-18.

[12] Zienkiewicz O.C and R.L.Taylor (1994), *The Finite Element Method*, Fourth Edition, [McGraw-Hill Book Company].

7 Authors:

Sri Mardiyati: Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Indonesia, Kampus Baru Depok 16424, Indonesia.
Phone/Fax: +6221 7863439
E-mail: sri_ mardiyati@hotmail.com

Peg Foo Siew:Department of Mathematics and Statistics, Curtin University of Technology, GPO Box U1987, Perth 6845, Western Australia
Phone: +61 8 9266 7663, Fax: 9266 3197.
E-mail: pfsiew@maths.curtin.edu.au

TUNNELING TIME AND TRANSMISSION COEFFICIENT OF AN ELECTRON TUNNELING THROUGH A NANOMETER-THICK SQUARE BARRIER IN AN ANISOTROPIC HETEROSTRUCTURE

Lilik Hasanah, Khairurrijal, Mikrajuddin, Toto Winata and Sukirno

Institut Teknologi Bandung, Bandung, Indonesia

Abstract. Analytical expressions of transmission coefficient and tunneling time of electrons incident on a heterostructure grown on an anisotropic material are derived by solving the effective-mass equation including off-diagonal effective-mass tensor elements. It is assumed that the direction of propagation of the electron makes an arbitrary angle with respect to the interfaces of the heterostructure and the effective mass of the electron is position dependent. The analytic expressions are applied to the Si(110)/$Si_{0.7}Ge_{0.3}$/Si(110) heterostructure, in which the SiGe barrier thickness is several nanometers. The calculated results shows that the transmission coefficient and the tunneling time are depend on the valley and it is not symmetric with the angle of incidence.

Key-words: Anisotropic material, heterostructure, nanometer-thick barrier, transmission coefficient, tunneling time.

1 Introduction

The tunneling phenomenon through a potential barrier has been discussed for last half century and also is of present day interest in the study of charge transport in a heterostructure. Paranjape has studied tunneling time and transmission coefficient of an electron in an isotropic heterostructure with different effective masses [1]. Kim and Lee have derived the electron tunneling time, post-tunneling position and transmission coefficient in a heterostructure barrier grown on anisotropic materials including off-diagonal effective-mass tensor elements [2]. In this paper, we report a theoretical study on the direct tunneling time and transmission coefficient of an electron in a heterostructure with nanometer-thick barrier grown on an anisotropic material with electron propagation direction making an arbitrary angle with respect to the interfaces of the heterostructure.

2 Theoretical model

In order to study the behavior of an electron in an anisotropic heterostructure, we must solve the Schrödinger equation :

$$H\psi(\mathbf{r}) = E\psi(\mathbf{r}), \quad (1)$$

where

$$H = \frac{1}{2m_o}\mathbf{p}^{\mathbf{T}}\alpha(\mathbf{r})\mathbf{p} + V(\mathbf{r}). \quad (2)$$

is the Hamiltonian, m_o is the mass of free electron, $\mathbf{p}$ is the momentum vector, $(1/m_o)\alpha$ is the inverse effective-mass tensor and $V(\mathbf{r})$ is the potential energy.

Figure 1 shows the potential profile in the normal direction (z direction) to the layer. The electron is incident from region I to potential barrier. The effective mass of the electron and potential are dependent only on the z direction. Φ is the potential barrier height due to band discontinuity of Si(110) and $Si_{0.7}Ge_{0.3}$ and d is the barrier width. The wave function of the effective-mass equation with the Hamiltonian of Eq. (1) is given as [2]:

$$\psi(\mathbf{r}) = \varphi(z)\exp(-i\gamma z))\exp(i(k_x x + k_y y)), \quad (3)$$

where

$$\gamma = \frac{k_x \alpha_{xz} + k_y \alpha_{yz}}{\alpha_{zz}} \quad (4)$$

is the wave number parallel to the interface. By substituting Eq (2) into Eq (1) it is found that $\varphi(z)$ satisfies the one dimensional Schrödinger-like equation:

$$-\frac{\hbar^2}{2m_o}\alpha_{zz,l}\frac{d^2\varphi(z)}{dz^2} + V(z)\varphi(z) = E_z\varphi(z) \quad (5)$$

where the subscript l in $\alpha_{zz,l}$ denotes each region in Fig. 1. Energy in the z direction can be then written as

$$E_z = E - \frac{\hbar^2}{2m_o}\sum_{i,j\in\{x,y\}}\beta_{ij}k_i k_j, \quad (6)$$

where

$$E = \sum_{i,j\in\{x,y,z\}}\frac{\hbar^2}{2m_o}\alpha_{ij,1}k_i k_j, \quad (7)$$

$$\beta_{ij} = \alpha_{ij} - \frac{\alpha_{iz}\alpha_{zj}}{\alpha_{zz}}, \quad (8)$$

and a_{ij} is the tensor elements associated with the inverse effective mass tensor.

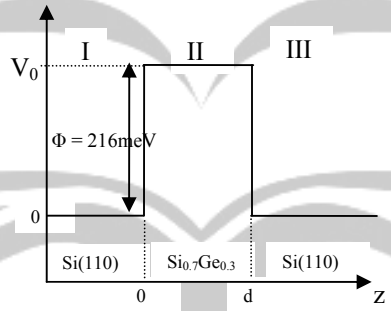


Fig 1. The model used in the numerical calculation

The time-independent electron wave function in each region is :

$$\Psi_I(z) = (Ae^{ik_1 z} + Be^{-ik_1 z})e^{-(i\gamma_1 z)}e^{-(ik_x x + ik_y y)} \quad \text{for } z \le 0, \quad (9)$$

$$\Psi_2(z) = (Ce^{-\int_0^z k_2(z)dz} + De^{\int_0^z k_2(z)dz})e^{-(i\gamma_2 z)}e^{-(ik_x x + ik_y y)} \quad \text{for } 0 < z < d, \quad (10)$$

$$\Psi_3(z) = Fe^{ik_3 z} e^{-(i\gamma_1 z)} e^{-(ik_x x + ik_y y)} \qquad \text{for } z \geq d. \tag{11}$$

The incident wave $A\exp(ik_1 z)$ has the wave number k_1 expressed as

$$k_1 = \left\{ \frac{2m_o E_z}{\hbar^2} \frac{1}{\alpha_{zz,I}} \right\}^{1/2}, \tag{12}$$

where E is smaller than the barrier height Φ. The wave numbers k_2 and k_3 are given as follows

$$k_2 = \left\{ \frac{2m_0}{\hbar^2} \frac{1}{\alpha_{zz,2}} \Phi - \frac{\alpha_{zz,1}}{\alpha_{zz,2}} k_1^2 - \frac{1}{\alpha_{zz,2}} \sum_{i,j\in(x,y)} (\beta_{ij,1} - \beta_{ij,2}) k_i k_j \right\}^{1/2}, \tag{13}$$

$$k_3 = \left\{ \frac{2m_o E_z}{\hbar^2} \frac{1}{\alpha_{zz,I}} \right\}^{1/2}, \tag{14}$$

with the continuity conditions of the wavefunction at z = 0 and z = d given by [2] :

$$\psi_I(z = 0^-) = \psi_2(z = 0^+), \tag{15a}$$

$$\frac{1}{m_o}\left[\alpha_{zx,I} \frac{d\psi_1}{dz} + \alpha_{zy,I} \frac{d\psi_1}{dz} + \alpha_{zz,I} \frac{d\psi_1}{dz} \right]_{z=0^-} = \frac{1}{m_o}\left[\alpha_{zx,2} \frac{d\psi_2}{dz} + \alpha_{zy,2} \frac{d\psi_2}{dz} + \alpha_{zz,2} \frac{d\psi_2}{dz} \right]_{z=0^+}, \tag{15b}$$

$$\psi_2(z = d^-) = \psi_3(z = d^+), \tag{15c}$$

$$\frac{1}{m_o}\left[\alpha_{zx,2} \frac{d\psi_2}{dz} + \alpha_{zy,2} \frac{d\psi_2}{dz} + \alpha_{zz,2} \frac{d\psi_2}{dz} \right]_{z=d^-} = \frac{1}{m_o}\left[\alpha_{zx,1} \frac{d\psi_3}{dz} + \alpha_{zy,1} \frac{d\psi_3}{dz} + \alpha_{zz,1} \frac{d\psi_3}{dz} \right]_{z=d^+}. \tag{15d}$$

With these boundary conditions we obtain the transmission amplitude T_a as :

$$T_a = G\exp(i\phi), \tag{16}$$

where

$$G = \frac{2k_1 k_2}{(P^2 \sinh^2(u) + Q^2 \cosh^2(u))^{1/2}}, \tag{17}$$

is the amplitude of T_a,

$$\varphi = \left[\tan^{-1}\left(\frac{P}{Q}\right) \tanh(u) \right] - k_3 d + (\gamma_1 - \gamma_2) d \tag{18}$$

is the phase of T_a,

$$P = \left(\frac{\alpha_{zz,I}}{\alpha_{zz,2}} k_1^2 - \frac{\alpha_{zz,2}}{\alpha_{zz,I}} k_2^2 \right), \tag{19}$$

$$Q = 2k_1 k_2, \tag{20}$$

$$u = k_2 d. \tag{21}$$

The transmission coefficient is easily obtained from

$$T = T_a^* T_a. \tag{22}$$

The direct tunneling time of an electron through the square barrier is [2]:

$$T_T = \frac{m_o}{\hbar k_3} \frac{1}{\alpha_{zz3}} \left(\frac{\partial \varphi(k_3)}{\partial k_3} + d \right), \tag{23}$$

Substituting Eq. (18) into Eq. (23), for energies lower than the potential barrier, we get

$$T_t = \frac{m_o}{\hbar k_3} \frac{1}{\alpha_{zzI}} \times \left\{ \frac{\left[k_1^2 k_2^2 \left(\frac{\alpha_{zz,1}}{\alpha_{zz,2}} + 1 \right) + k_1^4 \left(\frac{\alpha_{zz,1}}{\alpha_{zz,2}} \right)^2 + \frac{\alpha_{zz2}}{\alpha_{zz,1}} k_2^4 \right] \sinh(2k_2 d)}{k_2 [4k_1^2 k_2^2 \cosh^2(k_2 d) + P^2 \sinh^2(k_2 d)]} - \frac{2k_1^2 k_2 d \frac{\alpha_{zz,1}}{\alpha_{zz,2}} P}{k_2 [4k_1^2 k_2^2 \cosh^2(k_2 d) + P^2 \sinh^2(k_2 d)]} \right\}. \tag{24}$$

.

3 Calculated results and discussion

Referring to Fig. 1, a strained $Si_{0.7}Ge_{0.3}$ potential barrier (region II) is grown on Si (110) in region I or III. The width of the barrier is 50 Å and the band discontinuity is taken as 216 meV [2].

There are four equivalent valleys in the conduction band of Si (110). The effective mass tensor elements of these four valleys are not the same. There are two groups of valleys in Si (110) and $Si_{0.7}Ge_{0.3}$. The inverse effective masses used in our example are related to the tensor elements a_{ij} in Table 1 [3].

Table I. Tensor elements (a_{ij}) used in the numerical calculation.

Valley	Region I, III			Region II		
1	5.26	0	0	5.91	0	0
	0	3.14	2.12	0	3.86	2.45
	0	2.12	3.14	0	2.45	3.86
2	5.26	0	0	5.91	0	0
	0	3.14	-2.12	0	3.86	-2.45
	0	-2.12	3.14	0	-2.45	3.86

Figure 2 shows the chosen coordinate system. We take the position where the electron hits the barrier as the origin of the coordinate system. In the spherical coordinate system shown in Fig. 2, Eq. (7) becomes

$$E = \frac{\hbar^2}{2m_o} \times \left\{ \alpha_{xx1} k^2 \sin^2\theta \cos^2\varphi + \alpha_{yyx1} k^2 \sin^2\theta \sin^2\varphi + \alpha_{zz1} k^2 \cos^2\theta \right. \\ \left. 2\left(\alpha_{xy1} k^2 \sin^2\theta \cos\varphi \sin\varphi + \alpha_{yz1} k^2 \sin^2\theta \cos\theta \sin\varphi \right. \right. \\ \left. \left. + \alpha_{zx1} k^2 \sin^2\theta \cos\theta \cos\varphi \right) \right\} \quad (25)$$

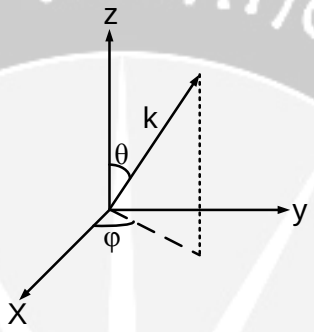


Fig.2. The coordinates used in the analysis.

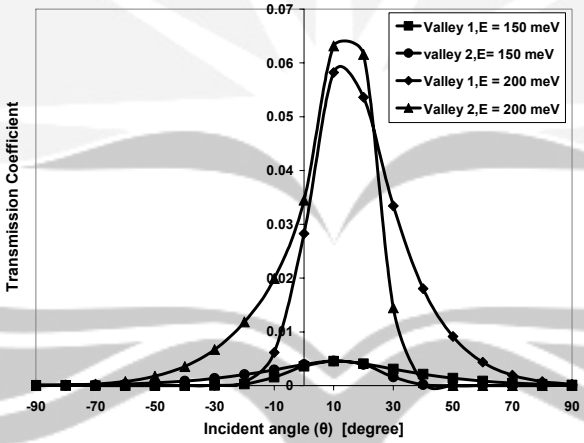


Fig 3. The transmission coefficients for the angle of incident varying from -90° to 90° with incident energies of 150 meV and 200 meV

We calculate the transmission coefficient for the incident angle of **k** (the wave vector of incident electron) varying from -90° to 90° with incident energies of 150 meV and 200 meV and the results are plotted in Fig. 3. Although the incidence angles are θ and φ, but we fix φ to π/2 for simplicity. It can be seen that the transmission coefficient for the incident energy of 200 meV is higher than that for the incident energy of 150 meV. This is because electrons have energy high enough

to tunnel the barrier. For all valleys, the transmission coefficient is maximum not for normal incident but at the incident angle of about 10°. This is due to fact that motions in the x and y directions are not independent of that in the z direction, but they are mutually coupled by the off-diagonal effective-mass tensor elements [4].

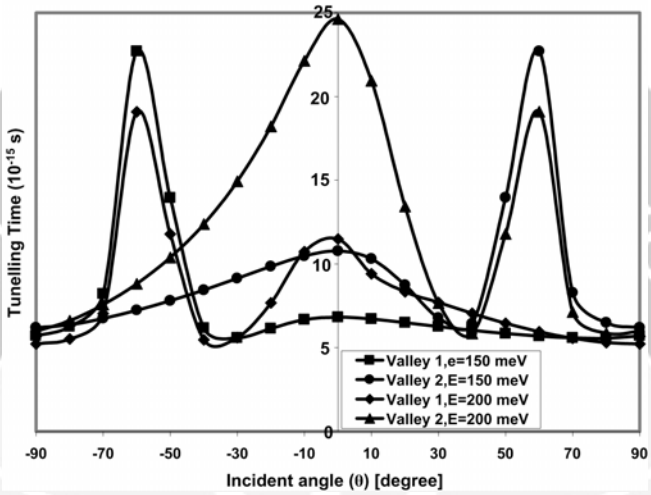


Fig 4. The tunneling time for the angle of incident varying from -90° to 90° with incident energies is 150 meV

The tunneling time versus incident angle is given in Fig.4. We see that the tunneling time depends on the valley where the electron belongs and the incident angle of **k**. It is noteworthy that, in all valleys, the tunneling time is not symmetric with the change of sign of the incidence angle (θ→-θ), which confirms the anisotropy of the material. For the valley 1, the tunneling time has a primary peak at the angle θ of -60° for the incident energy of 150 meV while the secondary peak occurs at the angle of about 0° for the incident angle of 200 meV. For the valley 2, electron with the incident energy of 200 meV have the longest tunneling time at θ = 0°. If the incident angle increases, the next peak of the tunneling time takes place at θ = 60° for the energy of 150 meV.

4 Conclusion

We have derived analytical expressions of the direct tunneling time and transmission coefficient of an electron in a nanometer-thick square barrier grown on anisotropic materials under non-normal incidence. We included the effect of different effective masses at heterojunction interfaces. The boundary condition for an electron wave function (under the effective-mass approximation) at a heterostructure anisotropic junction is suggested and included in the calculation. The calculation is done with a $Si_{0.7}Ge_{0.3}$ potential barrier grown on Si (110). The calculation shows that the transmission coefficient and the tunneling time depend on the valley and it is not symmetric with the angle of incidence.

References

[1] Paranjape, V.V. (1995), Transmission coefficient and stationary-phase tunneling time of an electron through a heterostructure, *Phys.Rev. B*, **52**, 10740-10743.
[2] Kim, K.-Y. and B. Lee (1998), Tunneling time and the post-tunneling position of an electron through a potential barrier in an anisotropic semiconductor, *Superlattice Microstruct.*, **24**, 389-397.
[3] Kim, K.-Y. and B. Lee (2001), Wigner function formulation in anisotropic semiconductor quantum wells, *Phys.Rev. B*, **64**, 115304.
[4] Kim, K.-Y. and B. Lee (1998), Transmission coefficient of an electron through a heterostructure barrier grown on anisotropic materials, *Phys.Rev. B*, **58**, 6728-6731.
[5] Khairurrijal, F. A. Noor, and Sukirno (2005), Electron direct tunneling time in heterostructure with nanometer-thick trapezoid barriers, *Solid-State Electronics*, **49**, 923-927.
[6] Lee, B. and W. Lee (1995), Electron tunneling through a potential barier under non normal incidence, *Superlattice Microstruct.*, **18**, 177-185.

LILIK HASANAH: Doctor student, Department of Physics, Institut Teknologi Bandung, Jalan Ganesa 10, Bandung 40132, Indonesia.. On leave from Physics Department, Universitas Pendidikan Indonesia, Jalan Setiabudi 225, Bandung, Indonesia.
E-mail: lilikhasanah@upi.edu

KHAIRURRIJAL: Department of Physics, Institut Teknologi Bandung, Jalan Ganesa 10, Bandung 40132, Indonesia. Phone: +62-22-250 0834. Fax: +62-22-250 6452.
E-mail: krijal@fi.itb.ac.id

MIKRAJUDDIN: Department of Physics, Institut Teknologi Bandung, Jalan Ganesa 10, Bandung 40132, Indonesia. Phone: +62-22-250 0834. Fax: +62-22-250 6452.
E-mail: din@fi.itb.ac.id

TOTO WINATA: Department of Physics, Institut Teknologi Bandung, Jalan Ganesa 10, Bandung 40132, Indonesia. Phone: +62-22-250 0834. Fax: +62-22-250 6452.

SUKIRNO: Department of Physics, Institut Teknologi Bandung, Jalan Ganesa 10, Bandung 40132, Indonesia. Phone: +62-22-250 0834. Fax: +62-22-250 6452.

GQ-CONSISTENT QUANTUM BOREL KINEMATICS

M.F. Rosyid

Gadjah Mada University, Yogyakarta Indonesia
Institute for Science (I-Es-Ye), Yogyakarta, Indonesia

Abstract. The structure of quantizable algebras in the sense of the $-(\frac{1}{2}+i\gamma)$-densities geometric quantization is studied. The so-called GQ-consistent kinematical algebras are extracted from the algebras. A suitable formulation of quantum Borel kinematics for symplectic manifolds with reducible polarizations is proposed. It is refered to as GQ-consistent quantum Borel kinematics for the simplectic manifolds. The formulation is tested by applying it to some examples of symplectic manifolds with polarizations : cotangent bundles with the vertical polarizations, cotangent bundles with polarizations induced from the vertical ones, cotangent bundle like symplectic manifolds, and homogeneous spaces of exponential groups.
Key-words: Mathematical Physics, quantization, differential geometry

1 Introduction

In classical mechanics, the traditional phase space for a dynamical system is a symplectic manifold (M,ω), i.e. a real differentiable manifold M equipped with a closed non-degenerate differential 2-form ω on M. Consider a dynamical system living on a configuration space[1] Q. The natural phase space for such system is the cotangent bundle T^*Q together with smooth 2-form $d\theta$, where θ is the canonical 1-form on T^*Q. We see that for every configuration space Q there exists a natural phase space associated to Q. But, does every phase space admits a natural configuration space? Let D be a smooth *real* distribution on T^*Q whose fibre D_m over $m\in T^*Q$ is defined by $D_m = T_m(T^*_{pr(m)}Q)$, where $pr : T^*Q \to Q$ is the canonical projection of T^*Q onto Q. It is easy to show that the distribution D is Lagrangian in the sense that D is of maximal rank so that $d\theta(X,Y)$ vanishes for all sections X,Y of D. Then the quotient space T^*Q/D of all maximal integral manifolds of D admits a differentiable structure such that T^*Q/D is diffeomorph to the configuration space Q. Therefore, the configuration space Q can be extracted from the phase space T^*Q by foliating it through the distribution D. However, such mechanism of deriving the configuration space from a phase space does not work in general as not every symplectic manifold admits a *real* distribution like D in the above situation.

In the framework of geometric quantization of symplectic manifolds [8, 15, 16, 18] there is a natural procedure generalizing the above mechanism of extracting configuration spaces from phase spaces. The general procedure is called *complex*

[1] i.e. a real differentiable manifold whose points describe the posible positions of the system under consideration.

polarization. In the procedure, the role of D is played by a complex Lagrangian distribution P [18, 15]. The configuration space yielded from a polarization of a symplectic manifold is a real differentiable manifold and refered to as the *generalized configuration space* associated to the polarization. The generalized configuration space associated to a polarization of a symplectic manifold M is of dimension greater than or equal to $\frac{1}{2}dim(M)$, whereas in the case of $M = T^*Q$ the dimension of Q is equal to $\frac{1}{2}dim(T^*Q)$.

Now it is posible that two symplectic manifolds which are not symplectomorphic[2] admit polarizations leading to the same configuration space. In other word, for every configuration space Q there is a family $\mathfrak{Spol}(Q)$ of symplectic manifolds with polarizations so that Q is the common generalized configuration space of all elements of $\mathfrak{Spol}(Q)$. Another interesting fact is that for the same symplectic manifold, different polarizations yield different configuration spaces. On the quantum level, it leads to equivalent representations only for certain phase spaces. For instance, consider the space $\mathbb{R}^6$ with the canonical closed 2-form $\sum_{i=1}^{3} dp_i \wedge dq^i$, where $(p_1, p_2, p_3, q^1, q^2, q^3)$ is the natural coordinates on $\mathbb{R}^6$: the so-called vertical polarization on $\mathbb{R}^6$ leads to the position representation of quantum mechanics, whereas the so-called horizontal polarization on $\mathbb{R}^6$ leads to momentum representation of quantum mechanics. At this point it is clear that the phase space description of a dynamical system is more essential than that of configuration space.

The relation between Borel-quantization (BQ) [1, 2, 7] of smooth manifolds and the geometric quantizations (GQ) of their cotangent bundles is studied concisely in [11, 5, 9]. In [9] $-(\frac{1}{2}+i\gamma)$-densities, where γ a real number, are suggested in the GQ of cotangent bundles. Still in the context of the relationship between BQ and GQ, the implementation of $-(\frac{1}{2}+i\gamma)$-densities in the geometric quantization of arbitrary symplectic manifolds is mathematically justified in [13] and applied in [14].

In [14] I have shown that for every orientable smooth manifold Q there exists a quantizable symplectic manifold (M, ω) with a real reducible polarization D so that $Q = M/D$ and the geometric $-(\frac{1}{2} + i\gamma)$-D-densities quantization of (M, ω) leads to an elementary differentiable *quantum Borel kinematics* on Q. On the other hand, given an arbitrary symplectic manifold (M', ω') with (real) reducible polarization P generally there does not exist an orientable smooth manifold Q' so that $Q' = M'/(\overline{P} \cap P \cap TM)$ and the geometric $-(\frac{1}{2} + i\gamma)$-P-densities quantization of (M', ω') yields elementary differentiable quantum Borel kinematics on Q'. There is a certain class $\mathfrak{Kas}^o$ of symplectic manifolds with real polarizations P on which the $-(\frac{1}{2}+i\gamma)$-P-densities quantization yields quantum Borel kinematics on the generalized configuration spaces.

The class $\mathfrak{Kas}^o$ contains also the so-called *cotangent bundle like* symplectic manifolds (see Sec. 6.2). In [14] it is also realized that the *size* of the leaves of a

[2]Two symplectic manifolds are said to be symplectomorphic if there exists a diffeomorphism from one onto the other which preserves the symplectic structure

cotangent bundle like symplectic manifold has no influence on the quantum Borel kinematic resulted from the $-(\frac{1}{2}+i\gamma)$-densities geometric quantization of the symplectic manifold.

In the present work we are going to consider symplectic manifolds with *complex* polarization. We suggest a formulation of differentiable quantum Borel kinematics suitable for symplectic manifolds with complex polarizations which will be refered to as **GQ-consistent quantum Borel kinematics**. Our starting point is the fact [13] that the quantizability of an observable in the $-(\frac{1}{2}+i\gamma)$-P-densities quantization depends only on the geometry of the symplectic manifold (M,ω,P) under consideration, i.e. it does not depend for instance on the real number γ which is nontopological in nature. Then we construct the so-called GQ-consistent kinematical algebra out of a certain subset of the space of all quantizable functions relative to the given polarization P. Furthermore a GQ-consistent quantum Borel kinematics for the quantizable symplectic manifold (M,ω,P) with complex polarization is actually defined as the restriction of a quantum Borel kinematics of the generalized configuration space M/D to a certain subset of the kinematical algebra which is the image of GQ-consistent kinematical algebra under a partial Lie homomorphism (see Definition 4.1 and Eq.(5)) into the kinematical algebra of the generalized configuration space. The real quantum number γ will be inherited in natural way from the restriction of the elementary quantum Borel-kinematics. It can be shown that the formulation leads to quantum Borel kinematics on the generalized configuration space when it is applied to a symplectic manifold (M,ω,P) contained in $\mathfrak{Kas}^o$.

2 Borel Quantization

Let Q be a smooth manifold and $\mathfrak{X}_c(Q)$ the set of all smooth complete vector fields on Q. The kinematical algebra of Q is the set $C^\infty(Q,\mathbb{R})\times\mathfrak{X}_c(Q)$ equipped with the bracket $[(f,X),(g,Y)]_s := (X(g)-Y(f),[X,Y])$. The kinematical algebra of Q is therefore the semidirect sum $C^\infty(Q,\mathbb{R})\oplus_s\mathfrak{X}_c(Q)$ of the abelian algebra $C^\infty(Q,\mathbb{R})$ as the ideal and the partial Lie-algebra $\mathfrak{X}_c(Q)$. In the sequel the kinematical algebra of Q will be denoted by $\mathcal{S}(Q)$

Let Φ be a closed integral real 2-form on Q. An elementary differentiable Φ-compactible quantum Borel kinematic is a quadruple $(\mathfrak{H},\mathbb{Q},\mathbb{P},\theta^\infty)$ with

1. $\mathfrak{H}=L^2(\eta,<\cdot,\cdot>,\nu)$ for a hermitean line bundle $(\eta,<\cdot,\cdot>)$ over Q and a smooth Borel measure ν auf Q ,
2. $\mathbb{Q}:C^\infty(Q,\mathbb{R})\to SA(\mathfrak{H})$ a map from $C^\infty(Q,\mathbb{R})$ into the set of all self-adjoint operators in $\mathfrak{H}$. $\mathbb{Q}$ is given by $\mathbb{Q}(\zeta)\sigma=\zeta\sigma$ for every $\zeta\in C^\infty(Q,\mathbb{R})$ and every $\sigma\in Dom(\mathbb{Q}(\zeta))=\{\sigma'\in\mathfrak{H}|\int_Q|\zeta(x)|^2\|\sigma'\|^2d\nu(x)<\infty\}$.
3. $\vartheta^\infty=Sec_0^\infty(\eta)$, where $Sec_0^\infty(\eta)$ is the set of all compact supported smooth sections of η.

4. $\mathbb{P} : \mathfrak{X}_c(Q) \to SA(\mathfrak{H})$ a map from $\mathfrak{X}_c(Q)$ into $SA(\mathfrak{H})$:

$$\mathbb{P}(X)\sigma = -i \bigtriangledown_X \sigma + (\frac{1}{2i} + \gamma)[div_\nu(X)]\sigma,$$

$\forall \sigma \in \vartheta^\infty$, $\forall X \in \mathfrak{X}_c(Q)$, where $\bigtriangledown$ is a hermitean connection with curvature $i\Phi$ und $\gamma \in \mathbb{R}$ is a real number.

Therefore, a quantum Borel-kinematics on Q is in a sense a representation of the kinematical algebra $\mathcal{S}(Q)$. Note that the kinematical part of the theory depends on the topology of the configuration space under consideration and on a nontopological part $\gamma \in \mathbb{R}$ as well.

The program is then completed by a compatible dynamics by making use of the so-called generalized Ehrenfest-relation ([6, 11]).

It is easy to shows that the kinematical algebra $\mathcal{S}(Q)$ is partially Lie-isomorphic to the subalgebra

$$\Sigma(T^*Q) = \{\zeta \circ \pi + (\zeta^i \circ \pi)p_i : \zeta, \zeta^i \in C^\infty(Q, \mathbb{R}) \ \ with \ \ \zeta^i \frac{\partial}{\partial q^i} \in \mathfrak{X}_c(Q)\}$$

of $C^\infty(T^*Q, \mathbb{R})$, where $(p_1, \cdots, p_n, q^1 \cdots, q^n)$ is a (local) canonical coordinat system on T^*Q. The isomorphism is given by

$$\zeta \longmapsto q_\zeta = \zeta \circ \pi \qquad X \longmapsto p_X = \theta_c(\overline{X}),$$

for every $(\zeta, X) \in \mathcal{S}(Q)$, where θ_c is the canonical symplectic potential and $\overline{X}$ is the natural lift of X to T^*Q.

3 Real Directions of A Complex Polarization and Generalized Configuration Spaces

Let (V, ω) be a $2n$-dimensional real symplectic vector space. A complex structure J on V is said to be compatible with the symplectic structure ω whenever J as an automorphism on V is canonic in the sense that $\omega(JX, JY) = \omega(X, Y)$ for all $X, Y \in V$. It is easy to show that

$$g(X, Y) = 2\omega(X, JY), \quad \forall X, Y \in V,$$

defines a nondegenerate symmetric bilinear form on V. Furthermore

$$< X, Y >_J = g(X, Y) + 2i\omega(X, Y)$$

defines a hermitean scalar product on $V_{(J)}$, where $V_{(J)}$ is the complex vector space induced by the complex structure J. The complex structure J is said to be positiv whenever $g(X, X) > 0$ for all $X \neq 0$ and $g(0, 0) = 0$.

Through the maping

$$i : V \ni X \mapsto \frac{1}{2}(X - iJX) \in V^{\mathbb{C}}$$

from V into $V^{\mathbb{C}}$ the complex vector space $V_{(J)}$ could be identified with the Lagrange subspace $P_J = \{X - iJX\} \subset V^{\mathbb{C}}$. We have $P_J \cap \overline{P_J} = 0$. Conversely, a complex Lagrange subspace $P \subset V^{\mathbb{C}}$ with $P \cap \overline{P} = 0$ determines uniquely a complex structure J' which is compatible with the symplectic structure ω so that $P = P_{J'}$. Therefore, a complex structure on (V, ω) is equivalent to a Lagrange subspace $P \subset V^{\mathbb{C}}$ with $P \cap \overline{P} = 0$. Let P be an arbitrary Lagrange subspace of $(V^{\mathbb{C}}, \omega)$. The subspace $D^{\mathbb{C}} = P \cap \overline{P}$ is isotropic and the coisotropic subspace $E^{\mathbb{C}}$ associated with $D^{\mathbb{C}}$ is given by $E^{\mathbb{C}} := (D^{\mathbb{C}})^{\perp} = P + \overline{P}$.

It can be proved that $V'^{\mathbb{C}} := E^{\mathbb{C}}/D^{\mathbb{C}}$ is a symplectic vector space and $P' = pr(P)$ is Lagrange subspace of $V'^{\mathbb{C}}$, where $pr : E^{\mathbb{C}} \to V'^{\mathbb{C}}$ is the natural projection. Since $P' \cap \overline{P'} = 0$, the subspace P' determines a complex structure J' on V'. By the use of an appropriate canonical transformation J' can be brought to a form in which the associated hermitean scalar product has the following matrix representation

$$diag(\overbrace{1, ..., 1}^{r}, \overbrace{-1, ..., -1}^{s}).$$

The number $n - r - s$, i.e. the dimension of D, is called the number of the real directions in P. This measures the degree of the reality of P.

Now consider a symplectic manifold (M, ω) with a strong integrable complex polarization P [18]. For every $m \in M$, $V'^{\mathbb{C}}(m) := E^{\mathbb{C}}(m)/D^{\mathbb{C}}(m)$ is a symplectic vector space and $P'(m) = pr(P(m))$ a Lagrange subspace of $V'^{\mathbb{C}}$ with $P'(m) \cap \overline{P'(m)} = 0$, where $pr : E^{\mathbb{C}}(m) \to V'^{\mathbb{C}}$ is the natural projection. For every $m \in M$, the space $P'(m)$ determines a complex structure $J'(m)$ on $V'^{\mathbb{C}}(m)$. Then, $J'(m)$ has a natural form for all $m \in M$ in which the hermitean scalar product $< \cdot, \cdot >_{J'}$ is represented by the matrix Eq.(3).

The Typ of P is determined by the pair (r, s). When the real direction vanish, i.e. $r + s = n$, then one obtains $P \cap \overline{P} = 0$ and P is called Kählerian. When $r = n$, P is said to be positive. If $s = 0$, P is said to be nonnegative. When r and s vanish simultaneously, namely $P = \overline{P}$, P is said to be real.

By the use of the polarization P, one extracts the generalized configuration space from the symplectic manifold M. The number of the real direction determines therefore the dimension of the generalized configuration space. The greater the number, the 'smaller' the space.

4 GQ-consistent Kinematical Algebras

Let (M, ω) be a quantizable symplectic manifold of dimension $2n$ and P a strong integrable complex polarization so that $0 \leq dim(D) \leq n$, where $D^{\mathbb{C}} = P \cap \overline{P}$ is the

isotropic distribution associated with P. In the $-(\frac{1}{2}+i\gamma)$-P-densities quantization, a funktion $f \in C^\infty(M,\mathbb{R})$ is said to be quatizable if the Hamiltonian vector field X_f generated by f preserves the polarization P in the sense that $[X_f, P] \subseteq P$. This is sufficient for $\mathcal{Q}_{i\gamma}(f)\mathfrak{W}_{P,i\gamma} \subseteq \mathfrak{W}_{P,i\gamma}$, where $\mathfrak{W}_{P,i\gamma}$ and $\mathcal{Q}_{i\gamma}(f)$ is the quantization pre-Hilbert space and the quantization maping respectively [13].

Let $\mathcal{F}_P(M,\mathbb{R})$ be the set of all real quantizable functions and $\mathcal{F}(M,\mathbb{R};D) \subseteq \mathcal{F}_P(M,\mathbb{R})$ a subset of $\mathcal{F}_P(M,\mathbb{R})$ defined by

$$\mathcal{F}(M,\mathbb{R};D) = \{f \in \mathcal{F}_P(M,\mathbb{R}) | [X_f, D] \subseteq D\}.$$

$\mathcal{F}(M,\mathbb{R};D)$ is therefore the set of all quantizable functions preserving the distribution D.

Proposition 4.1 *Let Z be a vector field on M. The maping*

$$\pi_{D*}Z : M/D \ni \pi_D(m) \mapsto \pi_{D*}|_m Z \in T_{\pi_D(m)}(M/D)$$

defines a smooth vector field on M/D if and only if Z preserves the distribution D, i.e. $[Z,D] \subseteq D$. A quantizable function f belongs to $\mathcal{F}(M,\mathbb{R};D)$ if and only if there exists a differentiable vector field X on M/D, so that X_f and X are π_D-related, i.e. $\pi_{D*}X_f$ is a differentiable vector field on M/D.

Proof. Let $g \in C^\infty(M/D,\mathbb{R})$ be an arbitrary smooth function on M/D. Then there is a smooth function $\tilde{g} \in C^\infty(M,\mathbb{R})$ with $\tilde{g} = g \circ \pi_D$. When Z preserves the distribution D, namely $[Z,D] \subseteq D$, we have $[Z,Y](\tilde{g}) = 0$ and $Y(Z(\tilde{g})) = 0$ for all $Y \in D$ because $Y'(\tilde{g})$ vanishes if Y' in D. Therefore $Z(\tilde{g})$ is constant on each leaf of D. Then there is a function $h \in C^\infty(M/D,\mathbb{R})$ so that $Z(\tilde{g}) = h \circ \pi_D$. For every $m \in M$, we have however

$$Z(\tilde{g})(m) = Z(\pi_D^* g)(m) = \pi_{D*}|_m Z(g) = h \circ \pi_D(m).$$

This means that $\pi_{D*}|_{m_1} Z = \pi_{D*}|_{m_2} Z$ for all $m_1, m_2 \in M$ with $\pi_D(m_1) = \pi_D(m_2)$, i.e. $\pi_{D*}Z$ depends differentiably only on the leaves of the distribution D. Thus, we have already shown that $\pi_{D*}Z$ constitutes a smooth vector field on M/D.

When $\pi_{D*}Z : M/D \ni \pi_D(m) \mapsto \pi_{D*}|_m Z \in T_{\pi_D(m)}(M/D)$ defines a differentiable vector field on M/D, then $\pi_{D*}Z(g)$ depends for every $g \in C^\infty(M/D,\mathbb{R})$ only on the leaves of D. It follows that

$$[\pi_{D*}Z(g)] \circ \pi_D = Z(g \circ \pi_D) = h \circ \pi_D$$

for a suitable function $h \in C^\infty(M/D,\mathbb{R})$. Furthermore Z satisfies

$$\begin{aligned} & Y(h \circ \pi_D) = Y(Z(\tilde{g})) = 0 \\ \iff \quad & [Z,Y](\tilde{g}) = 0 \\ \iff \quad & [Z,Y] \in D, \end{aligned}$$

for any $Y \in D$. We obtain therefore $[Z, D] \subseteq D$.

The second statement of the proposition follows automatically from the first. □

Let $\mathfrak{X}(M; D)$ be the set of all vector fields on M preserving the distribution D. Since for all $X, Y \in \mathfrak{X}(M; D)$ we have

$$[[X, Y], Z] = -[[Y, Z], X] - [[Z, X], Y] \in D, \qquad \forall Z \in D,$$

the set $\mathfrak{X}(M; D)$ forms a Lie subalgebra $\mathfrak{X}(M; D)$ of $\mathfrak{X}(M)$. By using of the next lemma, one can prove that the maping

$$\pi_{D}* : \mathfrak{X}(M; D) \to \mathfrak{X}(M/D)$$

is a Lie homomorphism, where $\mathfrak{X}(M/D)$ is the Lie algebra of all smooth vector fields on M/D.

Lemma 4.1 *Let $\varphi : M \to N$ be a smooth maping of differentiable manifolds. Furthermore, let X_1, X_2 and Y_1, Y_2 be vector fields on M and N respectively so that X_i and Y_i $(i = 1, 2)$ are φ-related. Then $[X_1, X_2]$ and $[Y_1, Y_2]$ are also φ-related, i.e. $\varphi_*[X_1, X_2] = [\varphi_* X_1, \varphi_* X_2]$.*

Proof. This is Theorem 7.9 of [4].

Now let $\mathfrak{X}_{\mathcal{F}}(M; D)$ denote the set of all Hamiltonian vector fields on M generated by the functions in the set $\mathcal{F}(M, \mathbb{R}; D)$,

$$\mathfrak{X}_{\mathcal{F}}(M; D) = \{X_h \in \mathfrak{X}(M) | h \in \mathcal{F}(M, \mathbb{R}; D)\}.$$

Define then $\pi_{D*}^{\mathcal{F}}$ as the restriction of π_{D*} to $\mathfrak{X}_{\mathcal{F}}(M; D)$,

$$\pi_{D*}^{\mathcal{F}} : \mathfrak{X}_{\mathcal{F}}(M; D) \to \mathfrak{X}(M/D).$$

The set $\mathfrak{X}_{\mathcal{F}}(M; D)$ is clearly a Lie subalgebra of $\mathfrak{X}(M; D)$ and $\pi_{D*}^{\mathcal{F}}$ is a Lie homomophism, since $[X_f, X_g] = X_{\{f,g\}} \in \mathfrak{X}_{\mathcal{F}}(M; D)$, where $\{\cdot, \cdot\}$ is the Poisson bracket induced by the symplectic structure ω.

Next we introduce an equivalent relation $\sim$ in $\mathfrak{X}_{\mathcal{F}}(M; D)$ defined by

$$X_h \sim X_f \Longleftrightarrow \pi_{D*} X_h = \pi_{D*} X_f.$$

However, this is equivalent to

$$X_h \sim X_f \Longleftrightarrow X_h \backsimeq X_f (Mod\ \ \mathcal{K}(\pi_{D*}^{\mathcal{F}})),$$

where $\mathcal{K}(\pi_{D*}^{\mathcal{F}})$ is the kernel of $\pi_{D*}^{\mathcal{F}}$. Let $\mathfrak{X}_{\mathcal{F}}^{\sim}(M; D)$ denote the set of all equivalent classes in $\mathfrak{X}_{\mathcal{F}}(M; D)$ relative to $\sim$ and $[\cdot, \cdot]^{\sim}$ be defined as a bracket in $\mathfrak{X}_{\mathcal{F}}^{\sim}(M; D)$ by

$$[[X_h], [X_g]]^{\sim} = [[X_{h'}, X_{g'}]], \qquad \forall [X_h], [X_g] \in \mathfrak{X}_{\mathcal{F}}^{\sim}(M; D),$$

where $X_{h'}$ and $X_{g'}$ is an arbitrary element of $[X_h]$ and $[X_g]$ respectively. The bracket $[\cdot,\cdot]^\sim$ is well defined, for it can be shown that $[X_f, X_k] \sim [X_h, X_g]$ for all $X_h, X_f \in [X_h]$ and $X_g, X_k \in [X_g]$. Therefore, the bracket $[\cdot,\cdot]^\sim$ is a Lie bracket and the pair $(\mathfrak{X}^\sim_{\mathcal{F}}(M;D), [\cdot,\cdot]^\sim)$ is a Lie algebra.

Let $\mathfrak{X}^\sim_{\mathcal{F},c}(M;D)$ be the subset of $\mathfrak{X}^\sim_{\mathcal{F}}(M;D)$ defined by

$$\mathfrak{X}^\sim_{\mathcal{F},c}(M;D) = \{[X_f] \in \mathfrak{X}^\sim_{\mathcal{F}}(M;D) | \pi_{D*}[X_f] \in \mathfrak{X}_c(M/D)\},$$

where $\pi_{D*}[X_f]$ is defined as $\pi_{D*}X_{f'}$ for arbitrary $X_{f'} \in [X_f]$ and $\mathfrak{X}_c(M/D)$ the set of all complete vector fields on M/D. Since $\mathfrak{X}_c(M/D)$ is in general not a Lie algebra, the set $\mathfrak{X}_c(M/D)$ does not constitute a Lie algebra. It is clear that $\pi_{D*}\mathfrak{X}^\sim_{\mathcal{F},c}(M;D) \subseteq \mathfrak{X}_c(M/D)$. In general, however, the identity

$$\pi_{D*}\mathfrak{X}^\sim_{\mathcal{F},c}(M;D) = \mathfrak{X}_c(M/D) \tag{1}$$

is not respected : whenever $X \in \mathfrak{X}_c(M/D)$ with $X = \pi_D * Y$ for an appropriate vector field Y on M, then Y is not necessary a Hamiltonian vector field. There exists, however, symplectic manifolds on which Eq.(1) is satisfied. We will consider such manifolds later.

Proposition 4.2 *The maping $\pi^c_{D*} : \mathfrak{X}^\sim_{\mathcal{F},c}(M;D) \to \mathfrak{X}_c(M/D)$ which is defined by $\pi^c_{D*}[X_f] = \pi_{D*}X_{f'}$ for arbitrary $X_{f'} \in [X_f]$ is an injective partial Lie homomorphism.*

Proof. That the maping is injective is already clear from the definition.
(a) Let $[X_f]$ and $[X_g]$ be two arbitrary elements of $\mathfrak{X}^\sim_{\mathcal{F},c}(M;D)$ and $\alpha \in \mathbb{R}$ a real number so that $[X_f]+\alpha[X_g]$ is also contained in $\mathfrak{X}^\sim_{\mathcal{F},c}(M;D)$. Then it follows that

$$\pi^c_{D*}([X_f] + \alpha[X_g]) = \pi^c_{D*}[X_f] + \alpha\pi^c_{D*}[X_g] \in \mathfrak{X}_c(M/D),$$

where the linearity of π_{D*} has been involved.
(b) Let $[X_k]$ and $[X_h]$ be two arbitrary elements of $\mathfrak{X}^\sim_{\mathcal{F},c}(M;D)$ so that the element $[[X_k],[X_h]]^\sim$ belongs also to the set $\mathfrak{X}^\sim_{\mathcal{F},c}(M;D)$. According to the definition, $\pi^c_{D*}[X_k] = \pi_{D*}X_{k'}$ is satisfied for every $X_{k'} \in [X_k]$ as well as $\pi^c_{D*}[X_h] = \pi_{D*}X_{h'}$ for every $X_{h'} \in [X_h]$. By the use of Lemma 4.1 we have

$$[\pi_{D*}X_{k'}, \pi_{D*}X_{h'}] = \pi_{D*}[X_{k'}, X_{h'}].$$

This is equivalent to

$$[\pi^c_{D*}[X_k], \pi^c_{D*}[X_h]] = \pi^c_{D*}[[X_{k'}, X_{h'}]].$$

From the definition of the bracket $[\cdot,\cdot]^\sim$ in $\mathfrak{X}^\sim_{\mathcal{F}}(M;D)$ it follows then

$$[\pi^c_{D*}[X_k], \pi^c_{D*}[X_h]] = \pi^c_{D*}[[X_k],[X_h]]^\sim \in \mathfrak{X}_c(M/D).$$

The proof is therefore complete. □

Define now $\mathfrak{X}$ as the assignment

$$\mathcal{F}(M,\mathbb{R};D) \ni f \mapsto X_f \in \mathfrak{X}_{\mathcal{F}}(M;D) \tag{2}$$

and consider the kernel $\mathcal{K}(\pi^{\mathcal{F}}_{D*})$ of the Lie homomorphism $\pi^{\mathcal{F}}_{D*}$ in $\mathfrak{X}_{\mathcal{F}}(M;D)$. Further, let $\mathcal{F}_{\mathcal{K}}(M,\mathbb{R};D)$ be the inverse image of $\mathcal{K}(\pi^{\mathcal{F}}_{D*})$ under the maping $\mathfrak{X}$. If we denote with $\mathfrak{X}_{\mathcal{K}}$ the restriction $\mathfrak{X}|\mathcal{F}_{\mathcal{K}}(M,\mathbb{R};D)$ of $\mathfrak{X}$ to $\mathcal{F}_{\mathcal{K}}(M,\mathbb{R};D)$, then we obtain the commutative diagram

$$\begin{array}{ccc} \mathcal{F}(M,\mathbb{R};D) & \xrightarrow{\mathfrak{X}} & \mathfrak{X}_{\mathcal{F}}(M;D) \\ \uparrow i & & \uparrow i \\ \mathcal{F}_{\mathcal{K}}(M,\mathbb{R};D) & \xrightarrow{\mathfrak{X}_{\mathcal{K}}} & \mathcal{K}(\pi^{\mathcal{F}}_{D*}). \end{array}$$

If g is contained in $\mathcal{F}_{\mathcal{K}}(M,\mathbb{R};D)$, then $\pi^{\mathcal{F}}_{D*}X_g = 0$. Therefore, for every $m \in M$, $X_g(m)$ is a vector tangent to the leaf of D containing the point m. For every $X \in D$, we have $X(g) = 0$, for $(X_g \lrcorner \omega)(X) + dg(X) = 0$ is respected. This means, g is constant on every leaf of D. Then, there exists uniquely a real smooth function $\zeta_g \in C^\infty(M/D,\mathbb{R})$ so that $g = \zeta_g \circ \pi_D$. However, the converse is in general not correct : If g is contained in $\mathcal{F}(M,\mathbb{R};D)$ so that $X(g) = 0$ for every $X \in D$, then

$$[X_g \lrcorner \omega](X) = -dg(X) = 0.$$

This is however equivalent to $X_g \in E$, where E is the coisotropic distribution $D^\perp$ associated to D [18]. The function g does not belong necessarily to the set $\mathcal{F}_{\mathcal{K}}(M,\mathbb{R};D)$.

Remarks 4.1 *Let $C^\infty_D(M,\mathbb{R})$ denote the set of all real smooth functions on M which are constant on every leaf of D and P be real so that $D = D^\perp$. Then one have*

$$\mathcal{F}_{\mathcal{K}}(M,\mathbb{R};D) = C^\infty_D(M,\mathbb{R}).$$

Since, $C^\infty_D(M,\mathbb{R})$ is equal to

$$\pi^*_D C^\infty(M/D,\mathbb{R}) := \{\zeta \circ \pi_D \in C^\infty(M,\mathbb{R}) | \zeta \in C^\infty(M/D,\mathbb{R})\},$$

it follows also that $\mathcal{F}_{\mathcal{K}}(M,\mathbb{R};D) = \pi^*_D C^\infty(M/D,\mathbb{R})$.

Now let Ξ denote the injection

$$\mathcal{F}_{\mathcal{K}}(M,\mathbb{R};D) \ni g \mapsto \zeta_g \in C^\infty(M/D,\mathbb{R}).$$

and define a linear operator $\mathcal{L}_{[X_f]}$ for every equivalent class $[X_f] \in \mathfrak{X}^{\sim}_{\mathcal{F},c}(M;D)$ on the algebra $\mathcal{F}_{\mathcal{K}}(M,\mathbb{R};D)$ through

$$\mathcal{L}_{[X_f]}g = X_{f'}(g), \qquad \forall g \in \mathcal{F}_{\mathcal{K}}(M,\mathbb{R};D)$$

for arbitrary $X_{f'} \in [X_f]$. The operators are well defined : if $X_{f'}$ and $X_{f''}$ in $[X_f]$, then for every $g \in \mathcal{F}_{\mathcal{K}}(M, \mathbb{R}; D)$ we have

$$\begin{aligned} X_{f'}(g) &= X_{f'}(\zeta_g \circ \pi_D) \\ &= \pi_{D*} X_{f'}(\zeta_g) \circ \pi_D \\ &= \pi_{D*} X_{f''}(\zeta_g) \circ \pi_D \\ &= X_{f''}(g), \end{aligned}$$

where the injection Ξ has been involved in the computation. For every $[X_f] \in \mathfrak{X}^{\sim}_{\mathcal{F},c}(M; D)$, $\mathcal{L}_{[X_f]}$ is therefore a derivation.

Furthermore, we build the semidirect sum

$$\mathcal{S}_{GQ}(M; D) := \mathcal{F}_{\mathcal{K}}(M, \mathbb{R}; D) \oplus_s \mathfrak{X}^{\sim}_{\mathcal{F},c}(M; D) \tag{3}$$

with Lie bracket $[\cdot, \cdot]^s$ defined by

$$[(g_1, [X_{f_1}]), (g_2, [X_{f_2}])]^s = (\mathcal{L}_{[X_{f_1}]} g_2 - \mathcal{L}_{[X_{f_2}]} g_1, [[X_{f_1}], [X_{f_2}]]^{\sim}) \tag{4}$$

for every $[X_{f_1}], [X_{f_2}] \in \mathfrak{X}^{\sim}_{\mathcal{F},c}(M; D)$ with $[[X_{f_1}], [X_{f_2}]]^{\sim} \in \mathfrak{X}^{\sim}_{\mathcal{F},c}(M; D)$ and all $g_1, g_2 \in \mathcal{F}_{\mathcal{K}}(M, \mathbb{R}; D)$.

Definition 4.1 *The pair $(\mathcal{S}_{GQ}(M; D), [\cdot, \cdot]^s)$ defined by Eq.(3) and Eq.(4) is called* **GQ-consistent kinematical algebra** *in (M, ω, P).*

Define a maping, denoted by $\Xi \oplus_s \pi^c_{D*}$, from the GQ-consistent kinematical algebra $\mathcal{S}_{GQ}(M; D)$ into the kinematical algebra $\mathcal{S}(M/D) = C^{\infty}(M/D, \mathbb{R}) \oplus_s \mathfrak{X}_c(M/D)$ through

$$\mathcal{S}_{GQ}(M; D) \ni (g, [X_f]) \mapsto (\Xi(g), \pi^c_{D*}[X_f]) \in \mathcal{S}(M/D). \tag{5}$$

The maping $\Xi \oplus_s \pi^c_{D*}$ is clearly a partial Lie homomorphism.

Definition 4.2 *The GQ-consistent kinematical algebra $\mathcal{S}_{GQ}(M; D)$ is said to be* **almost complete** *whenever the maping $\Xi \oplus_s \pi^c_{D*}$ is a partial Lie isomorphism. Furthermore, $\mathcal{S}_{GQ}(M; D)$ is said to be* **complete** *if it is almost complete and $\mathfrak{X}^{\sim}_{\mathcal{F},c}(M; D)$ is equal to $\mathfrak{X}^{\sim}_{\mathcal{F}}(M; D)$.*

5 GQ-consistent quantum Borel kinematics

Let Φ be a closed integral real 2-form on M/D and $(\mathfrak{H}, \mathbb{Q}, \mathbb{P}, \theta^{\infty})$ an elementary differentiable Φ-compactible quantum Borel kinematic on M/D.

Definition 5.1 The GQ-consistent quantum Borel kinematics subordinated to $(\mathfrak{H}, \mathbb{Q}, \mathbb{P}, \theta^{\infty})$ *for the symplectic manifold with polarization (M, ω, P) is the quadruple $(\mathfrak{H}, \mathbb{Q}_{GQ}, \mathbb{P}_{GQ}, \vartheta^{\infty})$, where*

1. $\mathbb{Q}_{GQ} := \mathbb{Q}|\Xi(\mathcal{F}_{\mathcal{K}}(M,\mathbb{R};D)) : \Xi(\mathcal{F}_{\mathcal{K}}(M,\mathbb{R};D)) \to SA(\tilde{\mathfrak{H}})$ *constitutes the restriction of* $\mathbb{Q}$ *to the set* $\Xi(\mathcal{F}_{\mathcal{K}}(M,\mathbb{R};D))$*, with*

$$\mathbb{Q}_{GQ}(\Xi(f))\sigma = \Xi(f)\sigma$$

on

$$Dom(\mathbb{Q}_{GQ}(\Xi(f))) = \{\psi \in \tilde{\mathfrak{H}} | \int_{M/D} |\Xi(f)(x)|^2 \|\psi\|^2 d\nu(x) < \infty\}$$

for every $f \in \mathcal{F}_{\mathcal{K}}(M,\mathbb{R};D)$*, and*

2. $\mathbb{P}_{GQ} := \mathbb{P}|\pi^c_{D*}(\mathfrak{X}^{\sim}_{\mathcal{F},c}(M;D)) : \pi^c_{D*}(\mathfrak{X}^{\sim}_{\mathcal{F},c}(M;D)) \to SA(\tilde{\mathfrak{H}})$ *constitutes the restriction of* $\mathbb{P}$ *to the set* $\pi^c_{D*}(\mathfrak{X}^{\sim}_{\mathcal{F},c}(M;D))$ *with*

$$\mathbb{P}_{GQ}(\pi^c_{D*}[X_f])\sigma = -i \bigtriangledown_{\pi^c_{D*}[X_f]} \sigma + (\frac{1}{2i} + \gamma)[div_\nu(\pi^c_{D*}[X_f])]\sigma,$$

$\forall \sigma \in \theta^\infty$, $\forall [X_f] \in \mathfrak{X}^{\sim}_{\mathcal{F},c}(M;D)$.

A GQ-consistent quantum Borel kinematics is said to be almost complete whenever the associated GQ-consistent kinematical algebra is almost complete. A GQ-consistent quantum Borel kinematics is said to be complete whenever the associated GQ-consistent kinematical algebra is complete.

The diagram Fig.1 describes the relation between ordinary quantum Borel kinematics and GQ-consistent quantum Borel kinematics.

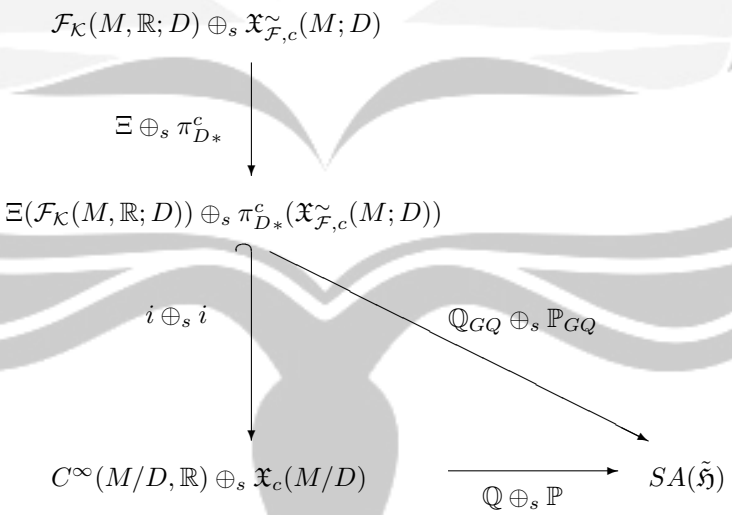


Figure 1

Definition 5.2 *If* $(\mathfrak{H}, \mathbb{Q}, \mathbb{P}, \vartheta^\infty)$ *and* $(\mathfrak{H}', \mathbb{Q}', \mathbb{P}', \vartheta'^\infty)$ *are two differentiable quantum Borel kinematics on* M/D *which are equivalent, then the GQ-consistent quantum Borel kinematics* $(\mathfrak{H}, \mathbb{Q}_{GQ}, \mathbb{P}_{GQ}, \vartheta^\infty)$ *and* $(\mathfrak{H}', \mathbb{Q}'_{GQ}, \mathbb{P}'_{GQ}, \vartheta'^\infty)$ *are said to be equivalent.*

6 Examples and Applications

In this section we mention some examples and applications of the above proposed formulation of quantum Borel kinematics for symplectic manifolds with polarizations. We will show that the proposed formulation leads to the old one on applying to the symplectic manifolds with polarizations on which the $-(\frac{1}{2} + i\gamma)$-densities geometric quantization yields quantum Borel kinematics.

6.1 Cotangent Bundles

6.1.1 Cotangent Bundles with The Vertical Polarizations

In the geometric quantization of the cotangent bundle T^*Q of an n-dimensional smooth manifold Q with the vertical polarization D^v it is shown that the set $\mathcal{F}_{D^v}(T^*Q, \mathbb{R})$ of all quantizable function on T^*Q is given by

$$\mathcal{F}_{D^v}(T^*Q, \mathbb{R}) = \{\xi + \zeta^i p_i | \xi, \zeta^i \in C^\infty(Q, \mathbb{R})\},$$

where (p_i, q^i) are local canonical coordinates on T^*Q. Since D^v is real with $D = D^v \cap \overline{D^v} = D^v$, the set $\mathcal{F}_{D^v}(T^*Q, \mathbb{R})$ equals the set $\mathcal{F}(T^*Q, \mathbb{R}; D)$ of all quantizable functions preserving the distribution D. For every $h = \xi + \zeta^i p_i \in \mathcal{F}(T^*Q, \mathbb{R}; D)$ one has

$$X_h = \zeta^i \frac{\partial}{\partial q^i} - (\frac{\partial \xi}{\partial q^i} + \frac{\partial \zeta^j}{\partial q^i} p_j) \frac{\partial}{\partial p_i}.$$

Therefore, the set $\mathfrak{X}_\mathcal{F}(T^*Q; D)$ can be writen as

$$\{\zeta^i \frac{\partial}{\partial q^i} - (\frac{\partial \xi}{\partial q^i} + \frac{\partial \zeta^j}{\partial q^i} p_j) \frac{\partial}{\partial p_i} \in \mathfrak{X}(T^*Q) | \xi, \zeta^i \in C^\infty(Q, \mathbb{R})\}.$$

Furthermore, we define an equivalent relation $\approx$ in $\mathcal{F}(T^*Q, \mathbb{R}; D)$ as

$$h \approx f \Longleftrightarrow h - f \in C^\infty(Q, \mathbb{R}),$$

for all $f, h \in \mathcal{F}(T^*Q, \mathbb{R}; D)$. It is clear that $h \approx f \Longleftrightarrow \pi_{D*} X_h = \pi_{D*} X_f$. The equivalent relation $\approx$ induces then the relation $\sim$ in $\mathfrak{X}_\mathcal{F}(T^*Q; D)$:

$$X_h \sim X_f \Longleftrightarrow h \approx f, \quad h, f \in \mathcal{F}(T^*Q, \mathbb{R}; D).$$

Let $\mathfrak{X}^\sim_\mathcal{F}(T^*Q; D) = \mathfrak{X}_\mathcal{F}(T^*Q; D)/\sim$ and

$$\mathfrak{X}^\sim_{\mathcal{F},c}(T^*Q; D) = \{[X_f] \in \mathfrak{X}^\sim_\mathcal{F}(T^*Q; D) | \pi_{D*}[X_f] \in \mathfrak{X}_c(Q)\}.$$

We have

Proposition 6.1 *The GQ-consistent kinematical algebra $\mathcal{S}_{GQ}(T^*Q;D)$ of the cotangent bundle (T^*Q,ω) with the vertical polarization D^v is partially Lie isomorphic to the kinematical algebra $\mathcal{S}(Q)=C^\infty(Q,\mathbb{R})\oplus_s\mathfrak{X}_c(Q)$ of Q.*

Proof. It is clear that $\pi_{D*}\mathfrak{X}^\sim_{\mathcal{F},c}(T^*Q;D)\subseteq\mathfrak{X}_c(Q)$. Therefore, it remains to show that $\mathfrak{X}_c(Q)\subseteq\pi_{D*}\mathfrak{X}^\sim_{\mathcal{F},c}(T^*Q;D)$. Let $X=\gamma^i\frac{\partial}{\partial q^i}\in\mathfrak{X}_c(Q)$ be a complete vector field. Then, there exists a class $[h]=[\gamma+\gamma^ip_i]\in\mathcal{F}(T^*Q,\mathbb{R};D)/\approx$ so that

$$[X_h]=[\gamma^i\frac{\partial}{\partial q^i}-(\frac{\partial\gamma}{\partial q^i}+\frac{\partial\gamma^j}{\partial q^i}p_j)\frac{\partial}{\partial p_i}]\in\mathfrak{X}^\sim_{\mathcal{F}}(T^*Q;D).$$

The class belongs even to $\mathfrak{X}^\sim_{\mathcal{F},c}(T^*Q;D)$ because $\pi_{D*}[X_h]=\gamma^i\frac{\partial}{\partial q^i}\in\mathfrak{X}_c(Q)$. Thus, one obtains $X=\gamma^i\frac{\partial}{\partial q^i}\in\pi_{D*}\mathfrak{X}^\sim_{\mathcal{F},c}(T^*Q;D)$, i.e. $\mathfrak{X}_c(Q)\subseteq\pi_{D*}\mathfrak{X}^\sim_{\mathcal{F},c}(T^*Q;D)$. Since the algebra $C^\infty_D(T^*Q,\mathbb{R})$ of all real smooth functions on T^*Q which are constant on every leaf of D^v is isomorphic to the algebra $C^\infty(Q,\mathbb{R})$ of all real smooth functions on Q, we may identify $C^\infty_D(T^*Q,\mathbb{R})$ with $C^\infty(Q,\mathbb{R})$. According to Remark 4.1, we have then $\mathcal{F}_\mathcal{K}(T^*Q,\mathbb{R};D)\cong C^\infty(Q,\mathbb{R})\cap\mathcal{F}(T^*Q,\mathbb{R};D)\cong C^\infty(Q,\mathbb{R})$. This is the end of the proof. □

The last proposition means, (T^*Q,ω,D^v) admits elementay almost complete GQ-consistent quantum Borel kinematics. Furthermore, when Q is compact, then (T^*Q,ω,D^v) admits elementary complete GQ-consistent quantum Borel kinematics.

Corollary 6.1 *Every GQ-consistent quantum Borel kinematics for (T^*Q,ω,D^v) is an elementary differentiable quantum Borel kinematics on $Q=T^*Q/D^v$.*

6.1.2 Cotangent bundles with polarizations induced from the vertical ones

Let ρ be a canonical transformation from (T^*Q,ω) into itself. We can build a real reducible polarization D^ρ from the vertical polarization D^v according to

$$D^\rho=\bigcup_{m\in T^*Q}\{\rho_*Z\in T_m(T^*Q)|Z\in D^v(\rho^{-1}(m))\}.$$

The transformation $\rho:(T^*Q,\omega,D^v)\to(T^*Q,\omega,D^\rho)$ is clearly a polarization preserving symplectomorphism. The GQ-consistent kinematical algebra $\mathcal{S}_{GQ}(T^*Q;D^\rho)$ is almost complete and admits an almost complete GQ-consistent quantum Borel kinematics. These facts follow directly from the results of the last subsection and Lemma 6.1 below (the proof of which is not presented here).

Lemma 6.1 *Let (M_i,ω_i) $(i=1,2)$ be symplectic manifolds with reducible polarizations P_i so that there is a polarization preserving symplectomorphism from M_1 onto M_2. Let D_i denote the isotropic distributions associated to P_i $(i=1,2)$. The GQ-consistent kinematical algebra $\mathcal{S}_{GQ}(M_1;D_1)$ is almost complete if and only if the GQ-consistent kinematical algebra $\mathcal{S}_{GQ}(M_2;D_2)$ is also almost complete.*

6.2 Cotangent Bundle Like Symplectic Manifolds

A cotangent bundle like symplectic manifold is a triple (M, ω, D), where (M, ω) is a symplectic manifold and D a real reducible polarization of M so that the leaves of D are convex[3] and $M \to M/D$ admits a global section. The cotangent bundle like symplectic manifold (M, ω, D) is called *maximal* if the leaves of D are geodesically complete. A maximal cotangent bundle like symplectic manifold is *symplectomorphic* to a cotangent bundle of a smooth manifold.

Also in this class of symplectic manifolds with polarizations we have

Proposition 6.2 *Let (M, ω, D) be a cotangent bundle like symplectic manifold. The GQ-consistent kinematical algebra $\mathcal{S}_{GQ}(M; D)$ is almost complete and the symplectic manifold (M, ω, D) admits therefore almost complete GQ-consistent quantum Borel kinematics.*

Proof. The proof of the proposition follows immediately from the definition of cotangent bundle like symplectic manifolds and Lemma 6.1 above. □

6.3 Homogeneous Spaces

A quantization procedure is a procedure to make a transition from the classical description of a physical system to the quantum description of the same system. In such procedure, the group of canonical transformations usually appears in quantum phase space as an invariant group [8, 15, 10, 18]. This leads to an access to quantum theory which bypasses the Hamiltonian mechanics. There, one does not need any quantization instructions in order to change to quantum level of the same system. Instead, one looks for an appropriate irreducible unitary representation of the classical invariant group.

However, there are unclarities in this context. One of them is related to the existence of the classical system of a quantum system. There exists quantum systems, e.g. elementary particles with spin, which have no classical counterparts. In order to solve such problem, A.A. Kirillov, B. Kostant, and J.M. Souriau have tried through the so-called orbit method for a certain Lie group G to find all classical phase space in which the Lie group G appears as transitive invariant group. The G-homogeneous classical phasen spaces are realised as the orbits of G in the dual space $\mathcal{G}^*$ of the Lie algebra $\mathcal{G}$ of G relative to the coadjoint representation of G in $\mathcal{G}^*$. It is remarkable that all classical G-homogeneous phase space arise essentially in this way : every classical phase space with transitive invariant group G is a cover of an orbit in $\mathcal{G}^*$.

Let Ad^* denote the coadjoint representation of G in $\mathcal{G}^*$. This means,

$$< Ad^*_g \mu', X > = < \mu', Ad_{g^{-1}} X >$$

[3] A leaf Λ of D is called convex if every two points of Λ can be connected precisely by a geodesic.

for every $X \in \mathcal{G}$, $g \in G$ and $\mu' \in \mathcal{G}^*$. Let $\mu \in \mathcal{G}^*$ and $\mathcal{O}_\mu = Ad^*_G\mu$ be the coadjoint orbit of G contains the point μ. If $\mathcal{O}_\mu$ is identified with G/G_μ, where G_μ denotes the isotropic group of μ in G, then $\mathcal{O}_\mu$ is submanifold of $\mathcal{G}^*$. For $\nu \in \mathcal{G}^*$ let $B_\nu : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ be a antisymmetric bilinear form defined by

$$B_\nu(X,Y) \ = < \nu, [X,Y] > .$$

Now define a symplectic form $\hat{B}_\nu : \mathcal{G}/\mathcal{G}_\nu \times \mathcal{G}/\mathcal{G}_\nu \to \mathbb{R}$ on $\mathcal{G}/\mathcal{G}_\nu$ through

$$(X + \mathcal{G}_\nu, Y + \mathcal{G}_\nu) \mapsto B_\nu(X,Y),$$

where $\mathcal{G}_\nu$ is the Lie algebra of G_ν. For every $\nu' \in \mathcal{O}_\mu$ let $\alpha_{\nu'} : \mathcal{G}/\mathcal{G}_{\nu'} \to T_{\nu'}\mathcal{O}_\mu$ be an isomorphism of vector spaces defined by

$$\alpha_{\nu'}(X + \mathcal{G}_{\nu'})(f)(\nu') = \frac{d}{dt}|_{t=0} f(Ad^*_{exp(tX)}\nu'),$$

where f is cantained in $C^\infty(\mathcal{O}_\mu, \mathbb{R})$.

Furthermore define a bilinear form $\omega^{\mathcal{O}}_{\nu'} : T_{\nu'}\mathcal{O}_\mu \times T_{\nu'}\mathcal{O}_\mu \to \mathbb{R}$ through

$$\omega^{\mathcal{O}}_{\nu'}(\alpha_{\nu'}(X + \mathcal{G}_{\nu'}), \alpha_{\nu'}(Y + \mathcal{G}_{\nu'})) = B_{\nu'}(X,Y).$$

The maping $\mathcal{O}_\mu \ni \nu' \mapsto \omega^{\mathcal{O}}_{\nu'}$ defines then a symplectic structure on $\mathcal{O}_\mu$ ([10, 15]). If $J_{\mathcal{O}} : \mathcal{G} \to C^\infty(\mathcal{O}_\mu, \mathbb{R})$ denotes the linear maping defined by

$$X \mapsto J_{\mathcal{O}}(X)(\nu) = < \nu, X >,$$

then $(\mathcal{O}_\mu, \omega^{\mathcal{O}}, J_{\mathcal{O}})$ is a Hamiltonian G-space [10].

6.3.1 Real Polarizations of Coadjoint orbits

Let $\mu \in \mathcal{G}^*$ and $\mathcal{H}$ be a subalgebra $\mathcal{G}$ which is maximal isotropic relative to the bilinear form B_μ and Ad_{G_μ}-invariant[4]. $\mathcal{H}$ is called real polarization in μ ([3]). Let H_0 be the analytical subgroup of G whose Lie algebra is $\mathcal{H}$; therefore $H = G_\mu H_0$. Then, H and H_0 are closed subgroup of G, and H_0 is the component of H containing the identity.

For $\nu = Ad^*_g\mu \in \mathcal{O}_\mu$, let $\mathcal{H}_\nu = Ad_g\mathcal{H}$, $H_\nu = gHg^{-1}$ and F_ν be the image of $\mathcal{H}_\nu/\mathcal{G}_\nu$ under the maping α_ν. Thus, $F_\mu = \alpha_\mu(\mathcal{H}/\mathcal{G}_\mu)$. If $\gamma(g) : \mathcal{O}_\mu \to \mathcal{O}_\mu$ denotes the maping $\nu \mapsto Ad^*_g\nu$, then we have $\gamma(g)_*F_\nu = F_{Ad^*_g\nu}$. The maping $F : \nu \mapsto F_\nu$ is a G-invariant real polarization of the symplectic manifold $(\mathcal{O}_\mu, \omega^{\mathcal{O}})$. The group H defines therefore a fibering of the orbit $\mathcal{O}_\mu \cong G/G_\mu$ over G/H with the projection $\pi_F : \nu' = Ad^*_g\mu \mapsto gH$. The fibers are to $H/G_\mu \cong Ad^*_H\mu$ diffeomorphic. The fiber $\mathcal{L}_\nu$ containing the point $\nu = Ad^*_g\mu$ is given by $\mathcal{L}_\nu = \{Ad^*_{gh}\mu | h \in H\}$.

[4] i.e. Ad_gX is in $\mathcal{H}$ for every $X \in \mathcal{H}$ and every $g \in G_\mu$.

6.3.2 Quantum Borel Kinematics for Coadjoint Orbits of Exponential Groups

Now we consider the case of an exponential, simply connected group. A group is said to be exponential, whenever it is solvable and the exponential maping exp reguler.

Since the polarization F is G-invariant, Ad_g^* induces according to Section 5 of [16] for every $g \in G$ a diffeomorphism of $\mathcal{O}_\mu/F$ into itself. Let φ_g denote the diffeomorphism and $\mathfrak{X}(\mathcal{O}_\mu)$ be the set of all smooth vector fields on $\mathcal{O}_\mu$. Define $\mathfrak{X}(\mathcal{O}_\mu; F)$ and $\mathcal{F}(\mathcal{O}_\mu, \mathbb{R}; F)$ as

$$\mathfrak{X}(\mathcal{O}_\mu; F) = \{V \in F | [V, F] \subset F\}$$

and

$$\mathcal{F}(\mathcal{O}_\mu, \mathbb{R}; F) = \{f \in C^\infty(\mathcal{O}_\mu, \mathbb{R}) | V_f \in \mathfrak{X}(\mathcal{O}_\mu; F)\}$$

respectively. Furthermore, let $\mathcal{F}^0(\mathcal{O}_\mu, \mathbb{R}; F)$ be the set of all real functions f contained in $C^\infty(\mathcal{O}_\mu, \mathbb{R})$, whose Hamiltonian vector fields are in F.

N.V. Pedersen ([12]) has obtained the following useful results. Suppose that $\mathcal{O}_\mu$ is of dimension $2n$.

Lemma 6.2 *There exists real functions $q_1, \cdots, q_n \in \mathcal{F}^0(\mathcal{O}_\mu, \mathbb{R}; F)$ and real functions $p_1, \cdots, p_n \in \mathcal{F}(\mathcal{O}_\mu, \mathbb{R}; F)$ so that $(p_1, \cdots, p_n, q_1, \cdots, q_n)$ are global canonical coordinates on $(\mathcal{O}_\mu, \omega^{\mathcal{O}})$.*

Let φ denote the global chart

$$\begin{aligned} \varphi \quad : \quad & \mathcal{O}_\mu \to \mathcal{O}'_\mu \subset \mathbb{R}^{2n} \\ & Ad_g^*\mu \mapsto \varphi(Ad_g^*\mu) = (p_1(Ad_g^*\mu), \cdots, p_n(Ad_g^*\mu), q^1(Ad_g^*\mu), \cdots, q^n(Ad_g^*\mu)), \end{aligned}$$

so that $\varphi^i = p_i$ $(1 \le i \le n)$ and $\varphi^i = q^i$ $(n+1 \le i \le 2n)$.

Lemma 6.3 *Let $\tilde{q} : G/H \to \mathbb{R}^n$ be a maping from G/H into $\mathbb{R}^n$ defined by $\tilde{q}(gH) = (q_1(Ad_g^*\mu), \cdots, q_n(Ad_g^*\mu))$ and Ω the image G/H in $\mathbb{R}^n$ under the maping $\tilde{q}$. Then, Ω is open in $\mathbb{R}^n$ and $\tilde{q}$ a global chart on $G/H \cong \mathcal{O}_\mu/F$. Furthermore, the set of all quantizable real functions is given by*

$$\mathcal{F}(\mathcal{O}_\mu, \mathbb{R}; F) = \{\zeta + \zeta^i p_i | \zeta, \zeta^i \in C^\infty(\Omega, \mathbb{R})\}.$$

Proof. See lemma 3.2.1 and 3.2.2 in [12]. The last statement is a special case of Lemma 3.2.1, where only real functions are regarded. □

Therefore, the diagram

$$\begin{array}{ccc} \mathcal{O}_\mu & \xrightarrow{\varphi} & \mathcal{O}'_\mu \\ {\scriptstyle \pi_F}\downarrow & & \downarrow{\scriptstyle pr_2|\mathcal{O}'_\mu} \\ G/H \cong \mathcal{O}_\mu/F & \xrightarrow[\tilde{q}]{} & \Omega \end{array} \tag{6}$$

is commutative, where pr_2 denotes the projection

$$\mathbb{R}^{2n} \ni (p_1, \cdots, p_n, q^1, \cdots, q^n) \mapsto (q^1, \cdots, q^n) \in \mathbb{R}^n.$$

Proposition 6.3 *With the same assumption of Lemma 6.3, $(\mathcal{O}_\mu, \omega^{\mathcal{O}}, F)$ admits an almost complete GQ-consistent quantum Borel kinematics.*

Proof. Let $\mathfrak{X}_{\mathcal{F}}(\mathcal{O}_\mu; F)$ denote the image of $\mathcal{F}(\mathcal{O}_\mu, \mathbb{R}; F)$ under the maping Eq.(2). From Lemma 6.2 and Lemma 6.3 we have

$$\mathfrak{X}_{\mathcal{F}}(\mathcal{O}_\mu; F) = \{V_{\zeta,\zeta^i} = \zeta^i \frac{\partial}{\partial q^i} - (\frac{\partial \zeta}{\partial q^i} + \frac{\partial \zeta^j}{\partial q^i} p_j) \frac{\partial}{\partial p_i} | \zeta, \zeta^i \in C^\infty(\mathcal{O}_\mu/F, \mathbb{R})\}.$$

Now, from diagram El.(6) it follows that

$$\pi_{F*} V_{\zeta,\zeta^i} = \zeta^i \frac{\partial}{\partial q^i} \in \mathfrak{X}(\mathcal{O}_\mu/F),$$

for every $V_{\zeta,\zeta^i} \in \mathfrak{X}_{\mathcal{F}}(\mathcal{O}_\mu; F)$. It is to show that $\pi_{F*}\mathfrak{X}^{\sim}_{\mathcal{F},c}(\mathcal{O}_\mu; F) = \mathfrak{X}_c(\mathcal{O}_\mu/F)$. Since $\pi_{F*}\mathfrak{X}^{\sim}_{\mathcal{F},c}(\mathcal{O}_\mu; F) \subset \mathfrak{X}_c(\mathcal{O}_\mu/F)$ (according to the definition), it remains to point out that

$$\mathfrak{X}_c(\mathcal{O}_\mu/F) \subset \pi_{F*}\mathfrak{X}^{\sim}_{\mathcal{F},c}(\mathcal{O}_\mu; F).$$

The complete proof then proceed like the proof of Proposition 6.1. □

Acknowledgment

The author would like to thank the Laboratory of Atomic and Nuclear Physics, Department of Physics, Gadjah Mada University, for wonderfull atmosphere and to Institut for Sains in Yogyakarta (I-Es-Ye) for so much support.

References

[1] Angermann, B., Doebner, H.D., and J. Tolar. (1983), Quantum Kinematics on Smooth Manifolds, in *Lecture Notes in Mathematics 1037*, Editor : Anderson, S. and H.-D. Doebner, Springer-Verlag, Berlin, 171 - 208.

[2] Angermann, B. (1983), *Über Quantisierungen lokalisierter Systeme - Physikalisch interpretierbare mathematische Modelle*, PhD thesis, Technische Universität Clausthal.

[3] Auslander, L., and B. Kostant (1971), Polarization and Unitary Representations of Sovable Lie Groups, *Invent. Math.*, **14**, 276 - 283.

[4] Boothby, W.M. (1975), *An introduction to differentiable manifolds and Riemannian geometry*, Academic Press, New York.

[5] Doebner, H.D., and P. Nattermann (1996), Borel Quantization : Kinematics and Dynamics, *Acta Physica Polonica B*, **27**(10).

[6] Doebner, H.D., and J.D. Hennig (1995), Quantum mechanical evolution equation for mixed states from symmetry and kinematics, in *Symmetry in Science VIII*, Editor : Gruber, B., Plenum Press, New York.

[7] Drees, M. (1992), *Zur Kinematik lokalisierter quantenmechanischer Systeme unter Berücksichtigung innerer Freiheitsgrade und äußerer Felder*, PhD thesis, Technische Universität Clausthal.

[8] Echeverría-Enríquez, A., Muñoz-Lecanda,M. C., Román-Roy, N., and C. Victoria-Monge (1998), Mathematical Foundations of Geometric Quantization, *Extracta Math.*, **13**, 135 - 235.

[9] Hennig, J.D. (1995), Nonlinear Schrödinger equations and Hamiltonian mechanics, in *Nonlinear, deformed and irreversible quantum systems*, Editor : Doebner, H.-D., Dobrev, V.K., and P. Nattermann, World Scientific, Singapore.

[10] Kostant, B. (1970), Quantization and Unitary Representations, in *Lecture Notes in Mathematics 170*, Editor : Taam, C.T., Springer-Verlag, Berlin.

[11] Nattermann, P. (1997), *Dynamics in Borel-Quantization : Non-linear Schrödinger Equations vs. Master Equations*, PhD thesis, Technische Universität Clausthal.

[12] Pedersen, N.V. (1988), On the Symplectic Structure of Coadjoint Orbits of (Solvable) Lie Groups and Applications, *Math. Ann.*, **281**, 633 - 669.

[13] Rosyid, M. F. (2003), On the Relation between Configuration and Phase Space Quantization I, *Journal of the Indonesian Mathematical Society*, **9**(1), 13 - 25.

[14] Rosyid, M. F., On the Relation between Configuration and Phase Space Quantization II, to appear.

[15] Simms, D.J., and N.M.J. Woodhouse (1976), Lectures on Geometric Quantizations, in *Lecture Notes in Physics 53*, Springer-Verlag, New York.

[16] Souriau, J.M. (1997), *The Structure of Dynamical Systems*, Birkhäuser, Basel.

[17] Weinstein, A. (1977), Lectures on symplectic manifolds, *Expository Lectures from CBMS Regional Conference at University of North Carolina*, American Mathematical Society.

[18] Woodhouse, N.M.J. (1992), *Geometric Quantization*, 2nd ed., Clarendon Press, Oxford.

M.F. Rosyid: Work Group on Mathematical Physics, Department of Physics, Gadjah Mada University Yogyakarta, SEKIP unit III Yogyakarta 55281, Indonesia.
Phone/Fax: +62 +274 545 939
Institute for Science in Yogyakarta.
E-mail: farchani@ugm.ac.id

GAUSSIAN DISTRIBUTION ANALYSIS OF STATISTICAL DOUBLE SLIT EXPERIMENT OF ELECTRONS

W.S.B. Dwandaru[a,b], D. Darmawan[b], Retno Subekti[c], N. Insani

[a] Computational and Theoretical Physics Laboratory, Physics Education Department, Yogyakarta State University, Karangmalang, Yogyakarta, 55281, Indonesia

[b]Yogyakarta Institute for Science, Sodanten, Yogyakarta, Indonesia

[c] Mathematics Education Department, Yogyakarta State University, Karangmalang, Yogyakarta, 55281, Indonesia

Abstract. Fascinating harmonious relationship occurs between modern physics and probability concept. This occurrence is mainly generated by the statistical property of modern physics which gives continuous effort in describing microscopic phenomena. One of these microscopic phenomena, which became a trademark of an ontological problem in modern physics is the famous double slit experiment of electrons. Having a particle-wave aspect, electrons are emitted one by one into a double slit until they reach a screen which then can be observed. In this paper, a statistical interpretation of the double slit experiment has been develop. The idea of the aforementioned interpretation lies in the heart of some modified Gaussian distribution of particle which forms a pattern on the screen. The present paper will direct its point view towards the derivation of the density function of electron formed in the screen, which is the result of wave-particle combination of electrons.

Key-words: double slit experiment, electrons, Gaussian density function, probability, modern physics (optics)

1 Introduction

1.1 Background

Applied probability theory has been widely recognized in the field of quantum physics. But not like common probability theory which deals with measurement on an ensemble of physical system or N trials of it, on the contrary, quantum mechanics exposes the probability of a state or position of a particle in one trial of position measurement. Despite the definition used is similar to that of the usual Kolmogoroff's axiom [2,5], quantum probability gives more exclusive interpretation to its definition.

Using the wave mechanics developed by Schrödinger [3,4,10], we start off with the so called Schrödinger differential equation, stated in one dimension as:

$$i\hbar\frac{\partial\Psi(x,t)}{\partial t}=-\frac{\hbar^2}{2m}\frac{\partial^2\Psi(x,t)}{\partial x^2}+V(x)\Psi(x,t), \quad (1)$$

with $\hbar$ is the reduced Planck constant and m is the mass of the physical system. From equation (1), the state of a quantum system characterized by the potential $V(x)$ is given by $\Psi(x,t)$, which gives all information concerning the physical system under consideration. Especially for free particle, $V(x)=0$, thus equation (1) can be modified as:

$$i\hbar\frac{\partial\Psi(x,t)}{\partial t}=-\frac{\hbar^2}{2m}\frac{\partial^2\Psi(x,t)}{\partial x^2}. \quad (2)$$

Now, here comes the probability interpretation. Max Born [4,10] postulated that $\Psi(x,t)$ has no physical interpretation unless it is squared and integrated over some region resulting the probability of the existence of the particle in some given area of $x+dx$, therefore:

Definition 1:

$\int|\Psi(x,t)|^2dx$ =The probability of finding a particle in time t in the region of $x+dx$. (3)

The form of $|\Psi(x,t)|^2dx$ in equation (3) is usually known as the probability density of finding a particle in time t over the area of $x+dx$.

Equation (3) is the main link between quantum mechanics and probability theory. This postulate has stood the test of time, especially in its mathematical aspects. It is applicable to various physical system which gives good agreement with experimental results.

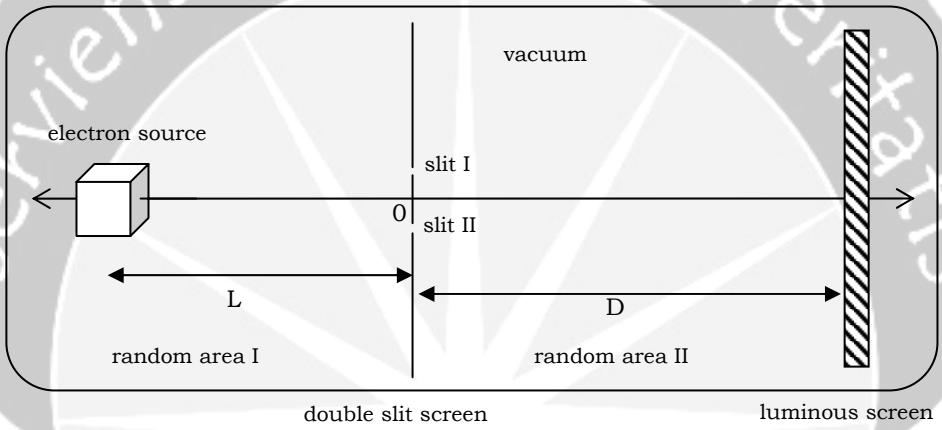
Quantum mechanics itself is a powerful tool to describe the behavior of the microworld. The application of quantum mechanics is almost inevitable in the rapid development of today's technology, which range from outerspace observation until nano-structure technology. Interestingly though, in spite of its enormous applicability, quantum mechanics is still lacking in its fundamental level. This is because various paradoxes have not yet been solved in a satisfactory manner. Most of paradoxes of quantum mechanics stream out as questions about the completeness and the validity of its concept. In a certain way, quantum mechanics is a big house, with all its luxurious furniture, security and entertainment, but built above a very weak foundation.

One of these paradoxes, which question the ontological notion of quantum mechanics lies in the heart of a thought experiment, viz. the double slit experiment of electrons, which eventually became a real experiment conducted by Giorgio Merli et.al. [5], and repeated by Tonomura et.al. [5] using advance device called electron biprism, yielding the same interference result.

1.2 A Brief Notion of the Double Slit Experiment of Electrons

Actually, the double slit experiment was a real experiment conducted about two centuries ago by Thomas Young [6,8]. But now, this experiment has become one of the trademarks in quantum mechanics to support the complementarity principle put forward by the Copenhagen interpretation of quantum mechanics.

Generally, the setting of a double slit experiment of electrons can be illustrated as observed in **picture 1**. In the picture, an electron source, a double slit screen and a luminous screen is placed parallel to each other. The source will emit electrons (continuously or one by one) in random area I, into the double slit screen, such that an interference phenomena occurs in random area II. Consequently, a certain pattern is observed in the luminous screen as a result of the interference between electrons which pass through slit I and slit II.



Picture 1: The double slit experiment set up using an electron source.

In the original Young double slit experiment, the sun is used as a source. First, sunlight is directed into a single slit screen. Then, light which passes through the single slit screen can be considered as a new source of light, which then goes through a double slit screen, hence an interference pattern occurs on a screen. The interference pattern consists of bright and dark fringes of parallel bands. These parallel bright and dark fringes forms some sort of an **intensity distribution**. The similar pattern should also be present if the sun is substituted by an electron source, as shown in **picture 1**.

Interference is the most fundamental nature of all wave phenomena. If one physical quantity exhibits interference, that quantity should have wave nature [8]. So, in the case above, light is a wave (more precisely, electromagnetic wave). Furthermore, because continuous emitting electrons shows interference, electrons also have wave nature. Now, here comes the quantum mechanics part. The experimental proof of de Broglie's wave-particle duality hypothesis suggests that electron wave is not necessarily a common mechanical wave. According to Born [4,1,7,10], electrons are probability wave which assign a probability value to find an electron over a certain area. Following [1], this means that the probability of finding an electron

within the bright bands (on the screen) is relatively high, and vice versa, the probability of finding an electron within the dark bands is relatively low.

Here where things get more interesting. Consider the electron source is being weakened so that electrons are emitted one by one from the source. In this continued experiment, one electron is emitted from the source into the double slits and observed where the electron hits the screen. After a while, another electron is released from the source and observed where it hits the screen, and so on. If this experiment proceeds in some finite interval of time, and it is known how many electrons that has been released, then for a large number of electrons hitting the screen, an interference pattern will appear. This result is interesting because electron which is known as a particle with mass, turns out to produce interference pattern, which is normally the property of wave. This result is absolutely different, if on the other hand, the source emits macroscopic objects such as balls or bullets. The result is obvious, no interference pattern occurs. The intensity distribution of balls or bullet which passes through two holes will be:

$$I = I_1 + I_2 , \tag{4}$$

with I is the total distribution of balls or bullets hitting the screen, I_1 is the total number of balls or bullets passing through the first hole, and I_2 is the total number of balls or bullets passing through the second hole. Equation (4) reveals no interference factor between I_1 and I_2, thus agrees with the intensity distribution of macroscopic objects such as balls or bullets. If one of the slits (holes) is closed, then the intensity distribution of balls, bullets or electrons is identical. But once both slits are open, balls or bullets tends to show its classical particle property, whereas electrons tends to show its quantum property with the occurrence of an interference pattern.

That is to say, **intensity distribution occurs naturally** from a double slit experiment of **waves**, thus **no probability aspects** are present. Instead, for a double slit experiment of **particles** that exhibits interference phenomena, **probability aspects occurs** in its **intensity distribution**, which then leads to the probability theory of quantum mechanics.

Therefore, this paper will discuss the intensity distribution of the double slit experiment of electrons, which is emitted one by one from its source, using a certain Gaussian density distribution. This paper attempts to make a clearer view towards the Gaussian density distribution which governs the intensity patterns of the double slit experiment of electrons.

2 Gaussian Probability Density of a Free Electron in Random Area I

As stated above, the discussion starts off from the 1-dimension Schrödinger differential equation of equation (1). In this case, we use the free particle scheme with $V(x) = 0$ in describing the electron movement.

An electron of mass m_e which is emitted from a source with a speed of v, satisfies a Schrödinger equation of:

$$i\hbar \frac{\partial \Psi(x,t)}{\partial t} = -\frac{\hbar^2}{2m_e} \frac{\partial^2 \Psi(x,t)}{\partial x^2}.$$

Using a variable separation method, $\Psi(x,t) = \psi(x)\varphi(t)$, the equation above can be solved analytically with $x_0 = -L$ (referring to **picture 1** as a set up of the double slit experiment), yielding a wavefuction, viz.:

$$\Psi(x,t) = \psi(x)\varphi(t) = \left[Ae^{-ik(x+L)} + Be^{ik(x+L)}\right] e^{-i\frac{Et}{\hbar}}, \quad (5)$$

with $k = \left(\frac{2mE}{\hbar^2}\right)^{\frac{1}{2}}$ and $E \geq 0$ is the energy of the electron. Now, $Ae^{-ik(x+L)}$ is interpreted as part of the wavefunction which is moving to the left, whereas $Be^{ik(x+L)}$ is the wavefunction moving to the opposite direction. Nevertheless, because the electron is moving to the right (according to **picture 1**), then the wavefunction moving to the right should be used, thus:

$$\Psi(x,t) = \psi(x)\varphi(t) = Be^{ik(x+L)} e^{-i\frac{Et}{\hbar}} = Be^{ik(x+L)-i\frac{Et}{\hbar}} = Be^{i\left[k(x+L)-\frac{Et}{\hbar}\right]}. \quad (6)$$

But, this kind of wavefunction cannot be normalized, viz.:

$$\int_{-\infty}^{+\infty} \left|\Psi(x.-L,t)\right|^2 dx = \int_{-\infty}^{+\infty} \left|Be^{i\left[k(x+L)-\frac{Et}{\hbar}\right]}\right|^2 dx = BB^* \int_{-\infty}^{+\infty} dx \to \infty.$$

So, following [1], in order to gain a normalized wavefunction for a moving free particle, a so called wavepacket must be introduced, that is:

$$\Psi(x,-L,t) = \int_{-\infty}^{+\infty} dk A(k) Be^{i\left[k(x+L)-\frac{Et}{\hbar}\right]}, \quad (7)$$

with $A(k)$ is a Fourier transform [technically the exponential Fourier transform of $\Psi(x,-L,t)$]. $A(k)$ can be filled with arbitrary function that satisfies equation (7). Not only that, $A(k)$ must have the ability to localize an infinite wave and normalize it. In this case, a Gaussian wavepacket in k-space [1] is used to complete equation (7), which is:

$$A(k) = e^{-\left[\frac{(k-k_0)}{\sqrt{2}\Delta k}\right]^2}, \quad (8)$$

where k_0 is the mean value and Δk is the uncertainty value. But it has to be understood that equation (8) has nothing to do yet with the probability density of an electron. Equation (8) is used to complete equation (7) because of its natural **bell-**

shaped curve. Gaussian function has the ability of damping a wave when x is relatively far from its peak value, producing the desired condition of particle localization. Not only that, a Gaussian function admits an uncertainty factor, viz. Δk, which usually called the uncertainty value in quantum mechanics term. So substituting equation (8) into equation (7) yields:

$$\Psi(x,-L,t)=\int_{-\infty}^{+\infty} e^{-\frac{(k-k_0)^2}{2(\Delta k)^2}} Be^{i\left[k(x+L)-\frac{Et}{\hbar}\right]}dk = B\int_{-\infty}^{+\infty} e^{\left[-\frac{(k-k_0)^2}{2(\Delta k)^2}+i\left[k(x+L)-\frac{Et}{\hbar}\right]\right]}. \quad (9)$$

Using some mathematical technique, a Gaussian wavepacket is gained in coordinate representation as:

$$\Psi(x,-L,t)=B\sqrt{2\pi}(\Delta k)e^{\left[i\left\{k_0(x+L)-\frac{Et}{\hbar}\right\}-\frac{1}{2}(x+L)^2(\Delta k)^2\right]}, \quad (10)$$

where B has to be determined by means of normalization, or

$$\int_{-\infty}^{+\infty}\Psi(x,-L,t)\Psi^*(x,-L,t)dx=1. \quad (11)$$

Inserting equation (10) into (11), and using again some mathematical procedure, finally, a normalized Gaussian function in coordinate representation is gained as:

$$\Psi(x,-L,t)=\left(\frac{\sqrt{\Delta k}}{\pi^{\frac{1}{4}}}\right)e^{\left\{i\left[k_0(x+L)-\frac{Et}{\hbar}\right]-\frac{1}{2}(x+L)^2(\Delta k)^2\right\}}. \quad (12)$$

In quantum mechanics, equation (12) has no physical realization. Instead, invoking **Definition 1**, a physical interpretation of finding an electron over a region of $x+dx$ is given by the probability density of:

$$|\Psi(x,-L,t)|^2dx=\Psi(x,-L,t)\Psi^*(x,-L,t)dx=\left[\frac{\sqrt{\Delta k}}{\pi^{1/4}}e^{\left\{i\left[k_0(x+L)-\frac{Et}{\hbar}\right]-\frac{1}{2}(x+L)^2(\Delta k)^2\right\}}\right]\times$$
$$\left[\frac{\sqrt{\Delta k}}{\pi^{1/4}}e^{\left\{-i\left[k_0(x+L)-\frac{Et}{\hbar}\right]-\frac{1}{2}(x+L)^2(\Delta k)^2\right\}}\right]dx.$$

Thus,

$$|\Psi(x,-L,t)|^2dx=\frac{1}{(1/\Delta k)\sqrt{\pi}}e^{-\frac{(x+L)^2}{(1/\Delta k)^2}}dx, \quad (13)$$

or

$$|\Psi(x,-L,t)|^2dx=\frac{1}{(1/\sqrt{2}\Delta k)\sqrt{2\pi}}e^{-\frac{1}{2}\frac{(x+L)^2}{(1/\sqrt{2}\Delta k)^2}}dx. \quad (14)$$

Equation (14) is the probability density function of an electron to be found somewhere in random area I, with

- mean value = $\mu = -L$;

- variance = $\sigma^2 = \left(1/\sqrt{2}\Delta k\right)^2$; and also
- standard deviation = $\sigma = 1/\sqrt{2}\Delta k$.

Then, we have,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx . \tag{15}$$

The support of equation (15) is $x \in (-\infty, +\infty)$. Now, one thing which differentiate between **Definition 1** and the usual Normal distribution density function is the notion of random variable. In the usual context, the support x is related to some variable random X . But, in **Definition 1**, no variable random is mentioned what so ever. In fact, the support x corresponds to the continuous position of the particle, which is not mentioned as a random variable. Therefore, the position of electron can be stated as $x \sim N\left(-L, \left[1/\sqrt{2}\Delta k\right]^2\right)$. The standard normal density function is:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz , \tag{16}$$

with $z = \frac{x-\mu}{\sigma} = \frac{x+L}{1/\sqrt{2}\Delta k} = \sqrt{2}\Delta k(x+L).$ (17)

Following [9] and referring also to **Definition 1**, a property of density function (16) can be stated as 99.7%, 95% and $2/3$ of finding an electron lies in the interval of $\mu \pm 3\sigma$, $\mu \pm 2\sigma$, and $\mu \pm \sigma$, respectively. Thus, the electron is maximally somewhere around $\mu - 3\sigma \le x \le \mu + 3\sigma$, or

$$2\Delta x_{\max} \approx |\mu + 3\sigma| - |\mu - 3\sigma| = 6\sigma \Rightarrow \Delta x_{\max} \approx 3\sigma . \tag{18}$$

with $\Delta x = -L + x$.

Likewise, $2/3$ of finding the electron lies in the interval of $\mu \pm \sigma$, thus $\mu - \sigma \le x \le \mu + \sigma$, or

$$2\Delta x \approx |\mu + \sigma| - |\mu - \sigma| = 2\sigma \Rightarrow \Delta x \approx \sigma . \tag{19}$$

Substitute $\sigma = \frac{1}{\sqrt{2}\Delta k}$ into (19), yields:

$$\Delta x \approx \frac{1}{\sqrt{2}\Delta k} \Rightarrow \Delta x \Delta k \approx \frac{1}{\sqrt{2}} . \tag{20}$$

Equation (20) is the minimum uncertainty of the Gaussian wavepacket between Δx and Δk . Thus, we have an uncertainty relation of $\Delta x \Delta k \ge 1/\sqrt{2}$.

Moreover, using De Broglie's wavelength relation $k = \frac{m_e v}{\hbar} \Rightarrow \Delta k = \frac{m_e}{\hbar}\Delta v$, thus,

$$\Delta x \Delta v \approx \frac{1}{\sqrt{2}} \frac{\hbar}{m_e}, \tag{21}$$

which is the minimum uncertainty between Δx and Δv. Again, we have an uncertainty relation of $\Delta x \Delta v \geq \hbar / \sqrt{2} m_e$.

According to quantum mechanics, the wavepacket tends to spread all over space, that is [1]:

$$\Delta x = \Delta x(t=0)\left[1 + \frac{t}{T}\right]^{\frac{1}{2}}. \tag{22}$$

where $T = \hbar / \left(m_e v_0^2\right)$ and $\Delta x(t=0) = \frac{\hbar}{m_e v_0}$. Substituting equation (22) into equation (21), yields:

$$\frac{\hbar}{m_e v_0}\left[1 + \frac{t}{T}\right]^{\frac{1}{2}} \Delta v \approx \frac{1}{\sqrt{2}} \frac{\hbar}{m_e} \Rightarrow \Delta v = \frac{v_0}{\sqrt{2 + \frac{2t}{T}}}. \tag{23}$$

3 Derivation of Gaussian Intensity Distribution of the Double Slit Experiment of Electrons in Random Area II

Now, here what will happen in the second random area, which is the area between the double slit screen and the luminous screen. Each slit is considered as a new source of electron. Using equation (12), now with $x_0 = 0$, a normalized Gaussian wavepacket which goes through slit 1 can be stated as:

$$\Psi(x_1, t) = \left(\frac{\sqrt{\Delta k}}{\pi^{\frac{1}{4}}}\right) e^{\left\{i\left[k_0 x_1 - \frac{Et}{\hbar}\right] - \frac{1}{2}(x_1)^2 (\Delta k)^2\right\}}. \tag{24}$$

Likewise, a normalized wavepacket which goes through slit 2 can be stated as:

$$\Psi(x_2, t) = \left(\frac{\sqrt{\Delta k}}{\pi^{\frac{1}{4}}}\right) e^{\left\{i\left[k_0 x_2 - \frac{Et}{\hbar}\right] - \frac{1}{2}(x_2)^2 (\Delta k)^2\right\}}. \tag{25}$$

Equation (24) and equation (25) are produced from the first Gaussian wavepacket which exist in random area I.

If the wavefunction (24) and (25) **interfere coherently** with each other, thus:

$$\Psi_{tot} = \Psi(x_1,t) + \Psi(x_2,t) = \left[\frac{\sqrt{\Delta k}}{\pi^{1/4}}\right]\left\{e^{i\left[k_0x_1-\frac{Et}{\hbar}\right]-\frac{1}{2}(x_2\Delta k)^2} + e^{i\left[k_0x_2-\frac{Et}{\hbar}\right]-\frac{1}{2}(x_2\Delta k)^2}\right\}. \quad (26)$$

Also, we have the complex conjugate of equation (26), which is:

$$\Psi^*_{tot} = \Psi(x_1,t) + \Psi(x_2,t) = \left[\frac{\sqrt{\Delta k}}{\pi^{1/4}}\right]\left\{e^{-i\left[k_0x_1-\frac{Et}{\hbar}\right]-\frac{1}{2}(x_2\Delta k)^2} + e^{-i\left[k_0x_2-\frac{Et}{\hbar}\right]-\frac{1}{2}(x_2\Delta k)^2}\right\}. \quad (27)$$

From equation (26) and (27), the intensity of the interference between $\Psi(x_1,t)$ and $\Psi(x_2,t)$ can be gained as follows:

$$I = \left|\Psi(x_1,t) + \Psi(x_2,t)\right|^2 = \left|\Psi_{tot}\right|^2 = \Psi_{tot}\Psi^*_{tot}$$

$$= \left[\frac{\Delta k}{\pi^{1/2}}\right]\left\{e^{i\left[k_0x_1-\frac{Et}{\hbar}\right]-\frac{1}{2}(x_2\Delta k)^2} + e^{i\left[k_0x_2-\frac{Et}{\hbar}\right]-\frac{1}{2}(x_2\Delta k)^2}\right\}\times$$

$$\left\{e^{-i\left[k_0x_1-\frac{Et}{\hbar}\right]-\frac{1}{2}(x_2\Delta k)^2} + e^{-i\left[k_0x_2-\frac{Et}{\hbar}\right]-\frac{1}{2}(x_2\Delta k)^2}\right\}$$

or

$$I = \left[\frac{\Delta k}{\sqrt{\pi}}\right]\left\{e^{-(x_1)^2(\Delta k)^2} + e^{-(x_2)^2(\Delta k)^2} + e^{-ik_0(x_2-x_1)-\frac{1}{2}(\Delta k)^2\left(x_1^2+x_2^2\right)} + e^{ik_0(x_2-x_1)-\frac{1}{2}(\Delta k)^2\left(x_1^2+x_2^2\right)}\right\}$$

$$= \left[\frac{\Delta k}{\sqrt{\pi}}\right]\left\{e^{-(x_1)^2(\Delta k)^2} + e^{-(x_2)^2(\Delta k)^2} + 2e^{-\frac{1}{2}(\Delta k)^2\left(x_1^2+x_2^2\right)}\cos\left(k_0\left[x_2-x_1\right]\right)\right\}.$$

Thus,

$$I = \left[\frac{\Delta k}{\pi^{1/2}}\right]e^{-\frac{1}{2}(\Delta k)^2\left(x_1^2+x_2^2\right)}\left\{\frac{e^{(x_1)^2}+e^{(x_2)^2}}{e^{\frac{1}{2}\left(x_1^2+x_2^2\right)}} + 2\cos\left(k_0\left[x_2-x_1\right]\right)\right\}. \quad (28)$$

Equation (28) is the Gaussian density distribution function of two wavefunctions which interfere each other. There are two important factors observed from equation (28) which points out the wave and matter properties of an electron all at once. First, $e^{-\frac{1}{2}(\Delta k)^2\left(x_1^2+x_2^2\right)^2}$ which shows the particle nature of the interference, and second $\left\{\frac{e^{x_1^2}+e^{x_2^2}}{e^{\frac{1}{2}\left(x_1^2+x_2^2\right)}} + 2\cos\left[k_0\left(x_2-x_1\right)\right]\right\}$ which indicates the wave nature.

Next, it has to be understood also that the density function (28) comes from the Gaussian wavepacket (12) which confine x_1 and x_2 only over the maximum area of three times its standard deviation value, that is:

$$-3\sigma \le x_1, x_2 \le 3\sigma \Rightarrow -3\frac{1}{\sqrt{2}\Delta k} \le x_1, x_2 \le 3\frac{1}{\sqrt{2}\Delta k}$$

$$\Rightarrow -\frac{3\hbar}{\sqrt{2}m_e}\frac{1}{\Delta v} \le x_1, x_2 \le \frac{3\hbar}{\sqrt{2}m_e}\frac{1}{\Delta v} \Rightarrow -\frac{0.2\times 10^{-3}}{\Delta v} \le x_1, x_2 \le \frac{0.2\times 10^{-3}}{\Delta v}. \quad (29)$$

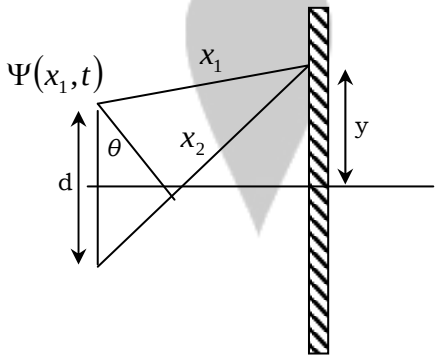
For an example, take the speed of the electron used by Hitachi [11] in their "electron biprism" experiment of the double slit experiment, which is 40% the speed of light or 12×10^7 m/s. Therefore, we would have:

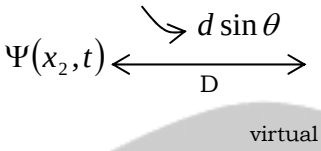
$$-\frac{0.2\times 10^{-3}}{12\times 10^7} \le x_1, x_2 \le \frac{0.2\times 10^{-3}}{12\times 10^7} \Rightarrow x_1, x_2 \le \left|2\times 10^{-12}\right|. \quad (30)$$

From equation (30), x_1, x_2 is in the order of 10^{-12} meter, which is a very small area.

4 GAUSSIAN PROBABILITY DENSITY FUNCTION OF ELECTRON INTERFERENCE

Equation (28) is a **master design** which will eventually produce the observed interference pattern on the luminous screen. The interference of two Gaussian wavefunction coming out of each slit, will interfere **only** in the interval which is given by equation (29). That is why, **before** reaching the real luminous screen, the two wavefunctions must interfere first, then it will show its particle property as a dot on the luminous screen. Now, the two wavefunctions will start to interfere if the distance between two electrons is about 3σ, which lies 99.7% of finding an electron. Then, we introduce a virtual screen, which is again not more than 3σ in length. This can be illustrated as follows:





Picture 2: Interference of two Gaussian wavefunctions.

Thus, from equation (28), $x_2 - x_1$ will be $d\sin\theta$, $x_2^2 = D^2 + \left(y + \frac{1}{2}d\right)^2$ and $x_1^2 = D^2 + \left(y - \frac{1}{2}d\right)^2$. $\sin\theta$ can be stated in terms of y, that is:

$$x_1^2 - (x_2 - d\sin\theta)^2 = d^2 - d^2\sin^2\theta$$

$$\Rightarrow x_1^2 - x_2^2 + 2dx_2\sin\theta = d^2$$

$$\Rightarrow \sin\theta = \frac{d^2 - x_1^2 + x_2^2}{2dx_2} = \frac{d + 2y}{2\sqrt{D^2 + (y + 1/2)^2}}$$

Thus, $k_0[x_2 - x_1] = \frac{1}{2}k_0 \frac{d^2 + 2dy}{\sqrt{D^2 + (y + 1/2\,d)^2}}$. (31)

Now, let us simplify the $\frac{e^{(x_1)^2} + e^{(x_2)^2}}{e^{\frac{1}{2}(x_1^2 + x_2^2)}}$ term in equation (28). This term can be stated as:

$$\frac{e^{x_1^2} + e^{x_2^2}}{e^{\frac{1}{2}(x_1^2 + x_2^2)}} = \frac{e^{D^2 + \left(y - \frac{1}{2}d\right)^2} + e^{D^2 + \left(y + \frac{1}{2}d\right)^2}}{e^{\frac{1}{2}\left(D^2 + \left(y - \frac{1}{2}d\right)^2 + D^2 + \left(y + \frac{1}{2}d\right)^2\right)}} = e^{-yd} + e^{yd}. \quad (32)$$

Using power series expansion of each term in equation (32), yields:

$$e^{-yd} + e^{yd} = \left[1 - (yd) + \frac{(yd)^2}{2!} - \frac{(yd)^3}{3!} + \ldots\right] + \left[1 + (yd) + \frac{(yd)^2}{2!} + \frac{(yd)^3}{3!} + \ldots\right]$$

$$= 2 + \frac{2(yd)^2}{2!} + \frac{2(yd)^4}{4!} + \frac{2(yd)^6}{6!} + \ldots = 2\left[\sum_{i=0}^{\infty} \frac{(yd)^{2i}}{2i!}\right] \approx 2. \quad (33)$$

Equation (33) is valid for small value of yd, that is $0 < yd << 1$. For a localized particle, equation (33) is satisfied within $t < T$ of equation (22).

So, by inserting equation (31) and (33) into equation (28), we gain

$$I = \left[\frac{\Delta k}{\pi^{1/2}}\right] e^{-\frac{1}{2}(\Delta k)^2\left(D^2+\left(y-\frac{1}{2}d\right)^2+D^2+\left(y+\frac{1}{2}d\right)^2\right)} \left\{2+2\cos\left(\frac{1}{2}k_0\frac{d^2+2dy}{\sqrt{D^2+\left(y+1/2\,d\right)^2}}\right)\right\}$$

$$= \left[\frac{\Delta k}{\sqrt{\pi}}\right] e^{-\frac{1}{2}(\Delta k)^2\left(2D^2+2y^2+\frac{1}{2}d^2\right)} 2\left\{1+\cos\left(\frac{1}{2}k_0\frac{d^2+2dy}{\sqrt{D^2+\left(y+1/2\,d\right)^2}}\right)\right\}$$

Because $\cos(2z)+1=2\cos^2(z)$, the above equation becomes:

$$I = |\Psi_{tot}|^2 = \left[\frac{1}{\sigma\sqrt{2\pi}}\right] e^{-\frac{1}{2}\frac{\left(D^2+y^2+\frac{1}{4}d^2\right)}{\sigma^2}} 4\cos^2\left(k_0\frac{d^2+2dy}{4\sqrt{D^2+\left(y+\frac{1}{2}d\right)^2}}\right), \tag{34}$$

with $\sigma = 1/\sqrt{2}\Delta k$. Following the uncertainty relation of equation (20), D and d are should be about σ, or $D^2 = d^2 \approx \sigma^2$. Finally, the Gaussian density function of electron interference is

$$I = |\Psi_{tot}|^2 \approx 2\left[\frac{1}{\sigma\sqrt{2\pi}}\right] e^{-\frac{1}{2}\frac{y^2}{\sigma^2}} \cos^2\left(k_0\sigma\frac{[y+(1/2)\sigma]}{2\sqrt{\sigma^2+[y+(1/2)\sigma]^2}}\right). \tag{35}$$

Equation (35) is the modified Gaussian density function with $\mu = 0$. An electron that has this distribution is written as $y \sim N(0,\sigma^2)\cos^2(0,\sigma)$. Equation (35) still contains its two parts which is different from the usual Gaussian density function of equation (15). First, the particle property, which is represented by $\left[1/\left(\sigma\sqrt{2\pi}\right)\right]\exp\left[-y^2/\left(2\sigma^2\right)\right]$ and second, the wave property, which is represented by the cosines periodic function. For the usual Young double slit experiment, the first factor (particle nature) does not occur. This is because its derivation comes from a pure wave consideration.

5 CONCLUSION

From the discussion above, we have derived a Gaussian probability density function of electron in random area I as observed in equation (16), including also equation (17). Then a modified Gaussian density function of electron interference is formulated in random area II that can be observed in equation (35). The modified Gaussian density function occurs because there is an extra factor of

$\cos^2\left(k_0\sigma\left[\{y+(1/2)\sigma\}^2/\left\{2\sqrt{\sigma^2+[y+(1/2)\sigma]^2}\right\}\right]\right)$ which shows the source of interference in the double slit experiment of electrons.

References

[1] Amit G. (1992), *Quantum Mechanics*, Prentice Hall, New Jersey.

[2] A. Papoulis (1992), *Probability, Random Variable and Stochastic Process*, Second Edition, Gadjah Mada University Press.

[3] Anton Z.C. (1985), *Nonrelativistic Quantum Mechanics (lecture Notes and Supplements in Physics*), The Benjamin/Cummings, Inc, California.

[4] David J.G (1995), *Introduction to Quantum Mechanics*, Prentice Hall, New Jersey.

[5] Dwandaru, W.S.B (2005), Towards the Statistical Interpretation of the Double Slit Experiment: Random Variable Analysis (Part I), *Proceedings of the National Seminar of Research, education and Applied Science and Mathematics*, Yogyakarta, Indonesia, Editors: H. Sutrisno, Dr. Ariswan, H. Kuswanto, H. Nurcahyo, Sugiman, F-118 to F-135.

[6] Eugene H., and Alfred, Z. (1974), *Optics*, Addison-Wesley, Reading, Massachusetts.

[7] F.K. Richtmyer, E.H. Kennard, and John N. Cooper (1976), *Introduction to Modern Physics*, Tata McGraw-Hill, New Delhi.

[8] Hirose A., and Lonngren, K.E.(1985), *Introduction to Wave Phenomena*, John Wiley & Sons, New York.

[9] J. B. Fitzpatrick and P.L. Galbraith (1990), *Reasoning and Data Heinemann Senior Mathematics*, Heinemann Educational Australia: a division of the Octopus Publishing Group Australia, Melbourne.

[10] Stephen G. (1996), *Quantum Physics*, John wiley&Sons, Inc., New York.
Ramamurti S. (1980), *Principle of Quantum Mechanics*, Plenum Press, new York.

[11] Hitachi Global: Research & Development: Advances in Research, (1994,2005), *Electron Phase Microscopy: Double-Slit Experiment* http://www.hqrd.hitachi.co.jp/em/doubleslit.cfm.

W.S.B. DWANDARU: Computational and Theoretical Physics Laboratory, Physics Education Department, Yogyakarta State University, Karangmalang,
Yogyakarta, 55281, Indonesia.
Yogyakarta Institute for Science, Sodanten, Yogyakarta, Indonesia.
E-mail: wipsarian@yahoo.com

D. DARMAWAN: Computational and Theoretical Physics Laboratory, Physics Education Department, Yogyakarta State University, Karangmalang,
Yogyakarta, 55281, Indonesia.
E-mail: darmawan@scientist.com

N. INSANI: Mathematics Education Department, Yogyakarta State University, Karangmalang, Yogyakarta, 55281, Indonesia.

R. Subekti: Mathematics Education Department, Yogyakarta State University, Karangmalang, Yogyakarta, 55281, Indonesia.

THE EFFECT OF IMMUNISATION RATE ON A MATHEMATICAL MODEL OF YELLOW FEVER EPIDEMIC

T. ADENIRAN AND R.O AYENI

Department of Pure and Applied Mathematics
Ladoke Akintola University of Technology, Ogbomoso.

ABSTRACT. We revisit the mathematical model of yellow fever which involves the interactions of two principal communities of hosts (humans) and vectors (aedes aegypti mosquitoes)[6]. We extend the model to a situation where immunisation is non – permanent. Using numerical technique, we evaluate S(t), R(t), I(t), N(t) and M(t) and show the immunisation and rate at which vaccine wanes have significant effects on the epidemic.

INTRODUCTION

We consider two interacting communities of hosts and vectors as in [1] where he host community is divided into three compartments of susceptible S(t), infected I(t), and recovered or immune R(t) while the vector community is partitioned into two compartments of susceptible N(t), and infective or virus carriers M(t) where $t \geq 0$ is the time.
Effective biting interaction between S(t) and M(t) results in the movement of members from S(t) into I(t), while a similar interaction between I(t) and N(t) leads to the flow of members of N(t) into M(t)[1].
Furthermore, the hosts are born into the system as susceptible as it takes about a year for a child to lose immunity after birth [6].
Akinwande [1] carried out the stability analysis of the equilibrium state of this same model using a modified fashion of implicit function theorem .
Unlike the paper considered, we assume that the effect/impact of the immunisation on the model particularly on the recovered population R(t) is not permanent(the immunisation wanes) .

MODEL EQUATIONS

The model equations are given as follows:

$$S' = \beta_1(S+I+R) - (\mu_1+\gamma)S - \alpha_1 MS + wR \quad \ldots\ldots 1$$
$$R' = -\mu_1 R + \gamma S + \alpha I \quad \ldots\ldots 2$$
$$I' = -(\mu_1+\alpha+\alpha_\delta)I + \alpha_1 MS \quad \ldots\ldots 3$$
$$N' = \beta_2(N+(1-\theta)M) - \mu_2 N - \alpha_2 NI \quad \ldots\ldots 4$$
$$M' = \theta\beta_2 M - \mu_2 M + \alpha_2 NI \quad \ldots\ldots 5$$
$$S(0) = S_0, R(0) = R_0, I(0) = I_0, N(0) = N_0, M(0) = M_0.$$

The parameters are defined as follows:
β_1 = natural birthrate for hosts.

β_2 = natural birthrate for vectors.
μ_1 = natural morality rate for hosts.
μ_2 = natural morality rate for vectors.
α = recovery rate.
α_1= effective biting interaction rate between S(t) and M(t) compartments.
α_2= effective biting interaction rate between N(t) and I(t) compartments.
α_δ= death rate from interaction.
θ = proportion of the offspring of M(t) that is infected vertically.
γ = Immunisation rate.
W = rate at which immunisation wanes

METHOD OF SOLUTION.

FINITE DIFFERENCE APPROXIMATION METHOD.

We solve equations (1) – (5) numerically,

$S' = \beta_1(S+I+R) - (\mu_1+\gamma)S - \alpha_1 MS + wR$ implies

$$\frac{dS}{dt} = \beta_1(S+I+R) - (\mu_1+\gamma)S - \alpha_1 MS + wR$$

where $\frac{dS}{dt} = \frac{S_{i+1} - S_i}{h}$

$\therefore S_{i+1} = ($

$\beta_{1_1}(S_i + I_i + R_i) - (\mu_1+\gamma)S_i - \alpha_1 MS + wR_i)h + S_i$ 6

likewise,

$R_{i+1} = (-\mu_1 R_i + \gamma S_i + \alpha I_i)h + R_i$ 7
$I_{i+1} = (-(\mu_1 + \alpha + \alpha_\delta)I_i + \alpha_1 MS_i)h + I_i$ 8
$N_{i+1} = (\beta_2(Ni + (1-\theta)Mi) - \mu_2 N_i - \alpha_2 N_i I_i)h + N_i$ 9
$M_{i+1} = (\theta\beta_2 M_i - \mu_2 M + \alpha_2 N_i I_i)h + M_i$ 10

The class of population that determines the effect of both immunisation (γ) and the rate at which immunisation wanes (w_R) is the Infected Population I(t).
We now assume and vary parameters for different instances as follows:
Assumptions:
h =0.02, β_1 = 0.8, β_2 = 0.5, μ_1 = 0.7, μ_2 = 0.3, α = 0.4, α_1= 0.65, α_2= 0.5, α_δ= 0.35, θ = 3.4.

Case 1:
γ = 0, 1, 10 and w = 0.

Case 2:
γ = 0, 1, 10 and w = 0.8.

Case 3:
γ = 0, 1, 10 and w = 1.5.

Case 4:
γ = 10 and w = 0, 0.8 and 1.5

the result is shown in figures (1) – (4)

Figure 1:

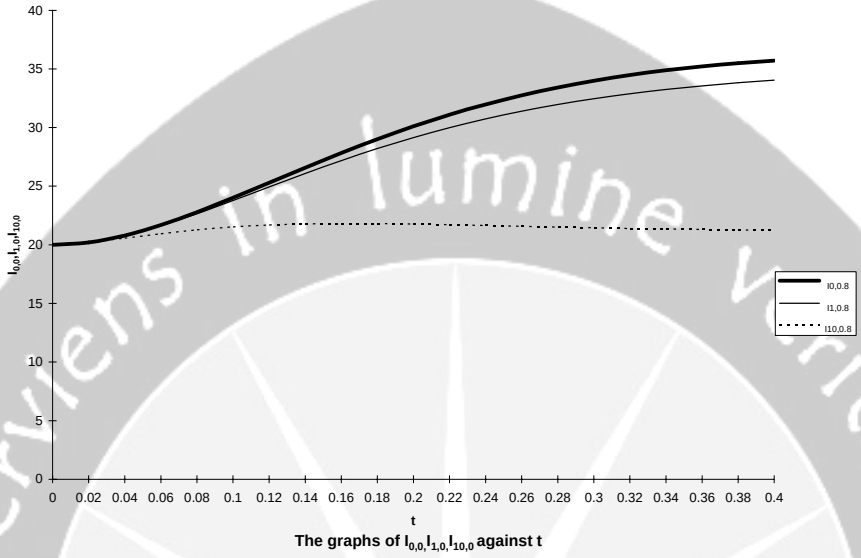


Figure 2:

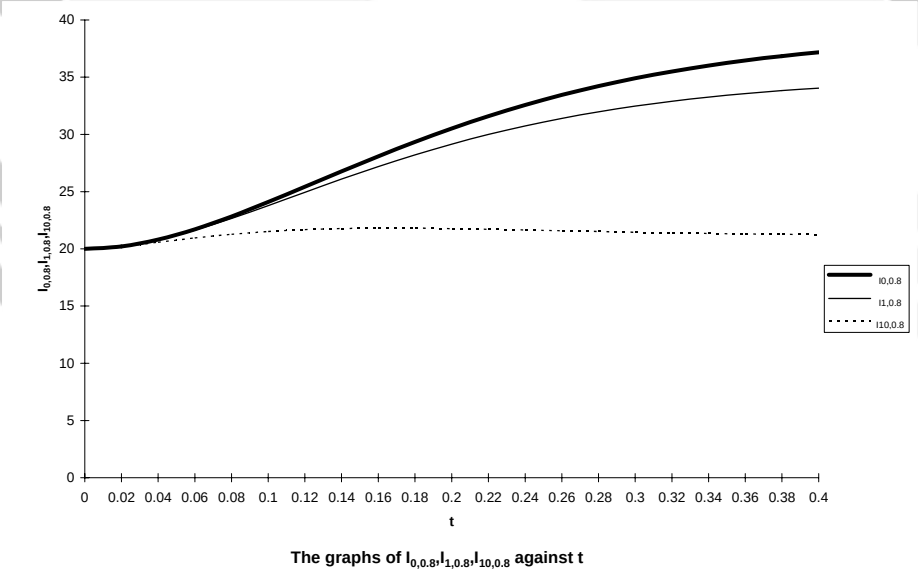
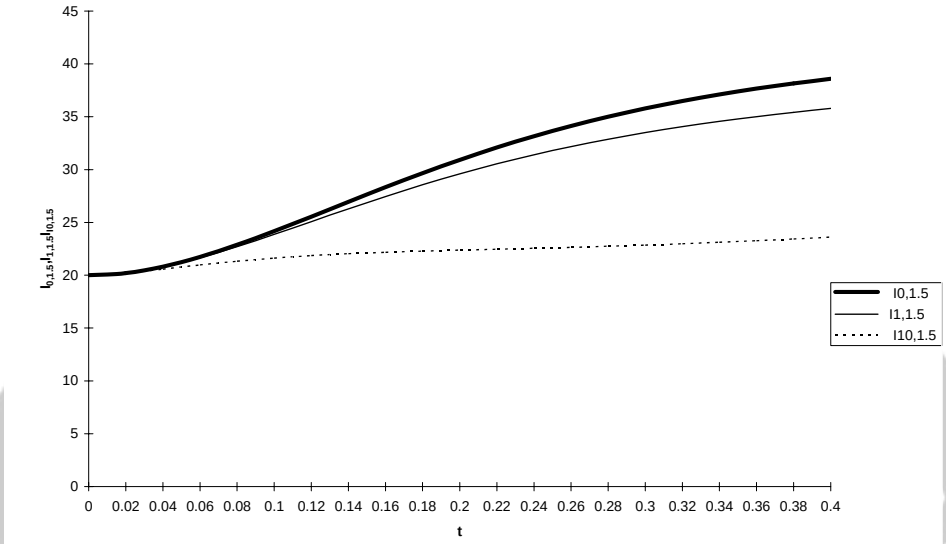
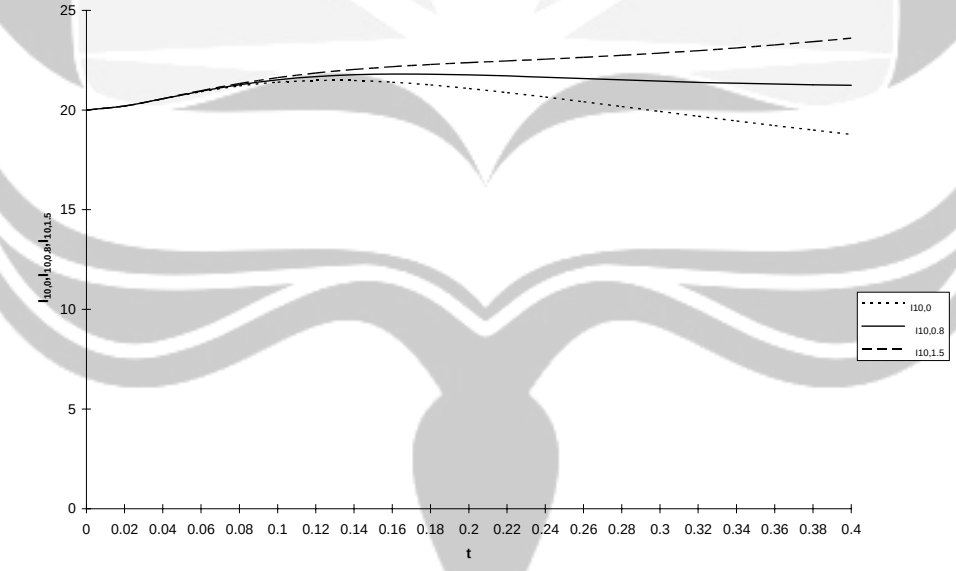


Figure 3:



The graphs of $I_{0,1.5}$,$I_{1,1.5}$, $I_{I0,1.5}$ against t

Figure 4:



The graphs of $I_{10,0}$,$I_{10,0.8}$,$I_{10,1.5}$ against t

DISCUSSION OF RESULTS

In figure 1, when the rate at which immunisation wanes is zero (w = 0), and varying the values of immunisation rate (γ = 0, 1, and 10), we discover that the infected population I(t), reduces as the immunisation rate increases with time. While the number of susceptible S(t) and the vector compartment of the susceptible N(t) reduces as time increases with the recovered, R(t) and the virus carriers, M(t) increases with time.

Figure 2, where the rate at which immunisation wanes is now a little greater than zero i.e. w = 0.8, we discover that we have the same outcome as in figure 1 but at a higher rate since he effect of the vaccine is no more permanent.

While in figure 3, where the impact of rate at which immunisation wanes is now significant and precisely at 1.5, it gives just the same result as in figures 1 and 2 but with an higher population of the infected.

Figure 4 shows the effect of rate at which immunisation wanes on the model by keeping the immunisation rate constant (γ = 10) and varying w (0, 0.8 and 1.5), we discover that the number of the infected I(t) increases as w increases.

CONCLUSION

We could see clearly from the figures that the infected population I(t) reduces as the immunisation rate increases, while the introduction of the rate at which the immunisation wanes retard the impact of the vaccine, the recovered R(t) increases with time as the immunisation rate increases while the susceptible S(t) reduces as the immunisation rate increases.

Conclusively, the epidemic outbreak may be prevented by keeping the immunisation rate far greater than zero, and if the outbreak occur at all, a vaccine that would permanently cure it should be used i.e when w = 0.

REFERENCES

1. Akinwande, N.I (1995): Local Stability Analysis of Equilibrium states of a Mathematical model of Yellow Fever Epidemics. JNMS Vol. 14/15 pp. 73 – 79
2. Beltrami, E (1989): "Mathematics for dynamics modeling", Academic Press Inc., London.
3. Gurtin, M.E and MacCamy (1974): Nonlinear age – dependent population dynamics, Arch. Rat. Mech. Anal. 54, 281 – 300.
4. Murray, J.D (1989): "Mathematical Biology", Springer – Verlag, New York.
5. Sowunmi, C.O.A (1987): On a set of sufficient conditions for the exponential asymptotic stability of equilibrium states of a female dominant model. J. Nig. Math. Soc. 6, pp. 56 – 69.
6. Tomori, O. (1988): Mathematical modeling and disease epidemics. Proceed. Internat. Workshop on Biomathematics, Univ. of Ibadan, Nigeria pp. 9 – 22.

T. ADENIRAN: Department of Pure and Applied Mathematics, Ladoke Akintola University of Technology, Ogbomoso
E-mail: easy775@yahoo.com

R.O AYENI: Department of Pure and Applied Mathematics, Ladoke Akintola University of Technology, Ogbomoso

Stochastic Models for the Spread of HIV in a Mobile Heterosexual Population

A. Sani [a,b], D. P. Kroese[b], P. K. Pollett[b]

[a] Dept. of Math., Universitas Haluoleo, Kendari, INDONESIA.
[b] Dept. of Math., University of Queensland, AUSTRALIA.

Abstract. An important factor in the dynamic transmission of HIV is the mobility of the population. We formulate various stochastic models for the spread of HIV in a heterosexual mobile population, under the assumptions of constant and varying population sizes. We also derive deterministic and diffusion analogues for these models, using a convenient rescaling technique, and analyze their stability conditions and equilibrium behavior. We illustrate the dynamic behavior of the models and their approximations via a range of numerical experiments.

Key-words: HIV/AIDS, Mobility, Multiple Patches, Epidemiology, Density Dependent Markov Process, Diffusion Approximation.

1 Introduction

One of the most urgent public-health problems in developing countries is the AIDS (Acquired Immune Deficiency Syndrome) epidemic, caused by the Human Immunodeficiency Virus (HIV). Since the first cases of AIDS were identified in 1981, the number of HIV infected people and AIDS deaths per year has continued to rise rapidly. In 2004, some 40 million people were living with HIV, which has killed over 20 million since 1981 and 3 million in 2003 alone [1]. The epidemic is not homogeneous within geographical regions. Some countries are more affected than others. Even at country level there are usually wide variations in infection levels between different provinces, states or districts, and between urban and rural areas. In reality, the national picture is made up of a series of epidemics with their own characteristics and dynamics.

The dynamic transmission of HIV is quite complex and there is no other human infection which has the same epidemiology characteristics with a similar mode of transmission. For instance, the incubation period after infection with HIV is known to be extremely long and is measured in years rather than days (such as in the case of measles, for example). During this period the individuals stay healthy and can unknowingly transmit the disease to others. In addition, although the disease is known as a sexually-transmitted disease, it is also passed on from infected mothers to their babies, and from sharing infected syringes, which is common among injected drug users. All these factors have made it difficult to understand how this epidemic spreads in the population. The growth of movement among populations further increases the contact between individuals in different patches and, consequently, it might trigger more epidemics. Thus, the migration of people among subgroups has many significant consequences for the outcome of epidemic

spread [20]. Indonesia in particular, as one of the most populous countries in the world, with a high population mobitttttity among its regions [10], seems to have a high risk for the spread of the epidemic [1]. The number of people infected with HIV/AIDS in this country shows to increase sharply. The prevalence of HIV/AIDS among provinces in this country varies widely.

Mathematical models based on the underlying transmission mechanism of HIV might help the medical and scientific community understand better how the disease spreads into the community. Even though the actual data needed for the models might not be accurate or available, such modelling is still vital in investigating how changes in the various assumptions and parameter values would effect the course of the epidemic [11]. Therefore, by developing such mathematical models, we can to some extend anticipate its spread in different populations and evaluate the potential effectiveness of different approaches for bringing the epidemic under control and help to devise effective strategies to minimize the destruction caused by this epidemic.

Mathematical models for the spread of the HIV/AIDS epidemic have been extensively studied since the first cases were recognized in the late 80's, considering many different aspects; see for example [8, 5, 16, 17, 12, 21, 22]. However, this area of study is still challenging, since so many different factors affect the transmission of HIV. Most of the articles have focused on only a single population of constant size, although some studies have stressed the importance of variable population size in epidemic dynamics [9, 16, 8]. In addition, many models have only focused on a single homosexual population [21], whereas in much of the world, heterosexual contact is the predominant mode of transmission [1]. Finally, the spatial aspect of the epidemic and, related with this, the *mobility* of the population, is often ignored. All these assumptions might limit the application of such models in describing the complex dynamics.

The purpose of this paper is to develop new mathematical models for the spread of HIV that incorporate factors such as mobility, heterosexual transmission and varying population size, which are crucial for countries such as Indonesia, with its many distinct regions. The models will be stochastic in nature, as opposed to the more common deterministic models. However, we will show that the more natural stochastic approach can be approximated well with the traditional deterministic approach, which can be analyzed in more detail, in particular with respect to their equilibrium behavior. In addition we derive stochastic diffusion approximations, which show that the original process around the equilibrium can be approximated well via an Ornstein-Uhlenbeck process. Both the deterministic and diffusion approximations are based on the theory of density dependent processes [13, 18].

Our models are mostly motivated by the work of [8] and [16]. Both [8] and [16] formulate deterministic models of HIV spread in a heterogeneous population. They consider the female and male subpopulations separately (i.e., individuals are well-mixed only in their subpopulation) and assume that HIV transmission is possible only through sexual contact between female and male. There are some differences between these two models: [8] assumes that the rate of new recruits of susceptibles (for both males and females) is constant, whereas in [16] this rate is assumed to be proportional to the total population, which varies in time. In [8] only males choose

partners from the female subpopulation. Thus, susceptible males and females get the infection at a rate which is proportional to the size of the total female population. On the other hand, [16] assumes that also females choose partners from the male subpopulation. Therefore, susceptible males get infected relative to the total female population and susceptible females get the infection relatively to the total male population. Consequently, the models [8] and [16] have slightly different formulations for the infection rate of susceptibles. Both [8] and [16] study the situation under the assumption of a varying population.
The rest of the paper is organized as follows. In Section 2, we describe the various stochastic models. We start with a single, constant (i.e., a closed system) or varying (i.e., an open system) population with a female and male subpopulation, and then look at the case of a multiple-patch population, incorporating the mobility of people. In Section 3, we present various results from Kurtz [13, 14] concerning density dependent processes. In particular, we review under which conditions and in what manner such a stochastic process converges to its deterministic and diffusion counterpart. In Section 4 we will use the results from Section 3 to study the dynamics of our stochastic models. Numerical experiments are presented in Section 5. Finally, in Section 6, we summarize our findings and give direction for future research.

2 Models

In this section we formulate various stochastic models for the spread of HIV in both a single population and in multiple populations, under the assumption of either a constant or varying population size.

2.1 Model with a Closed Single Population

We consider first a closed (constant) single heterosexual population of size N in which all individuals, both females and males, are well mixed in the population. We assume, as in [8] and [16], that a susceptible female gets infected only from an infected male (via sexual contact) and, similarly, a susceptible male gets the infection only from an infected female. A single female or male selects her/his partner (of different sex) randomly from the whole population.
Let the random variables $S_F(t), I_F(t)$, $S_M(t)$, $I_M(t)$, and $A(t)$ represent the number of susceptible females, infected females, susceptible males, infected males, and the number of AIDS cases at time t, respectively. We assume that a susceptible female (male) will get infected from an infected male (female) at a rate that is proportional to the fraction of infected males (females):

$$\lambda_F = \beta \frac{I_M(t)}{N} \qquad \Big(\lambda_M = \beta \frac{I_F(t)}{N}\Big), \tag{1}$$

where λ_F and λ_M are called the *forces of infection* (see also Remark 2.1). We assume that all individuals, including AIDS people, die at a natural death rate μ. In addition, AIDS people also die due to the disease, at rate δ. All deaths are replaced (balanced) by births of susceptibles, at a proportion α for females

and $(1-\alpha)$ for males. Thus, the birth rates for susceptible females and males are $B_F = \alpha\,(\mu\,N + \delta\,A)$ and $B_M = (1-\alpha)\,(\mu\,N + \delta\,A)$, respectively. The infected individuals develop AIDS at rate γ. This situation can be viewed as a stochastic Susceptible-Infected-Removed (SIR) model; see for example [2]. The scheme is illustrated in Figure 1.

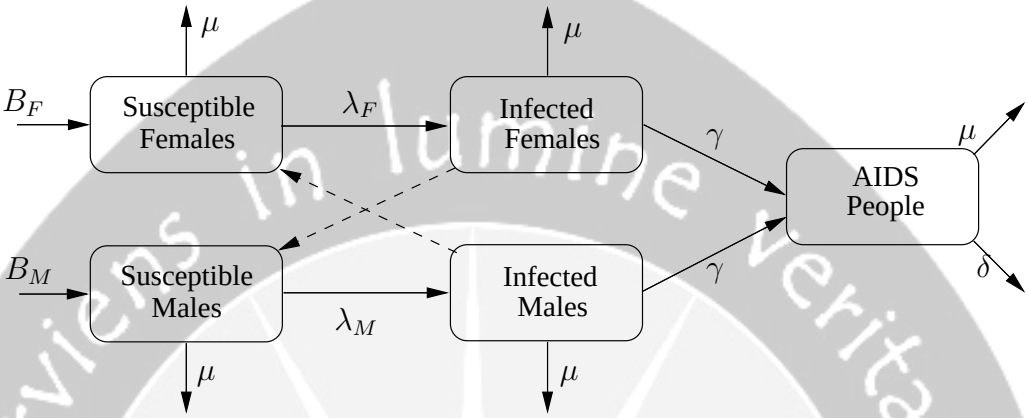


Figure 1: The scheme of the model. Susceptible females (males) are infected by infected males (females) via sexual contact only, indicated by the dashed lines.

Remark 2.1 (Force of Infection) The parameter β is defined in [8] as the product of the contact rate κ per unit time and the probability p that a successive number of contacts leads to infection. The constants κ and p are given as follows: $\kappa = \frac{1}{T}$ and $p = 1-(1-h)^{cT}$, where T is the time interval between two encounters with new partners, c is the average number of sexual contacts between partners, and h is the probability that one sexual contact between a susceptible and an infected individual leads to infection.

Consider the process $(X(t), t \geq 0)$, with

$$X(t) = \Big(S_F(t), I_F(t), S_M(t), I_M(t)\Big),$$

which takes values in $E \subset \mathbb{N}^4$, where $\mathbb{N}$ is the set of positive integers (including zero). We model $(X(t), t \geq 0)$ as a Continuous Time Markov Chain (CTMC) (see, e.g., [19]), where the transition rates are chosen according to the description above. Thus, we assume that given the whole history $X(s), s \leq t$, a future state of the system, $X(t+\Delta t)$, depends only on the current state $X(t)$. In the formulation of the model we can ignore $A(t)$, since the population size, $N = S_F(t) + I_F(t) + S_M(t) + I_M(t) + A(t)$, is constant for all t. If one is interested in the number of AIDS cases, one can find it from $A(t) = N - S_F(t) - I_F(t) - S_M(t) - I_M(t)$.

Transition Rates

We now have a closer look at the transition rates of the CTMC $(X(t), t \geq 0)$. In a small time interval Δt we assume that one of the following events occurs: (1) a new susceptible female enters the group of single females, (2) a susceptible female gets infected, (3) a susceptible female dies, (4) an infected female is removed (develops AIDS or dies), (5) a new susceptible male enters the group of males, (6) a susceptible male becomes infected, (7) a susceptible male dies, or (8) an infected male is removed (due to AIDS or natural death). The other possible events are ignored.
Suppose that the system at time t is in state $\mathbf{k} = (s_F, i_F, s_M, i_M)$, $\mathbf{k} \in E$. The transition scheme of the process is described in Figure 2 (ignoring boundary effects).

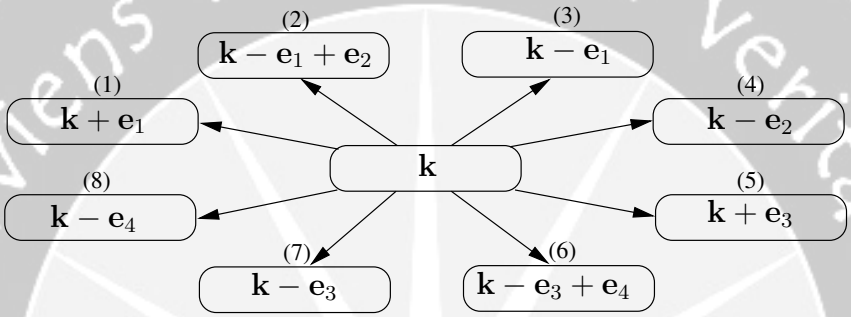


Figure 2: The transition scheme from state $\mathbf{k}$ to other states, where $\mathbf{e}_i$ represents the $i-$th unit row vector in $\mathbb{N}^4$.

Thus, in any small time interval of length Δt the process jumps from state $\mathbf{k}$ to $\mathbf{k}+l$ with probability $q_{\mathbf{k},\mathbf{k}+l}\Delta t$, where the rates $q_{\mathbf{k},\mathbf{k}+l}$ follow from the formulation above, and are given by

$$q_{\mathbf{k},\mathbf{k}+l} = \begin{cases} \alpha\,(\mu\,N + \delta\,A), & l = \ \ \mathbf{e}_1, \\ \beta\,\frac{i_M}{N}\,s_F, & l = -\mathbf{e}_1 + \mathbf{e}_2, \\ \mu\,s_F, & l = -\mathbf{e}_1, \\ (\mu+\gamma)\,i_F, & l = -\mathbf{e}_2, \\ (1-\alpha)\,(\mu\,N + \delta\,A), & l = \ \ \mathbf{e}_3, \\ \beta\,\frac{i_F}{N}\,s_M, & l = -\mathbf{e}_3 + \mathbf{e}_4, \\ \mu\,s_M, & l = -\mathbf{e}_3, \\ (\mu+\gamma)\,i_M, & l = -\mathbf{e}_4, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Note that the process $(X(t), t \geq 0)$ has an absorbing state $\mathbf{0}$, and once the process reaches a state where no infection is present (i.e., $I_F(t) = I_M(t) = 0$), it will remain infection free forever, and will eventually end up in $\mathbf{0}$.

2.2 Model with an Open Single Population

In this model, we consider a population size $N(t)$ which varies with time. We have now a slightly different interpretation for the population size. In the constant population case, we include AIDS people in the total population, which makes it possible to formulate the situation as a type of SIR model. With a varying population size, both the female and male subpopulation are simply divided into two groups of susceptibles and infectives, as in the case of the standard SI model. We no longer explicitly consider AIDS people as a part of the population, that is, $N(t) = S_F(t) + I_F(t) + S_M(t) + I_M(t)$. However, if one is interested in the number of AIDS cases at time t, $A(t)$, one can find it from the number of infectives who eventually develop AIDS, that is, $A(t) = \int_0^t \gamma(I_F(s) + I_M(s))ds$. We assume as in [8] that the number of new susceptibles of both females and males arrive into the system at a constant rate $B_F = B_M = B$ (that is, according to a Poisson process with rate B). Thus, the transition scheme is similar to the previous model, but the transition rates of the process are given as follows:

$$q_{\mathbf{k},\mathbf{k}+l} = \begin{cases} B, & l = \mathbf{e}_1, \\ \beta \frac{i_M(t)}{N(t)} s_F(t), & l = -\mathbf{e}_1 + \mathbf{e}_2, \\ \mu\, s_F(t), & l = -\mathbf{e}_1, \\ (\mu+\gamma)\, i_F(t), & l = -\mathbf{e}_2, \\ B, & l = \mathbf{e}_3, \\ \beta \frac{i_F(t)}{N(t)} s_M(t), & l = -\mathbf{e}_3 + \mathbf{e}_4, \\ \mu\, s_M(t), & l = -\mathbf{e}_3, \\ (\mu+\gamma)\, i_M(t), & l = -\mathbf{e}_4, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

Similar to the previous case, this process has an absorbing state $\mathbf{0}$, and once the process reaches the state with no infected individuals, it will remain infection free and will eventually go to $\mathbf{0}$.

2.3 Multiple Patch Models with Varying Population Size

In order to incorporate mobility effects, we consider individuals residing in many *patches* or *regions*. The population sizes of the patches need not be equal and may vary with time. Individuals may get the infection or transmit the disease during their visit to other patches. People might visit the same patches several times and spend a varying length of time in the visited patches. Suppose v_{rj} denotes the immigration rate of individuals from patch R_r to R_j. The following diagram illustrates the mobility of people among patches.

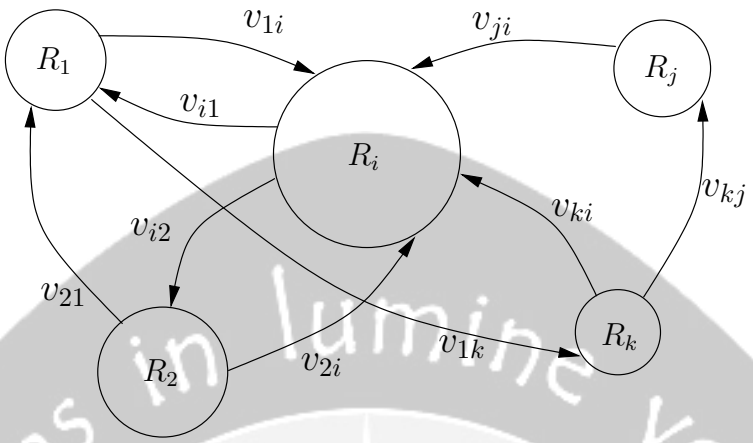


Figure 3: The scheme for the mobility of people among patches. The size of a circle corresponding to the total population size in that patch.

We formulate two types of model, assuming that each patch (as in the previous model for a single population) contains a female and a male subpopulation. In the first type of model, we assume that individuals do not actually leave their home patches but that there is an infection force from other patches. In the second type of model, we assume that individuals do leave their home patches and spend a considerable amount of time in the visited patches before they return. They might emigrate and stay permanently in a visited patch. We call the first model the *model with a force of infection* and the second model the *model with actual mobility*. We consider both constant and varying population sizes.
In both models there are K patches and each patch contains a female and male subpopulation. Let $S_F^{(r)}(t), I_F^{(r)}(t), S_M^{(r)}(t), I_M^{(r)}(t)$ represent the number of susceptible (infected) females and the number of susceptible (infected) males at time $t \geq 0$ in patch $r, r = 1, \ldots, K$, respectively. Define a CTMC $(X(t), t \geq 0)$ with

$$\begin{aligned} X(t) \quad = \quad & \Big(S_F^{(1)}(t), I_F^{(1)}(t), S_M^{(1)}(t), I_M^{(1)}(t), \ \ldots, \ S_F^{(r)}(t), I_F^{(r)}(t), S_M^{(r)}(t), I_M^{(r)}(t), \ \ldots, \\ & S_M^{(K)}(t), I_F^{(K)}(t), S_M^{(K)}(t), I_M^{(K)}(t) \Big). \end{aligned}$$

The state of this process is a $4K-$dimensional row vector with elements in $\mathbb{N}$, that is, the state is an element of $\mathbb{N}^{4K}$.

Model with a Force of Infection

To formulate the first model, let β_{rj} denote the infection rate of susceptibles in patch r by infected individuals from patch j and $\beta_r = \beta_{rr}$ the infection rate within patch r. Then, the transition rates for this situation $(r = 1, 2, \ldots, K)$ are given as follows: For a constant population size

$$q_{\mathbf{k},\mathbf{k}+l} = \begin{cases} \alpha\,(\mu N^{(r)} + \delta\, A^{(r)}), & l = \ \ \mathbf{e}_{4r-3}, \\ \sum_{j=1}^{K} \beta_{rj}\, \frac{s_F^{(r)}}{N^{(r)}}\, i_M^{(j)}, & l = -\mathbf{e}_{4r-3} + \mathbf{e}_{4r-2}, \\ \mu\, s_F^{(r)}, & l = -\mathbf{e}_{4r-3}, \\ (\mu+\gamma)\, i_F^{(r)}, & l = -\mathbf{e}_{4r-2}, \\ (1-\alpha)\,(\mu N^{(r)} + \delta\, A^{(r)}), & l = \ \ \mathbf{e}_{4r-1}, \\ \sum_{j=1}^{K} \beta_{rj}\, \frac{s_M^{(r)}}{N^{(r)}}\, i_F^{(j)}, & l = -\mathbf{e}_{4r-1} + \mathbf{e}_{4r}, \\ \mu\, s_M^{(r)}, & l = -\mathbf{e}_{4r-1}, \\ (\mu+\gamma)\, i_M^{(r)}, & l = -\mathbf{e}_{4r}, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

with a constant $N^{(r)} = S_F^{(r)}(t) + I_F^{(r)}(t) + S_M^{(r)}(t) + I_M^{(r)}(t) + A^{(r)}(t)$, and $\mathbf{e}_m$ the $m-$th unit vector in $\mathbb{N}^{4K}$. For the case of varying population size

$$q_{\mathbf{k},\mathbf{k}+l} = \begin{cases} B, & l = \ \ \mathbf{e}_{4r-3}, \\ \sum_{j=1}^{K} \beta_{rj}\, \frac{s_F^{(r)}}{N^{(r)}}\, i_M^{(j)}, & l = -\mathbf{e}_{4r-3} + \mathbf{e}_{4r-2}, \\ \mu\, s_F^{(r)}, & l = -\mathbf{e}_{4r-3}, \\ (\mu+\gamma)\, i_F^{(r)}, & l = -\mathbf{e}_{4r-2}, \\ B, & l = \ \ \mathbf{e}_{4r-1}, \\ \sum_{j=1}^{K} \beta_{rj}\, \frac{s_M^{(r)}}{N^{(r)}}\, i_F^{(j)}, & l = -\mathbf{e}_{4r-1} + \mathbf{e}_{4r}, \\ \mu\, s_M^{(r)}, & l = -\mathbf{e}_{4r-1}, \\ (\mu+\gamma)\, i_M^{(r)}, & l = -\mathbf{e}_{4r}, \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

with a varying size $N^{(r)}(t) = S_F^{(r)}(t) + I_F^{(r)}(t) + S_M^{(r)}(t) + I_M^{(r)}(t)$. Note that with these notations, if there is only one patch $(r, j = 1)$, the transition rates have the same form as those in the previous models for an open and closed single population.

Model with Actual Mobility

In this model, we assume that people physically visit other patches. During their visit the infected individuals can transmit the disease to the susceptibles in the visited patches, and susceptibles visiting a patch might get the infection from an infected individuals in a visited patch. This situation is modelled by considering people moving from one patch to another without any forces of infection from outside of patch; however we do have a force of infection within patch. The force of infection within a patch may differ from patch to patch. We consider for this

situation a varying population size only, since it is more realistic. The transition rates of the process are given by

$$q_{\mathbf{k},\mathbf{k}+l} = \begin{cases} B, & l = \quad \mathbf{e}_{4r-3}, \\ \beta_i \frac{s_F^{(r)}}{N^{(r)}} i_M^{(r)}, & l = -\mathbf{e}_{4r-3} + \mathbf{e}_{4r-2}, \\ \mu\, s_F^{(r)}, & l = -\mathbf{e}_{4r-3}, \\ \rho_{rj} \frac{U^{(r)}}{N^{(r)}} s_F^{(r)}, & l = -\mathbf{e}_{4r-3} + \mathbf{e}_{4j-3}, \\ \rho_{rj} \frac{U^{(r)}}{N^{(r)}} i_F^{(r)}, & l = -\mathbf{e}_{4r-2} + \mathbf{e}_{4j-2}, \\ (\mu+\gamma)\, i_F^{(r)}, & l = -\mathbf{e}_{4r-2}, \\ B, & l = \quad \mathbf{e}_{4r-1}, \\ \beta_i \frac{s_M^{(r)}}{N^{(r)}} i_F^{(r)}, & l = -\mathbf{e}_{4r-1} + \mathbf{e}_{4r}, \\ \mu\, s_M, & l = -\mathbf{e}_{4r-1}, \\ \rho_{rj} \frac{U^{(r)}}{N^{(r)}} s_M^{(r)}, & l = -\mathbf{e}_{4r-1} + \mathbf{e}_{4j-1}, \\ \rho_{rj} \frac{U^{(r)}}{N^{(r)}} i_M^{(r)}, & l = -\mathbf{e}_{4r} + \mathbf{e}_{4j}, \\ (\mu+\gamma)\, i_M, & l = -\mathbf{e}_{4r}, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

3 Density Dependence and Diffusion Approximation

To study the dynamic behavior of the stochastic models formulated previously, we present some results developed by Kurtz [13, 14]. These results also justify to some extent the use of deterministic models, which is quite common in modelling the epidemic spread, whereas the real situation is in fact a random processes.

Definition 3.1 *A one-parameter family of CTMCs* $(X^{(N)}(t), t \geq 0)$ *with state space* $E \subset \mathbb{Z}^d$ *and transition rates* (q_{ij}) *is called* density dependent *if there exists a continuous function* $f(x,l) : \mathbb{R}^d \times \mathbb{Z}^d \to \mathbb{R}$, *such that*

$$q_{k,k+l} = N f\Big(\frac{k}{N}, l\Big), \quad l \neq 0 \quad \textit{and} \quad k, l \in \mathbb{Z}^d.$$

Suppose $(X(t) = X^{(N)}, t \geq 0)$ is a density dependent process (from now on we drop the superscript N). By rescaling with N we obtain another a CTMC $(X_N(t), t \geq 0)$ called the *density process*. Thus,

$$X_N(t) = \frac{1}{N} X(t).$$

It turns out that under certain mild conditions $(X_N(t))$ converges to a deterministic process that is the solution of a system of first order ODEs that is governed

by the following function F:

$$F(x) = \sum_{l \in \mathbb{Z}^d} l\, f(x, l). \tag{7}$$

Theorem 3.1 (Deterministic Approximation). *Suppose that there exists (1) an open set $E \subset \mathbb{R}^d$ where the function $f(x, l)$ is bounded for each l and (2) the function F is Lipschitz continuous on E. Then, for every trajectory $(x(\tau, x_0),\ \tau \geq 0)$ satisfying the following system of ODEs*

$$\frac{d}{d\tau} x(\tau, x_0) = F(x(\tau, x_0)),$$
$$x(0, x_0) = x_0, \quad x(\tau, x_0) \in E \qquad 0 \leq \tau \leq t,$$

$\lim_{N\to\infty} X_N(0) = x_0$ implies for every $\delta > 0$,

$$\lim_{N\to\infty} \mathbb{P}\Big(\sup_{\tau \leq t} \Big| X_N(\tau) - x(\tau, x_0) \Big| > \delta \Big) = 0, \text{ for every } \quad t \geq 0.$$

The proof is given in [13].
Theorem 3.1 implies that the process $(X_N(t))$ can be approximated to first order by a deterministic process, for large N. If the density process $(X_N(t))$ is initially closed to x_0, it will tend to stay closed to the trajectory $(x(\tau, x_0),\ \tau \leq t)$ in some appropriate time-interval, subject to small random oscillations about the path.
It is even possible to describe the behavior of the random fluctuations of the density process $(X_N(t), t \geq 0)$ around its deterministic approximation. This is done via a diffusion approximation, which is governed by two $d \times d$ matrices $G = G(x) = (g_{ij}(x))$ and $H = H(x) = (h_{ij}(x))$ defined by

$$g_{ij}(x) = \sum_{i=1}^{d} \sum_{j=1}^{d} l_i\, l_j f(x, l), \quad \text{where} \quad l = (l_1, \ldots, l_d) \in \mathbb{Z}^d,$$

and

$$h_{jk}(x) = \frac{\partial F_j(x)}{\partial x_k}.$$

Note that $H(x)$ is simply the Jacobian matrix of $F(x)$.

Theorem 3.2 (Diffusion Approximation). *Suppose $F(x)$ is bounded and Lipschitz continuous on E. Suppose $G(x)$ is also bounded and uniformly continuous on E. Suppose that*

$$\lim_{N\to\infty} \sqrt{N}\Big(X_N(0) - x_0 \Big) = z.$$

Then, as $N \to \infty$, the family of processes $(Z_N(t), t \geq 0)$, defined by

$$Z_N(t) = \sqrt{N}\, \Big(X_N(t) - x(t, x_0) \Big), \quad 0 \leq t \leq s,$$

converges weakly in $D[0,t]$ *to a Gaussian diffusion* $(Z(t), t \geq 0)$ *with initial value* $Z(0) = z$ *and with characteristic function* $\mathbb{E}\, \mathbf{e}^{i\theta \cdot Z(t)} \equiv \psi(t,\theta)$ *that satisfies*

$$\frac{\partial \psi}{\partial t}(t,\theta) = -\frac{1}{2} \sum_{j=1}^{d} \sum_{k=1}^{d} \theta_j\, \theta_k\, g_{jk}(x(t,x_0))\, \psi(t,\theta) + \sum_{j=1}^{d} \sum_{k=1}^{d} \theta_j\; h_{jk}(x(t,x_0))\; \frac{\partial \psi}{\partial \theta_k}(t,\theta). \tag{8}$$

Only in special cases can one obtain an explicit expression for the characteristic function. However, using (8) one can easily determine the *mean vector* and *covariance matrix* of $Z(t)$. In particular, the mean vector of $Z(t)$ is given by

$$\mu = \mathbb{E}Z(t) = M(t)\, z, \tag{9}$$

where $M(t) = \mathbf{e}^{\int_0^t H_s\, ds}$, that is, the unique solution to

$$\frac{dM(t)}{dt} = H(t)\, M(t), \quad \text{with} \quad M(0) = I. \tag{10}$$

On the other hand, the covariance matrix of $Z(t)$, say $\Sigma(t)$, is given by

$$\Sigma(t) = M(t) \left(\int_0^t M(s)^{-1}\, G(x(s,x_0)) \left(M(s)^{-1} \right)^T ds \right) M(t)^T, \tag{11}$$

which is the unique solution to

$$\frac{d\Sigma(t)}{dt} = H(t)\, \Sigma(t) + \Sigma(t)\, H(t)^T + G(x(t,x_0)), \quad \text{with} \quad \Sigma_0 = \Sigma(0) = 0. \tag{12}$$

If $X_N(0)$ and x_0 are chosen to be equal to an equilibrium point x^* of the ODE system in Theorem 3.1, one can be far more precise in specifying the approximating diffusion. Namely, in that case $(Z(t))$ is an Ornstein-Uhlenbeck (OU) process (i.e., a stationary, Gaussian, and Markovian process), with local drift matrix $H(x^*)$ and local covariance matrix $G = G(x^*)$. In particular, $Z(t)$ has a Gaussian/normal distribution with zero mean and a covariance matrix Σ which is given by the solution of (12) with $\frac{d\Sigma_t}{dt} = 0$, see [4]. It follows that $X_N(t)$ has an approximate Gaussian distribution with

$$\mathrm{Var}(X_N(t)) \approx \frac{1}{N} \Sigma, \tag{13}$$

and the mean, obtained by setting $z = \sqrt{N}(X_N(0) - x_0)$, is given by

$$\mathbb{E}X_N(t) \approx x^*. \tag{14}$$

Therefore, we can approximate the equilibrium distribution of the process $(X(t), t \geq 0)$ by a multivariate normal distribution with mean vector $\mu = N\, X_2^*$ and covariance matrix $N\Sigma$. For more general results density dependent processes and theorems we refer to [18] and [3], and the discussion of the diffusion approach and its application in epidemic models can be found, for example, in [6, 7].

4 Analysis

In this section we analyze the stochastic models formulated in Section 2 by using the results in Section 3, and predict their dynamic behavior via their deterministic and diffusion counterparts.

4.1 Closed Single Population

To study the behavior of $(X(t), t \geq 0)$ with the transition rates $q_{\mathbf{k},\mathbf{k}+l}$ as given in (2), we show that it is a density-dependent Markov process, parameterized by the population size N. By scaling with N, we obtain a scaled Markov process $(X_N(t), t \geq 0)$ with $X_N(t) = \frac{1}{N}X(t) = \frac{1}{N}(S_F(t), I_F(t), S_M(t), I_M(t))$. Define the function f as follows

$$f(\mathbf{x}, l) = \begin{cases} \alpha\,(\mu + \delta\,z), & \text{if} \quad l = \ \mathbf{e}_1, \\ \beta\,y_2\,x_1, & \text{if} \quad l = -\mathbf{e}_1 + \mathbf{e}_2, \\ \mu\,x_1, & \text{if} \quad l = -\mathbf{e}_1, \\ (\mu + \gamma)\,x_2, & \text{if} \quad l = -\mathbf{e}_2, \\ (1-\alpha)\,(\mu + \delta\,z), & \text{if} \quad l = \ \mathbf{e}_3, \\ \beta\,x_2\,y_1, & \text{if} \quad l = -\mathbf{e}_3 + \mathbf{e}_4, \\ \mu\,y_1, & \text{if} \quad l = -\mathbf{e}_3, \\ (\mu + \gamma)\,y_2, & \text{if} \quad l = -\mathbf{e}_4, \\ 0, & \text{otherwise}, \end{cases} \tag{15}$$

with $\mathbf{x} = \frac{\mathbf{k}}{N} = (x_1,\, x_2,\, y_1,\, y_2)$ and $z = 1 - (x_1 + x_2 + y_1 + y_2)$. Then, one can check that $q_{\mathbf{k},\mathbf{k}+l} = N\,f(\mathbf{x}, l)$. Therefore, $(X(t), t \geq 0)$ is, by Definition 3.1, a density dependent process. The corresponding function F is derived from (15) and (7):

$$F(\mathbf{x}) = \begin{pmatrix} \alpha\,(\mu + \delta\,z) - \beta\,y_2\,x_1 - \mu\,x_1, \\ \beta\,y_2\,x_1 - (\mu + \gamma)\,x_2 \\ (1-\alpha)\,(\mu + \delta\,z) - \beta\,x_2\,y_1 - \mu\,y_1 \\ \beta\,x_2\,y_1 - (\mu + \gamma)\,y_2 \end{pmatrix}. \tag{16}$$

The function F is Lipschitz continuous. So, the dynamic behavior of the process $(X_N(t), t \geq 0)$, see Theorem 3.1, can be approximated by a system of first order ODEs

$$\mathbf{x}'(t) = F(\mathbf{x}), \tag{17}$$

as $N \longrightarrow \infty$.

Equilibria and Their Stability

From now on we assume for simplicity that $\alpha = \frac{1}{2}$ (i.e., the ratio of females and males entering the population is 50:50). Solving $F(X) = 0$ in (16) gives three equilibrium points, two of which fall in the positive quadrant: the *disease-free* equilibrium and the *positive endemic* equilibrium. Let $X^* = (x_1^*, x_2^*, y_1^*, y_2^*)$ denote a generic equilibrium of the system (17).

Disease-free Equilibrium

The disease-free equilibrium is given by

$$X_1^* = (x_1^* = \frac{1}{2},\, x_2^* = 0,\, y_1^* = \frac{1}{2},\, y_2^* = 0). \tag{18}$$

In the absence of the disease ($x_2 = y_2 = 0$), the fraction of susceptibles of both females and males will reach a constant number: $x_1 = x_1^* = \frac{1}{2}$ and $y_1 = y_1^* = \frac{1}{2}$, respectively. We are interested in whether in the early epidemic spread (after a few infected people are present) the number of infectives will grow or die out. The following result sheds some light onto this. Here, the basic quantity R_0 serves the same role as the *basic reproduction rate* in epidemiology.

Theorem 4.1 *Let $R_0 = \frac{\beta}{2(\mu+\gamma)}$. The disease-free equilibrium X_1^* in (18) is locally asymptotically stable if $R_0 < 1$ and unstable if $R_0 > 1$.*

Proof. The Jacobian matrix of (16) is given by

$$H(\mathbf{x}) = \begin{pmatrix} -\frac{\delta}{2} - \beta\, y_2 - \mu & -\frac{\delta}{2} & -\frac{\delta}{2} & -\frac{\delta}{2} - \beta\, x_1 \\ \beta\, y_2 & -(\mu+\gamma) & 0 & \beta\, x_1 \\ -\frac{\delta}{2} & -\frac{\delta}{2} - \beta\, y_1 & -\frac{\delta}{2} - \beta\, x_2 - \mu & -\frac{\delta}{2} \\ 0 & \beta\, y_1 & \beta\, x_2 & -(\mu+\gamma) \end{pmatrix}. \tag{19}$$

Evaluating (19) at X_1^* yields

$$H(X_1^*) = \begin{pmatrix} -\frac{\delta}{2} - \mu & -\frac{\delta}{2} & -\frac{\delta}{2} & -\frac{1}{2}(\delta+\beta) \\ 0 & -(\mu+\gamma) & 0 & \frac{\beta}{2} \\ -\frac{\delta}{2} & -\frac{1}{2}(\delta+\beta) & -\frac{\delta}{2} - \mu & -\frac{\delta}{2} \\ 0 & \frac{\beta}{2} & 0 & -(\mu+\gamma) \end{pmatrix}. \tag{20}$$

If the real parts of all the eigenvalues of this matrix are negative, then the disease-free steady-state is locally asymptotically stable. The matrix (20) has four eigenvalues

$$r_1 = -\mu, \quad r_2 = -(\mu+\delta), \quad r_3 = -(\mu+\gamma) - \frac{\beta}{2}, \quad r_4 = -(\mu+\gamma) + \frac{\beta}{2}. \tag{21}$$

Therefore, the stability of this equilibrium is determined by the last eigenvalue r_4, since the other eigenvalues are always negative for the non-negative parameters $\beta, \gamma, \mu, \delta$. Thus, the disease-free equilibrium is *stable* if and only if $r_4 = -(\mu+\gamma) + \frac{\beta}{2} < 0$ ($R_0 < 1$) and it is *unstable* if and only if $r_4 = -(\mu+\gamma) + \frac{\beta}{2} > 0$ ($R_0 > 1$). □

Positive Endemic Equilibrium

The endemic equilibrium is given by

$$X_2^* = (x_1^* = \rho,\, x_2^* = \eta,\, y_1^* = \rho,\, y_2^* = \eta), \tag{22}$$

where $\eta = (1-2\rho)\Delta = (1-\frac{1}{R_0})\Delta$, with $\rho = \frac{\mu+\gamma}{\beta} = \frac{1}{2R_0}$ and $\Delta = \frac{\mu+\delta}{2(\mu+\gamma+\delta)}$. It is clear from (22) that the system (17) has a positive-endemic equilibrium if and only if $(1-2\rho) > 0$ (or equivalently $R_0 > 1$). The Jacobian matrix for the positive-endemic equilibrium is

$$H(X_2^*) = \begin{pmatrix} -\frac{\delta}{2} - \beta\,\eta - \mu & -\beta\,\eta & -\frac{\delta}{2} & 0 \\ -\frac{\delta}{2} & -\beta\rho & 0 & \frac{\beta}{2} \\ -\frac{\delta}{2} & 0 & -\frac{\delta}{2} - \beta\,\eta - \mu & \beta\,\eta \\ 0 & \frac{\beta}{2} & 0 & -\beta\rho \end{pmatrix}. \tag{23}$$

This matrix has four eigenvalues

$$r_1 = \frac{1}{2}\Big(B_1 + \sqrt{\Theta_1}\Big), \quad r_3 = \frac{1}{2}\Big(B_2 + \sqrt{\Theta_2}\Big),$$

$$r_2 = \frac{1}{2}\Big(B_1 - \sqrt{\Theta_1}\Big), \quad r_4 = \frac{1}{2}\Big(B_2 - \sqrt{\Theta_2}\Big),$$

where

$$\begin{aligned} B_1 &= -\beta\,\eta - \mu - 2\,\beta\,\rho, \\ B_2 &= -\beta\,\eta - \mu - \delta, \\ \Theta_1 &= 4\,\beta^2\Delta^2\rho^2 - 4\,\beta^2\Delta^2\rho - 4\,\beta\,\Delta\,\rho\,\mu + \beta^2\Delta^2 + 2\,\beta\,\Delta\,\mu + \mu^2 - 4\,\mu\,\beta\,\rho + 4\,\beta^2\rho^2, \\ \Theta_2 &= 4\,\beta^2\Delta^2\rho^2 - 4\,\beta\,\Delta\,\rho\,\mu - 4\,\beta^2\Delta^2\rho + 4\,\delta\,\beta\,\Delta\,\rho + \mu^2 + 2\,\beta\,\Delta\,\mu + 2\,\delta\,\mu + \beta^2\Delta^2 \\ &\quad -2\,\delta\,\beta\,\Delta + \delta^2 - 4\,\beta^2\Delta\,\rho + 8\,\beta^2\Delta\,\rho^2. \end{aligned}$$

If $R_0 > 1$ (or $2\rho < 1$), it follows that $B_1, B_2 < 0$. Therefore, $\mathrm{Re}(r_2)$ and $\mathrm{Re}(r_4)$ are always negative. We need to show that for some $\beta, \mu, \gamma, \delta > 0$, $\mathrm{Re}(r_1)$ and $\mathrm{Re}(r_3)$ are also negative. If $\Theta_1 \le 0$ and $\Theta_2 \le 0$, $\mathrm{Re}(r_1) = B_1 < 0$ and $\mathrm{Re}(r_3) = B_2 < 0$. Now, suppose that $\Theta_1 > 0$ and $\Theta_2 > 0$. Let $C_1 = -B_1 > 0$ and let $C_2 = -B_2$. Then, we obtain

$$\Theta_1 - C_1^2 = -8\,\mu\,\beta\,\rho + 4\,\beta^2\Delta\,\rho\,(2\rho - 1) < 0. \tag{24}$$

and

$$\Theta_2 - C_2^2 = 4\,\delta\,\beta\,\Delta\,(2\,\rho - 1) + 4\,\beta^2\Delta\,\rho\,(2\,\rho - 1) < 0. \tag{25}$$

¿From (24), we have $\Theta_1 - C_1^2 < 0 \iff 0 < \Theta_1 < C_1^2 \iff 0 < \sqrt{\Theta_1} < C_1$. Thus, $-C_1 + \sqrt{\Theta_1} = B_1 + \sqrt{\Theta_1} < 0$, which implies $\mathrm{Re}(r_1) < 0$. From (25), we have $\Theta_2 - C_2^2 < 0 \iff 0 < \Theta_2 < C_2^2 \iff 0 < \sqrt{\Theta_2} < C_2$. Thus, $-C_2 + \sqrt{\Theta_2} = B_2 + \sqrt{\Theta_2} < 0$, which implies $\mathrm{Re}(r_3) < 0$. We summarize these findings in the following theorem.

Theorem 4.2 *The endemic equilibrium X_2^* exists iff $R_0 > 1$, and it is locally asymptotically stable.*

Diffusion Approximation

The approximating OU process $(Z(t), t \geq 0)$ around the equilibrium point X_2^* has local drift matrix $H(X_2^*)$ in (23), and local covariance matrix $G(X_2^*)$, defined in Theorem 3.2, as follows

$$G(X_2^*) = \begin{pmatrix} g_{11} & g_{12} & 0 & 0 \\ g_{21} & g_{22} & 0 & 0 \\ 0 & 0 & g_{33} & g_{34} \\ 0 & 0 & g_{43} & g_{44} \end{pmatrix}, \tag{26}$$

where

$$\begin{aligned} g_{11} &= \frac{1}{2}(\mu + \delta A) + \beta x_1^* y_2^* + \mu x_1^*, \\ g_{12} &= g_{21} = \beta x_1^* y_2^*, \\ g_{22} &= \beta x_1^* y_2^* + (\mu + \gamma) x_2^*, \\ g_{33} &= \frac{1}{2}(\mu + \delta A) + \beta x_2^* y_1^* + \mu y_1^*, \\ g_{34} &= g_{43} = \beta x_2^* y_1^*, \\ g_{44} &= \beta x_2^* y_1^* + (\mu + \gamma) y_2^*. \end{aligned}$$

Therefore, we can approximate the equilibrium distribution of the process $(X(t), t \geq 0)$ by a multivariate normal distribution, see (14) and (13), with mean $\mu = N X_2^*$ and covariance matrix $N\Sigma$.

4.2 Open Single Population

To derive a deterministic analogue, as in the previous model, we show that the process $(X(t), t \geq 0)$ with the transition rates $q_{\mathbf{k},\mathbf{k}+l}$ as given in (3) is a density-dependent Markov process parameterized by $V = \frac{2B}{\mu}$. We will see shortly that this constant corresponds to the total population size in the disease-free equilibrium. Define $\mathbf{x} = \frac{\mathbf{k}}{V} = (x_1(t), x_2(t), y_1(t), y_2(t))$. Then, we can write

$$q_{\mathbf{k},\mathbf{k}+l} = V f(\mathbf{x}, l),$$

where $f(\mathbf{x}, l)$ is given by

$$f(\mathbf{x}, l) = \begin{cases} \frac{\mu}{2}, & l = \mathbf{e}_1, \\ \beta \frac{y_2}{v} x_1, & l = -\mathbf{e}_1 + \mathbf{e}_2, \\ \mu x_1, & l = -\mathbf{e}_1, \\ (\mu + \gamma) x_2, & l = -\mathbf{e}_2, \\ \frac{\mu}{2}, & l = \mathbf{e}_3, \\ \beta \frac{x_2}{v} y_1, & l = -\mathbf{e}_3 + \mathbf{e}_4, \\ \mu y_1, & l = -\mathbf{e}_3, \\ (\mu + \gamma) y_2, & l = -\mathbf{e}_4, \\ 0, & \text{otherwise}, \end{cases} \tag{27}$$

with $v = x_1 + x_2 + y_1 + y_2$. Therefore, the process $(X(t), t \geq 0))$ is a density dependent Markov process. As the parameter $V \longrightarrow \infty$, by Theorem 3.1, the dynamic behavior of the scaled Markov process $(X_V(t), t \geq 0)$ can be approximated by a system of first order ODEs $\mathbf{x}' = F(\mathbf{x})$, with $F(\mathbf{x})$ defined as follows:

$$F(x) = \begin{pmatrix} \frac{\mu}{2} - \beta \frac{y_2}{v} x_1 - \mu x_1 \\ \beta \frac{y_2}{v} x_1 - (\mu + \gamma) x_2 \\ \frac{\mu}{2} - \beta \frac{x_2}{v} y_1 - \mu y_1 \\ \beta \frac{x_2}{v} y_1 - (\mu + \gamma) y_2 \end{pmatrix}. \tag{28}$$

Again, we examine the dynamic behavior of the deterministic model around its equilibrium points.

Equilibrium Points and Analysis

This system also has two equilibrium points: the disease-free and the endemic equilibrium. As in the previous model, the disease-free equilibrium is

$$X_1^* = \Big(x_1^* = \frac{1}{2},\, x_2^* = 0,\, y_1^* = \frac{1}{2},\, y_2^* = 0\Big). \tag{29}$$

The Jacobian matrix of (28) is in the form

$$H(X_1^*) = \begin{pmatrix} \Lambda - \mu & \Lambda - \Theta & \Lambda & \Psi + \Lambda \\ -\Lambda & -\Lambda + \Theta - \mu - \gamma & \Psi - \Lambda & -\Lambda \\ \Lambda & \Lambda - \Theta & -\Psi + \Lambda - \mu & \Lambda \\ -\Lambda & -\Lambda + \Theta & \Psi - \Lambda & -\Lambda - \mu - \gamma \end{pmatrix}, \tag{30}$$

with $\Lambda = \frac{\beta\, y_1\, x_2}{v^2}$, $\Theta = \frac{\beta\, y_1}{v}$, and $\Psi = \frac{\beta\, x_2}{v}$. Evaluated at the disease-free equilibrium (29), we obtain

$$H(X_1^*) = \begin{pmatrix} -\mu & -\frac{\beta}{2} & 0 & 0 \\ 0 & \frac{\beta}{2} - \mu - \gamma & 0 & 0 \\ 0 & -\frac{\beta}{2} & -\mu & 0 \\ 0 & \frac{\beta}{2} & 0 & -\mu - \gamma \end{pmatrix}. \tag{31}$$

This matrix (31) has four eigenvalues (two of which are equal)

$$r_1 = r_2 = -\mu, \quad r_3 = \frac{1}{2}\beta - \mu - \gamma, \quad \text{and} \quad r_4 = -\mu - \gamma. \tag{32}$$

Thus, the stability of this equilibrium is determined by r_3, since the other eigenvalues are always negative for the non-negative parameters $\beta, \gamma, \mu, \delta$. Hence, the disease-free equilibrium is *stable* if and only if $\frac{1}{2}\beta - \mu - \gamma < 0$ $(R_0 = \frac{\beta}{2(\mu+\gamma)} < 1)$ and it is *unstable* if and only if $r_3 = \frac{1}{2}\beta - \mu - \gamma > 0$ $(R_0 > 1)$.
Next, we analyze the endemic equilibrium. The endemic equilibrium is of the form

$$X_2^* = \Big(x_1^* = \xi_1,\, x_2^* = \xi_2,\, y_1^* = \xi_1,\, y_2^* = \xi_2\Big), \tag{33}$$

where $\xi_1 = \frac{\mu}{\beta - 2\,\gamma}$, $\xi_2 = \frac{\mu\,\beta\,(1-2\rho)}{2\,\rho\,\beta\,(\beta-2\,\gamma)}$ and $\rho = \frac{\mu+\gamma}{\beta}$. So, a positive endemic equilibrium occurs if $(1-2\,\rho) > 0$, that is, $R_0 > 1$. The Jacobian matrix evaluated around this positive endemic equilibrium X_2^* has four eigenvalues:

$$r_1 = -\mu, \quad r_2 = -\mu - \gamma, \quad r_3 = \frac{1}{4\,\beta}\left(B + \sqrt{\Theta}\right), \quad r_4 = \frac{1}{4\,\beta}\left(B - \sqrt{\Theta}\right), \qquad (34)$$

with

$$\begin{aligned} B &= \beta\,(2\,\gamma - \beta), \\ \Theta &= 36\,\gamma^2\,\beta^2 - 12\,\beta^3\,\gamma + \beta^4 - 64\,\mu\,\gamma^2\beta - 32\,\mu^2\,\gamma\,\beta + 48\,\mu\,\gamma\,\beta^2 \\ &\quad -32\,\gamma^3\beta + 16\,\mu^2\beta^2 - 8\,\mu\,\beta^3. \end{aligned}$$

Since $R_0 > 1 \Leftrightarrow \beta > 2\,(\mu+\gamma) > 2\,\gamma$, we have $B < 0$. Let $C = -B > 0$, then

$$\begin{aligned} \Theta - C^2 &= 32\,\gamma^2\beta^2 - 8\,\beta^3\gamma - 64\,\mu\,\gamma^2\beta - 32\,\mu^2\gamma\,\beta \\ &\quad +48\,\mu\,\gamma\,\beta^2 - 32\,\gamma^3\beta + 16\,\mu^2\beta^2 - 8\,\mu\,\beta^3 \\ &= 8\,\mu\,\beta^2\,(2\,\gamma + 2\,\mu - \beta) - 8\,\gamma\,\beta\,(2\,\gamma + 2\,\mu - \beta)^2\,. \end{aligned} \qquad (35)$$

Since $R_0 > 1$, that is, $2(\mu+\gamma) - \beta < 0$ (for a positive-endemic equilibrium), where $\Theta - C^2 < 0$. Therefore, $-C + \sqrt{\Theta} = B + \sqrt{\Theta} < 0$, which implies $\mathrm{Re}(r_3) < 0$. We summarize these results in the following theorem.

Theorem 4.3 *The disease-free equilibrium X_1^* (29) is locally asymptotically stable if $R_0 < 1$ and unstable if $R_0 > 1$. A stable positive endemic equilibrium X_2^* (5) exists iff $R_0 > 1$.*

Thus both the open and closed population models, under the assumption of both constant and variable population size, have the same stability conditions: the disease-free equilibrium is stable if $R_0 < 1$, otherwise, it is unstable, and the endemic equilibrium occurs when $R_0 > 1$ and it is stable. The differences are only in the size of the endemic equilibrium and the eigenvalues of the corresponding Jacobian matrix.

4.3 Multiple Patch Models

To study the dynamic behaviour of the multiple patch models presented in Section 2, we also apply the deterministic and diffusion approach as in the case of a single population. We construct a density Markov process by scaling with a certain parameter, and derive a deterministic model to approximate the scaled process. The deterministic analogues of those two multiple patch models are given next.

Model with a Force of Infection

For the multiple patch model with constant population size; if all patches have equal size N, we can use this parameter as a scale factor for all random variables in the process. However, for the case where the patches have unequal size, all

random variables are scaled by the total population size $N = \sum_{r=1}^{K} N^{(r)}$ and we define an extra constant $c^{(r)} = \frac{N}{N^{(r)}}$ for each r. Thus, one can obtain $q_{k,k+l} = N\, f(\mathbf{x}, l), \quad r = 1, \ldots, K$ where f is given as follows

$$f(\mathbf{x},l) = \begin{cases} \alpha\left(\frac{\mu}{c^{(r)}} + \delta\, z^{(r)}\right), & l = \ \mathbf{e}_{4r-3}, \\ \sum_{j=1}^{K} \beta_{rj}\ c^{(r)}\ x_1^{(r)}\, y_2^{(j)}, & l = -\mathbf{e}_{4r-3} + \mathbf{e}_{4r-2}, \\ \mu\, x_1^{(r)}, & l = -\mathbf{e}_{4r-3}, \\ (\mu+\gamma)\, x_2^{(r)}, & l = -\mathbf{e}_{4r-2}, \\ (1-\alpha)\left(\frac{\mu}{c^{(r)}} + \delta\, z^{(r)}\right), & l = \ \mathbf{e}_{4r-1}, \\ \sum_{j=1}^{K} \beta_{rj}\ c^{(r)}\ y_1^{(r)}\, x_2^{(j)}, & l = -\mathbf{e}_{4r-1} + \mathbf{e}_{4r}, \\ \mu\ s_M^{(r)}, & l = -\mathbf{e}_{4r-1}, \\ (\mu+\gamma)\ y_2^{(r)}, & l = -\mathbf{e}_{4r}, \\ 0, & \text{otherwise}, \end{cases} \tag{36}$$

with $z^{(r)} = 1 - \left(x_1^{(r)} + x_2^{(r)} + y_1^{(r)} + y_2^{(r)}\right)$.

As $N \longrightarrow \infty$, we can apply again the results of Kurtz and derive the following deterministic analogue, for the process with transition rates (4):

$$\begin{aligned} \frac{dx_1^{(r)}}{dt} &= \alpha\left(\frac{\mu}{c^{(r)}} + \delta\, z^{(r)}\right) - \sum_{j=1}^{K} \beta_{rj}\ c^{(r)}\ x_1^{(r)}\, y_2^{(j)} - \mu\, x_1^{(r)}, \\ \frac{dx_2^{(r)}}{dt} &= \sum_{j=1}^{K} \beta_{rj}\, c^{(r)}\ x_1^{(r)}\, y_2^{(j)} - (\mu+\gamma)\, x_2^{(r)}, \\ \frac{dy_1^{(r)}}{dt} &= (1-\alpha)\left(\frac{\mu}{c^{(r)}} + \delta\, z^{(r)}\right) - \sum_{j=1}^{K} \beta_{rj}\ c^{(r)}\ y_1^{(r)}\, x_2^{(j)} - \mu\, y_1^{(r)}, \\ \frac{dy_2^{(r)}}{dt} &= \sum_{j=1}^{K} \beta_{rj}\ c^{(r)}\ y_1^{(r)}\, x_2^{(j)} - (\mu+\gamma)\ y_2^{(r)}, \end{aligned}$$

with $z^{(r)} = \frac{1}{c^{(r)}} - \left(x_1^{(r)} + x_2^{(r)} + y_1^{(r)} + y_2^{(r)}\right)$.

For the case of varying population case, the ODEs version of the stochastic model is derived by parameterizing each random variables of the process $(X(t), t \geq 0)$, with the transition rates (5), with a parameter $V = \frac{2B}{\mu}$ (as in the single varying population model). The deterministic system is given by the following equations:

$$\begin{aligned}
\frac{dx_1^{(r)}}{dt} &= \frac{\mu}{2} - \sum_{j=1}^{K} \beta_{rj} \frac{x_1^{(r)}}{n^{(r)}} y_2^{(j)} - \mu\, x_1^{(r)}, \\
\frac{dx_2^{(r)}}{dt} &= \sum_{j=1}^{K} \beta_{rj} \frac{x_1^{(r)}}{n^{(r)}} y_2^{(j)} - (\mu+\gamma)\, x_2^{(r)}, \\
\frac{dy_1^{(r)}}{dt} &= \frac{\mu}{2} - \sum_{j=1}^{K} \beta_{rj} \frac{y_1^{(r)}}{n^{(r)}} x_2^{(j)} - \mu\, y_1^{(r)}, \\
\frac{dy_2^{(r)}}{dt} &= \sum_{j=1}^{K} \beta_{rj} \frac{y_1^{(r)}}{n^{(r)}} x_2^{(j)} - (\mu+\gamma)\, y_2^{(r)},
\end{aligned}$$

with $n^{(r)} = x_1^{(r)} + x_2^{(r)} + y_1^{(r)} + y_2^{(r)}$.

Model with Actual Mobility

As explained previously that for the model with actual mobility we only consider the case under a varying population size. The ODE analogue of this model is given, after scaling the process $(X(t), t \geq 0)$ (with transition rates in (6)) with $V = \frac{2B}{\mu}$, by the following system:

$$\begin{aligned}
\frac{dx_1^{(r)}}{dt} &= \frac{\mu}{2} - \beta_r \frac{x_1^{(r)}}{n^{(r)}} y_2^{(r)} - \mu\, x_1^{(r)} + \sum_{j=1}^{K} \rho_{rj} \frac{u^{(j)}}{n^{(j)}} x_1^{(j)} - \frac{u^{(r)}}{n^{(r)}} x_1^{(r)}, \\
\frac{dx_2^{(r)}}{dt} &= \beta_r \frac{x_1^{(r)}}{n^{(r)}} y_2^{(r)} - (\mu+\gamma)\, x_2^{(r)} + \sum_{j=1}^{K} \rho_{rj} \frac{u^{(j)}}{n^{(j)}} x_2^{(j)} - \frac{u^{(r)}}{n^{(r)}} x_2^{(r)}, \\
\frac{dy_1^{(r)}}{dt} &= \frac{\mu}{2} - \beta_r \frac{y_1^{(r)}}{n^{(r)}} x_2^{(r)} - \mu\, y_1^{(r)} + \sum_{j=1}^{K} \rho_{rj} \frac{u^{(j)}}{n^{(j)}} y_1^{(j)} - \frac{u^{(r)}}{n^{(r)}} y_1^{(r)}, \\
\frac{dy_2^{(r)}}{dt} &= \beta_r \frac{y_1^{(r)}}{n^{(r)}} x_2^{(r)} - (\mu+\gamma)\, y_2^{(r)} + \sum_{j=1}^{K} \rho_{rj} \frac{u^{(j)}}{n^{(j)}} y_2^{(j)} - \frac{u^{(r)}}{n^{(r)}} y_2^{(r)},
\end{aligned}$$

where $n^{(r)} = x_1^{(r)} + x_2^{(r)} + y_1^{(r)} + y_2^{(r)}$.
Here, we have not proved analytically the existence and the stability of their equilibrium points. However, we consider the endemic equilibria numerically and use them to derive the diffusion counterparts.

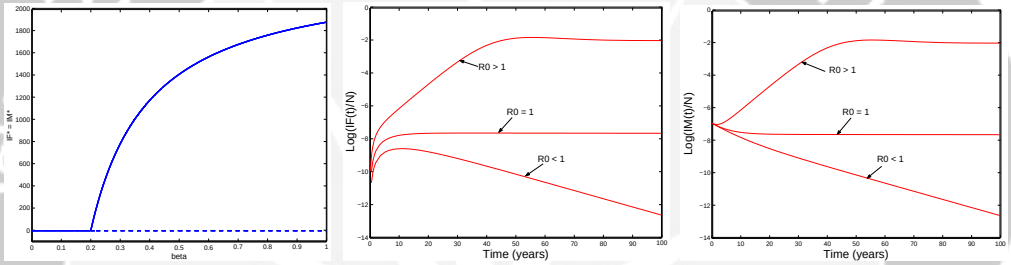
5 Numerical Experiments and Discussion

In this section we illustrate the behavior of the various population models and their deterministic and diffusion approximations via a number of numerical experiments.

The following parameters are the same in each experiment: The natural death rate is $\mu = 0.02$ (which corresponds to the life expectancy 50 years), the death rate due to AIDS is $\delta = 0.05$ (which means a life expectancy for AIDS people of only 20 years), and the removal rate is $\gamma = 0.08$ (which corresponds to a 12 year infectious period of HIV before AIDS sets in). We always assume $\alpha = \frac{1}{2}$, which implies a $50 : 50$ ratio of females and males in the recruitment of new susceptibles. The other parameter settings are explained in each individual experiment.

5.1 Models for a Single Population

In these experiments the important parameter is β, since it determines the stability of the disease-free equilibrium (see Section 4 for the threshold condition assuming the parameters μ and γ are fixed). The numerical results in Figure 4, for the deterministic model with a constant single population, illustrate how crucial the parameter β is.



(a) Bifurcation diagram. The *solid* and *dashed* lines denote *stable* and *unstable* equilibria, respectively.

(b) The dynamics of the number of infected females.

(c) The dynamics of the number of infected males.

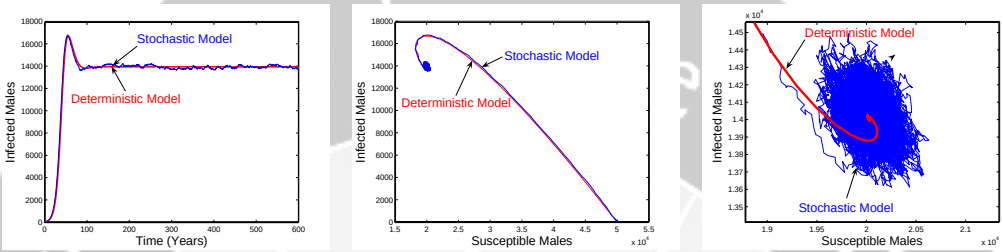
Figure 4: (a). The stability of the disease-free equilibrium, and the birth of the endemic equilibrium as the parameter β varies. (b) and (c) illustrate how the disease-free equilibrium of the deterministic model behaves for three different values of β, ($0.5(R_0 > 1)$; $0.2(R_0 = 1)$; $0.1(R_0 < 1)$).

It can be seen from the two logarithmic plots in Figure 4(b) and 4(c) that when R_0 is below the threshold ($R_0 < 1$) the proportion of infectives of both females and males, after a few infectives are introduced in the population, returns to no infection, but it grows away from the disease-free equilibrium if R_0 is above the threshold ($R_0 > 1$).
The value of the parameter β can be set by using the formula in Remark 2.1. For the purpose of our numerical study, we choose the parameter $\beta = 0.5$ so that $R_0 > 1$ which results in a positive endemic equilibrium. We then look at how the stochastic processes converge to their deterministic and diffusion approximation around the equilibrium.

Model with a Closed Single Population

For the numerical experiments, we apply the parameter settings above and use the following initial values: $50,000$ susceptible females and $50,000$ susceptible males, 100 infected males, no infected females, and no AIDS cases. So, the total population size, $N = 100100$.

Figure 5 describes the dynamic behaviour in the male subpopulation (similar results hold for the female subpopulation).



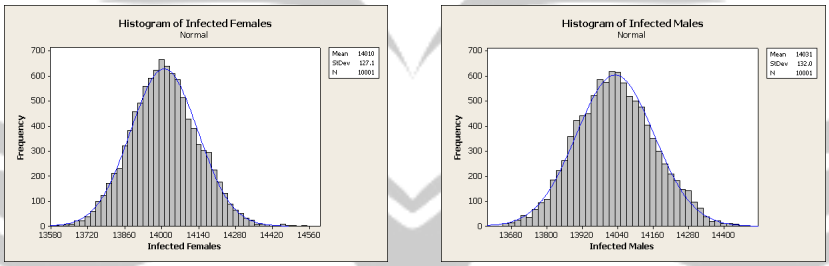
(a) The number of male infectives versus time.

(b) The dynamic behavior in the male subpopulation.

(c) The graph of (b) around the endemic equilibrium.

Figure 5: Stochastic and Deterministic Model for Male Subpopulation.

We can see that the stochastic process remains close to the trajectory of its deterministic analogue during a finite time interval. We should note that the process will eventually leave the trajectory and be absorbed in the disease-free equilibrium. The following histograms describe the empirical distribution of the number of infectives based on a simulation of the CTMC with transition rates (2) around the equilibrium point of the deterministic process.



(a) The distribution of the number of the infected females.

(b) The distribution of the number of the infected males.

Figure 6: Equilibrium distributions around the endemic equilibrium.

These numerical results illustrate that the "stationary" distribution of the process can be approximated by a normal distribution. The empirical means and standard deviations for the number of infected females (males) are 14010 and 127.1 (14031 and 132), respectively. ¿From the diffusion approximation, the exact form of

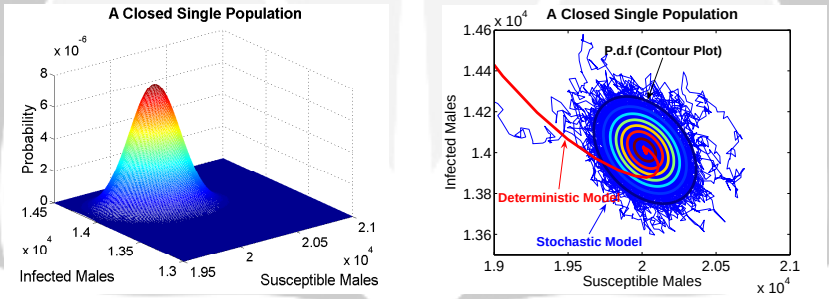
the mean $\mu = N X_2^*$ and covariance matrix $N\Sigma$ of $X(t)$ can be calculated from Equations (9) and (11), which numerically can be found to be

$$\mu = (20020, 14014, 20020, 14014),$$

and

$$N\Sigma = N \begin{pmatrix} 0.2933721451 & -0.08025078370 & 0.04948499776 & -0.1197492163 \\ -0.08025078370 & 0.1685788924 & -0.1197492163 & 0.08475444096 \\ 0.04948499776 & -0.1197492163 & 0.2933721451 & -0.08025078370 \\ -0.1197492163 & 0.08475444096 & -0.08025078370 & 0.1685788924 \end{pmatrix}.$$

The means and standard deviations obtained from the diffusion approximation for the number of infected females (males), which are 14014 and 129.9 for both females and males, are close to the experiment results.
To illustrate the accuracy of this diffusion approximation, we plot the dynamic behavior of the male subpopulation around the equilibrium point, together with its diffusion analogue, see Figure 7.



(a) The p.d.f of male population corresponding to the diffusion approximation.

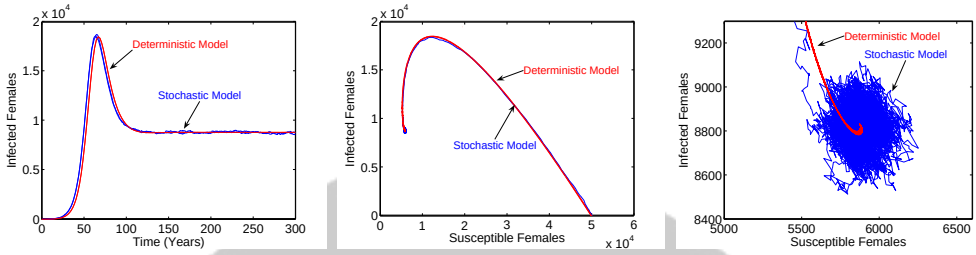
(b) The stochastic process with its deterministic limit, and the contour lines corresponding to the p.d.f in (a).

Figure 7: The stochastic model and its deterministic and diffusion analogues for a constant population size for male subpopulation.

We see that the equilibrium distribution of infectives around the endemic equilibrium is closely approximated by a two-dimensional Gaussian distribution derived from the diffusion process.

Model with an Open Single Population

In this experiments we use the same parameters as in the model with a single closed (constant) population. In addition, we set $B = 1000$. The following figure illustrates that the stochastic process for an open single population converges to its deterministic counterpart.



(a) The number of female infectives versus time.

(b) The dynamic behavior in the female subpopulation.

(c) The graph of (b) around the endemic equilibrium.

Figure 8: Stochastic and Deterministic Model for the Female Subpopulation.

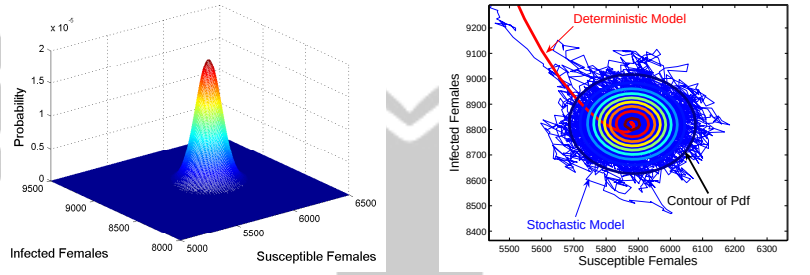
From the diffusion approximation, the mean and covariance of the process $(X(t), t \geq 0)$ can be approximated by using Equations (14) and (13). The Gaussian distribution derived from the diffusion process (with $V = 100,000$) has the mean vector

$$\mu = V\, X_2^* = (5882, 8823, 5882, 8823),$$

and the covariance matrix $V\,\Sigma$, with

$$\Sigma = \begin{pmatrix} 0.08565685893 & -0.001747475767 & 0.01428207858 & -0.02308460238 \\ -0.001747475767 & 0.08496559914 & -0.02308460237 & 0.02484510713 \\ 0.01428207858 & -0.02308460237 & 0.08565685893 & -0.001747475777 \\ -0.02308460238 & 0.02484510713 & -0.001747475777 & 0.08496559913 \end{pmatrix}.$$

The following figures illustrate the accuracy of the diffusion approach in approximating the distribution of susceptibles and infectives around the equilibrium point.



(a) The p.d.f of female population corresponding to the diffusion approximation.

(b) The stochastic process with its deterministic limit, and the contour lines corresponding to the p.d.f in (a).

Figure 9: The stochastic model and its deterministic and diffusion analogue for varying population size for the female subpopulation.

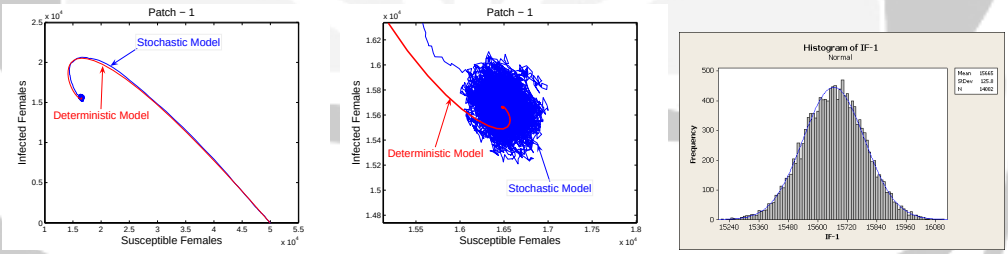
5.2 Multiple Patch Models

In these numerical experiments, we carry out the simulations with $M = 10$ patches, in all multiple patch models. We set the initial values to 50,000 susceptibles females and 50,000 susceptibles males for each patch. The infected people — set to be 100 infected males — are assumed to be initially concentrated only in patch 1. Thus, initially no infected individuals are present in other patches. All parameters (except β) have the same values as specified previously.

Models with Force of Infection

In this model, we assumed that within-patch mixing is stronger (and often considerably stronger) than between-patch mixing, and hence that the between-patch transmission parameters β_{rj} (for $r \neq j$) are small compared to the within-patch transmission parameters β_r (or β_{rr}), see [15]. In addition, the force of infection from the patch where the infection is initially concentrated to the other patches is assumed to be stronger than the forces among other patches. We might consider this patch, for example, as a big city where people from other small cities come visit more often. With these assumptions, we set the values of β as follows: $\beta_r = 0.5$, $\beta_{1j} = 0.5$, $j = 1, \ldots, 10$, and $\beta_{rj} = 0.01$, $r = 2, \ldots, 10$, $j = 1, \ldots, 10$.

The following numerical results (see Figure 10) describe the dynamic behavior of the process and its deterministic analogue in patch 1 for the female subpopulation. This behaviour is similar to that in other patches, for each subpopulation.



(a) The number of female infectives versus time.

(b) The dynamic behavior of in the female subpopulation.

(c) The graph of (b) around the endemic equilibrium.

Figure 10: Stochastic and Deterministic Model.

We conclude that the stochastic process in the multiple patch model, at least from the numerical evidence, converges to its deterministic version.

To obtain the diffusion approximation, we evaluate the equilibrium points by solving the deterministic counterparts numerically. Then, we determine numerically the mean vectors and the covariance matrices around these equilibrium for their multivariate Gaussian distribution. These results can be seen in Table 1 and Table 2 for the case of constant population sizes and varying population sizes, respectively.

Table 1: Means and standard deviations of the Diffusion approximation; for the closed population model.

Patches	**Infected Females**		**Infected Males**	
	μ	σ	μ	σ
1	15666	126	15666	129
2-10	18667	129	18667	129

Table 2: Means and standard deviations of the diffusion approximation; for the open population model.

Patches	**Infected Females**		**Infected Males**	
	μ	σ	μ	σ
1	9080	91.0	9080	91.0
2-10	9550	93.4	9550	93.4

These calculation are in close agreement with the empirical means and standard deviations obtained by simulating the stochastic process and collecting data after equilibrium has been reached. We summarize in Table 3 and Table 4 the sample means and sample standard deviations obtained from a Monte Carlo simulation for a closed and open multiple population, respectively, with the force of infection.

Table 3: Sample means and standard deviations for the model of a constant population size.

Patches	**Infected Females**		**Infected Males**	
	$\tilde{\mu}$	$\tilde{\sigma}$	$\tilde{\mu}$	$\tilde{\sigma}$
1	15665	126	15660	126
2	18675	129	18661	127
3	18668	126	18665	126
4	18669	126	18659	133
5	18670	131	18661	130
6	18658	128	18660	139
7	18661	122	18665	132
8	18672	131	18673	132
9	18673	125	18659	132
10	18668	134	18669	128

Table 4: Sample means and standard deviations for the model of a varying population size.

Patches	**Infected Females**		**Infected Males**	
	$\tilde{\mu}$	$\tilde{\sigma}$	$\tilde{\mu}$	$\tilde{\sigma}$
1	9085.1	91.2	9076.0	91.2
2	9559.7	90.7	9554.6	94.3
3	9554.1	96.2	9554.0	89.0
4	9552.2	91.7	9550.1	95.6
5	9552.6	90.1	9547.3	91.8
6	9561.1	91.1	9547.4	91.9
7	9554.3	92.2	9551.2	95.3
8	9551.6	89.4	9545.9	93.8
9	9546.1	94.8	9544.1	93.6
10	9553.1	93.2	9553.8	86.9

Model with Actual Mobility

For the model with actual mobility, we assume that the forces of infection within a patch are the same for all patches, which is set at $\beta_i = 0.5$. The initial numbers of susceptibles and infective in each patch are as in the model with force of infection. Here, we assume that the number of people leaving their home patches is equal for all patches ($u_r = 10$) and they will visit other patches with the same probability.

The mean vector and standard deviation of the multivariate Gaussian distribution corresponding to the diffusion approximation are obtained in the same way as before and the corresponding mean vector and standard deviation of the stochastic models are presented in Table 5 and Table 6, respectively. Again there is close agreement with the sample means and variances obtained by Monte Carlo simulation.

Table 5: Means and standard deviations of Multivariate Gaussian distribution.

Patches	**Infected Females**		**Infected Males**	
	μ	σ	μ	σ
1-10	8824	92.19	8824	92.19

Table 6: Sample means and sample standard deviations from numerical experiments.

Patches	**Infected Females**		**Infected Males**	
	$\tilde{\mu}$	$\tilde{\sigma}$	$\tilde{\mu}$	$\tilde{\sigma}$
1	8809.4	93.34	8782.0	95.01
2	8823.6	95.84	8795.9	90.79
3	8813.5	93.62	8802.5	91.72
4	8817.0	96.04	8827.5	93.34
5	8835.8	95.63	8822.3	97.57
6	8824.5	93.92	8834.1	92.97
7	8811.5	89.68	8820.4	89.19
8	8830.8	98.32	8821.9	98.94
9	8824.7	96.90	8835.8	98.94
10	8838.4	86.97	8807.3	84.99

Thus, the deterministic and diffusion approach can be applied to study the dynamic behavior of the stochastic multiple patch model with the actual mobility.

6 Conclusion and Future Research

The dynamic behavior of the stochastic models for the spread of HIV presented in this paper are well approximated by their deterministic and diffusion counterparts. We find the same threshold conditions $R_0 = 1$ for a disease-free equilibrium in the case of both an open and closed single population. As $R_0 > 1$ (above threshold), this equilibrium loses its stability and a stable endemic state occurs. The numerical results also indicate that there exists a positive-stable endemic equilibrium in the multiple patch models, although we have not proved this analytically. Although the stochastic models presented in this paper are perhaps too simple to describe the actual spread of HIV, they provide some clues how e.g., more realistic models can be formulated. Moreover, for future research, it should be feasible to use the deterministic and diffusion approaches to study more complex stochastic models of HIV/AIDS spread; for example, stochastic models in a mobile heterogeneous population, classified according to age and sexual behavior, or (since the disease is primarily a sexually transmitted disease) models that include partnership pattern formation. Another possible direction for future research is to consider how control strategies may be devised. For example, to find a strategy that provides the greatest reduction in the endemic level of the disease for a given cost, or to find the cheapest strategy that guarantees a upper level of prevalence of HIV in all patches. Finally, taking into account the available statistical data and control strategies into the models will further improve our understanding how the disease spread into the heterogenous population. However, as many factors of consideration are included in the models, the complexity of the models increases.

Acknowledgments

The first author is grateful to ADS-AUSAID for funding his PhD scholarship. We thank also Joshua Ross for his helpful discussions regarding density dependent processes. This research was supported by the Australian Research Council, via the Discovery Grant DP0558957.

References

[1] *Report on the Global Epidemic.* UNICEF, UNDP, 2004.

[2] Linda J. S. Allen. *An Introduction to Stochastic Processes with Applications to Biology.* Pearson Education, Inc., 2003.

[3] F. Arrigoni and A. Pugliese. Limits of a multi-patch SIS epidemic model. *J. Mathematical Biology*, 45:419–440, 2002.

[4] Andrew D. Barbour. Quasi-stationary distribution in Markov population processes. *Advance Applied Probability*, 8:296–314, 1976.

[5] Mode C. J. and Sleeman C. K. A new design of stochastic partnership models for epidemic of sexually transmitted diseases with stages. *Mathematical Biosciences*, 156:95–122, 1999.

[6] D. Clancy. A stochastic SIS infection model incorporating indirect transmission. *To Appear.*

[7] D. Clancy, P. D. O'Neill, and P. K. Pollett. Approximations for the long-term behavior of an open-population epidemic model. *Methodol. Comput. Appl. Probab.*, 3(1):75–95, 2001.

[8] Klaus Dietz. On the transmission dynamics of HIV. *Math. Biosci.*, 90(1-2):397–414, 1988. Nonlinearity in biology and medicine (Los Alamos, NM, 1987).

[9] W. Huang, K.L. Cooke, and C. Castillo-Chavez. Stability and bifurcation for a multiple-group model for the dynamics of HIV/AIDS transmission.

[10] G. Hugo. Indonesia, internal and international population mobility: Implications for the spread of HIV/AIDS. Technical report, Adelaide University, Australia, November 2001.

[11] V. Isham. Mathematical modelling of the transmission dynamics of HIV infection and AIDS. a review. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 151:5–30, 1988.

[12] M. Kremer and C. Morcom. The effect of changing sexual activity on HIV prevalence. *Mathematical Biosciences*, 151:99–122, 1998.

[13] T. G. Kurtz. Solutions of ordinary differential equations as limits of pure jump Markov processes. *J. Applied Probability*, 7:49–58, 1970.

[14] T. G. Kurtz. Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *J. Applied Probability*, 8:344–356, 1971.

[15] A. L. Lloyd and V. A. A. Jansen. Spatiotemporal dynamics of epidemics: synchrony in metapopulation models. *Mathematical Biosciences*, 2004.

[16] R. M. May, R. M. Anderson, and A. R. McLean. Possible demographic consequences of HIV/AIDS epidemics. I. assuming HIV infection always leads to AIDS. *Mathematical Biosciences*, 90:475–505, 1988.

[17] C. J. Mode and C. K. Sleeman. *Stochastic Processes in Epidemiology, HIV/AIDS, Other Infectious Diseases and Computers*. World Scientific, 2000.

[18] P. K. Pollett. On a model for interference between searching insect parasites. *J. Austral. Math. Soc. Ser. B*, 32:133–150, 1990.

[19] Sheldon M. Ross. *Stochastic Processes*. John Wiley & Sons, Inc., second edition, 1996.

[20] L. Sattenspiel and K. Dietz. A structured epidemic model incorporating geographic mobility among regions. *J. Mathematical Biosciences*, 123:71–91, 1995.

[21] W. Y. Tan and X. Zhu. A stochastic model for the HIV epidemic in homosexual populations involving age and race. *J. Mathematical Computational Modelling*, 24(12):67–105, 1996.

[22] W. Y. Tan and X. Zhu. A stochastic model of the HIV epidemic for heterosexual transmission involving married couples and prostitutes: I. the probabilities of HIV transmission and pair formation. *J. Mathematical Computational Modelling*, 24(11):47–100, 1996.

Authors

A. Sani: Department of Mathematics, FMIPA, Haluoleo University, Kendari, Indonesia. PhD Student at Department of Mathematics, University of Queensland, St. Lucia, Brisbane, Australia.
E-mail: asani@maths.uq.edu.au, saniasrul2001@yahoo.com

D. P. Kroese: Department of Mathematics, University of Queensland, St. Lucia, Brisbane, Australia.
E-mail: dpk@maths.uq.edu.au

P. K. Pollett: Department of Mathematics, University of Queensland, St. Lucia, Brisbane, Australia.
E-mail: pkp@maths.uq.edu.au

Epidemiological Model for the Spread of Anti-Malarial Resistance and Its Economics Aspect

Lusia Krismiyati Budiasih

Departement of Mathematics, Institut Teknologi Bandung, Indonesia

Abstract: The spread of anti-malarial resistance is making malaria control increasingly difficult. Mathematical models for the transmission dynamics of resistant and sensitive strains of the parasite can be used as a tool to help to understand the factors that influence the spread of anti-malarial resistance, and they can help in thinking the strategic steps to control the spread of resistance.
Now it can be found that malaria parasite resistant to chloroquine. With increasing the resistance, attention has begun to be pressed in treatment strategies by using some new drugs such as *sulfadoxine-pyrimethamine* (SP), and artemisinin-based combination treatments (ACT). Although there are strong theoretical arguments in favor of switching to ACT, the validity of these arguments in the economics aspects has not been previously analyzed.
This paper presents an epidemiological framework to investigate the spread of anti-malarial resistance. Several mathematical models, based on Macdonald-Ross model of malaria transmission, enable to examine the processes and parameters that influenced in the spread of resistance and also can be used to examine the optimal treatment strategies.
In the simplest model, the resistance does not spread if the fraction of infected individuals treated is less than a threshold value. The threshold value is determined by the rates of infection and level of anti-malarial effectiveness. Using the developing of this model, it can be showed that in the appropriate level of treatment, the treatment strategy by using ACT early needs less cost than by SP in a few time and then replaced with ACT.

Keywords: malaria epidemiological model, spread of resistance, ACT, SP, treatment strategy

THE EFFECT OF A DIAGNOSIS MECHANISM ON SARS EPIDEMIC

Hermayanti, Ponidi, Arie Wibowo, Rahmi Rusin

Department of Mathematics, University of Indonesia, Depok, Indonesia

Abstract. In this paper, we discuss the effect of a diagnosis mechanism for SARS epidemic. First, we derive the mathematical model of SARS epidemic with and without diagnosis mechanism, and then from that model, a Basic Reproduction Ratio Number (R_0) is obtained using the next generation operator approach. By comparing the dynamics of the two models, we will see that diagnosis mechanism has large effects in SARS epidemic.
Key-words: Basic Reproduction Ratio Number, epidemic, equilibrium, SEIJR, SARS.

1 Introduction

SARS is a respiratory disease caused by *coronavirus* (SARS-CoV). The first case of SARS happened in Guangdong Province, China in November 2002. The virus has the ability to spread rapidly on global scale. But, although there is a very rapid increase in the number of SARS, it has not known yet how the virus is transmitted. There was some hypothesis about the mode of transmission that is: the virus transmitted by droplet, airborne transmission or person-to-person contact. [4]. If an individual lives in a population with patient of active SARS then the possibility to be infected is very high.

The transmission of SARS disease is not only a threat to a local region where SARS disease occurs but also to international community since it spreads rapidly. If there is no immediate action, the disease will be transmitted to numerous numbers of other people, for example a 26-year old airport worker; have transmitted the disease to 112 people [4]. Another negative impact of SARS disease to the country is on tourism and economic side where people from other countries will avoid traveling to the epidemic region; hence it will reduce the state income.

To control the SARS epidemic, two diagnosis mechanisms can be conducted: isolation and quarantine. Both of mechanisms have the same aim, dissociating the patient with his community in order to prevent the transmission per contact. The difference between those two mechanisms is on the object. Isolation is emphasized on infected individuals and has the symptom; while quarantine is emphasized on infected individuals but have no symptom yet. These mechanisms are expected to help government reducing the number of SARS cases and hopefully reducing deaths as a result of SARS.

In this paper, we will discuss about the effect of diagnosis mechanism on SARS epidemic. In the next section we will derive a mathematical model of SARS epidemic with a diagnosis mechanism, and from the model we will derive the basic reproduction ratio number (R_0) and finally to see the effect of the diagnosis

mechanism we will compare the model with a model of SARS epidemic without diagnosis mechanism.

2 Modeling

In this section, we will formulate a mathematical model for SARS epidemic, in this formulation we are modifying a model derived by G. Chowell et a.l [2]. We introduce 6 classes of individuals in a population: high-risk susceptible (S_1), low risk susceptible (S_2), infected individuals but not yet infectious or exposed (E), infectious (I), recovery (R) and diagnosed (J). The total population is denoted by N.

We assumed that: the size of population is affected only from death and birth; every infectious and diagnosed individual has the same probability to spread SARS to susceptible individuals; the number of natural birth is the same constant for classes S_1 and S_2; recovery individuals cannot acquire a new infection from infectious individuals; an infectious individuals can be a diagnosed individual after experiencing the diagnosis mechanism; if the diagnosed individuals are handled well, they might not become an infectious one but a recovery individuals.

If Λ is the constant recruitment rate, β is the transmission rate, p is the reduction in risk of SARS infection for S_2, q is the relative measure of infectiousness for class E, l is the relative infectiousness after isolation , k is the rate of an exposed individuals become an infectious individuals, α is the rate of progression from infective to diagnosed per day, γ_1 is the rate at which infectious individuals recover per day, γ_2 is the rate at which diagnosed individuals recover per day, μ is the natural death rate and δ is the SARS-induced mortality per day.

Hence, the SARS transmission can be described completely as the following figure:

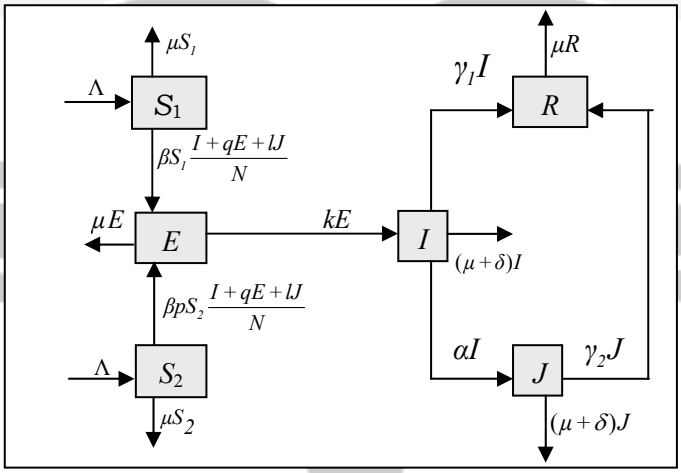


Figure 1. Compartment Diagram of SARS transmission with a diagnosis mechanism

From Figure 1, the SARS transmission with a diagnosis mechanism can be modeled mathematically by the following autonomous nonlinear system of differential equations:

$$\begin{aligned}
\frac{d}{dt}S_1 &= \Lambda - \beta S_1 \frac{I + qE + lJ}{N} - \mu S_1 \\
\frac{d}{dt}S_2 &= \Lambda - \beta S_2 p \frac{I + qE + lJ}{N} - \mu S_2 \\
\frac{d}{dt}E &= \beta (S_1 + pS_2) \frac{I + qE + lJ}{N} - (\mu + k)E \\
\frac{d}{dt}I &= kE - (\gamma_1 + \alpha + \delta + \mu)I \\
\frac{d}{dt}R &= \gamma_1 I + \gamma_2 J - \mu R \\
\frac{d}{dt}J &= \alpha I - (\gamma_2 + \delta + \mu)J
\end{aligned} \tag{2.1}$$

Next, we will present the derivation of *basic reproduction ratio number* (R_0) from system (2.1) using the next generation operator approach [4]

3 Basic Reproduction Ratio Number (R_0)

To understanding the dynamic of SARS transmission, we can use a number known as a basic reproduction ratio number (R_0) to determine the stability of the equilibria. R_0 is defined as the average number of secondary cases produced by a "typical" infected (assumed infectious) individual during his/her entire life as infectious (infectious period) when introduced in a population of susceptibles. [1]

Most of the epidemiological models have two equilibria: disease free equilibrium and endemic equilibrium. If $R_0 > 1$, an epidemic occur and if there is no treatment the will be an endemic, or mathematicaly we say that the endemic equilibrium is asymptotically stable. On the contrary, if $R_0 < 1$, the epidemic cannot occur and for a long period of time there will be a free disease situation. Or mathematically we say that disease free equilibrium is asymptotically stable.

Using the next generation operator approach we divide the population into three classes, that is

$$X = (S_1, S_2, R),\ Y = (E),\ Z = (I, J)$$

where X denotes the number of susceptibles, recovered and non infectious individuals, Y denotes the number of infected individuals who do not transmit the disease and Z denotes the number of infectious individuals that has the ability of transmitting the disease.

Hence the system (2.1) can be written as follows:

$$f_{S_1}(X, Y, Z) = f_{S_1}(S_1, S_2, E, I, J, R) = \beta S_1 \frac{I + qE + lJ}{N} - \mu S_1$$

$$f_{S_2}(X, Y, Z) = f_{S_2}(S_1, S_2, E, I, J, R) = \beta p S_2 \frac{I + qE + lJ}{N} - \mu S_2$$

$$f_R(X, Y, Z) = f_R(S_1, S_2, E, I, J, R) = \gamma_1 I + \gamma_2 J - \mu R$$
$$g(X, Y, Z) = g(S_1, S_2, E, I, J, R) = \beta(S_1 + pS_2)\frac{I + qE + lJ}{N} - (\mu + k)E$$
$$h_I(X, Y, Z) = h_I(S_1, S_2, E, I, J, R) = kE - (\alpha + \mu + \gamma_1 + \delta)I$$
$$h_J(X, Y, Z) = h_J(S_1, S_2, E, I, J, R) = \alpha I - (\gamma_2 + \mu + \delta)J$$

Using the steps on the method (see [3] for details), we obtain the formula of basic reproduction ratio number (R_0) for SARS epidemic with a diagnosis mechanism.

$$R_0 = \frac{k\beta\, l\alpha(1+p)}{(\gamma_2 + \delta + \mu)[(2(k+\mu) - \beta q(1+p))(\alpha + \gamma_1 + \delta + \mu) - k\beta(1+p)]} \tag{3.1}$$

From the formula (3,1), it can be seen that there are some factors that have large effect on SARS epidemic which can be controlled, they are the effectiveness of diagnosis mechanism to reduce the number of SARS cases, the rate of infection and the SARS-induced mortality. If the diagnosis mechanism (l) is more effective then it is likely that the possibility to be recovered is higher. Also, the smaller the rate of infection (k, β), the less number of SARS cases happen.

4 Simulations

From the system (2.1), we obtained two equilibrium points: disease free equilibrium and endemic equilibrium. Since the system is nonlinear, we use linearization around each equilibrium. If $E = I = J = 0$ then the following disease free equilibrium is obtained

$$U_0 = (S_1^*, S_2^*, R^*, 0, 0, 0) = \left(\frac{\Lambda}{\mu}, \frac{\Lambda}{\mu}, 0, 0, 0, 0\right)$$

and if $E \neq 0$, $I \neq 0$, $J \neq 0$, we will obtain endemic equilibrium $U_1 = (S_1^*, S_2^*, E^*, I^*, J^*, R^*)$ where each value of $S_1^*, S_2^*, E^*, I^*, J^*, R^*$ is in [3]. Since for endemic equilibrium has a very complicated form, we will analyze its stability using simulation for given parameter values.

Jacobian Matrix for the system (2.1) is

$$\frac{\partial f}{\partial x}(\bar{x}) = \begin{pmatrix} -\frac{\beta(K+qE+lJ)}{N} - \mu & 0 & \frac{-\beta S_1 q}{N} & \frac{-\beta S_1}{N} & \frac{-S_1\beta l}{N} & 0 \\ 0 & -\frac{\beta(K+qE+lJ)}{N} - \mu & \frac{-\beta S_2 pq}{N} & \frac{-\beta S_2 p}{N} & \frac{-\beta S_2 lp}{N} & 0 \\ \frac{\beta(K+qE+lJ)}{N} & \frac{\beta p(K+qE+lJ)}{N} & \frac{\beta(S_1+pS_2)q}{N} - \mu - k & \frac{-\beta(S_1+pS_2)}{N} & \frac{\beta(S_1+pS_2)l}{N} & 0 \\ 0 & 0 & k & -\alpha - \gamma_1 - \mu - \delta & 0 & 0 \\ 0 & 0 & 0 & \alpha & -\mu - \gamma_2 - \delta & 0 \\ 0 & 0 & 0 & \gamma_1 & \gamma_2 & -\mu \end{pmatrix} \tag{3.2}$$

and for disease free equilibrium $U_0 = (\dot{S}_1, \dot{S}_2, \dot{R}, 0, 0, 0) = \left(\frac{\Lambda}{\mu}, \frac{\Lambda}{\mu}, 0, 0, 0, 0\right)$, we obtained the following Jacobian matrix

$$B=\begin{pmatrix} -\mu & 0 & \frac{-\Lambda\beta q}{\mu N} & \frac{-\beta q}{\mu N} & \frac{-\Lambda\beta q}{\mu N} & 0 \\ 0 & -\mu & \frac{-\Lambda\beta pq}{\mu N} & \frac{-\beta pq}{\mu N} & \frac{-\Lambda\beta l}{\mu N} & 0 \\ 0 & 0 & \frac{-\beta\left(\frac{\Lambda}{\mu}+p\frac{\Lambda}{\mu}\right)q}{N}-\mu-k & \frac{-\beta\left(\frac{\Lambda}{\mu}+p\frac{\Lambda}{\mu}\right)}{N} & \frac{-\beta\left(\frac{\Lambda}{\mu}+p\frac{\Lambda}{\mu}\right)l}{N} & 0 \\ 0 & 0 & k & -\alpha-\gamma_1-\mu-\delta & 0 & 0 \\ 0 & 0 & 0 & \alpha & -\gamma_2-\mu-\delta & 0 \\ 0 & 0 & 0 & \gamma_1 & \gamma_2 & -\mu \end{pmatrix}$$

Characteristics equation of the matrix is

$$(\mu+\lambda)^3(\lambda+\gamma_2+\mu+\delta)(\lambda+\mu+\delta+\alpha+\gamma_1)\left(\lambda+\beta\frac{\left(\frac{\Lambda}{\mu}+p\frac{\Lambda}{\mu}\right)}{N}+(\mu+k)\right)=0,$$

and the roots are $\lambda_1=-\mu, \lambda_2=-(\gamma_2+\mu+\delta), \lambda_3=-(\mu+\delta+\alpha+\gamma_1), \lambda_4=-\beta\frac{\frac{\Lambda}{\mu}+p\frac{\Lambda}{\mu}}{N}-\mu-k$

Since all eigenvalues of B have negative real part then $U_0 = \left(\frac{\Lambda}{\mu}, \frac{\Lambda}{\mu}, 0, 0, 0, 0\right)$ is asymptotically stable, that is for a long period of time the only individuals left is susceptible.

For the following values of parameter:

$\Lambda = 20;$ $p = 0.1;$ $q = 0.1;$ $\beta = 0.15;$ $l = 0.05;$

$\alpha = 0.167;$ $\mu = 0.0001;$ $\delta = 0.0016;$ $\gamma_1 = 0.125;$ $k = 0.3;$ $\gamma_2 = 0.2;$

we obtained $R_0 = 0.13 < 1$. Hence, disease free equilibrium $U_0 =$ $(2000, 2000, 0, 0, 0, 0)$ is asymptotically stable. Using MAPLE software, the graphs of $S_1(t), S_2(t), E(t), I(t), R(t)$ and $J(t)$ are obtained with initial values:

$$S_1(0) = 2000, S_2(0) = 1000, E(0) = 0, \quad I(0) = 1000, \quad R(0) = 0, \quad J(0) = 800$$

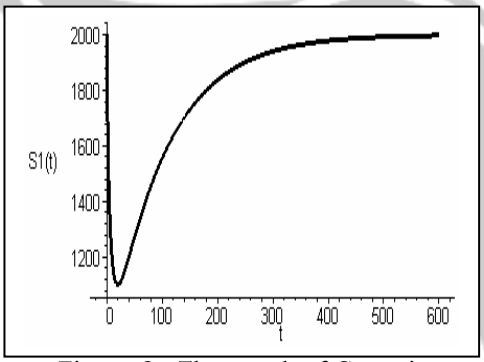


Figure 2. The graph of S_1 vs time

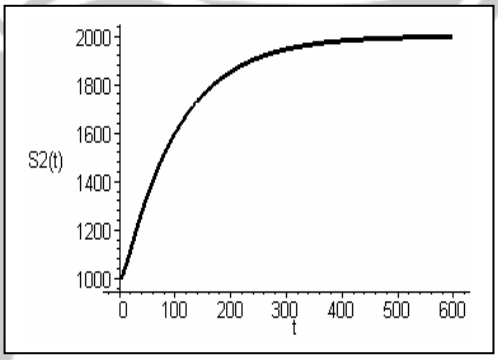


Figure 3. The graph of S_2 vs time

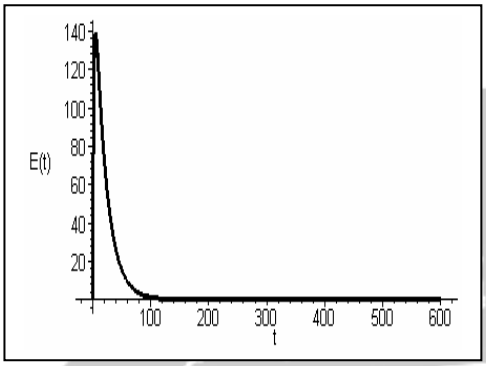


Figure 4. The graph of E vs time

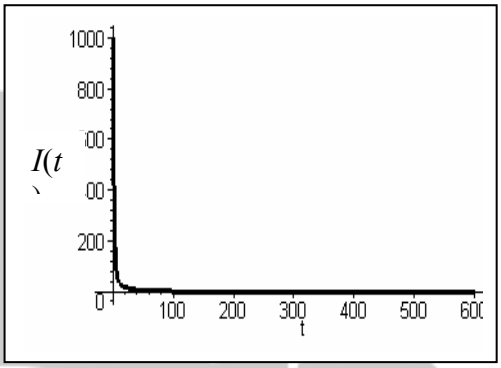


Figure 5. The graph of I vs t

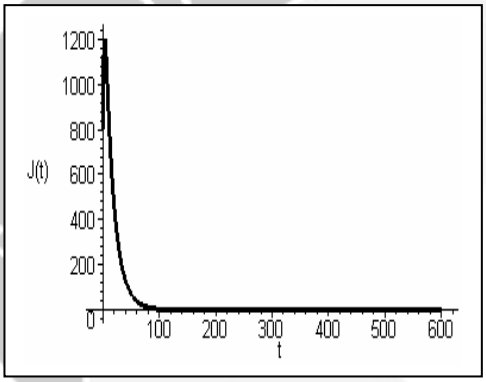


Figure 6. The graph J of vs t

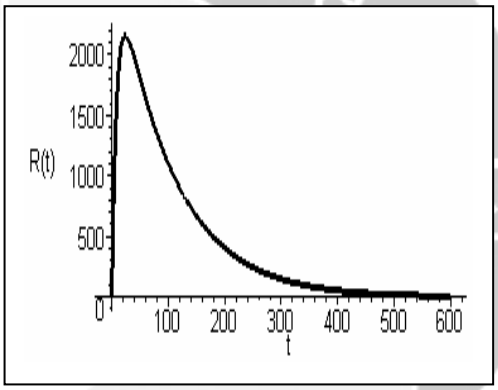


Figure 7. The graph of R vs time

From Figure 2 and 3, it can be seen that at first, the number of high risk susceptible individuals decreasing, it is probably because of the contact with the infected individuals, and then continue increasing until a long period of time to the equilibrium value. From Figure 4 – figure 7, it is shown that the number of infected individuals (*E, I,* and *J)* continue decreasing and disappear after a long period of time while the number of recovery individuals at first increasing because of the decreasing of number infected individuals and then after a long period of time continue decreasing, this is probably because of natural death.

Next, for given parameter values for $R_0 < 1$ and $R_0 > 1$, we will analyze the stability of endemic equilibrium.

Case 1: $R_0 < 1$

Given parameter values as follows:

$\beta = 0.15;$ $q = 0.1;$ $l = 0.05;$ $N = 4000000;$ $\mu = 0.0001;$ $k = 0.3;$

$\gamma_1 = 0.125;$ $\gamma_2 = 0.2;$ $\alpha = 0.167;$ $\delta = 0.0016;$ $p = 0.1;$

Using the formula (3.1), we obtained $R_0 = 0.13 < 1$. If those parameter values above are substituted to Jacobian matrix (3.2) then we get a characteristic equation:

$$(\lambda+0.2932478226)(\lambda+0.2012069457)(\lambda+0009778747247)$$
$$(\lambda+0.0006000000623)(\lambda+0.00035404165992)(\lambda-0.0009728301097)=0 \quad (3.3)$$

From (3.3) we have the following eigenvalues:

$\lambda_1 = -0.2932478226$ $\lambda_2 = -0.2012069457$ $\lambda_3 = -0.0009778747247$

$\lambda_4 = -0.006000000623$ $\lambda_5 = -0.0035404165992$ $\lambda_6 = 0.0009728301097$

Since λ_6 is positive, then the endemic equilibrium $U_1 = (S_1^*, S_2^*, E^*, I^*, J^*, R^*)$ is not stable.

Case 2: $R_0 > 1$

Given parameter values as follows:

$\beta = 0.75;$ $q = 0.1;$ $l = 0.8;$ $N = 4000000;$ $\mu = 0.0001;$ $k = 0.75;$

$\gamma_1 = 0.125;$ $\gamma_2 = 0.2;$ $\alpha = 0.3;$ $\delta = 0.0016;$ $p = 0.1;$

Using the formula (3.1), we obtained $R_0 = 1.78 > 1$. If those parameter values above are substituted to Jacobian matrix (3.2) then we get a characteristic equation:

$$0.9999999997(\lambda+0.4259380046)(\lambda+0.2010076511)(\lambda+0.0003410883103)$$
$$(\lambda+0.0003400886408)(\lambda+0.0003087256892)(\lambda+0.0004160978019)=0 \quad (3.4)$$

From (3.4) we have the following eigenvalues:

$\lambda_1 = -0.4259380046$ $\lambda_2 = -0.2010076511$ $\lambda_3 = -0.0003410883103$

$\lambda_4 = -0.0003400886408$ $\lambda_5 = -0.0003087256892$ $\lambda_6 = -0.00004160978019$

Hence, the endemic equilibrium $U_1 = (S_1^*, S_2^*, E^*, I^*, J^*, R^*)$ is asymptotically stable.

To plot the solutions for case 2, we take $\Lambda = 20$, hence we get the endemic equilibrium, $U_1 = (2000.58, 3000.67, 11000.23, 12.04, 53.89, 2600.77)$. This equilibrium is asymptotically stable. Using MAPLE software, the graphs of $S_1(t), S_2(t), E(t), I(t), R(t)$ and $J(t)$ are obtained with initial values:

$$S_1(0) = 8000,\ S_2(0) = 6000,\ E(0) = 0,\ I(0) = 1,\ R(0) = 0 \text{ and } J(0) = 0$$

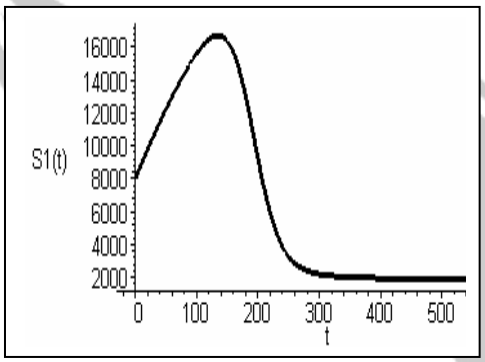


Figure 8. The graph of S_1 vs time

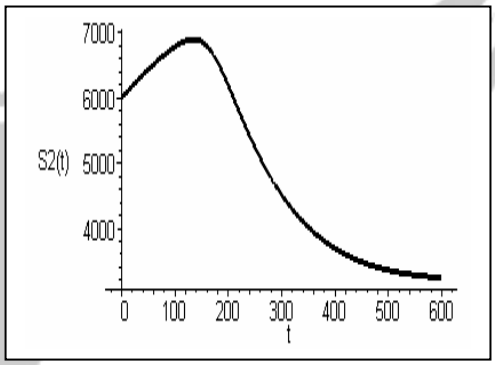


Figure 9. The graph of S_2 vs time

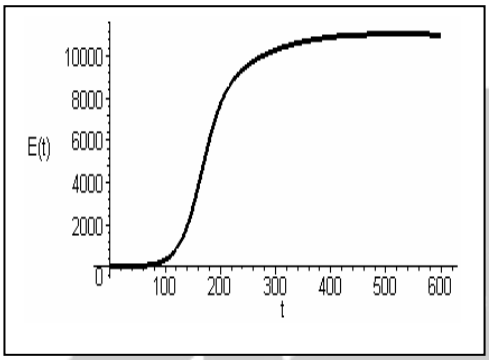


Figure 10. The graph of E vs time

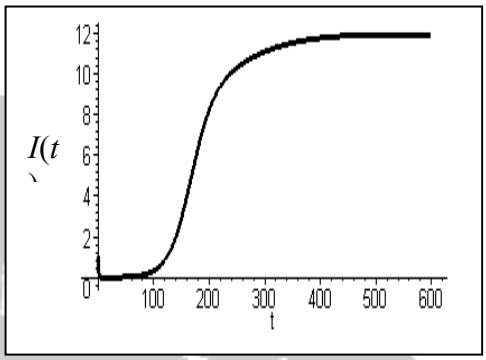


Figure 11. The graph of I vs time

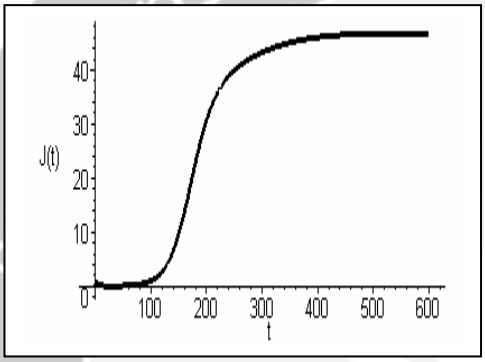


Figure 12. The graph of J vs time

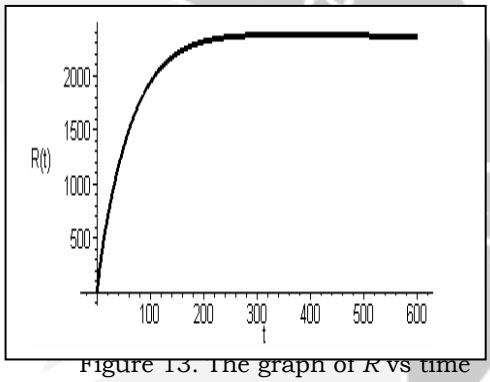


Figure 13. The graph of R vs time

From Figure 8 – Figure 11, it can be seen that after reaching a certain value, the number of high risk susceptible individuals (S_1) and low risk susceptible individuals (S_2) will decrease to a constant values for a long period of time, this is because the increasing number of exposed individuals (E) (Figure 10). While the number of infectious individuals (I) at first decreases as a result of death, but since $R_0 > 1$, then the number of infectious still increasing and keep continue to the equilibrium after a long period of time. (Figure 11).
The number of diagnosed individuals (J) can be seen from Figure 10. It increases until reach a constant for a long period of time because many infectious individuals are handled well by diagnosis mechanism. As a result, recovery individuals keep increasing after a long period of time (Figure 13).

To see the effect of a diagnosis mechanism, next we will compare the model (2.1) with model of SARS epidemic without diagnosis mechanism. First we derived the model without diagnosis mechanism and then obtained the basic reproduction ratio number for the system.

The following is the compartment diagram for the system without diagnosis mechanism

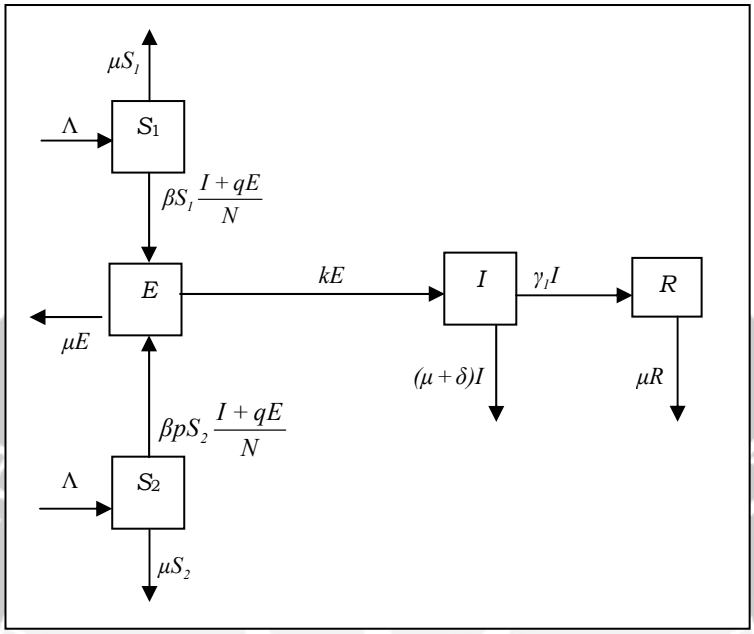


Figure 14. Compartment diagram of SARS transmission without diagnosis mechanism

From the diagram above, we obtained the following mathematical model of SARS transmission without a diagnosis mechanism.

$$\begin{aligned}
\frac{d}{dt}S_1 &= \Lambda - \beta S_1 \frac{I+qE}{N} - \mu S_1 \\
\frac{d}{dt}S_2 &= \Lambda - \beta S_2 p \frac{I+qE}{N} - \mu S_2 \\
\frac{d}{dt}E &= \beta (S_1 + pS_2)\frac{I+qE}{N} - (\mu + k)E \\
\frac{d}{dt}I &= kE - (\gamma_1 + \delta + \mu)I \\
\frac{d}{dt}R &= \gamma_1 I - \mu R
\end{aligned} \tag{3.5}$$

Similarly with the section before, we can get the formula of basic reproduction ratio number (R_0) of SARS epidemic model without a diagnosis mechanism that is:

$$R_0 = \left(\frac{k\beta(1+p)}{-\beta q(1+p) + 2(\mu + k)} \right) \frac{1}{(\gamma_1 + \delta + \mu)}$$

In the next section, we will compare the simulation between the model with diagnosis mechanism and a model without a diagnosis mechanism in order to see the effect of the mechanism: isolation and quarantine.

5 Comparison system with and without a diagnosis mechanism

Using the same parameter values, we will compare the graph of each individual for disease free equilibrium. The graphs are from Figure 15 – Figure 19.

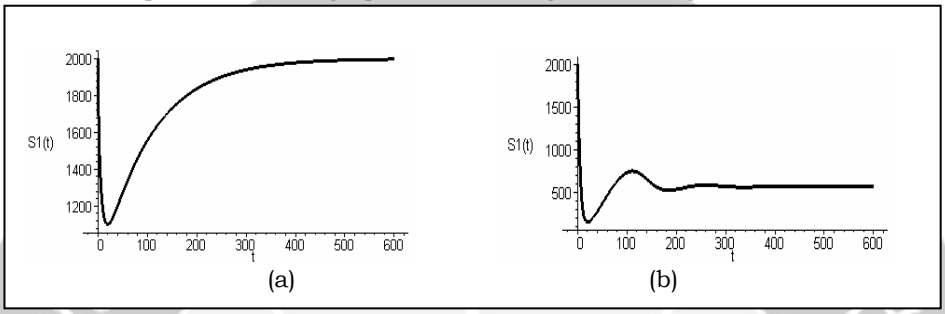


Figure 15. (a) The graph of S_1 vs time (days) with a diagnosis mechanism
(b) The graph of S_1 vs time (days) without a diagnosis mechanism

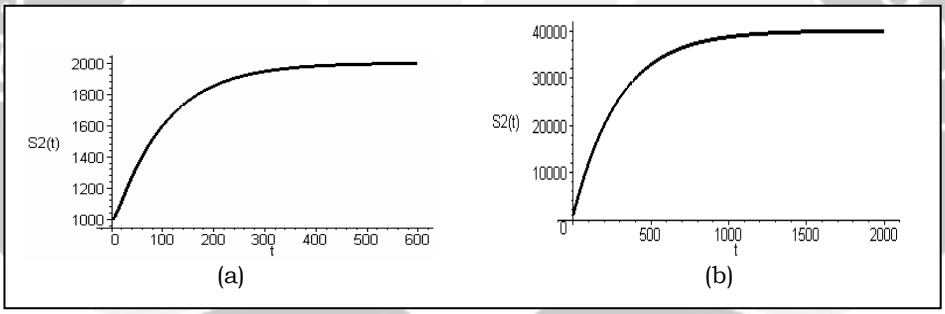


Figure 16. (a) The graph of S_2 vs time (days) with a diagnosis mechanism
(b) The graph of S_2 vs time (days) without a diagnosis mechanism

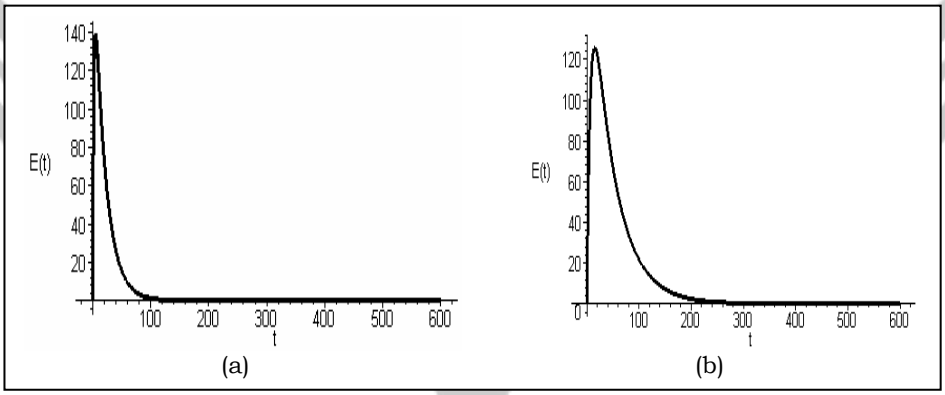


Figure 17. (a) The graph of E vs time (days) with a diagnosis mechanism
(b) The graph of E vs time (days) without a diagnosis mechanism

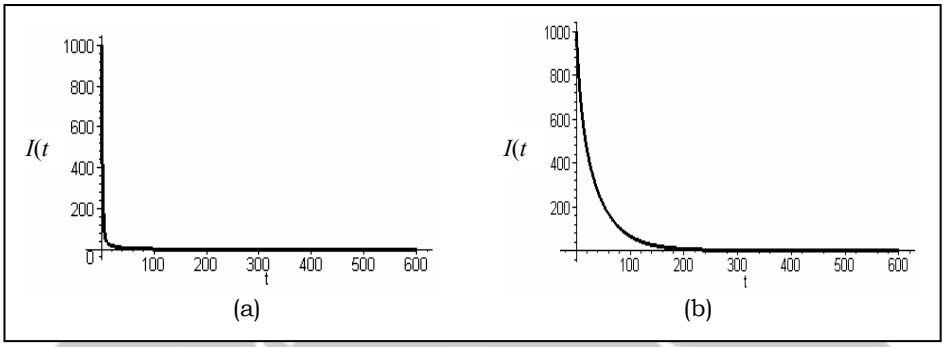


Figure 18. (a) The graph of I vs time (days) with a diagnosis mechanism
(b) The graph of I vs time (days) without a diagnosis mechanism

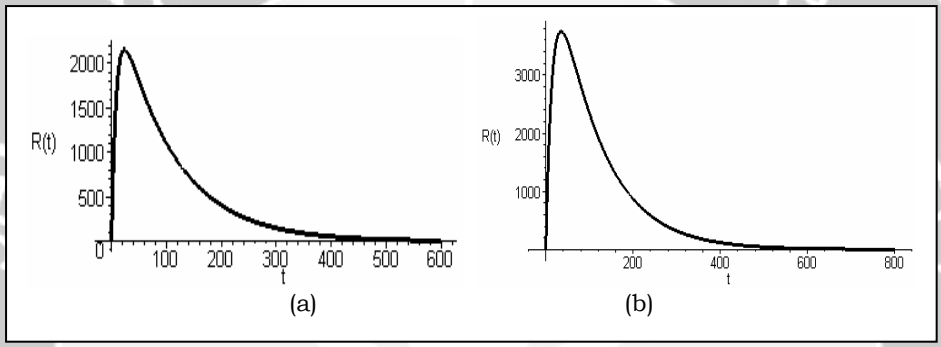


Figure 19. (a) The graph of I vs time (days) with a diagnosis mechanism
(b) The graph of I vs time (days) without a diagnosis mechanism

From the figures above, we can see the differences between the dynamics of SARS epidemic between model with (a) and without (b) diagnosis mechanism. It can be seen that the diagnosis mechanism has large effects on SARS epidemic. For example in Figure 15, the increasing number of low-risk susceptible in Figure 15(b) is lower than the number of low-risk susceptible in Figure 15(a). This is because of the intensive handling through diagnosis mechanism. It also can be seen that it needs more time for high-risk susceptible in Figure 15(b) to go to stable condition than the high-risk susceptible in Figure 15(a). It is a result of mechanism where it makes the time for cure is faster.

Similar for other classes of individuals, we can say that the diagnosis mechanism is reducing the time of cure and also quicken the time to go to stable condition.

For endemic equilibrium, using the same parameter values on preceding section, we obtained the following plots of the solutions.

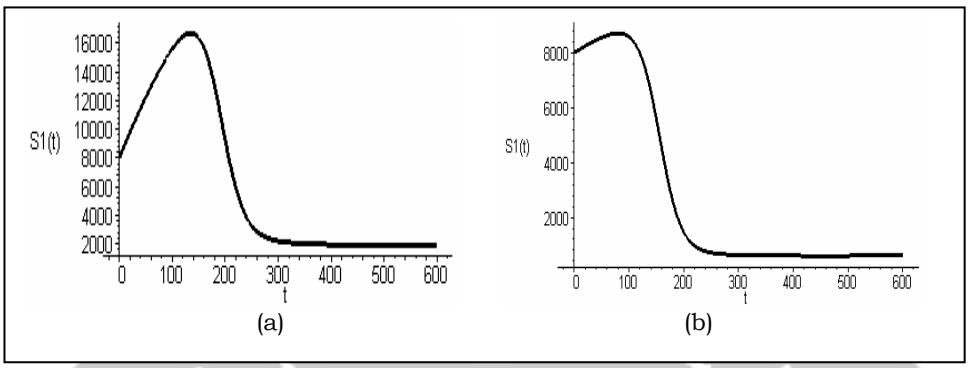


Figure 20. (a) The graph of S_1 vs time (days) with a diagnosis mechanism
(b) The graph of S_1 vs time (days) without a diagnosis mechanism

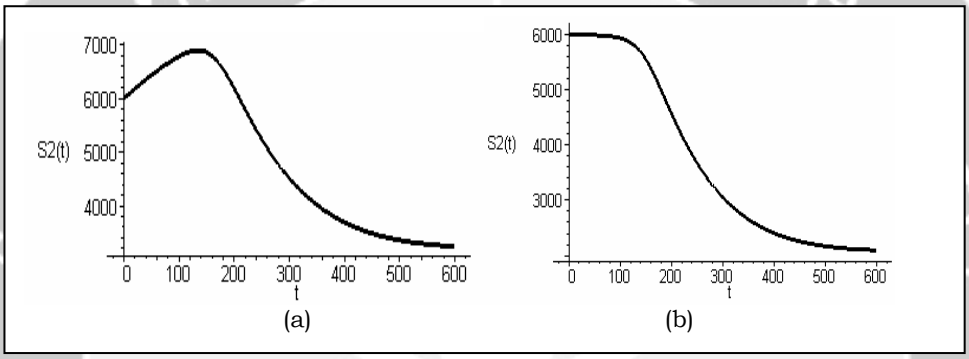


Figure 21. (a) The graph of S_2 vs time (days) with a diagnosis mechanism
(b) The graph of S_2 vs time (days) without a diagnosis mechanism

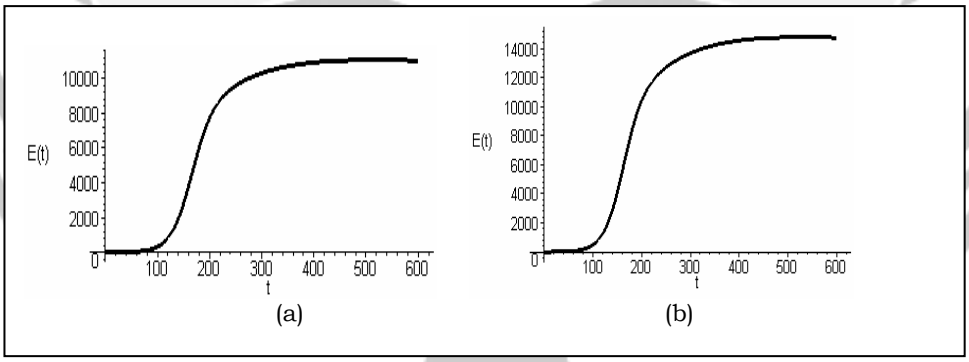


Figure 22. (a) The graph of E vs time (days) with a diagnosis mechanism
(b) The graph of E vs time (days) without a diagnosis mechanism

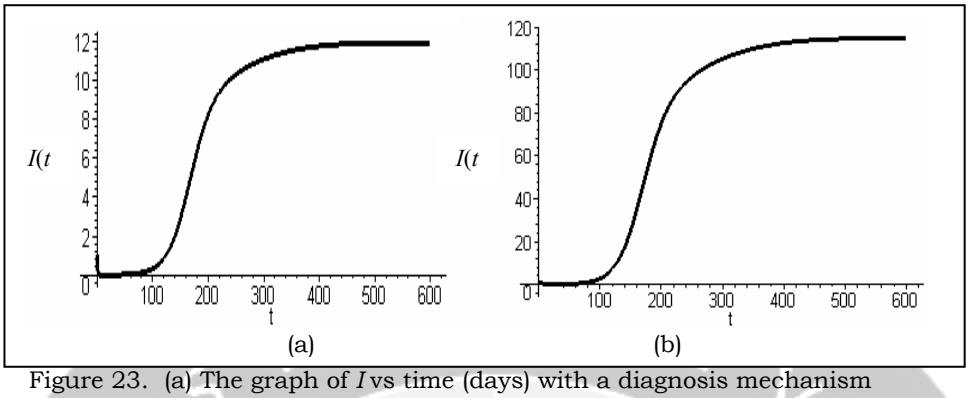


Figure 23. (a) The graph of I vs time (days) with a diagnosis mechanism
(b) The graph of I vs time (days) without a diagnosis mechanism

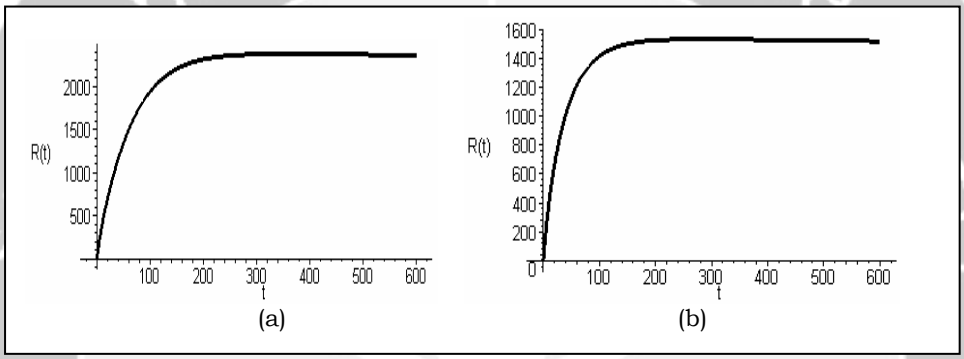


Figure 24. (a) The graph of R vs time (days) with a diagnosis mechanism
(b) The graph of R vs time (days) without a diagnosis mechanism

From Figure 20 – Figure 24 above, we can see the differences between the system with and without a diagnosis mechanism. The diagnosis mechanism make the transmission of SARS disease slower, so the number of susceptible in the system with diagnosis is more than the number of susceptible in the system without diagnosis, hence the number of exposed in the system with diagnosis is less than the number of exposed in the system without diagnosis (Figure 20 – Figure 22). Since the isolation mechanism prevents person-to-person contact, the number of infectious is less than if the mechanism is not conducted and finally it makes the number or recovery higher (Figure 23- Figure 24).

6 Conclusions

The diagnosis mechanism has large effects on the transmission of SARS disease in a population. It can be seen from the simulation, that the diagnosis mechanism is effective and efficient in order to reduce the transmission of the virus and also making the time of healing faster.

References

[1] Diekmann, O and J.A.P Heesterbeek (2000), *Mathematical Epidemiology of Infectious diseases*, John Wiley & Sons Ltd. New York.

[2] Grimshaw, R (1990), *Nonlinear Ordinary Differential Equations*, Blackwell Scientific Publications. Melbourne.

[3] Hermayanti (2005), *Pengaruh Mekanisme Diagnosa terhadap Penyebaran Epidemi Severe Acute Respitory Syndrome (SARS)*, UI, Depok

[4] Hale, J and Kocak,H. (1991), *Dynamics and bifurcations*, Springer-Verlag. New York.

[5] G. Chowell, P.W, Fenimore M.A, and Castillo-Chavez, C (2003), SARS outbreaks in Ontario, Hongkong and Singapore, *Journal of Theoretical Biology*, **224**, 1 – 8.

[6] Castillo-Chavez, C., Z Feng and W.Huang (2002), On the computational of R_0 and its role on global stability, *Journal of Theoretical Biology*, **125**, 229 – 250.

HERMAWATI: Department of Mathematics, University of Indonesia, Ged D FMIPA UI Kampus Baru Depok, Indonesia 16424.

PONIDI: Department of Mathematics, University of Indonesia, Ged D FMIPA UI Kampus Baru Depok, Indonesia 16424.
E-mail: ponidiui@yahoo.com.

ARIE WIBOWO: Department of Mathematics, University of Indonesia, Ged D FMIPA UI Kampus Baru Depok, Indonesia 16424.
E-mail: kak_arie@yahoo.com.

RAHMI RUSIN: Department of Mathematics, University of Indonesia, Ged D FMIPA UI Kampus Baru Depok, Indonesia 16424.
E-mail: rahmirusin@yahoo.com.

MATHEMATICAL MODEL OF SKIN COLOUR FOR FACE DETECTION

Setiawan Hadi[a], Adang Suwandi[b], Iping Supriana[b], Farid Wazdi[b]

[a] Universitas Padjadjaran, Bandung, Indonesia
[b] Institut Teknologi Bandung, Indonesia

Abstract. Skin colour in digital image is important element that is very useful for preprocessing task in skin-based automatic image detection. In this paper, skin colour was investigated and its representation in several colour spaces has been explored. A mathematical model for the representation has been developed and successfully implemented for automatic single frontal face detection conjuncted with mathematics morphological and segmentation image processing filters and simple 4-neigbourhood image measurement algorithm. The experiment has been conducted based from our proposed method using face database images that have been collected from several sources, such as standard FERET Colour face database and local native face database.

Key-words: Skin, color, face, detection

1 Introduction

Skin colour is an important element in detecting image that contain skin or skin-like region. Skin colour can be used to detect faces [5, 8, 9] or hands [13, 28], in dynamic images as well as in still images. Skin colour has also been used to detect images of naked people for Internet content altering [3, 6, 7]. In the field of health and disease, skin colour can also be used to analyse medical images. For example, the ability to segment an image using skin colour can aid in the diagnosis of skin cancer [17].

Skin becomes an interesting feature in the field of image detection especially in face detection due to (1) it covers most of the face image area, (2) skin of different people appears to vary over a wide range, however the differ is much less in colour (chromaticity) than brightness [18]. In the computational perspective, detection of skin area in digital image are more practical and easy to implement. In addition, skin color chromaticity distribution from different ethnic groups lie in the same Gaussian distribution [12]. In our research, skin model is investigated in three colour spaces, *rg* space, HSB space and YCbCr space, and represented in histogram and skin colour distribution.

Skin colour detection will give result of skin region and skin-like region. False detection of skin-like colour region can degrade the performance of skin colour detection. To cover this problem, skin colour detection must be followed with

other detection techniques. In our approach, skin color detection in three colour spaces was conducted and compared using skin detector that has been created semi-manually based on the statistical skin model that has been developed from the previous step. The result of skin color detection is followed with two morphological filters, those are erosion, for thinning the image and removing the noise with salt and pepper technique, and dilation, for compacting the binary image representation.

In the following pages, we will describe skin color distribution in the explored color spaces. Description about colour in digital images can be read in [4], meanwhile elaboration of colour spaces and their comparison can be found in detail in [27].

2 Skin Colour Distribution

A color histogram is a distribution of colors in the color space and has long been used by the computer vision community in image understanding. For example, analysis of color histogram has been a key tool in applying physics-based models to computer vision. It has been shown that color histograms are stable object representations unaffected by occlusion and changes in view, and that they can be used to differentiate among a large number of objects. In the mid-1980s, it was recognized that the color histogram for a single in homogeneous surface with highlights will have a planar distribution in color space. It has since been shown that the colors do not fall randomly in a plane, but form clusters at specific points. It has been further observed that (1) human skin colors cluster in a small region in a color space; (2) human skin colors differ more in intensity than in colors, and (3) under a certain lighting condition, a skin- color distribution can be characterized by a multivariate normal distribution in the normalized color space [25]. The figure 1 shows a face image and the skin color distribution in the RGB, HSI and YCbCr colour spaces.

3 Mathematical Model of Skin Colour

Literature research [6, 20, 12] shown that there are three approaches of skin colour modelling, those are empirical approach, statistical approach and adaptive approach. A simple **empirical approach** to colour modelling in general is to use a representative sample of pixels for some target colour to generate a histogram in the selected colour space. A threshold is then selected. Any pixel belonging to a histogram region above that threshold is then classified as being the target colour. The main disadvantages of the histogram approach, when compared to the statistical approach, are that the resulting model is not as compact, and classifying pixels is more computationally demanding [13]. In addition, the size of the sampling bins used to create the histogram affects the performance of the model. The main advantages of the histogram approach are that it is relatively simple to implement, can be applied to non-Gaussian distributions, and does not require detailed knowledge of the skin colour distribution. In the **statistical models**,

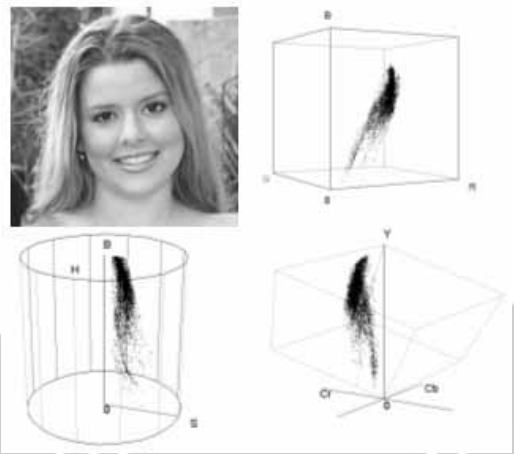


Figure 1: Sample image and its skin colour distribution in colour spaces

the normalised *rg* colour space and various hue and saturation based colour spaces result in a skin colour distribution that, is well modelled with a bivariate Gaussian distribution [19, 25, 23]. There is contradiction about the best statistical model to use for skin colour. Most researchers use a single distribution, while some [24, 26] claim that a single multivariate distribution is inadequate, and a Gaussian mixture model is more appropriate. Other researchers use a Gaussian mixture model, yet observe that a single Gaussian distribution is often adequate [14, 15]. Some researchers have used **adaptive skin colour models** to overcome the limitations between fixed and statistical colour models. Jordao et al. [8] used the variance of the skin colour distribution in the HSV colour space to detect regions in an image as being skin coloured based on the variance of colour within the region. Yang and Waibel [22] used a statistical model in the normalised *rg* colour space, and allowed the model parameters to adapt from one frame to the next in a video sequence. Raja, McKenna, and Gong [15] used an adaptive Gaussian mixture model that modified the model parameters from frame to frame by resampling a window of the image, while ignoring frames that were too different to the current model. Similarly, Rowley [16] used a Gaussian model in the normalised *rg* colour space that resampled areas around the centre of the face to adapt the model parameters from frame to frame.

In our approach, skin color is modelled using mean and covariance of chrominant color. If we use normalize rg space, then the value of r and g are calculated. If HSB, h element and s element are considered. Similar with that, Cb and Cr components are used for modelling skin colour in YCbCr colour space.

3.1 Mean chromaticity calculation

In the HSB colour space, hue ranges from $[0 \ldots 1)$, and wraps around so that $0 \equiv 1$. The mean hue is therefore not calculated in the usual way. First, the hue is mapped onto the range $[0 \ldots 2\pi)$ radians, and then treated as a point on the unit circle in polar coordinates. Hence, it is represented by the polar coordinates $(\theta, 1)$, where $\theta = 2\pi h$, and h is the hue. This point is then converted into Cartesian coordinates, so that for every hue, there is a Cartesian point (x, y) where $x = cos(\theta)$ and $y = sin(\theta)$. The mean x and y values for the sample of n pixels are then calculated in the usual way:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{1}$$

The mean (x, y) location is then used to derive an angle in polar coordinates which is then mapped onto the range $[0..2\pi)$, by adding or subtracting multiples of 2π as necessary, and then mapped onto the range [0, 1) by dividing by 2π. Saturation ranges from [0..1], and does not wrap around, so the mean saturation is calculated in the usual way. The calculation of the angle θ, mean hue ans saturation are

$$\theta = \arctan(\frac{\overline{y}}{\overline{x}}) \qquad \overline{h} = \frac{\theta}{2\pi} \qquad \overline{s} = \frac{1}{n}\sum_{i=1}^{n} s_i \tag{2}$$

For calculations in the normalised rg colour space, the chromaticity is defined as (r, g) instead of (h, s), and the mean chromaticity is calculated in the usual way:

$$\overline{r} = \frac{1}{n}\sum_{i=1}^{n} r_i \qquad \overline{g} = \frac{1}{n}\sum_{i=1}^{n} g_i \tag{3}$$

3.2 Covariance chromaticity calculation

After calculating the mean chromaticity, the covariance matrix is calculated, with the additional constraint that the distance from any hue to the mean hue cannot exceed 0.5. This takes into account the fact that hue wraps around and consequently the maximum distance between any hue and the mean hue is 0.5. For a sample containing n pixels, the covariance matrix for chromaticity is

$$C = \begin{bmatrix} C_{hh} & C_{hs} \\ C_{hs} & C_{ss} \end{bmatrix} \tag{4}$$

where

$$C_{hh} = \frac{1}{n}\sum_{i=1}^{n} \Delta h_i \Delta h_i \qquad C_{hs} = \frac{1}{n}\sum_{i=1}^{n} \Delta h_i \Delta s_i \qquad C_{ss} = \frac{1}{n}\sum_{i=1}^{n} \Delta s_i \Delta s_i \tag{5}$$

and the distances to the mean hue and saturation is defined as

$$\Delta h_i = \begin{cases} h_i - \overline{h} & \text{iff } |(h_i - \overline{h})| < 0.5, \\ h_i - \overline{h} + 1 & \text{iff } (h_i - \overline{h}) \leq -0.5, \\ h_i - \overline{h} - 1 & \text{iff } (h_i - \overline{h}) \geq 0.5 \end{cases} \tag{6}$$

and

$$\Delta s_i = s_i - \overline{s}. \tag{7}$$

Chromaticity is defined as (r, g) for the normalised rg colour space instead of (h, s). The covariance matrix is calculated in the same way as for the HSB colour space by substituting r for h, g for s, $\overline{r}$ for $\overline{h}$, and $\overline{g}$ for $\overline{s}$. The absolute value of $\overline{h}$ is not constrained when using the normalised rg colour space because this space does not wrap around.

3.3 Measuring Distance from the Population

The Mahalanobis distance metric D can be used to measure the distance between a point and a population of points in a way that is sensitive to changes in the variance along the principal axes of the distribution for the population. It is defined [10] as

$$D_{(\vec{p})} = \sqrt{(\vec{p} - \vec{m})^T C^{-1} (\vec{p} - \vec{m})} \tag{8}$$

where $\vec{p} = (h, s)$ is the chromaticity of the pixel being classified, $\vec{m} = (\overline{h}, \overline{s})$ is the mean chromaticity for the population and C^{-1} the inverse of the covariance matrix of chromaticity for the population.
As stated earlier, the covariance matrix is

$$C = \begin{bmatrix} C_{hh} & C_{hs} \\ C_{hs} & C_{ss} \end{bmatrix} \tag{9}$$

If $(C_{hh}C_{ss} - C_{hs}C_{hs}) \neq 0$ (which in practise is usually the case), the covariance matrix is invertible, and the invers [2] is

$$C^{-1} = \frac{1}{C_{hh}C_{ss} - C_{hs}C_{hs}} \begin{bmatrix} C_{ss} & -C_{hs} \\ -C_{hs} & C_{hh} \end{bmatrix} \tag{10}$$

This calculation is the same for the rg colour space using the appropriate chromaticity values and covariance matrix.

4 Face Detection

Face detection can be considered as part of face recognition implementation. Literature shows many researches have been conducted in face detection area that have been reported in survey papers [11, 21]. Skin colour has been used in research as a preprocessing step in face detection. Many methods have been developed and reported in the literature that use skin colour to detect skin and skin-like region in digital image.

Our approach in skin detection has been implemented in three colour spaces, normalized rg, HSB and YCbCr colour spaces. Figure 2 shows the schematic approach of our proposed face detection based on the skin colour representation.

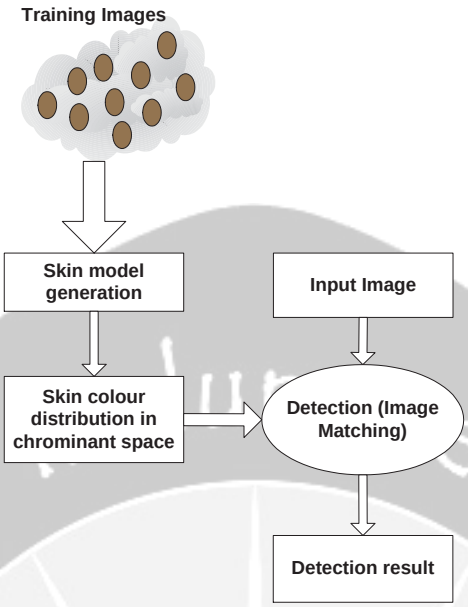


Figure 2: Face Detection Scheme

4.1 Generating skin colour model

To generate skin colour model, mean chromaticity of each training image was calculated for every colour space. Combination of all mean chromaticities for every colour space was calculated using formula 11 and assumed as a general skin model.

$$M_k = \sum_{i=1}^{n} \frac{1}{\delta_i} T_i \tag{11}$$

$$M_k = \frac{1}{\delta_1} T_1 + \frac{1}{\delta_2} T_2 + \cdots + \frac{1}{\delta_{n-1}} T_{n-1} + \frac{1}{\delta_n} T_n \tag{12}$$

M_k is skin model-k, meanwhile T_i is tripixel RGB of skin image-i (training image). δ_i is parameter is interval between 1 and n. If $\sum \delta_i = n$ the process is *averaging*. The amout of training image is presented by n.

Images that will be used as training images are segmented and cropped semi-manually. Figure 3 and figure 4 illustrated sample images before segmentation and the result of segmentation. Figure 5 show combined images of segmented images and their statistical measurement characteristics (mean chromaticities) and visualization in histogram.
Figure 6 shows distribution of skin color in an image for every colour space. It is clearly seen that skin cluster is located in a small area or region (the white small region) of the whole skin distribution (the black large region).

Figure 3: Sample images BEFORE segmentation

Figure 4: Sample images AFTER segmentation

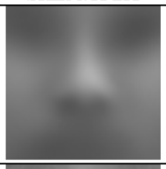
DB set	Facial Skin Model	R Component	G Component	B Component
01				
		μ=138.82 δ=24.39	μ=104.28 δ=19.93	μ=71.02 δ=13.24
02				
		μ=169.70 δ=10.76	μ=120.99 δ=9.01	μ=95.74 δ=7.53

Figure 5: Mean Chromaticities

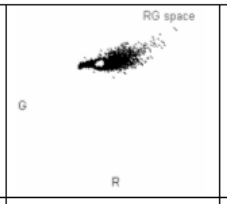
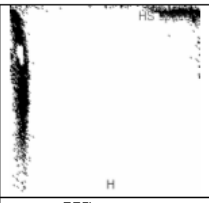
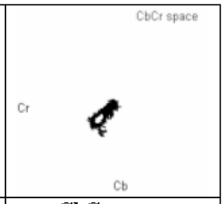
RGB Image	rg space	HS space	CbCr space
			

Figure 6: Skin Colour Distribution in Color Spaces

4.2 Improved skin-based face detection

Our proposed algorithm is based on the skin colour representation of face images. However, this method is insufficient due to pixel that is false detected and supposed as skin pixel but it is not part of the skin. In addition, our focus is on the face detection, so the skin pixels that are detected but not part of face, has to be eliminated. Our basic detection algorithm is

$$P_{skin}(i,j) = P_{skin}(i,j) \in D_{M_k} \quad \forall\, P(i,j) \wedge \forall\, R^n \tag{13}$$

where $P_{skin}(i,j)$ is probability of pixel P as skin pixel if included in distribution skin model D_{M_k} for every colour spaces R^n.

To localize face-only image, two morphological filters are utilized. First, Erosion filter, which is mainly used to remove skin-like detected pixels that can be considered as noise. After image is cleaned, we have to compact the image, by filling the holes in image using dilation filter. The result of this implementation of two filters are image that has clear and compact representation. Detail description on mathematical foundation and its practical implementation of these filters can be found a lot in literatures. After face can be localized, a 4-neighbourhood ellipse generator algorithm is implemented to mark the face region. Figure 7 shows step by step our proposed face detection.

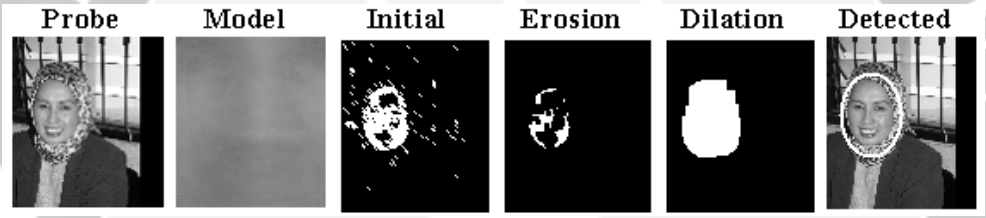


Figure 7: Proposed face detection, step by step

5 Result of experiment

Experiment has been conducted using several face databases that have been collected in various ways. FDB-08 is obtained officially from FERET face database NIST which consists of two DVD-ROM (10 Gigabytes); for this research we just use small number of pictures. FDB-02 was created using 3.3 megapixels digital camera Minolta S304, consists of single photograph of academic and administrative staffs of Mathematics Department UNPAD. FDB-07 is set of face images which have been taken from extension student candidates of UNPAD using Fuji 3.3 megapixels MPIX digital camera. Other database sets have been collected from various multimedia resources and internet. Table 1 shows the detail specification of databases used in the experiment.

Table 1: Face Databases

Group	Image Size	Number of images	Graphic Type	Colour Bits
FDB-01	640x480	754	JPEG	24
FDB-02	1280x960	70	JPEG	24
FDB-03	Vary	285	JPEG	24
FDB-04	Vary	150	JPG, GIF	24, 8, 7
FDB-05	Vary	123	JPG, GIF	24, 8, 4
FDB-06	Vary	32	JPG, GIF	24, 8
FDB-07	140x160	2475	JPG, GIF	24
FDB-08	512x768	114	PPM	24

Table 2: Result of Experiment

Dataset	*rg* space			HSB space			YCbCr space		
	Ins	Outs	%Acc	Ins	Outs	%Acc	Ins	Outs	%Acc
FDB-01	103	521	17	127	503	20	160	513	24
FDB-02	52	18	74	48	22	69	57	13	81
FDB-03	149	114	57	135	128	51	153	110	58
FDB-04	103	46	69	95	54	64	92	57	62
FDB-05	72	50	59	67	55	55	74	48	61
FDB-06	21	5	81	21	5	81	18	8	69
FDB-07	1371	1021	57	1135	1221	48	1376	897	61
FDB-08	59	55	52	64	50	56	68	46	60

The goal of this experiment is to measure the detection performance of our proposed face detection algorithm. In this report, we use simple qualitative performance measurement by using human-eye visual examination. The referenced skin model was generated from FDB-02 face database (see figure 7). We define the result into two categories, (i) the target face is in ellipse and (ii) target face is outside the ellipse. The result can be shown in table 2. In advance measurement, a face detection system system makes two types of errors [1]: (i) mistaking measurement of non face region which is detected as face (called false match or false accept), and (ii) mistaking in measurement of a face region which is not detected as face (called false non-match or false reject). There is a trade-off between false match rate (FMR) and false non-match rate (FNMR) in every face detection system. In fact, both FMR and FNMR are functions of the system threshold t; if t is decreased to make the system more tolerant to input variations and noise, then FMR increases. On the other hand, if t is raised to make the system more secure, then FNMR increases accordingly. The system performance at all the operating points (thresholds, t) can be depicted in the form of a Receiver Operating Characteristic (ROC) curve.

6 Concluding Remarks

We have presented the theoretical background of mathematical skin colour modelling and its practical implementation in the area of face detection. The proposed approach utilizes of a skin-color detector to detect the skin region, as the first step (initial) detection. To localize the face region, morphological filters, erosion and dilation, are used, conjuncted with 4-neigbourhood ellipse generation algorithm. Experiment has been performed in three colour spaces using more than 1 Gigabytes data of face databases.

It is shown that the result was not very accurate in general. This problem arised due to the use of only one skin model (from FDB-02) which perhaps not appropriate for other face database sets. Nonetheless, the future research will involve the use of other techniques to enhance and assist in face localization and also to perform face tracking such as: detection of symmetricity and skewness of faces(by using feature point detection), multiple face detection (using clustering), general skin detector generation , non frontal face detection, advance measurement of face detection accuracy using ROC curve, interactive detection of face(s), 3D generation of detected face(s), and other enhancement.

References

[1] Anil K. Jain, Arun Ross and Salil Prabhakar. An Introduction to Biometric Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, Jan 2004.

[2] Anton, H. *Elementary Linear Algebra*. 4th ed. John Wiley & Sons, Inc, New York, USA, 1984.

[3] Forsyth, D. A. and Fleck, M. M. Automatic detection of human nudes. *International Journal of Computer Vision*, vol. 32 63–77, Aug 1999.

[4] Setiawan Hadi. Laporan Kemajuan Penelitian - Pengembangan Metode 1. *Departemen Informatika ITB*, Jun 2005.

[5] Hunke, M. and Waibel, A. Face locating and tracking for human computer interaction. *In Proceedings of the 28th Asilomar Conference on Signals, Systems & Computers*, Nov 1994.

[6] Jones, M. J. and Rehg, J. M. Statistical color models with application to skin detection. *Tech. Rep. CRL-98/11, Compaq - Cambridge Research Laboratory*, Dec 1998.

[7] Jones, M. J. and Rehg, J. M. Statistical color models with application to skin detection. *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 3656, 1999.

[8] Jordao L., Perrone M., Costeira J. P. and Santos-Victor J. Active face and feature tracking. *In Proceedings of the 10th International Conference on Image Analysis and Processing*, Sep 1999.

[9] Kim, S.-H. and Kim, H.-G. Face detection using multi-modal information. *In Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 14–19, Mar 2000.

[10] Manly, B. F. J. *Multivariate Statistical Methods: A Primer*. Chapman and Hall, 1986.

[11] Ming-Hsuan Yang, David J. Kriegman, Narendra Ahuja. Detecting Faces in Images: A Survey. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, Jan 2002.

[12] David Thomas John O'Mara. Automated Facial Metrology. *PhD Dissertaion Univ. of Western Australia*, Feb 2002.

[13] Pavlovic V. I., Sharma R. and Huang T. S. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 7, no. 19, pp. 677–695, 1997.

[14] Raja Y., McKenna S. J. and Gong S. Segmentation and tracking using colour mixture models. *In Proceedings of the Asian Conference on Computer Vision*, vol. 1 607–614, 1998.

[15] Raja Y., McKenna S. J. and Gong S. Tracking and segmenting people in varying lighting conditions using colour. *In Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 228–233, Apr 1998.

[16] Rowley, H. A. Neural network-based face detection. *Tech. Rep. CMU-CS-99-117, Carnegie Mellon University*, May 1999.

[17] Schmid, P. *Segmentation & symmetry measure for image analysis: Application to digital dermatoscopy*. PhD thesis, École Polytechnique Fédérale de Lausanne, 1999.

[18] Mariań Sedlaćek. Evaluation of RGB and HSV models in Human Faces Detection. *Slovak University of Technology*.

[19] Terrillon, J. and Akamatsu, S. Comparative performance of different chrominance spaces for color segmentation and detection of human faces in complex scene images. *In Vision Interface '99*, 1999.

[20] Terrillon J., David M. and Akamatsu S. Detection of human faces in complex scene images by use of a skin color model and of invariant Fourier-Mellin moments. *In Proceedings of the 14th International Conference on Pattern Recognition*, vol. 2 1350–1355, Aug 1998.

[21] W Zhao, R Chellappa, A Rosenfeld and P J Phillips. Face Recognition: A Literature Survey. *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, Dec 2003.

[22] Yang, J. and Waibel, A. A real-time face tracker. *In Proceedings of the IEEE Workshop on Applications of Computer Vision*, pp. 142–147, 1996.

[23] Yang, M. and Ahuja, N. Detecting human faces in color images. *In Proceedings of the 1998 IEEE International Conference on Image Processing*, vol. 1 127–130, Oct 1998.

[24] Yang, M. and Ahuja, N. Gaussian mixture model for human skin color and its applications in image and video databases. *In Proceedings of SPIE-the International Society for Optical Engineering: Conference on Storage and Retrieval for Image and Video Databases VII*, vol. 3656 458–466, 1999.

[25] Yang J., Lu W. and Waibel A. Skin-color modeling and adaptation. *Tech. Rep. CMU-CS-97-146, Carnegie Mellon University*, May 1997.

[26] Yang M., Ahuja N. and Kriegman D. Face detection using mixtures of linear subspaces. *In Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 70–76, 2000.

[27] Zarit B. D., Super B. J. and Quek F. K. H. Comparison of five color models in skin pixel classification. *In Proceedings of the IEEE International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-time Systems*, 1998.

[28] Zhu X., Yang J. and Waibel A. Segmenting hands of arbitrary color. *In Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 448–455., 2000.

Setiawan Hadi: S3 student at Department of Informatics, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia.
Lecturer at Dept. of Mathematics UNPAD, Phone/Fax: +62 +22 779 4696.
E-mail: setiawanhadi@ieee.org

Adang Suwandi Ahmad: Professor at Department of Electrical Engineering, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia.
E-mail: asa@isrg.itb.ac.id

Iping Supriana Suwardi: Senior Lecturer at Department of Informatics, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia.
E-mail: iping@informatika.org

Farid Wazdi: Senior Lecturer at Department of Informatics, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia.
E-mail: farid@informatika.org

A SIMULATION STUDY FOR ESTIMATION OF TORQUES AND BODY SEGMENT PARAMETERS FROM BIOMECHANIC DISPLACEMENT DATA USING OPTIMAL CONTROL

S. Munzir[a], L.S. Jennings[a], M.T. Koh[b]

[a] School of Maths and Stats, The University of Western Australia
[b] National Institute of Education, Singapore

Abstract. This work is a preliminary simulation study of the use of dynamic optimization for biomechanical systems to simultaneously obtain joint torques and body segment parameters, while smoothing kinematic raw data of biomechanic subjects obtained from film or video. Angular displacement data is generated through forward integration, by intuitively assigning an appropriate set of torques to segments over a time horizon. The raw data smoothing approach is implemented in the form of an additional inequality constraint similar to the splines smoothing. The formulation is tested on a two segment rigid body with the assumption that the mass is uniformly distributed as a rod. The simulation results show the formulation capabilities and weaknesses in computing torques and body segment parameters from noisy simulated data.
Key-words: biomechanics, optimal control, dynamic optimization, BSP estimation.

1 Introduction

The study of biomechanical system dynamics consists of research to obtain an accurate model of biomechanical systems and to find appropriate torques or forces that reproduce motions of a biomechanic subject. The latter part of the study usually uses inverse dynamics or dynamic optimization as the tool to compute torques or forces that produce the angular coordinates of each body segment that is relatively close to kinematic data. Inverse dynamics tends to produce noisy resultant joint moment (RJM) estimates due to inherent systematic and random errors associated with displacement-time data obtained from film or video [6].

The displacement data usually consists of the real signal and low amplitude-high frequency noise, which, for arguements sake, can be assumed to have a mathematical form as sinusoidal term $A \sin \omega t$ with a small A and a large ω value. This noise term does not give significant differences to the measured displacement data since the amplitude is assumed to be small. However, the amplitude ωA of its derivative is very large for a sufficiently large value of ω, and this becomes larger for the higher derivatives. Hence, this noise gives very significant change to the velocity and acceleration values which are used to compute torques or forces. Therefore, several techniques for kinematic raw data smoothing have been proposed to solve

this problem. As mentioned in [13], among the most famous methods are: Butterworth digital filter [10, 14], cubic spline [7, 11], quintic spline [15], and Fourier series [1].

On the other hand, dynamic optimization/optimal control method may include estimation of forces and torques of each body segment with the cost function in the form of minimize the jerk or maximize the smoothness [8]. Therefore, the noisy RJM estimates could be avoided through several mechanism in the objective or constraints of the optimal control problem. This approach does not require a split between kinematic data smoothing, velocity and acceleration computation, and RJM estimates. Instead, these could be done simultaneously within the optimal control computation.

The research using optimal control becomes more interesting and challenging because most of biomechanic optimal control problem also need to estimate body segment parameters (BSP) whose values may not be accurately known. Usually these parameters are assumed to be known and fixed prior to the computation of the optimal control problems. The practical process of body parameter estimation alone is a complex process, involving approximated volume computation of each segment and segment density estimation using sophisticated equipment. There are several techniques to estimate the BSP as a separate process prior to the torque or force estimation using optimal control. For instance, in [8], mass, center of mass, and moment of inertia for each body segment are assumed as for an adult male using a regression equation from literature [16, 12]. In [6] BSP were obtained from elliptical zone mathematical modeling technique due to [4], which uses total body mass (TBM) as the basis for BSP estimation.

All available techniques for BSP estimation are subject to error, because of the existence of variation in BSP based on race, sex, body type and age. The error in BSP may lead to a significant change in biomechanic dynamic analysis. Pearsall and Costigan [9] examined the effect of BSP error on gait analysis results, and found that comparisons between six different predictive formulae led to a 40 percent difference in mass and inertia predictions. They also stated that the need for accurate BSP values is greatest in movements involving high accelerations or in open chain movements.

This work is aimed as a preliminary study of using dynamic optimization for biomechanical systems to simultaneously obtain joint torques and BSP, while smoothing kinematic raw data of biomechanic subjects obtained from film or video. This work is different from the previous work [6] where the 'closeness' of the motion and smoothed kinematic data is treated as the objective of the optimization. Here this closeness is between the motion and kinematic raw data, and is treated as inequality constraints where the measurement of closeness of fit could be decided. The smoothing approach is similar to the cubic and quintic splines presented in [2]. In this way, not only the torques or forces which reproduce similar motion can be estimated, also smaller or smoother torques reproducing similar motion can be

estimated in the case where degrees of freedom exist.

Here, this approach is applied to a two segment rigid body where the 'exact' historical data for torques is given and historical angular displacement data is generated through forward integrations of the multibody dynamic equations. The angular displacement data is then transformed into distal end coordinates of the first and second segment, and this data is then rounded to create small amplitude-high frequency noise in (x, y) data. The inequality constraints for the optimization are the coordinate differences between this 'noisy' data and the actual distal end coordinate produced by the optimal control computation (integrating the ode from initial condition).

2 Problem Formulation

Given the $\{(t_i, \hat{x}_i^j, \hat{y}_i^j); i = 1, \ldots, N\}$ as the 2-D displacement data of the j-th body segment obtained from video or film, the purpose is to find accurate torques and BSP that reproduce a smoothed motion close to the time-space history of the joints. This closeness is represented as the inequality constraints:

$$h_j = \sum_{i=1}^{N} \left(\frac{x_i^j - \hat{x}_i^j}{\sigma_i} \right)^2 + \left(\frac{y_i^j - \hat{y}_i^j}{\sigma_i} \right)^2 - S \leq 0,$$

where σ_i is the standard error in $\hat{x}_i^j$ and $\hat{y}_i^j$, assumed known. This is used in time series spline smoothing. For an n segment biomechanical model, being an open chain, $j \in \{1, \ldots, m\}$, where $m = n + 1$ for the free-flight case, and $m = n$ for the case of one fixed contact at the proximal end of segment one. The parameter S controls the overall balance between smoothing and closeness of fit. The values of (x_i^j, y_i^j) are obtained as the result of translational and rotational motion of the whole segmented body. The augmented equation of motion for 2-D segmented bodies resulting from the translational and rotational equation of motion with one contact at the proximal end of segment one is given by

$$\boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{\alpha} = \boldsymbol{T}\boldsymbol{\tau} + \boldsymbol{D}_c\boldsymbol{n} + (\boldsymbol{D}_c\boldsymbol{E}\boldsymbol{D}_s - \boldsymbol{D}_s\boldsymbol{E}\boldsymbol{D}_c)\boldsymbol{\omega}^2.$$

where $\boldsymbol{\omega} = \dot{\boldsymbol{\theta}}$, $\boldsymbol{\alpha} = \dot{\boldsymbol{\omega}}$

$$\boldsymbol{Q} = \boldsymbol{J} + \boldsymbol{D}_s\boldsymbol{E}\boldsymbol{D}_s + \boldsymbol{D}_c\boldsymbol{E}\boldsymbol{D}_c,$$

and

$$\begin{array}{ll} \boldsymbol{E} = \boldsymbol{L}^t\boldsymbol{D}_m\boldsymbol{L}, & \\ \boldsymbol{J} = diag(I_1, I_2, \ldots, I_n), & \boldsymbol{D}_c = diag(\cos\theta_1, \cos\theta_2, \ldots, \cos\theta_n), \\ \boldsymbol{D}_m = diag(m_1, m_2, \ldots, m_n), & \boldsymbol{D}_s = diag(\sin\theta_1, \sin\theta_2, \ldots, \sin\theta_n). \end{array}$$

Here $\boldsymbol{L}$ is a lower triangular matrix with i-th row representing the distance of the CoM (center of mass) of the i-th segment to the proximal end of the first segment.

The complete derivation of the model is found in [5], where the modification for free flight can be found also.

The optimal control statement for this problem can be written in the following form:

$$\text{minimize}_{\boldsymbol{\tau},\boldsymbol{z}}\colon \tilde{G}(\boldsymbol{\tau},\boldsymbol{z}) = \int_0^1 ((\boldsymbol{\tau}_1)^2 + (\boldsymbol{\tau}_2)^2)dt. \tag{1}$$

subject to the equations of motion for an n segmented body in state space form

$$\dot{\boldsymbol{\theta}} = \boldsymbol{\omega}\boldsymbol{Q}(\boldsymbol{\theta},\boldsymbol{z})\dot{\boldsymbol{\omega}} = \boldsymbol{f}(\boldsymbol{\theta},\boldsymbol{\omega},\boldsymbol{\tau},\boldsymbol{z}),$$

with initial condition

$$\boldsymbol{\theta}(0) = \boldsymbol{\theta}^0, \boldsymbol{\omega}(0) = \boldsymbol{\omega}^0.$$

subject to the inequality constraints for closeness of fit:

$$h_j = \sum_{i=1}^{N} \left(\frac{x_i^j - \hat{x}_i^j}{\sigma_i}\right)^2 + \left(\frac{y_i^j - \hat{y}_i^j}{\sigma_i}\right)^2 - S \leq 0, \tag{2}$$

for $j \in \{1,\ldots,n\}$.

The objective of the optimization in (1) is to minimize the magnitude of both joint torques, and this is usually related to the cost/energy of the motion. Objectives of the optimization for this kind of problem can be set up in many ways depending on the necessity. For instance, since most of biomechanic subjects moves in a smooth motion, the objective could be chosen as minimizing the change of angular acceleration. It is also possible to set the objective as smoothing torque or the derivative of torque. In [3] this is implemented through first order or second order regularization. In fact, by choosing the later form of objective, the process of kinematics data smoothing to obtain accurate and smooth angular velocity and acceleration and then using inverse dynamics to obtain accurate and smooth torque has been unified into a single computation process. There is also the possibility of choosing zero objective, if the purpose is just to find a feasible torque for a specific motion.

The system parameters $\boldsymbol{z}$ represent the BSP, such as: moment of inertia, mass, length to center of mass, and length of all segments of the rigid body. These parameters are to be estimated within reasonable upper and lower limits. In a real case, the anthropometric estimated BSP values could be used as the initial value of this parameter, with the expectation that the optimization could overcome its inaccuracy.

Integration of the equation of motion produces the value of $\boldsymbol{\theta}_i$, angular position vector, of each segment at every sample time i where $i \in \{1,\ldots,N\}$. The value of x_i^j and y_i^j are obtained from

$$x_i^j = \sum_{k=1}^{j} l_k \cos\theta_i^k, y_i^j = \sum_{k=1}^{j} l_k \sin\theta_i^k,$$

where l_k is the length of segment k and θ_i^k is the angular position of segment k at the i-th sample time. This result is substituted into inequality (2) for the constraints evaluation.

With proper choice of the value of S, the inequality constraints (2) function as smoothing the computed data to the closest feasible value to the noisy historical data over all sampling times. This could be achieved if the optimization choose accurate value of BSP and torques which contribute to the evaluation of the position coordinate of distal end of each segment. However, too small value of S (for example less than N) may lead to infeasibility for the optimization routine via the constraints (2). On the other hand if the value of S is too big, the solution may be oversmoothed, allowing computed BSP and torques to not represent the actual biomechanic subject motion. Since these inequality constraints (2) involve each i-th sampling time, this requires the multiple characteristic time implementation (details can be found in [3]) in these constraints. In the case of uniform sampling time, this could simplify the constraint implementation in the optimization software.

3 Results and Discussion

To investigate the effectiveness of the simultaneous kinematic data smoothing and BSP estimation using dynamic optimization, the formulation is tested on a two segment rigid body. The data is generated using MISER3.3 for two segments of mass 1 and 0.5 kgs, length 1 and 0.5 metres repectively, with the assumption that mass is uniformly distributed along the segment (center of mass 0.5 and 0.25 from proximal end of each segment). The initial position of the segment is horizontally straight with initial angular position $\theta_1(0) = \theta_2(0) = 0$. The generated torques are carefully chosen to prevent body segments from falling rapidly due to gravitation. The generated data which is in the form of angular displacement is transformed into coordinates of the distal end of body segments. This is in order that data resemble the real data that is obtained using video or camera, where the motion of a kinematic subject is captured through limited indication marks pasted at the distal/proximal end of each body segment. In addition, the noise should also be imposed on this data to make it more realisticly represent the actual digitised data and this could be done by rounding the data to a certain digit of significance. This noisy data represent the $\hat{x}_i^j$ or $\hat{y}_i^j$ in inequality constraint (2).

As the total mass M of a biomechanic subject is usually known, an additional equality constraint for the total mass can easily be formulated as:

$$\sum_{i=1}^{n} m_i - M = 0 \tag{3}$$

At the beginning, the problem was computed using only two constraints (2) and (3). However, the result was not convincing although the solution obtained in the optimization is a regular point. The solution indicates that the moment of inertia

of the first segment always hit the upper bound on this parameter, even though the upper bound has been set unrealisticly large. Similarly, the center of mass for the second segment also hit the lower bound even though it is set as close as possible to zero. This condition should be avoided, because if the solution of the parameter value is within the bounds, then the active constraints at the solution should be only the total mass equation and the smoothing constraints. This motivates the consideration for other possible constraints to prevent BSP bounds from becoming active.

The first idea is to put constraints between the individual masses and the moment of inertia of each segment as these are proportional. However, this constraint depends on mass distribution on each segment. Two extreme cases are; the case where all the mass is located in balance at each end of the segment which gives the maximum value for moment of inertia ($mL^2/4$) and the case where all the mass is located at the center of mass which give the lower bound value of zero. The case of uniform mass distribution is in the middle of the two. Assuming uniform mass distribution, the following constraints are added

$$12I_i - m_i L_i^2 = 0, \quad i = 1, 2, \ldots, n \tag{4}$$

The radius to the center of mass and length are related by mass distribution too. With all the mass at the center of mass, the radius can be anything from zero to L. On the other hand, if all the mass is located at the two ends, the amounts of mass at each end determine exactly the position of r. With a uniform mass distribution, the center of mass is exactly at L/2 and this could be represented in constraints as

$$2r_i - L_i = 0, \quad i = 1, 2, \ldots, n \tag{5}$$

This means that the value of r_i will depend on L_i, whose values are determined from the quadratic inequality constraints. The value of I is determined from the sum of masses constraint. This gives the optimal control an additional interesting work to accurately distribute the mass and the torques to each segment. As for the real case, the distribution of mass in each body segment becomes another challenging formulation. After introducing all constraints in the optimization, the two segment problem has two inequality constraints and five equality constraints (all 7 constraints). Hence, for an n segment problem, the overall number of constraints is $3n+1$, with n inequality constraints and $2n+1$ equality constraints. Constraints (3)-(5) represent $2n + 1$ equality constraints on $4n$ parameters.

The results of optimal control computation for this complete formulation are shown in Figures 1-3. Figure 1 shows a comparison between the generated torque (assumed as the actual torques) and torques computed via optimal control. As expected, the torques for the second joint follow the actual torques closely since they are uniquely determined by the quadratic inequality constraint of the distal end of the second segment. However, the computed torques of the first joint are further from the actual torques because they are not only determined by the inequality constraint to be satisfied, but also depend on the computed torques on the second

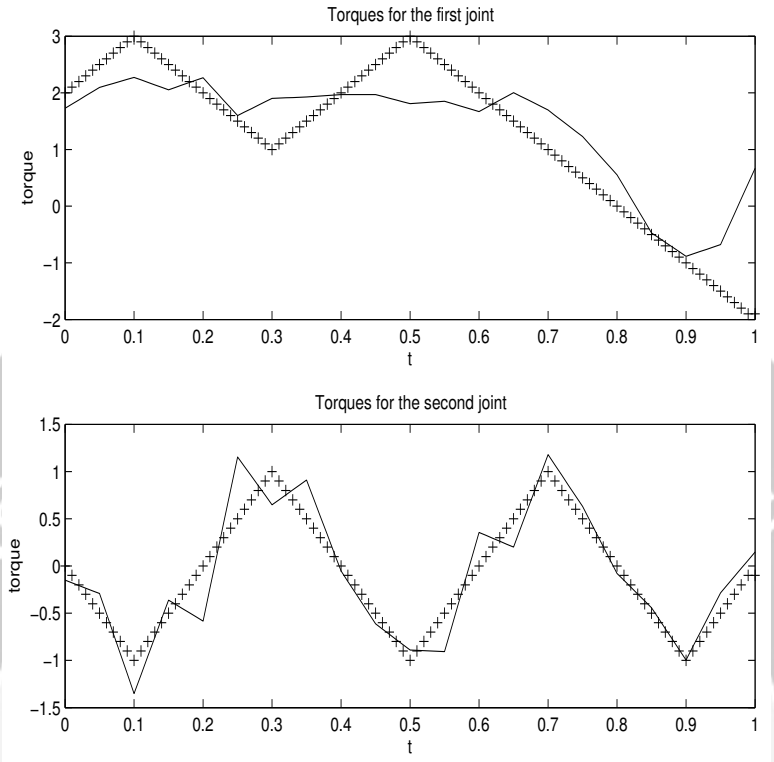


Figure 1: Graph of comparison between smoothed and generated torques for 21 control parameters

link. Since the magnitude of the torque for the first segment is significantly larger than for the second segment, the optimization does not seem to do much work to minimize the torque for the second segment. Indeed, the figure show that the minimization is only done for the first torque and it let the second torque vary within relatively smaller amplitude than the first torque. All of this result is obtained using 21 piece-wise linear control for each control variable. In simulation, it is also found that much better torques representation of generated torque could be obtained by setting control derivative continuity exactly the same for both torques(7 parameter piece-wise linear control), as shown in Figure 2.

Figure 3 shows that the optimization has computed torques and BSP which produces the motion that is close to the 'real' generated data. The smoothed distal end coordinate of both segment one and two fluctuates with small amplitude and low frequency near the real coordinate from generated data. With all constraints active at the solution of the optimization (including the inequality constraints) and none of the BSP parameter at their bounds, this state that the optimization has found a regular point that smooths the motion with the current setting of $\sigma^2 S$.

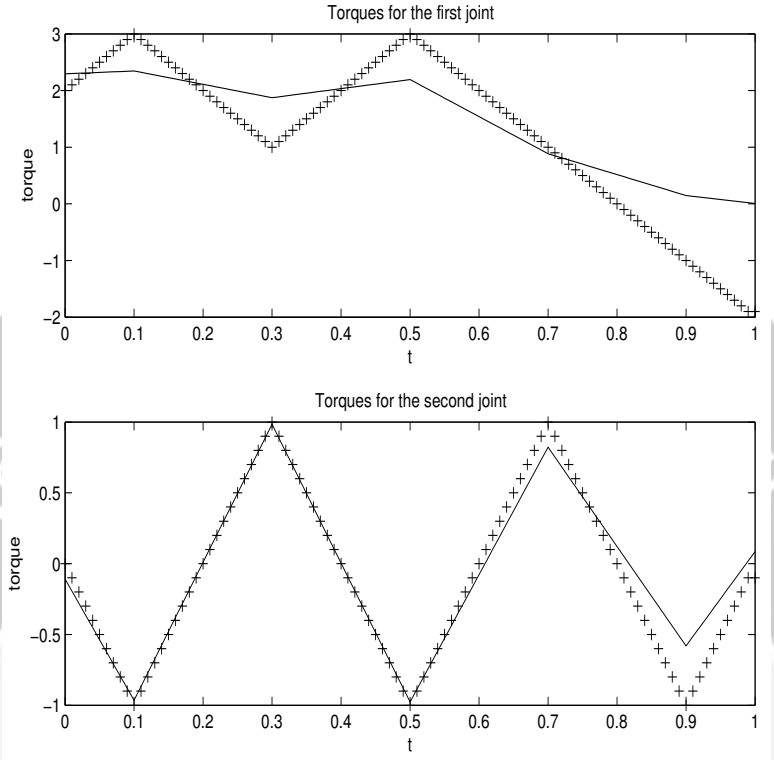


Figure 2: Graph of comparison between smoothed and generated torques for 7 control parameters

Table 1: Comparison of body segment parameters

BSP	*First Segment*		*Second Segment*	
	Actual	*Computed*	*Actual*	*Computed*
Mass	1	0.9209	0.5	0.579087
Segment length	1	1	0.5	0.499722
Center of mass	0.5	0.500034	0.25	0.249861
Moment of inertia	0.08333	0.0767532	0.0104125	0.0120509

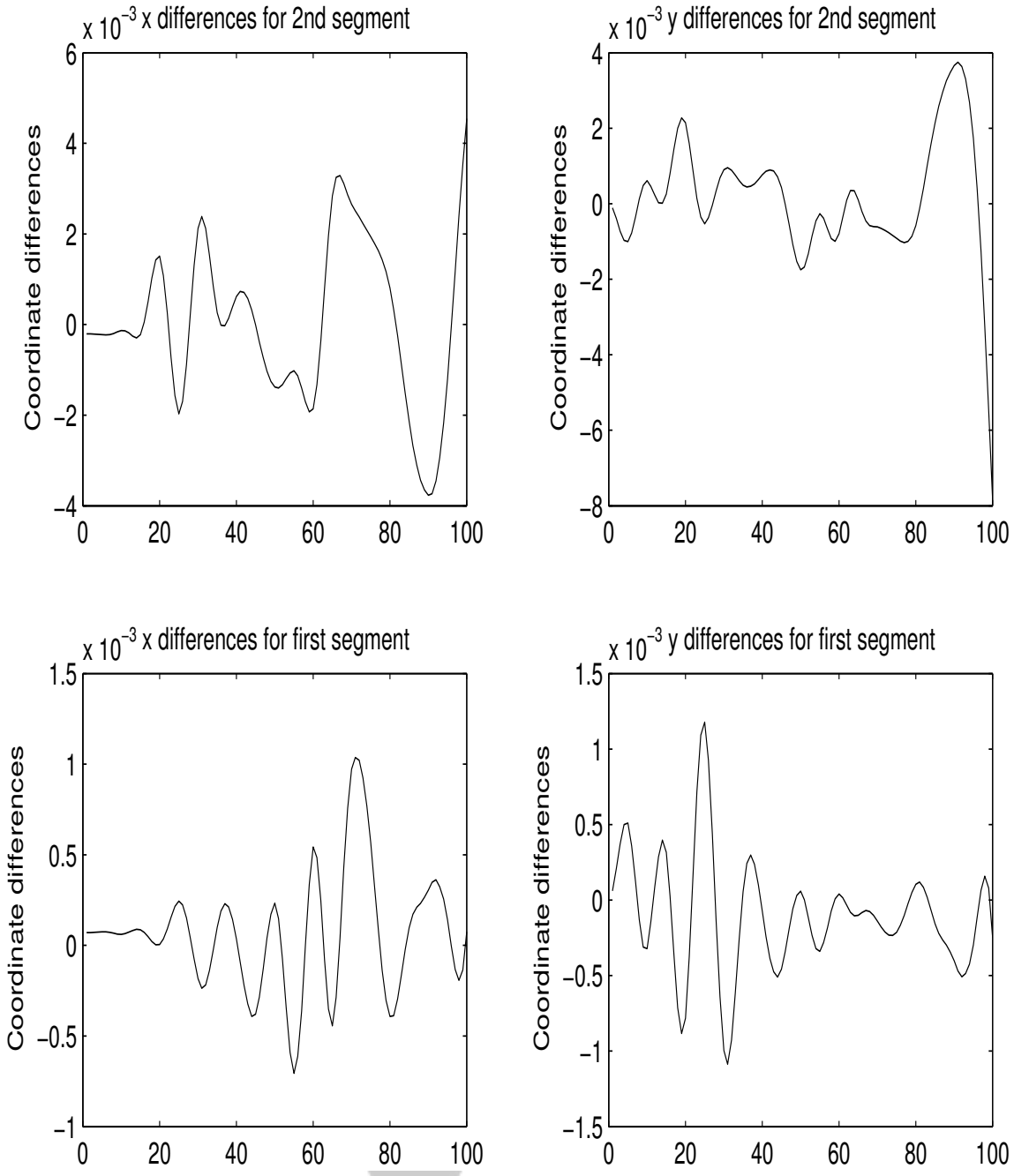


Figure 3: Graph of distal end coordinate difference between smoothed and generated data

This parameter is set as 0.000045 for the first segment and 0.000145 for the second segment. The bigger value has to be assigned to the second segment inequality constraint since it probably carry inherently the deviation (error) of the first segment in the left hand side of the constraint. This tendency has been obtained during running of the problem, where small decrease for the S value of the 2nd segment inequality constraints lead to infeasibility of this constraint. Increasing either one or both of the S parameter values can lead to a solution which allows the smoothed data to more deviate from the generated data and, if the increase is too large, the BSP could hit their upper or lower bounds.

The difference between actual and computed length and center of mass of both first and second segment is relatively small as shown in Table (1). However, this is not the case for mass and moment of inertia. Mass of the first segment is significantly smaller than the actual value because the optimization tend to choose smaller torques at the cost of smaller mass and also moment of inertia. As the result, because of the total mass constraint, it decrease significantly the mass of the second segment. This happened because the magnitude of the torques on the first segment is significantly larger than the second one.

Introducing noise to data was done through rounding to an accuracy of 0.005. This created data with the maximum standard deviation between rounded and generated data σ of 0.0025. With 100 data points, this gives the maximum acceptable $\sigma^2 S$ is 0.000625 (assuming that $S = N$). To make the optimal control reasonably smooth the noisy-rounded data while choosing acceptable BSP, the value of $\sigma^2 S$ should be slightly smaller than 0.000625.

From the active constraints we should be able to work out whether there is any degeneracy between minimizing torques and the body parameter constraints. Since the Jacobian of all active constraints (including body parameter constraints) is of full rank (without rank deficiency) at the solution, it show that the solution is a non-degenerate solution. At the moment the restriction on body parameters bears no relation to the physics of the human body or of a robot. However for the real experiment, this formulation may become an interesting phenomena to be explored.

4 Conclusion

From the simulation result, it is shown that the problem of simultaneous estimation of BSP together with torques that reproduce a biomechanic subject motion is an under-determined constrained problem unless additional constraints on BSP are added. It mean that, without additional BSP constraints, the optimization has the flexibility to choose different kinds of mass distribution for each segment. This provide a very significant clue to the implementation of this method to the real biomechanic systems, where the formulation of BSP constraints become necessary and yet challenging. It is also found that the objective function has a significant

influence on the BSP selected in the optimization, especially the selected mass and moment of inertia.

References

[1] Anderssen, R.S., and P. Bloomfield (1974), Numerical differentiation procedures for non-exact data, *Numer. Math.*, **22**,1157 – 1182.

[2] Jennings, L.S. (2004), *CSS: Constrained Simultaneous Smoothing of Time Series Using Cubic and Quintic Splines*, CADO Reports - UWA .

[3] Jennings, L.S., M.E. Fisher, K.L. Teo, and C.J. Goh (2004), *MISER3: Optimal Control Software, Theory and User Manual, Version 3.*, http://www.maths.uwa.edu.au/u/les .

[4] Jensen, R.K.(1978),Estimation of the biomechanical properties of three body types using a photogrammetric method, *Journal of Biomechanics*, **11**,349 – 358.

[5] Koh, M.T. and L.S. Jennings (2002), Dynamic optimization: A solution to the inverse dynamic problem of biomechanics using MISER3, *Dynamics of Continuous, Discrete and Impulsive Systems, Series B, Applications and Algorithms*, **9B**, 3, 369 – 386.

[6] Koh, M.T., L.S. Jennings, B. Elliott and D. Lloyd (2003), A predicted optimal performance of the Yurchenko layout vault in women's artistic gymnastics, *Journal of Applied Biomechanics*, **19**, 187 – 204.

[7] McLaughlin, T.M., C.J. Dillman, and T.J. Lardner (1977), Biomechanical analysis with cubic splines, *Res. Q.*, **48**, 569 – 582.

[8] Menegaldo, L.L., A.D.T. Fleury and H.I. Weber (2003), Biomechanical modeling and optimal control of human posture, *Journal of Biomechanics*, **36**, 1701 – 1712.

[9] Pearsall, D.J. and P.A. Costigan (1999), The effect of segment parameter error on gait analysis results, *Gait and Posture*, **9**, 3, 173 – 183.

[10] Pezzack, J.C., R.W. Norman and D.A. Winter (1977), An assessment of derivative determining techniques used for motion analysis, *Journal of Biomechanics*, **10**, 377 – 382.

[11] Soudan, K. and P. Dierckx (1979), Calculation of derivatives and Fourier-coefficients of human motion data while using spline functions, *Journal of Biomechanics*, **12**, 21 – 26.

[12] Vaughan, C.L., B.L. Davis and J.C. O'connor (1992), *Dynamics of Human Gait*, Human Kinetics Publisher, Chicago.

[13] Vint, P.F. and R.N Hinrichs (1996), Endpoint error in smoothing and differentiating raw kinematic data: an evaluation of four popular methods, *Journal of Biomechanics*, **29**, 12, 1637 – 1642.

[14] Winter, D.A., H.G. Sidwell and D.A. Hobson (1974), Measurement and reduction of noise in kinematics of locomotion, *Journal of Biomechanics*, **7**, 157 – 159.

[15] Wood, G.A. and L.S. Jennings (1979), On the use of spline functions for data smoothing, *Journal of Biomechanics*, **12**, 477 – 479.

[16] Yeadon, M.R. and M. Morlock (1989), The appropriate use of regression equations for the estimation of segmental inertial parameters, *Journal of Biomechanics*, **22**, 683 – 689.

S. Munzir: PhD student at School of Mathematics and Statistics, The University of Western Australia, 35 Stirling Highway, Crawley 6009, Western Australia.
Phone: +61 8 64887035, Fax: +61 8 64881028
E-mail: msaid@maths.uwa.edu.au

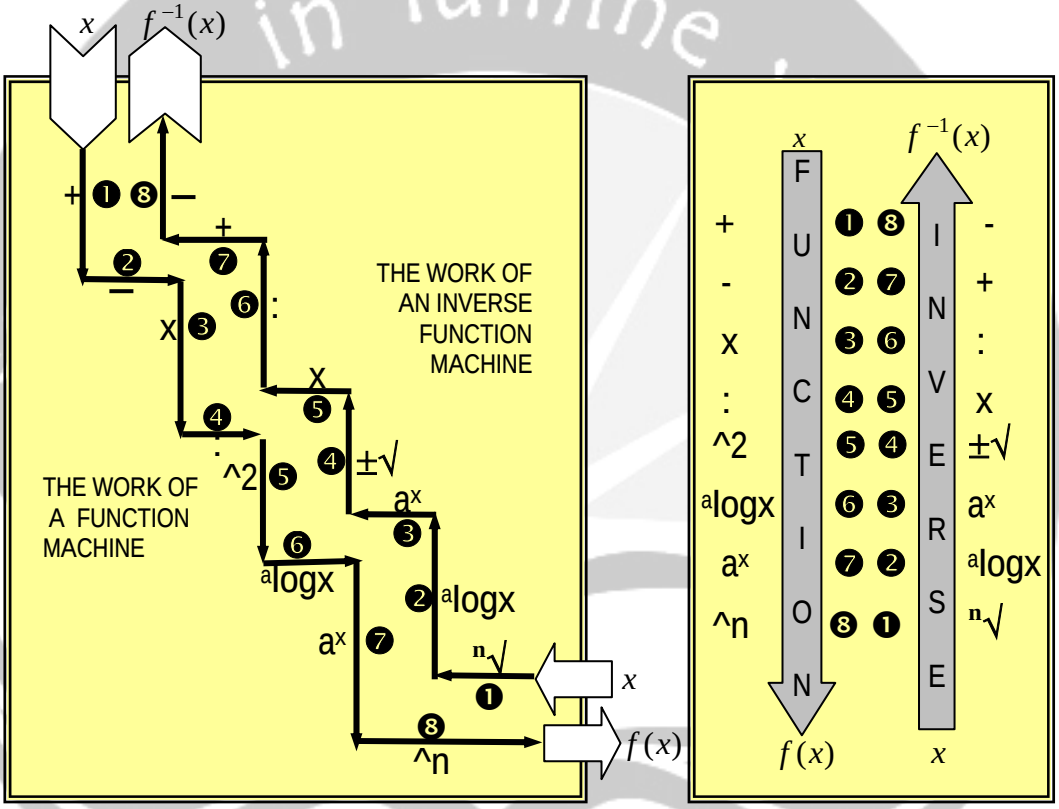
L.S. Jennings: School of Mathematics and Statistics, The University of Western Australia, 35 Stirling Highway, Crawley 6009, Western Australia.
Phone: +61 8 64883361, Fax: +61 8 6488 1028
E-mail: les@maths.uwa.edu.au

M.T. Koh:Physical Education and Sports Science, National Institute of Education,1 Nanyang Walk, Singapore.
E-mail: thmkoh@nie.edu.sg

REALISTIC MATHEMATICS EDUCATION (RME) IN THE SUB-TOPIC OF DECIDING AN INVERSE FUNCTION (*THE REVERSIBLE JOURNEY METHOD*)

SUDARMOYO, S.Pd.
SMA Negeri 2 Kuningan, Indonesia

Abstract. Realistic Mathematics Education (RME) in The Sub-Topic of Deciding An Inverse Function (*The Reversible Journey Method*) is an innovation in education that interlinks a function as journey to go and an inverse function as a journey to come back. The thinking framework can be simplified through this sketch :



According to the writer's experience in SMA Negeri 2 Kuningan, The result of applying Realistic Mathematics Education (RME) in The Sub-Topic of Deciding An Inverse Function (*The Reversible Journey Method*) are as follows : (1) *The teaching process proceeds more interactive, students are active and feel the teaching approach is more realistic.* (2) *Students feel that deciding an inverse function by "The Reversible Journey Method" is easier and faster.* (3) *Students' success in deciding an inverse function by "The Reversible Journey Method" is good.*

1. INTRODUCTION

1.1. BACKGROUND

Curriculum of 2004 (*Competency Based Curriculum*) has been valid nationally since 2004 – 2005 academic year. But in fact, the process in mathematics education in school is still using the conventional approach (*Active theacher and passive student*).

Some mathematicians think that the active and realistic mathematics educations are the requarements for the effective education. According to Prof. Hans Freudenthal's that was

quoted by Zulkardi (2000), the basic phylosopies of realistic mathematics education are : (1) *mathematics is a human activity*, and (2) *mathematics should be interlinked with someting real for the students.*

Therefore, the writer intends to share his experience and create the active and realistic mathematics educations, especially in the sub-topic of *"deciding the inverse function formula"* that is the mathematics material for class XI in the second semester. The title of this article is **REALISTIC MATHEMATICS EDUCATION IN THE SUB-TOPIC OF DECIDING AN INVERSE FUNCTION (THE REVERSIBLE JOURNEY METHOD).**

1.2. PURPOSE

The purposes of writing this article are : (1) to give the contribution of thinking (sharing experience) with SMA Mathematics teachers in choosing the alternative of the realistic mathematics education, especially in the sub-topic of deciding the general formula of an inverse function, (2) to persuade SMA Mathematics theachers for having an innovation for the realistic mathematics education, and (3) to take part in mini-symposium that is held by International Conference on Applied Mathematics (ICAM '05) ITB.

1.3. TARGET

Through this articles, it is hoped that it can give the concrete illustration for the SMA Mathematics theacher about the steps of the realistic mathematics education, especially in the sub-topic of deciding the general formula of an inverse function.

1.4. SCOPE

This article, explains about :
1. The definition of an inverse function.
2. The explanation of sense of an inverse function.
3. Deciding the general formula of an inverse function using the ordinary method.
4. Deciding the general formula of an inverse function using the reversible journey method.
5. The limitation in deciding an inverse function using the reversible journey method.
6. The advantages of deciding an inverse function using the reversible journey method.

2. MAIN PROBLEMS

2.1. SENSE OF AN INVERSE FUNCTION

2.1.1. The Definition of An Inverse Function

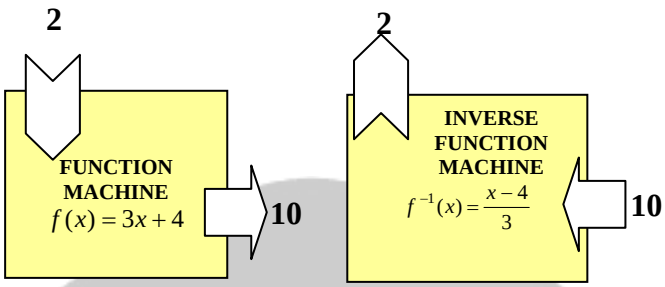
a. If a function $f : A \to B$ is stated by $f : \{(a,b) | a \in A \text{ and } b \in B\}$, the inverse of f function is $f^{-1} : B \to A$ will be decided by $f : \{(b,a) | b \in B \text{ and } a \in A\}$.

b. A function $f : A \to B$ has the inverse function $f^{-1} : B \to A$ that is a function, if only f function is a bijective function.

c. The inverse of a function is not always a function. If the inverse of a function is a function, so the inverse will be called an inverse function.

2.1.2. The Explanation of The Definition of An Inverse Function

Look at the f function with the formula $f(x) = 3x + 4$ with the result that $f(2) = 3(2) + 4 = 10$. Therefore, that f function maps 2 to 10 (*picture 1*). How if we know the result, in the case above the result is 10, how do we decide the input? If there is a function that maps its f(x) return to x, it will be called inverse function (picture 2).



Pict.1 **Pict.2**

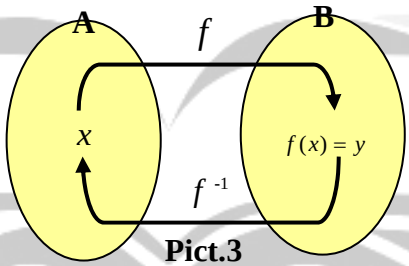
The problem is *how to decide the formula of* $f^{-1}(x)$ *if the formula of* $f(x)$ *function has been known.*

2.2. DECIDING THE GENERAL FORMULA OF AN INVERSE FUNCTION.

To decide the inverse of a function can be done by using same kinds of method. In this case, it will be explained deeply how to decide the inverse of a function using the reversible journey method. However, as the comparison, the first will be explained how to decide the inverse of a function using the ordinary method.

2.2.1. Deciding The Inverse Function Using The Ordinary Method

If f and f^{-1} are function that are inverse one another, it means that $f(x) = y \Leftrightarrow f^{-1}(y) = x$ (Picture.3).



Pict.3

Deciding the inverse function formula using the ordinary method can be done by the steps as follows :

a) *Step 1* : think that $f(x) = y$.

b) *Step 2* : state "x in y", the form of it is $f^{-1}(y)$.

c) *Step 3*: substitute y in $f^{-1}(y)$ with x to get $f^{-1}(x)$ that is the inverse of $f(x)$.

☺ ***Example 1.***
Decide the inverse of this function : $f(x) = 3x - 6$

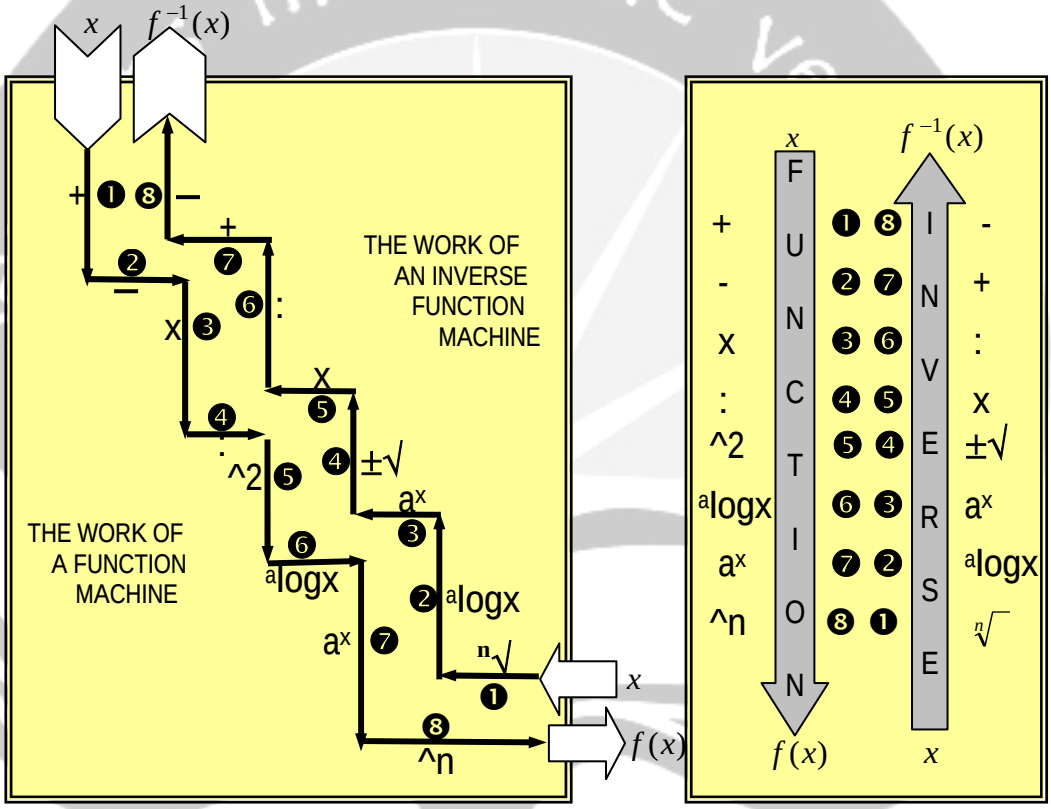
☺ ***Answer :***

$$f(x) = y = 3x - 6$$
$$\Rightarrow y + 6 = 3x$$
$$\Leftrightarrow \frac{y+6}{3} = x$$
$$\Leftrightarrow \frac{y+6}{3} = f^{-1}(y)$$

So, the inverse of $f(x) = 3x - 6$ is $f^{-1}(x) = \frac{x+6}{3}$.

2.2.2. Deciding The Inverse Function Using The Reversible Journey Method

To decide the inverse of a function using the reversible journey method, look at picture 4 below :

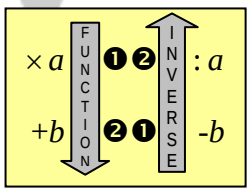


Pict.4

From picture 4, it can be explained that $f^{-1}(x)$ can be got by doing the operation that is reversible with $f(x)$. This is called *deciding the inverse using the reversible journey method.*

a). Deciding The General Formula of The Inverse of A Linear Function

If : $f(x) = ax + b$,



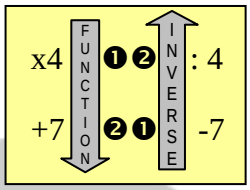
so $f^{-1}(x) = \frac{x-b}{a}$

☺ ***Example 2.***

Decide the inverse of this function : $f(x) = 4x + 7$

☺ ***Answer :***

$f(x) = 4x + 7$



$f^{-1}(x) = \dfrac{x-7}{4}$

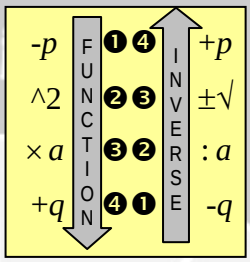
b). Deciding The Inverse of A Square Function.

if $f(x) = ax^2 + bx + c$

$= a(x-p)^2 + q$

and :

$p = \dfrac{-b}{2a}$, $q = \dfrac{D}{-4a}$



so $f^{-1}(x) = \pm\sqrt{\dfrac{x-q}{a}} + p$

☺ ***Example 3.***

Decide the inverse of these functions :

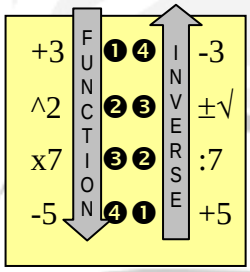
(i) $f(x) = 7(x+3)^2 - 5$

(ii) $f(x) = x^2 + 2x - 3$

(iii) $f(x) = 4x^2 - 16x + 25$

☺ ***Answer :***

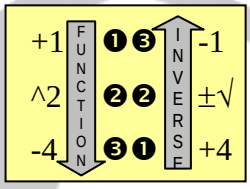
(i) $f(x) = 7(x+3)^2 - 5$



$f^{-1}(x) = \pm\sqrt{\dfrac{x+5}{7}} - 3$

(ii) $f(x) = x^2 + 2x - 3$

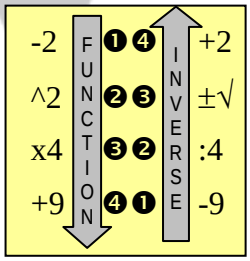
$f(x) = (x+1)^2 - 4$



$f^{-1}(x) = \pm\sqrt{x+4} - 1$

(iii) $f(x) = 4x^2 - 16x - 25$

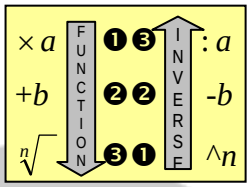
$f(x) = 4(x-2)^2 + 9$



$f^{-1}(x) = \pm\sqrt{\dfrac{x-9}{4}} + 2$

c). Deciding The Inverse of A Rational Function

If $f(x)=\sqrt[n]{ax+b}$



so $f^{-1}(x)=\dfrac{x^n-b}{a}$

☺ ***Example 4.***

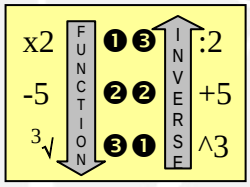
Decide the inverse of these functions :

(i) $f(x)=\sqrt[3]{2x-5}$

(ii) $f(x)=\sqrt[5]{\dfrac{3x-4}{7}}$

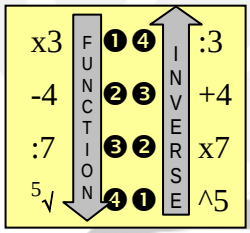
☺ ***Answer :***

(i) $f(x)=\sqrt[3]{2x-5}$



$f^{-1}(x)=\dfrac{x^3+5}{2}$

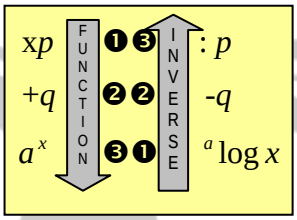
(ii) $f(x)=\sqrt[5]{\dfrac{3x-4}{7}}$



$f^{-1}(x)=\dfrac{7x^5+4}{3}$

d). Deciding The Inverse Function In Exponent Function

If $f(x)=a^{px+q}$



so $f^{-1}(x)=\dfrac{\left({}^{a}\log x\right)-q}{p}$

☺ ***Example 5.***

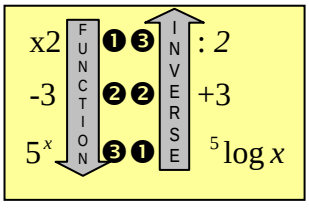
Decide the inverse of these functions :

(i) $f(x)=5^{2x-3}$

(ii) $f(x)=\dfrac{2^{3x}-4}{5}$

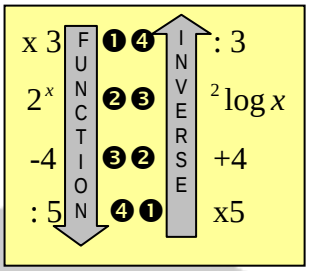
☺ ***Answer :***

(i) $f(x)=5^{2x-3}$



$f^{-1}(x)=\dfrac{\left({}^{5}\log x\right)+3}{2}$

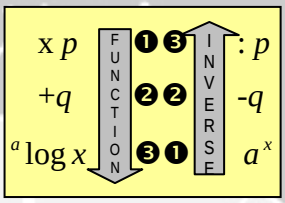
(ii) $f(x)=\dfrac{2^{3x}-4}{5}$



$f^{-1}(x)=\dfrac{^2\log(5x+4)}{3}$

e). Deciding The Inverse Function In Logarithm Function

If $f(x)=^a\log(px+q)$



so $f^{-1}(x)=\dfrac{a^x-q}{p}$

☺ ***Example 6.***

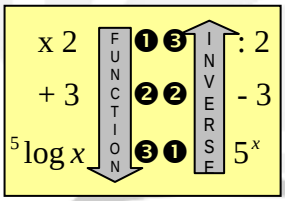
Decide the inverse of these functions :

(i) $f(x)=^5\log(2x+3)$

(ii) $f(x)=3.^2\log(5x-1)+7$

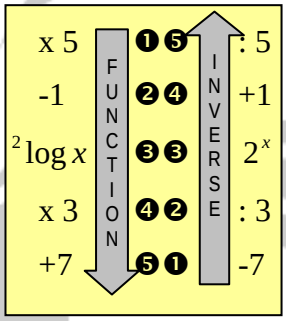
☺ ***Answer :***

(i) $f(x)=^5\log(2x+3)$



$f^{-1}(x)=\dfrac{5^x-3}{2}$

(ii) $f(x)=3.^2\log(5x-1)+7$



$f^{-1}(x)=\dfrac{2^{\frac{x-7}{3}}+1}{5}$

f). Exercise

Decide the inverse of these functions :

1). $f(x)=6x+3$

2). $f(x)=\frac{1}{3}x-2$

3). $f(x)=2(x+1)^2+3$

4). $f(x)=\frac{1}{3}(x-2)^2+5$

5). $f(x)=x^2+4x+6$

6). $f(x)=4x^2-6x+7$

7). $f(x)=\frac{1}{4}x^2+2x-3$

8). $f(x)=\sqrt[5]{3x+2}$

9). $f(x)=(4x-3)^{\frac{1}{3}}$

10). $f(x)=\sqrt{\dfrac{2x+3}{6}}$

11). $f(x)=\sqrt[3]{\dfrac{(3x+2)^5}{4}}$

12). $f(x)=3^{\frac{1}{2}x+5}$

13). $f(x)=\dfrac{5^{2x+4}-6}{3}$

14). $f(x)=\dfrac{3^{(2x-5)^2}+2}{5}$

15). $f(x)={}^{2}\log(5x-6)$

16). $f(x)={}^{3}\log\sqrt[5]{2x-1}$

17). $f(x)=3\left({}^{5}\log(2x-3)\right)-4$

18). $f(x)=4\left({}^{3}\log(x+4)^2-5\right)+7$

2.3. THE LIMITEDNESS IN DECIDING AN INVERSE FUNCTION USING THE REVERSIBLE JOURNEY METHOD

The limitedness in deciding the inverse of a function using the reversible journey method is that it cannot be used for deciding the inverse formula of $f(x)=\dfrac{ax+b}{px+q}$ (*rational function*). So that, it is suggested to use the ordinary method in deciding the general function of $f(x)=\dfrac{ax+b}{px+q}$. By using the ordinary method, we can get the general inverse function of $f(x)=\dfrac{ax+b}{px+q}$, it is $f^{-1}(x)=\dfrac{-qx+b}{px-a}$.

2.4. THE ADVANTAGES OF DECIDING AN INVERSE FUNCTION USING THE REVERSIBLE JOURNEY METHOD

According to the writer's experience in trying to teach how to decide the inverse function using the reversible journey method in class XI second semester in SMAN 2 Kuningan, the advantages are :

1. The process of mathematics education run more interractive, the students are active and feel that the teaching approach is more realistic.
2. The students feel that deciding the inverse of a function using the reversible journey method is easier and faster.
3. The students' succes in deciding the inverse of a function using the reversible journey method is good. More than 75 % of students can decide the invers of the function that are equivalent with the exercise in this article using the reversible journey method. So, the process of mathematics education is more effective.

3. CLOSING

3.1. CONCLUSIONS

According to the explanation in chapter II, we can get the conclussion, as follows :

2. Deciding an inverse of a function using the reversible journey method can be one of the alternatives in creating the active and the realistic education, so that it will be more effective.
3. It is suggested to use the ordinary method because of the limittedness in deciding the inverse of a function using the reversible journey method for deciding the inverse formula of $f(x)=\dfrac{ax+b}{px+q}$ (rational function)
4. The advantages of deciding the inverse of a function using the reversible journey method : (1) The process of education run more interactively, the students are active, and feel that the teaching approach is more realistic; (2) Students feel that deciding the invers of a function using the reversible journey method is easier and faster, and (3) The students' miscarriage of studying increases.

3.2. SUGGESTIONS

According to the explanation in chapter II and the conclussion above, the writer would like to suggest as follows :

1. In deciding the inverse of a function using the reversible journey method, it is suggested to the SMA Mathematics teachers to deliver the explanation using the following order : (1) give t*he Definition of Inverse Function*; (2) *explain the sense of an invers function*; (3) *decide the inverse function using the ordinary method*; (4) *decide the inverse function using the reversible journey method*; (5) *explain the limitedness*; and (6) *explain the advantages*.
2. Mathematics teacher, is better to act as the facilitator and make the condition in studying mathematics is active and realistic so that the process of education can be more effective.

References

Bob Foster Harlin, 2004. *1001 Soal dan Pembahasan Matematika*. Erlangga, Jakarta.

Herman Hudojo, 1998. *Pembelajaran Matematika Menurut Pandangan Kuntruktivistik*. Pasca Sarjana IKIP Malang, Malang.

Leithold Louis, 1992. *Kalkulus dan Geometri Analitis Jilid I* (*translasion*) Erlangga, Jakarta.

Murray R. Spiegel, 1989. *Matematika Dasar* (*translasion*). Erlangga, Jakarta.

Pursel Edwin J, 1993. *Kalkulus dan Geometri Analitis Jilid I* (*translasion*). Erlangga, Jakarta.

Sartono Wirodikromo, 2004. *Matematika SMA Kelas XI Semester 2*. Erlangga, Jakarta.

Seymour Lipschutz, 1988. *Matematika Hingga* (*translasion*). Erlangga, Jakarta.

Zulkardi, 2000, *Efektivitas Lingkungan Belajar Berbasis Kuliah Singkat dan Situs WEB sebagai suatu Inovasi dalam Menghasilkan Guru RME di Indonesia*. National Symposium Article, Yogyakarta.

SUDARMOYO, S.Pd.: The Mathematics Teacher of SMA Negeri 2 Kuningan, Jl. Aruji Kartawinata 16. Kuningan 45511. Phone +62(0)232 871063, +62 (0)8122409191

E-mail: smanda.kuningan@gmail.com

Changing Instructional Approach of Indonesian Mathematics Teachers with RME

Sugiman

Mathematics Education Department of Yogyakarta State University

Abstract. Most of Indonesian mathematics teachers use expository method in teaching mathematics and depends on books. Their students are passive receivers, they listen to teacher's explanation/demonstration, practice some exercise, and take evaluation/ examination. As a result, student have no enough chance to express their original idea, use mathematics to solve daily problems, discus with his tablemates, show his genuine result in his class, and think alternatively. Situation of instructional as preceding mention have to change continuously and gradually. One of the good alternative to make mathematics instructional more effective and useful is Realistic Mathematics Education (RME).

To change instructional approach from traditional tends to RME, we have to improve quality and knowledge of teacher and provide appropriate teaching material. The main programs of RME implementation are (1) teachers and materials preparation, (2) collaboration between teachers and lectures, and (3) exchange experience among participants. Based on the experience, it needs at least one year to make teacher understand about RME principles. Basically, teacher faces many obstacles in the first half year and the longer it takes the better the result.

Keywords: RME, instructional approach

1. Why do we need RME?

Suyono (1996), in Dian Armanto (2002), found the weakness of teachers due to teachers' competency in pedagogical aspect. The weakness are teachers (1) have low ability in using variety of teaching methods, (2) teach the basic skills only for answering test (teaching for test), and (3) teach using conventional methods with less of considering the logical and critical thinking. Generally, this result agrees with the ultimate experience below.

In the middle of July 2005, forties elementary school teacher from Grobrogan District of Middle Java visit Mathematics Education of Yogyakarta State University and all of them never hear RME before. Through discussion, the conclusion was the orientation of many elementary mathematics teachers is as follows.

(1) Teacher tends to act their students as an object instead of subject in instructional mathematics process. Student is a passive receiver.

(2) Teachers have very strong authority in developing mathematics students. In that case, there is a less of developing student idea. Student only imitates teacher's strategy and they rarely use different way because of fear factors.

(3) Teacher orientation in instructional process is subject mater (subject oriented).

(4) Teacher opinion in teaching mathematics is student is as an empty bottle where teacher can pour his knowledge.

(5) Mathematics is only as a tool not as a human activity.

The preceding problem due to the final goal of mathematics education is success in examination. On the other hand, Turmudi (2001) said learning is a process not result. To change the culture of teaching learning, we do need new paradigm of mathematics education, as like Realistic Mathematics Education (RME). In RME, students learn to understand contextual problems and solve them, develop communication skill, to exchange idea, to develop mathematical knowledge, and to formulate appropriate strategies. R.K. Sembiring (2003) has wrote that these characteristics of RME are called *Democratic Teaching* and meet with democratic principle of Indonesia nation. Education that emphasizes student provides the controlled freedom. In that case, the student will get much learning experiences compare with if teacher arranges all student-learning routes in high discipline (Djohar, 1999). Since four years ago, RME have been tried out in limited elementary school. The main problem is how to disseminate and implement RME on broad scale. This paper will not discuss/ overcome this very big challenge.

Indonesia government, through Education Ministry, has launched the new mathematics curriculum since 2004. The curriculum is based on student competency and urges mathematics teacher to use environment context. In RME, context becomes starting point in every mathematics lesson. Jan de Lange (1987) classified/categorized the context in three orders, i.e. (1) context involves mathematical operations, (2) context motivates student to organize idea, to find relevant mathematics, and to solve problem solving, and (3) context becomes a tool of introducing or developing mathematics model or concept.

In order to gain the main aims of the 2004 curriculum, the conventional teaching and learning process that disregard process and students' creativity needs to be changed to another approach where the teacher challenge the students with contextual mathematics problems and to another classroom culture that encourages and facilitates learning.

Through using various role of context in teaching learning of mathematics with implementing RME in school, student can reach not only soft knowledge of mathematics but also mathematical literacy. The PISA (2003) defines that mathematical literacy is an individual's capacity to identify and understand the role that mathematics plays in the world, to make well-founded judgments and to use and engage with mathematics in ways that meet the needs of that individual's life as a constructive, concerned and reflective citizen.

2. Changing Instructional Approach

Almost all of mathematics teacher in early grade of primary school is educated in old way or in traditional approach. This evidence causes obstacles in running new instructional approach, especially in RME. The obstacles relate to make general plan of lesson, look for meaningful and useful contexts, create manipulative teaching aids, open the instructional in class, encourage student to compose idea, accept alternative notion of pupils, understand informal mathematics language of students, and evaluate process and result of instructional activity. Prerequisite of RME implementation is as below.

The main results of Sutarto's dissertation (2002) are the success of RME in Indonesia influenced by these aspect: (1) the provision of RME curriculum materials, (2) the implementation RME to mathematics instruction in micro scale (in class), (3) the change of teachers' belief that teaching mathematics means guiding students to learn and doing mathematics, and (4) the change of students' attitude from passive receiver to active learners who have ability in thinking mathematically and to do mathematics. From all aspects mentioned above, teacher position is strategic, prominent, and important to influence the successful of RME implementation. Therefore teachers should become agents of the innovation, they can make plan, organize system, prepare instructional aid, create manipulative tool, support and maintenance role of student in class, raise student creative idea, and evaluate process by them self. Gravemeijer (1994) divides the levels of teacher changing in three grades, i.e. (1) use of material, (2) educational activities, and (3) beliefs.

For educational activities in teaching mathematics, there are five RME tenets/principles (Leen Streefland, 1990). These are

(1) Constructions stimulated by concreteness

(2) Developing mathematical tools to move from concreteness to abstraction.

(3) Stimulating free productions and reflection.

(4) Stimulating the social activity of learning by interaction.

(5) Intertwining learning strands in order to get mathematical material structured.

As mention above, teacher is an important person. These pictures illustrate some tenets of RME in introducing second concept for 2nd grade students, see figure 1. Teacher asks student to measure how long they can open his eyes without wink (figure 2) and use one hand to support weight books without fall down (figure 3), they learn by interaction in small group, and ask student to make a report (figure 4).

Fig. 1. Starting Point with Daily Evidence

Fig. 2. Wink of His Eyes

Fig. 3. Learning by Interaction

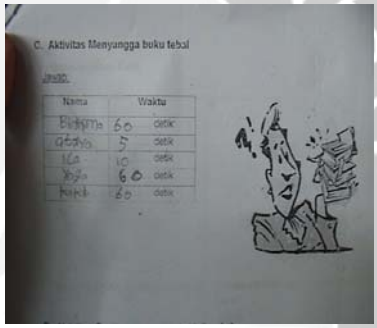


Fig. 4. Student's Report

The activity using manual clock can be related or intertwined with some others concepts such as follows.

(1) Circle and its properties

(2) Number factor (of 60)

(3) Conversion concept between minute and second

(4) Basis notion (of 60)

(5) Algebra operations

(6) Fractional number

The teaching learning activities above show that changing instructional approach begin move from traditional to RME. Furthermore, some teachers brave in taking his initiates not only depends on script. One of the teacher said "I try to understand content of teacher's book, some time I do not follow all due do situation, facilities, student ability, and local context." For an example, he done mathematics lesson in open hall. He

asked 42 second grade students to make small circle where every small circle consist of exactly 4 pupils.

Teacher: "How many small circle we have?"

Student: (Counting) "ten"

Teacher: "How many pupils are not in circle?

Student: (Watching around) "two, sir!"

Teacher: "It means, how many students are in circle?"

Student: (Thinking) "Forty!"

Teacher: "How do you get 40?"

Student: "Because there are 10 circle and every circle consist of 4 pupils. Because of that 10 × 4 = 40."

Teacher: "Therefore, from 42 students we make circle. Every circle consists of exactly 4 students. How many small circle we have?"

Student: "10."

The activity was actually very meaningful and useful. Students not only though mathematically but also doing a fun activity. But teachers did not explore meaningfully and effectively in order to get student's genuine strategy/idea. Student's answer were very formal mathematics, no students used his in-formal mathematics language. Some examples of advanced question that can be presented by teacher are as follows.

(1) "Please you present how to get the answer, you can use anyway like drawing, adding, jumping in empty number line, or others."

(2) "If every circle consists of exactly 5 pupils, how many circles do you have and how many pupils are not in any circle?"

(3) "If every circle consists of exactly 6 pupils, how many circles do you have and how many pupils are not in any circle?"

(4) "All of you will go to festival by *becak*. The load of every *becak* is no more than 4 pupils. How many *becak* you need?"

In the teaching multiplication teacher used many context. Various contexts that were used by teacher in multiplication lesson can be seen in the picture 5, 6, 7, and 8.

Fig. 5. Environment Context

Fig 6. Vendor Context

Fig. 7. Box Context

Fig. 8. Book Formation Context

It indicates that teacher realizes and knows that student cannot trace many routes of learning mathematics from real world problem to mathematics world. There are a lot of student's learning trajectories. One of the student's routes is illustrated in figure 9 (Turmudi, 2001). Letter A means real world and B represents mathematics world. Frans Moerlands describes iceberg model in four stages: (1) mathematical world orientation, (2) model material, (3) building stones (number relation), and (4) formal notation.

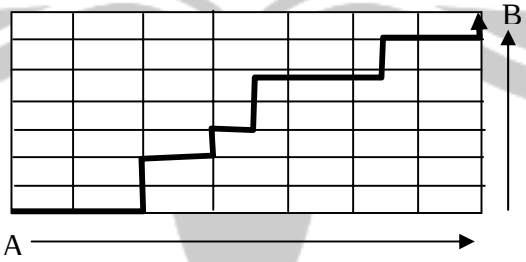


Figure 5. One Example of Student's Learning Route in RME

In Indonesia, RME becomes an embryo of Pendidikan Matematika Realistik Indonesia (PMRI). The main programs of PMRI (RME) implementation are:

(1) Teachers and materials preparation. Until 2005 years, it has been written seven PMRI (RME) books by scriptwriters from four different universities. The books are for 1st grade until 4th grade of elementary school. The books are as sources for all attached PMRI (RME) school. Futhermore, some teachers brave to create others activities and or others worksheets for their lesson.

(2) Collaboration between teachers and lectures in the forms (a) lecturers goes to class to observe and (b) meeting periodically among the teachers and lecturers to discuss the implementation of PMRI (RME) principles. Based on the experience, it needs at least one year to make teacher understand about RME principles, it depends on the teacher himself, the more active is the faster. Generally, almost all teachers face many obstacles in first half year and the longer it does the better the result..

(3) Exchange experience among participants that conducted by every university or Team of PMRI. The good slogan is "*form teacher, by teacher, and for teacher.*"

3. Conclusion

RME is good approach in teaching learning of mathematics and can supports implementation of "Curriculum 2004." Basically, teachers have ability to change from their traditional approach to RME. Among five tenets of RME, the most difficult principle is stimulating free production and the easiest one is stimulating social activity of learning by interaction. It needs at least one year to master all of the RME tenets.

References

The PISA 2003 Assessment Framework- Mathematics, Reading, Science and Problem solving Knowledge and Skill.

Dian Armanto. 2002. *Teaching Multiplication and Division Realistically in Indonesian Primary Schools: A Prototype of Local Instructional Theory*. Disertasion. Enschede: PrintPartners Ipskamp.

Djohar. 1999. Reformasi dan Masa Depan Pendidikan di Indonesia. Yogyakarta: IKIP Yogyakarta.

Gravemeijer.1994. Developing Realistic Mathematics Education. Utrech: CDβ Press.

Jan de Lange. 1987. *Mathematics, insight and Meaning*. Utrecht The Netherland: OW&OC.

Leen Streefland. 1990. *Realistic Mathematics Education (RME) What does it mean?* In the book "Context Free Productions Test and Geometry in Realistic Mathematics Education". Editor: E.J. Hanepen. Utrecht The Netherland: OW & OC.

Sembiring, R.K. 2003. *PMRI, Usaha ke Arah Reformasi Pendidikan Matematika di Indonesia*. Bulletin of PMRI (Pendidikan Matematika Realistik Indonesia). First Edition, June 2003.

Sutarto Hadi. 2002. *Effective Teacher Professional Development for Implementation of Realistic Mathematics Education in Indonesia*. Disertasion. Enschede: PrintPartners Ipskamp

Turmudi. 2001. Matematika Kontekstual dalam Pembelajaran Matematika dan Contoh Pengembangannya di Tingkat Mikro. Procidding of Mathematics National Seminar in Mathematics Education Department of UNY Yogyakarta, 21st April 2001.

SUGIMAN: Mathematics Education Department of Yogyakarta State University

E-mail: sugiman_uny@yahoo.com

EVALUATION OF TEACHING AT A UNIVERSITY: A FUZZY SET APPROACH

[a] Sabri Ahmad, [b] Mohd Lazim Abdullah and [a] Abu Osman Md Tap.

[a] Department of Mathematics,
University College of Science and Technology Malaysia,
Kuala Terengganu, Terengganu, Malaysia

[b] MARA Junior Science College – Terengganu Foundation (MRSM – YT),
Besut, Terengganu, Malaysia.

Abstract: In this paper, the authors will evaluate the excellence of lecturers in teaching by using the fuzzy set decision making approach. The approach outlined here is based on the fuzzy sets theory with the objective of putting the data available on teaching evaluation scores to decide the most excellent faculty and department. The study employed a questionnaire which consists of five major constructs of teaching and learning for evaluating the teaching scores of lecturers. The questionnaires were distributed to students by general staff at the end of every semester for three consecutive semesters. Fuzzy decision making approach was used to analyze the data from one hundred and twenty three lecturers from a public university in the east coast of the Malaysian Peninsula. The evaluation procedures and the use of decision making approach pertaining to the fuzzy word 'excellent' will be discussed in this paper.

1. Introduction

In daily working environments, the superior always stresses the important of excellence in all aspect works tendering to produce the maximum productivity or working outputs. Every single worker stretches their effort to fulfil organisational objectives gearing toward excellence. In spite of positive attributes derived from the fruits for an excellence works, the word of excellence is very hard to define and does not have clear boundaries. Some people would suggest excel in job might be measured in term of mass volume of productivity and for the others it is just a measurement of quality. Still, the excellence would be varied in their definitions and measurements but undoubtedly it plays an enormous role in shaping organisations.

Comes to narrow the excellence in the organisation of higher learning. Excellence appears to be one of the most commonly used words among modern higher education institution. There are many attributes governing the excellence at a university or institute of higher learning. One of the much-talked issues recently was the quality of academic staff. Academic staff must propel excellent in delivering knowledge and

information to students. They are expected not only to give their best in teaching but more importantly they must give a sense satisfaction to their main customers i.e. students. Centra (1993) provided a comprehensive list about the characteristics of excellent in teaching. These include good organisation of subject matter, effective communication and positive attitude toward students. These are among the very common attributes in determining excellent in teaching at a university. In view of laymen, these attributes are very clear in definition and understandable. Putting in a different perspective, what attributed to the excellence does not have a clear cut definition, vague and very subjective. Logically, this implied that the word excellence is also vague and very difficult to give in an exact definition. Despite the subjectivity of explaining excellence, there was a mathematical theory which can suits with the unclear boundaries and subjective in nature. The fuzzy sets theory was created in response to the need to have a mathematical measurement of excellence. Indeed, it was very fortunate that the fuzzy sets theory provides a framework that cope with uncertainty in language, that is, subjective uncertainty (Mukaidono, 2001)

Zadeh (1965) attempted to provide a mathematical model that would better suits to these situations. Fuzzy set theory has become a branch of mathematics that generalises the concepts of sets to provide better tool for dealing with the sort of situations that looks fuzzy. Zadeh (1965) proposed the idea of a fuzzy set *A*. A fuzzy set is one to which objects can belong to different degrees called grades of membership. Grade of membership represents the level of confidence that descriptions of words are true.

A collection of objects (universe of discourse) U has a fuzzy set A described by a membership function μA that takes the value in interval [0, 1], $\mu A \in [0,1]$. Thus, A can be represented as: $A \rightarrow \mu A(u)/u$, where $u \in U$. The degree to which u belongs to A is given by membership function μA. The range of membership function is [0, 1] and this degree of U shows its confidence level.

If $\mu A(x) = 1$, then x is completely in the set A and if $0 < \mu A < 1$, then x is in the set A with $\mu A(x)$ as degree of memberships. If $\mu A(x)$ is nearer to 1, then the higher degree of membership element x in set A. Conversely, if $\mu A(x)$ is nearer to 0, then the degree of membership of element x in set A is lower.

It is always said that the formal mathematical knowledge comes from the very exactness of the science of mathematics. Hence, there is no possible space in mathematics for any lack of definition or vagueness. But the excellence of teaching that the lecturers possessed and assessed by the other parties is usually subjective, characterised by a different degree of depth of the excellence. This suggests the application of the fuzzy sets theory in measuring the excellence in teaching and could be very promising tool to search the highest degree of excellence. Therefore, the purpose of this study attempts to apply a fuzzy set decision making approach to measure the excellence of teaching in a public university.

2. Decision Making by Intersection of Fuzzy Goals and Constraints.

Decision making is characterised by selection or choice from alternative which are available (Bojadziev, 1999). In the process of decision making, specified goal have to be attained and specified constraint have to be fulfilled. For the purpose of making decision, consider a simple model consisting of a goal described by a fuzzy set $\check{G}$ with membership function $\mu_{\check{G}}(x)$ and a constraint described by a fuzzy set $\check{C}$ with membership function $\mu_{\check{C}}(x)$, where x is element of crisp set of alternatives $\check{A}_{alt}$. By definition (Bellman and Zadeh, 1970) the decision is a fuzzy set $\check{D}$ with membership function $\mu_{\check{D}}(x)$ expressed as intersection of $\check{G}$ and $\check{C}$.

$$\check{D} = \check{G} \cap \check{C} = \{ x. \mu_{\check{D}}(x)/ x \in [d_1, d_2]. \quad \mu_{\check{D}}(x) \in [0, h \leq 1] \qquad (2.1)$$

It is a multiple decision resulting in selection the crisp set $[d_1, d_2]$ from the set alternatives $\check{A}_{alt}$.: $\mu_{\check{D}}(x)$ indicates the degree to which any $x \in [d_1, d_2]$ belongs to the decision $\check{D}$. A schematic representation is shown on Fig. 1 when $x \in \check{A}_{alt} \subset R$ and $\check{G}$ and $\check{C}$ have monotone continuous membership functions.

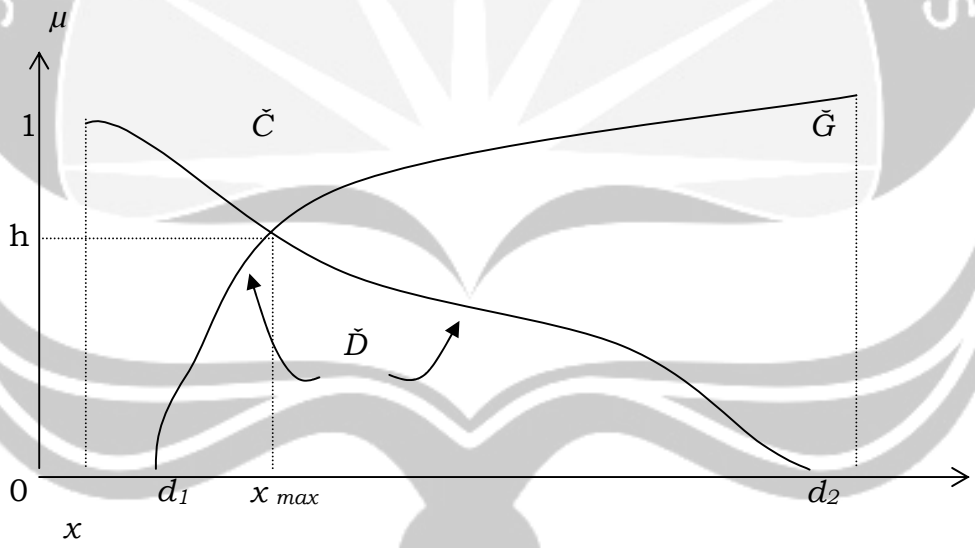


Fig.1: Fuzzy goal $\check{G}$, constraint $\check{C}$, decision $\check{D}$, max decision x_{max}.

Using membership functions and operation intersection on fuzzy sets, formula (2.1) gives

$$\mu_{\check{D}}(x) = \min(\mu_{\check{G}}(x), \mu_{\check{C}}(x)), \quad x \in \check{A}_{alt} \tag{2.2}$$

The operation intersections is commutative, hence the goal and constraint in (2.1) can be formally interchanged, i.e. $\check{D} = \check{G} \cap \check{C} = \check{C} \cap \check{D}$. According to Bojadzied (1999), in real situation, goal can be considered as constraint and vice-versa. Sometimes there is no need to specify the goal and constraint. It can simply be called as objectives or aspects of a problem.

Usually the decision makers want to have crisp result, a value among the elements of set
$[\, d_1, d_2] \subset \check{A}_{alt}$ which best or adequately represents the fuzzy set $\check{D}$. That requires defuzzification of $\check{D}$. It is natural to adopt for that purpose the value x from the selected set $[\, d_1, d_2]$ with highest degree of membership in the set $\check{D}$. Such a value x maximises $\mu_{\check{D}}(x)$ and is called maximising decision (Fig. 1) . It is expressed by

$$x_{max} = \{x / \max \mu_{\check{D}}(x) = \max \min (\mu_{\check{G}}(x), \mu_{\check{C}}(x))\} \tag{2.3}$$

The process of decision making is shown as a block diagram on Fig.2.

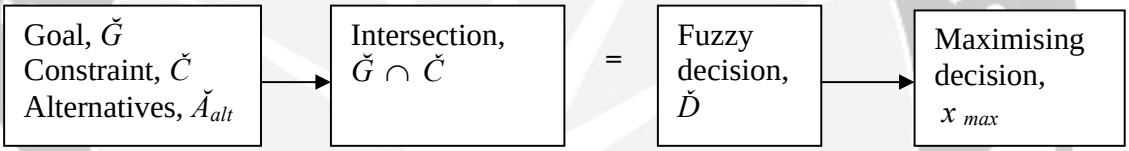


Fig. 2: Process of decision making by intersection

Formulas (2.1) - (2.3) have been generalised for decision making model with many goals and constraints (Bellman and Zadeh, 1987). For n goals $\check{G}_i$,$i = 1, \ldots\ldots n$, and m constraints, $\check{C}_j$, $j= 1, \ldots\ldots m$, the decision is

$$\check{D} = \check{G}_1 \cap \ldots\ldots\ldots \check{G}_n \cap \check{C}_1 \cap \ldots\ldots\ldots \cap \check{C}_m, \tag{2.4}$$

the membership function of $\check{D}$ is

$$\mu_{\check{D}}(x) = \min (\mu_{\check{G}1}(x), \ldots\ldots\ldots \mu_{\check{G}m}(x), \mu_{\check{C}1}(x), \ldots\ldots\ldots \mu_{\check{C}m}(x)), \tag{2.5}$$

and the maximising decision is given by

$$x_{max} = \{x / \mu_{\check{D}}(x) \text{ is max}\} \tag{2.6}$$

A case study of evaluation of teaching at a university is presented by using the process of decision making in fuzzy environment. The decision process was adopted from Case

Study 9 based on material in the book by Li and Yen (1995) on the evaluation of learning performance.

3. Measuring the Excellence in Teaching: A Numerical Example

An application of fuzzy decision making is sought to decide the highest degree of excellence in a case study of teaching evaluation at a small yet young university in the East Coast of Peninsular Malaysia. As a newly established university, there were two main faculties in operating. In this study, a fuzzy decision making will be conducting to decide the better performer of excellence in teaching between two faculties. To avoid discriminations and prejudices, these two faculties were named as f_1 and f_2. In the faculty of f_1, there were seven departments while the faculty of f_2 was roofing five departments. Also, the decision will be made to decide the most excellent department in teaching at the faculty of f_1 and f_2.

Scores of the evaluation of teaching were collected from a questionnaire. This questionnaire was distributed to students by general (non-academic) staff at the end of every semester for three consecutive semesters. The questionnaire focuses on five major constructs of teaching and learning categorized as,
a_: Planning and preparation of teaching,
b_: Lecture room/lab teaching,
c_: Preparation of teaching resources,
d_: Course evaluation system, and
e_: Relationship between student-lecturer.

Answers given by respondents were on a four-point scale (4=most agreeable to 1=least agreeable). A total of 123 courses/lecturers were sampled and all the data were analysed using a fuzzy set decision making approach. Since this paper focuses on the methods of decision making hence details of five constructs were not to be elaborated. A total score for every construct was calculated by adding sub-scores of items before converted to percentages. The total score of every construct represents the excellence of samples in teaching. Excellence in teaching must be defined prior to the computation procedures.

3.1 Definition of excellence in teaching

Excellent in teaching was measured based on five constructs. Excellent is a linguistic label which described separately into two trapezoidal numbers. Excellent in constructs of a_, d_ and e_ (**E a_d_e_**) is defined as Fig. 3(a) while excellent in constructs of b_ and c_ (**E b_c_**) is defined as Fig. 3(b) using part of trapezoidal numbers on the universe [0, 100] of scores.

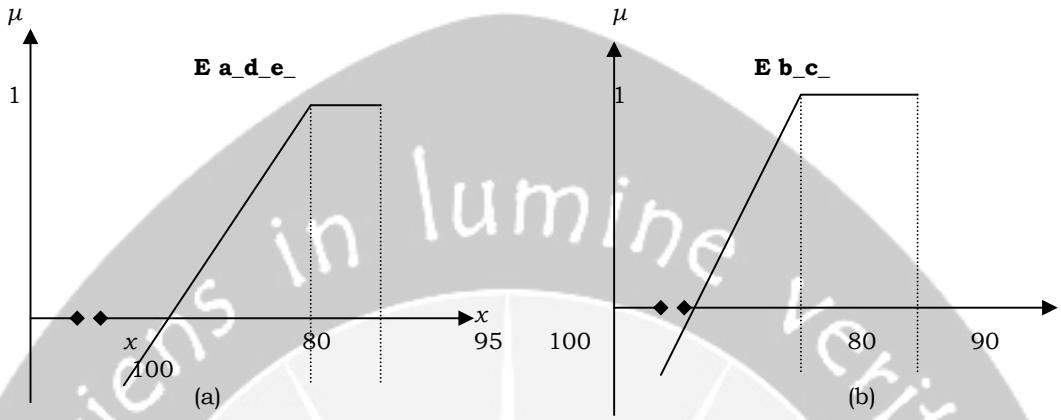


Fig. 3 (a) *Excellent* in constructs of a_, d_ and e_ , Fig. 3 (b) *Excellent* in constructs of b_ and c_

Using the trapezoidal fuzzy numbers, the membership functions of excellent in constructs a_, d_, and e_ is defined as

$$\mu_{E\,a\_d\_e\_}(x) = \begin{cases} 0 & \text{for } 0 \le x \le 80 \\ \frac{x-80}{15} & \text{for } 80 \le x \le 95 \\ 1 & \text{for } 95 \le x \le 100 \end{cases} \qquad \text{.....(3.1)}$$

while trapezoidal number is defined on excellent in construct b_ and c_ by

$$\mu_{E\,b\_c\_}(x) = \begin{cases} 0 & \text{for } 0 \le x \le 80 \\ \frac{x-80}{10} & \text{for } 80 \le x \le 90 \\ 1 & \text{for } 90 \le x \le 100 \end{cases} \qquad \text{.....(3.2)}$$

These two definitions definitely offer a different grade of membership for every construct. For instance, if total score in evaluation of teaching of a faculty is 85 in constructs a_ , d_, and e_, then it has grade of membership 0.33 in the set **E a_d_e_** while the same score in construct b_ and c_ has grade of membership 0.50 in the set **E**

b_c_. Details of the computation procedures and results will be explained in the following section.

3.2 Results and Computation Procedures.

As stated in the first paragraph of this section, there are three decisions will be made. Computation procedures to reach the decisions will be presented in three sub-sub-sections as follow.

3.2.1 Which faculty has performed excellent in teaching better than the other?

Total scores of five constructs in every faculty were converted into percentages. The conversion must be made to conform to the definition of excellent (section 3.1). The total scores in every faculty presented on Table 1.

Table 1: Total Scores in Five Constructs of Teaching at Faculty of f_1 and f_2.

Constructs/ Faculty	**a_**	**b_**	**c_**	**d_**	**e_**
f_1	87.84	86.01	85.88	87.61	88.84
f_2	93.14	87.98	89.75	90.62	90.85

The set of alternatives is $A_{alt} = \{ f_1, f_2\}$.
Substituting scores in constructs a_, d_ and e_ into (3.1) and those in constructs b_ and c_ into (3.2) gives the degree of excellence corresponding to the scores. They are shown in Table 2.

Table 2: Degree of Excellent in Five Constructs of Teaching at Faculty of f_1 and f_2

Construct/ Faculty	**a_**	**b_**	**c_**	**d_**	**e_**
f_1	0.52	0.60	0.59	0.51	0.59
f_2	0.88	0.79	0.98	0.71	0.72

The degree of excellence attached to each faculty produce the fuzzy sets of excellence in five constructs of teaching evaluation.
Excellent in a_
$\check{G}_1 = \{(f_1, 0.52), (f_2, 0.88)$
Excellent in b_
$\check{G}_2 = \{(f_1, 0.60), (f_2, 0.79)\}$
Excellent in c_
$\check{G}_3 = \{(f_1, 0.59), (f_2, 0.98)\}$
Excellent in d_
$\check{G}_4 = \{(f_1, 0.51), (f_2, 0.71)\}$

Excellent in e_
$\check{G}_5 = \{(f_1, 0.59), (f_2, 0.72)\}$

The decision formula (2.4) gives

$$\check{D} = \check{G}_1 \cap \check{G}_2 \cap \check{G}_3 \cap \check{G}_4 \cap \check{G}_5 ,$$
$$= \{(f_1, 0.51), (f_2, 0.71)\}.$$

Hence the conclusion is that f_2 i.e. faculty of f_2 with the degree of membership o.71 in $\check{D}$ is excellent in teaching better than f_1.

Similar computation procedures could be used to decide the best department at the faculty of f_1 and f_2.

3.2.2 Which department at the faculty of f_1 has performed the most excellent in teaching.

Total scores of five constructs segregated to the seven departments denoted as s_1, s_2, s_3, s_4, s_5, s_6, s_7 and converted to percentages as shown in Table 3.

Table 3: Total Scores in Five Constructs of Teaching at Departments of f_1

Construct/ Departments	**a_**	**b_**	**c_**	**d_**	**E_**
s_1	89.28	88.46	88.77	90.03	90.57
s_2	89.87	87.16	89.30	89.36	85.78
s_3	88.35	87.15	85.32	82.79	87.61
s_4	86.95	85.30	82.90	87.75	89.64
s_5	88.45	84.15	85.64	88.56	90.60
s_6	85.85	84.02	87.92	86.03	89.41
s_7	87.39	86.81	86.07	86.35	87.02

Substituting scores in constructs a_, d_ and e_ into (3.1) and those in constructs b_ and c_ into (3.2) gives the degree of excellence corresponding to the scores. They are shown in Table 4.

Table 4: Degree of Excellent in Five Constructs of Teaching at Departments of f_1

Constructs/ Departments	**a_**	**b_**	**c_**	**d_**	**e_**
s_1	0.62	0.85	0.88	0.67	0.70
s_2	0.66	0.72	0.93	0.62	0.39
s_3	0.56	0.72	0.53	0.19	0.51
s_4	0.46	0.53	0.29	0.52	0.64
s_5	0.56	0.42	0.56	0.57	0.71

s_6	0.39	0.40	0.79	0.40	0.70
s_7	0.49	0.85	0.61	0.42	0.39

The decision formula (2.4) gives

$$= \{(s_1, 0.62), (s_2, 0.39), (s_3, 0.19), (s_4, 0.29), (s_5, 0.42), (s_6, 0.39), (s_7, 0.42)\}$$

Hence the conclusion is that s_1 i.e. department of s_1 with the degree of membership 0.62 in $\check{D}$ is the most excellent in teaching.

3.2.3 Which department at the faculty of f_2 has performed the most excellent in teaching.

Total scores of five constructs in departments of m_1, m_2, m_3, m_4, m_5 were formulated into percentages. The total scores of each department at faculty of f_2 are presented in Table 5.

Table 5: Total Scores in Five Constructs of Teaching at Departments of f_2

Constructs/ departments	**a_**	**b_**	**c_**	**d_**	**e_**
m_1	90.91	86.45	89.85	91.28	90.78
m_2	92.66	85.48	86.01	89.83	88.27
m_3	96.66	93.49	94.03	95.06	95.37
m_4	92.80	86.87	90.02	88.90	90.68
m_5	92.78	89.16	91.07	88.92	90.75

Substituting scores in constructs a_, d_ and e_ into (3.1) and those in constructs b_ and c_ into (3.2) gives the degree of excellence corresponding to the scores. They are shown in Table 6.

Table 6: Degree of Excellent in Five Constructs of Teaching at Departments of f_2

Constructs/ faculty	**a_**	**b_**	**c_**	**d_**	**e_**
m_1	0.73	0.65	0.98	0.75	0.72
m_2	0.84	0.55	0.60	0.66	0.55
m_3	1.00	1.00	1.00	1.00	1.00
m_4	0.85	0.69	1.00	0.59	0.71
m_5	0.85	0.92	1.00	0.59	0.72

The decision formula (2.4) gives

$\check{D} = \{(m_1, 0.65), (m_2, 0.55), (m_3, 1), (m_4, 0.59), (m_5, 0.59)\}$

Hence the conclusion is that m_3 i.e. department of m_3 with the degree of membership 1 in $\tilde{D}$ is the best score in teaching evaluation at the faculty of f_2.

4. Remarks and Conclusions

Look at a glance of the final results that presented in this paper, one might think of other approaches which might easier rather simpler methods to decide the best performer. Normally in any decision involving integers, research people tend to find the mean scores and standard deviations. Value of mean as a central value of the whole data sometimes is not very accurate to reflect the distribution of data. This is especially more complex if each group of data carry a different weight and definition. This paper has shown an alternative method in finding the best performer. Limitations about the nature of a mean value and the unclear definition of excellence give a chance to fuzzy decision making approach to be applied in. Decisions do not only stopped at the highest degree of membership but also can be ranked accordingly. Ultimately, decision making in fuzzy environment offers an alternative mean to highlight the best among the rest. Another good example of the fuzzy synthetic decision in assessing the performance of university teachers' in Taiwan can be looked further from Ying and Ling (2002).

From the computations of the previous section, it becomes evident that the use of fuzzy sets theory in decision making environment leads to useful numerical hierarchy which can give an effective indicator to the decision makers. It becomes also evident that the same method with appropriate goals and constraints could be used in many other decision making environments. An analogous application in selection for building construction can be retrieved from Novak (1989).

References

Bellman, R.E. and Zadeh. L.A. (1987). *Fuzzy Sets and Application: Selected Papers by L.A. Zadeh.* John Wiley & Sons, New York. pp. 53-79.

Bojadziev, G & Bojadziev, M.(1999). *Fuzzy Logic for Bussiness, Finance and Managemant.* Danvers, MA : World Scientific Publishing.

Centra, J. (1993). *Reflective faculty evaluation*, San Francisco, Josey Bass.

Li, H. X. and Yen, V. C. (1995). *Fuzzy Sets and Fuzzy Decision Making*, CPC Press, Boca Raton, Florida.

Mukaidono, M. (2001). *Fuzzy Logic for Beginners*. Singapore: World Scientific Publishing.

Novak, V. (1989). *Fuzzy Sets and their Application*, Techno House, Bristol.

Ying, F.K and Ling, S.C.(2002). Using the fuzzy synthetic decision approach to assess the performance of university teachers in Taiwan. *International Journal of Management*, Vol. 19 (4), pp. 593- 604.

Zadeh, L.A. (1965). *Fuzzy Set, Information and Control* **8**: 338 – 353.

SABRI AHMAD: Department of Mathematics, University College of Science and Technology Malaysia, Kuala Terengganu, Terengganu, Malaysia

E-mail: sba@kustem.edu.my

MOHD LAZIM ABDULLAH: MARA Junior Science College – Terengganu Foundation (MRSM – YT), Besut, Terengganu, Malaysia.

ABU OSMAN MD TAP: Department of Mathematics, University College of Science and Technology Malaysia, Kuala Terengganu, Terengganu, Malaysia