40th
ANNIVERSARY
MATHEMATICS
WISDOM TO SOCIETY FOR ALL MANKIND

KING MONGKUT'S
UNIVERSITY OF TECHNOLOGY
THONBURI

55th ANNIVERSARY
1960 - 2015
Defining Futures. Embracing Changes. Finding Solutions.

# PROCEEDINGS
# ICAS 2015

## International Conference on Applied Statistics 2015

“Statistics for Global Evolution Vision in the 21st Century”

July 15-17, 2015

Siam Bayshore Resort & Spa, Pattaya, Thailand

Organized by

Department of Mathematics, Faculty of Science
King Mongkut’s University of Technology Thonburi (KMUTT)

เครือข่ายวิจัยสถิติศาสตร์

ICAS 2015

15-17 July 2015 Pattaya, Thailand

# Proceedings of

# International Conference on Applied Statistics 2015

*Statistics for Global Evolution Vision in the 21st Century*

July 15 – 17, 2015

Siam Bayshore Resort & Spa, Pattaya, Thailand.

Organized by

Department of Mathematics, Faculty of Science
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
Tel: +66(0)-2470-8822
Email: icas2015@mail.kmutt.ac.th
URL: http://icas2015.kmutt.ac.th

# Message from President of King Mongkut's University of Technology Thonburi

Associate Professor Dr. Sakarindr Bhumiratana

It is my great pleasure to welcome you all to the International Conference on Applied Statistics 2015 (ICAS 2015) which is being hosted by King Mongkut's University of Technology Thonburi. This conference will focus on the Statistics for Global Evolution Vision against the backdrop of the 21st Century.

Not only is this conference setting a stage to promote new knowledge and research in Statistics from academics and researchers from around the world, but it also provides a platform to build a network of researchers in both theoretical and applied statistics. The progress in the subject is indeed vital as it plays an important role in all areas of research. In this conference, there will be invited talks, oral and poster presentations. I believe that the exchange of ideas during the conference will indeed make connections fit for the theme of the conference.

I would like to take this opportunity to express my gratitude to the organizers, committees and all who make this possible. I wish to sincerely thank all honorable speakers who have helped in the preparation and organizing the conference from the start and they are here with us to see it through, despite being occupied with their other obligations. I would like to thank our sponsors, especially Minitab, SPSS and AIS, for their generosity and interest in the conference.

I wish the conference to be a great success and hope that you will find it fruitful with exciting new ideas for teaching and research in many areas. I declare the official opening of ICAS 2015 and hope you have a pleasant and enjoyable time in Pattaya and our hospitality from King Mongkut's University of Technology Thonburi throughout this conference.

Associate Professor Dr. Sakarindr Bhumiratana
President of King Mongkut's University of Technology Thonburi

# Message from President of the Thai Statistical Association

Associate Professor Adisak Pongpullponsak

On behalf of the Chairperson of the International Conference on Applied Statistics 2015 (ICAS 2015), I would like to warmly welcome all of you to the conference. I am particularly delighted that there are over 120 participants from 9 other countries apart from Thailand. We will have talks from a keynote speaker and 11 invited speakers, as well as 94 oral and 16 poster presentations. I am confident that during the course of the conference we will come across many interesting and insightful works in Statistics. I would like to take this opportunity to thank our keynote, invited speakers including all those contributing their works to the conference.

Pattaya, a seaside city on the Eastern Gulf Coast of Thailand, is a very attractive and famous tourist destination where you will enjoy sightseeing, absorbing unique culture and tradition, food and way of living. For those who wish to explore the region, we arrange a post-conference excursion which will take you to Pattaya Floating Market and the Sanctuary of Truth. I wish you all have a great time enjoying what we have prepared for you during the conference and have a pleasant time in Pattaya.

Professor Adisak Pongpullponsak
President of the Thai Statistical Association

# Message from Dean of Faculty of Science, King Mongkut’s University of Technology Thonburi

Assistant Professor Dr. Woranut Koetsinchai

On behalf of Faculty of Science, King Mongkut's University of Technology Thonburi, I would like to welcome each and every one of you to the International Conference on Applied Statistics 2015 (ICAS 2015). The ICAS 2015 aims to bring together the national and international statisticians, other scientists, educators and students from academia, industries, government, and research institutes to exchange and share their experiences and research results about all aspects of statistics and applied statistics. It also provides a platform to mutually interact and share their thoughts on the recent innovations, practical challenges encountered and the solutions adopted in the difference areas of the subject. With a deep sense of pride, Department of Mathematics, Faculty of Science, KMUTT receives the great honor from the Thai Statistical Association to host ICAS 2015. It is our first international conference in the area of statistics which is an integral part of our celebrating activities to commemorate the 55th year of KMUTT and the 40th year of Department of Mathematics. For the past 40 years, the Department of Mathematics, Faculty of Science, KMUTT has been focused on committing excellent teaching and research in the area of mathematics, statistics and computer science based on international standard. In this conference, clearly, we will witness lively and productive sharing of experiences, creative ideas and valuable information at the oral presentations and poster session. I would like to express our deep appreciation for all supports and would like to thank all the staffs and referees without their contributions, this conference would not be possible. We are gratified by the quality contributions all participants have brought to the proceeding. We look forward with great anticipation to all the sessions of our conference.

Thank you everyone for being here with us at this important event and, for those from abroad, do enjoy your =stay in Thailand.

Worant Koetsinchai

Assistant Professor Dr. Woranuch Koetsinchai
Dean, Faculty of Science, King Mongkut’s University of Technology Thonburi

# Contents

# Keynote Speakers and Invited Speakers

Professor Neerchal, Nagaraj
University of Maryland, Baltimore County, USA
Topic Talk: Estimating Risk from Incidence Data

Dr. Nagaraj Neerchal obtained his Ph.D. in Statistics from the Iowa State University in 1986, under the guidance of Wayne A. Fuller. Dr. Neerchal joined the faculty of Department of Mathematics and Statistics at UMBC in 1986, where is currently a Professor of Statistics and the department chair. He is currently the co-director of an NSF funded Research Experience for Undergraduates Site on High Performance Computing, which is in its 5th summer. His research interests are modeling and analysis of correlated observations. Most recently has been involved in a number of applications in Engineering, Biological sciences involving high performance computing. He has over 50 publications, mentored 3 Masters and 10 PhD students.

Dr. Neerchal received the Distinguished Achievement Award and Medal from the American Statistical Association's Section on Statistics and the Environment in 2000. He was elected Fellow of American Statistical Association in 2010. In 2011, he was appointed Faculty Fellow of Alex Brown Center for Entrepreneurship at UMBC.

Dr. Neerchal is this year's recipient of University System of Maryland Board of Regents Award for Excellence in Innovation.

Professor Bikas K Sinha
Retired Professor of Statistics
Indian Statistical Institute, Kolkata, INDIA
Topic Talk: SuDoKu Puzzles : A Combinatorial Marvel

After earning Ph.D. in Statistics from Calcutta University in 1972, Bikas K Sinha was inducted as the youngest faculty member in the Department of Statistics, Calcutta University. Professor Sinha was awarded "PC Mahalanobis Gold Medal: 1980" by the Sadharan Brahma Samaj for his contributions in Statistics. He was appointed an "Expert on Mission" for United Nations Statistics Training Program on Survey Methodology in 1991. He has served as a Visiting Faculty in a number of universities in USA & Canada and has visited a host of universities as Research Collaborator/ Invited Speaker in all the continents. He has also co-authored (i) one graduate-level text book published by John Wiley on Finite Population Sampling, and (ii) one research-level book on Social Networks and Applications published by Sage Publications, USA. Besides these monographs and books, Professor Sinha has authored / co-authored more than 130 research articles published in National & International Journals.

Professor Sinha is an elected member of the International Statistical Institute since 1985. He has served on the Editorial Board of several National & International Journals in Statistics. He also served as the Editor of Calcutta Statistical Association Bulletin for 8 years. Of late, Professor Sinha had been nominated as a Member of the National Statistical Commission [NSC], GoI, set-up as an Apex Body by an Act of Indian Parliament and operationalized on July 12, 2006 for a 3-year term. This is the highest policy-making body for Indian Statistical System.

Professor N. Rao Chaganty
Old Dominion University
Norfolk, USA
Topic Talk: Models and Estimation Methods for Repeated Ordinal Data

Professor N. Rao Chaganty is the Statistics Program Director in the Department of Mathematics and Statistics at Old Dominion University in Norfolk, VA, USA. He obtained his PhD and MS degrees in Statistics from Florida State University, USA. Prior to that he was a student at the Indian Statistical Institute, India, where he obtained B. Stat (Hons) and M. Stat degrees. He is a Fellow of the American Statistical Association, and an active member of the International Indian Statistical Association, served as its president in 2014. His current research interests are generalized linear models, estimating equations, modeling and analysis of repeated, clustered binary, ordinal, and count data using copulas.

Professor John P. Morgan
Virginia Tech Blacksburg, USA
Topic Talk: Row-Column Designs for Factorial Experiments

J.P. Morgan, professor of statistics, is the Assistant Dean for Graduate Studies and Strategic Initiatives in the College of Science at Virginia Tech. Morgan, of Bedford, Va., has worked in the Department of Statistics since his arrival at Virginia Tech in 2000. As the dean for graduate studies and strategic initiatives, Morgan will be responsible for managing graduate studies administration in the college, coordinating departmental reviews, spearheading faculty development initiatives, meeting with prospective faculty, and assisting the dean of curriculum and instruction with graduate-level course proposals.

Prior to his arrival at Virginia Tech, Morgan was a professor in the Department of Mathematics and Statistics at Old Dominion University. He has twice been a Visiting Fellow at the Isaac Newton Institute for Mathematical Sciences, Cambridge; a Visiting Researcher at the College of Mathematical Sciences, Queen Mary University of London; and a Visiting Researcher, Department of Mathematics and Statistics, Goldsmith's College, University of London. He is a member of the American Statistical Association, Institute of Mathematical Statistics, the Institute for Combinatorics and its Applications, and has been a three-time President of the Virginia Chapter of the ASA.

Morgan holds a bachelor's in mathematics and statistics, summa cum laude, from Roanoke College; and a master's and doctorate degrees in statistics from the University of North Carolina at Chapel Hill. He has more than 60 published articles, many as part of projects supported by the National Science Foundation, and has presented more than 60 papers at professional meetings. He is currently associate editor of The American Statistician, and past associate editor of the Journal of the American Statistical Association and the Journal of Statistical Planning and Inference.

Professor Dankmar Böhning
University of Southampton, UK
Topic Talk: Capture-Recapture Without Tears – the Size of the Homeless Population in Utrecht (NL)

Professor Dankmar Böhning is Chair in Medical Statistics with a joint appointment between the Southamption Statistical Sciences Research Institute, of which he is also Deputy Director, and the Faculty of Medicine at the University of Southampton. Previously, had professorial positions in Berlin, Penn State, Munich, Reading and Southampton.

Professor Böhning has an international reputation in several key areas in medical statistics, biostatistics and biometry including meta-analysis and review research, nonparametric random effects (mixed) modelling and capture-recapture applications for the life sciences. He has published about 100 papers in leading journals in statistics and medical statistics including Biometrika, Biometrics, Biostatistics, Statistics in Medicine, the Biometrical Journal or the Journal of the Royal Statistical Society A, B, and C.

Professor Böhning has been Associate Editor of Biometrics, Computational Statistics and Data Analysis and Statistical Modelling for more than 10 years and is Associate Editor of the Biometrical Journal and Statistical Methods in Medical Research,. He has supervised successfully about 30 PhD students and is currently supervising 4 PhD students on different topics. Active collaborations exist with partners in Berlin and Münster (Germany), Utrecht (Netherlands), Rome (Italy), Izmir (Turkey), Bangkok (Thailand), Cornell (USA), Rennes (France) and Graz (Austria).

In his academic life span, Professor Böhning has attracted numerous grants including 8 grants from the German Research Foundation. He has written several books including books on Epidemiology, meta-analysis and mixture models.

Professor Cathy W.S. Chen
Feng Chia University, TAIWAN
Topic Talk: Dynamic Quantile Forecasting: Application to Pairs Trading

Professor Cathy W.S. Chen is a distinguished Professor at Department of Statistics, Feng Chia University, Taiwan. She has made significant contributions to the areas of Bayesian inference, diagnostics and model comparison techniques for time series, as well as forecasting. She contributed to developing Bayesian methodology in econometrics and to applications of statistics to epidemiology research.

Professor Chen has an internationally recognized record of research and scholarship as is demonstrated by the 75 papers published or accepted for publication in high quality academic journals. She became an Elected Member of International Statistical Institute (ISI) in 2008. In 2010, she was awarded with the Chartered Statistician (CStat) title which is the Royal Statistical Society's highest professional award.

Professor Chen is also making a large contribution to the profession through her editorial work. She is an Associate Editor of a number of prominent journals, including the following highly regarded journals: Journal of Business and Economic Statistics, Computational Statistics & Data Analysis, the Australian and New Zealand Journal of Statistics, PLOS ONE, and Computational Statistics.

Professor Sangyeol Lee
Seoul National University, Seoul, KOREA
Topic Talk: Change Point Test for Time Series Models

Professor Sangyeol Lee has been professor at Department of Statistics, Seoul National University since 1997. He earned Ph.D from Department of Mathematics, University of Maryland, in August 1991. His research interests are time series analysis, change point test and goodness of fit test. He received – Korea Gallup Prize in 2010 and has been an elected member of ISI (International Statistical Institute) in 2011. He is research director of Korean Statistical Society from 2009 to 2011 and has served as associate editor a number of journals including Sequential Analysis, Annals of Institute of Statistical Mathematics, Asia and Pacific Financial Markets, Journal of Korean Statistical Society, Statistica Sinica and Computational Statistics.

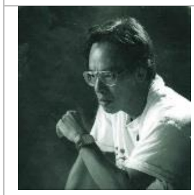
Professor Samruam Chongcharoen
Graduate School of Applied statistics
National Institute of Development Administration, Thailand
Topic Talk: A New Test for the Mean Vector in High-Dimensional Data

Professor Samruam Chongcharoen earned his Ph.D. in Statistics from Department of Statistics, University of Missouri-Columbia, U.S.A. in 1998 with master degree in Mathematics Actuarial Science, Central Connecticut State University, U.S.A. in 1994. His research interests include order restricted statistical inference, Statistical modeling and currently in high dimensional statistics and modeling in actuarial science. He has published about 25 papers in international journals. He is a reviewer of both international journal and national journal such as journal of Applied Statistics, computational Statistics and Data Analysis etc. He has supervised about 10 Ph.D. students and is currently supervising 4 Ph.D. students, 2 Master students in statistics and 2 Master students in actuarial science. He wrote 2 Thai textbooks in applied nonparametric method and linear models.

# Invited Speakers for the Session of Cooperative Statistical Research Network (CRN)

Professor S. Rao Jammalamadaka
University of California, California, USA
Topic Talk: Statistical Consulting—the Need for Theory

Professor Sreenivasa Rao Jammalamadaka, known to many as J.S. Rao, is among the leading researchers in statistical science during the last four decades. His research and scholarship has had a great influence in theory and applications in goodness of fit testing and directional statistics. He earned his Ph.D., M.Stat. and B.Stat. from the Indian Statistical Institute, Calcutta, India. He started his academic career at Indiana University before joining University of California, Santa Barbara in 1976 where he is now a Distinguished Professor. He also holds Adjunct Professor appointments at the Curtin University in Australia and the Advanced Institute of Mathematics, Statistics, and Computer Science in India. He is the author of three books and close to 200 research publications. He produced more than 40 Ph.D.'s. He achieved international reputation for pioneering work in developing new tools to analyze nonlinear phenomenon like directions and circularity. For his pioneering contribution to statistical theory and applications, Professor Jammalamadaka received various awards and honors. He has been elected a Fellow of the American Statistical Association, Institute of Mathematical Statistics, Institute of Combinatorics and its Applications, the C. Y. O'Connor Fellow of the Curtin University of Technology, and an Elected Member of the International Statistical Institute. He has been awarded an Honorary Doctorate by the Swedish University of Agricultural Sciences in 2012, and received Honorary Fellowship of the Indian Society for Probability and Statistics in 2014. He has been an Invited Speaker at more than 250 national and international statistical conferences and served as a member of the Editorial Boards of Statistics and Probability Letters, Stochastic Modeling and Applications, Journal of Nonparametric Statistics and the Journal of American Statistical Association.



Professor Hung T. Nguyen
New Mexico State University, USA and Chiang Mai University, Thailand
Topic Talk: Why & when we should use Bayesian Statistics?

Prof. Hung T. Nguyen received his Ph.D. degree in Mathematics in1975. He is Emeritus Professor of Mathematical Sciences, New Mexico State University, USA. and Adjunct Professor of Economics, Chiang Mai University, Thailand. His current research interests include Statistics for Econometrics, Copulas, heavy tailed distributions, maximum entropy inference. He has received numerous awards and recognition including:

- Westhafer Award for Excellence in Research and Creativity (2000): New Mexico State University, Las Cruces.
- Distinguished Lukacs Professor of Statistics (Spring 2002): Bowling Green State University, Ohio.
- Distinguished Faculty Fellow (Summers 2002 and 2003): American Association of Engineering Education (ASEE), Summer Faculty Research Programs.
- Fellow of the International Fuzzy Systems Association (April 2007).
- Life Fellow of Japan Society of Management Engineers (October 2012).

Associate Professor Jirawan Jitthavech
Graduate School of Applied statistics
National Institute of Development Administration, Thailand
Topic Talk: A Statistical Method for Variable Elimination in Nested DEA Models

Jirawan Jitthavech is an associate professor at School of Applied Statistics, National Institute of Development Administration. She earned a Bachelor degree of Science in Education (Mathematics, first class honors) from Prince of Songkla University in 1974, M.S. in Applied Statistics (Honors) in 1976 from School of Applied Statistics, National Institute of Development Administration and Ph.D. in Statistics in 1984 from University of Georgia, USA. Her research interests are statistical modeling, sampling theory and techniques, data mining, experimental designs and including data analysis.

Associate Professor Montip Tiensuwan
Mahidol University, Thailand

Montip Tiensuwan is an Associate Professor at Department of Mathematics, Faculty of Science, Mahidol University. She earned a Ph.D. from Department of Mathematics, the University of Western Australia. She is an IDP association membership of Australia.

Her research interests include statistical modeling applied to environment biology and medicine such as, for example, air pollution, water pollution, cancer, malaria, HIV/AIDS, etc. She is also interested in Ranked Set Sampling, Multiple Criteria Decision Making (MCDM), Statistical methods for assessing agreement, Bayesian analysis, and non-linear mixed effects models.

She served on the editorial board of Science Asia (2010-present), vice president and editor in chief Thailand Statistician Journal of Thai Statistical Association (2012-present). She is also a membership of society such as the international biometric society (1985-present), the statistical society Australia (1988-1994), the Thai statistical association (1979-present), and the science society of Thailand under the Royal Patronage of H.M. the King (1994-present).

# Committee

## *Advisory Board*

Adisak Pongpullponsak, Prof.
Bundit Fungtammasan, Ph.D., Assoc. Prof.
Khunying Suchada Kiranandana, Ph.D., Prof.
Sakarindr Bhumiratana, Ph.D., Assoc. Prof.
Sivawan Phoolphundh, Ph.D., Asst. Prof.
Woranut Koetsinchai, Ph.D., Asst. Prof.

## *Steering Committee*

Adisak Pongpullponsak, Prof.
Boobphachard Chansawang, Lecturer
Chunchom Pongchavalit, Ph.D., Asst. Prof.
Khajonpong Akkarajitsakul, Ph.D.
Sukuman Sarikavanij, Ph.D., Asst. Prof.
Tasanee Tantipisankul, Asst. Prof.
Thiradet Jiarasuksakun, Ph.D., Asst. Prof.
Waraporn Chatanin, Ph.D.
Yupaporn Areepong, Ph.D., Assoc. Prof.

## *Technical Program Committee*

Adisak Pongpullponsak, Prof.
Anirut Luadsong, Ph.D., Asst. Prof.
Anuwat Sae-tang, Ph.D.
Chatchawan Watchararuangwit, Ph.D.
Chukiat Worasucheep, Asst. Prof.
Chunchom Pongchavalit, Ph.D., Asst. Prof.
Nitima Aschariyaphotha, Ph.D., Asst. Prof.
Poom Kumam, Ph.D., Assoc. Prof.
Puchong Praekhaow, Assoc. Prof.
Settapat Chinviriyasit, Ph.D., Assoc. Prof.
Sukuman Sarikavanij, Ph.D., Asst. Prof.
Teerapol Saleewong, Ph.D.
Thanet Chitsuphaphan, Lecturer
Usa Humphries, Ph.D., Assoc. Prof.
Wibulsak Wattayu, Ph.D.
Wirawan Chinviriyasit, Ph.D., Assoc. Prof.

## *Registration and Financial Committee*

Chunchom Pongchavalit, Ph.D., Asst. Prof.
Daow Saguanrangsirikul, Lecturer
Duenpen Kirdnark, Ms.
Peangpit Rodpan, Mrs.
Pirun Dilokpatpongsa, Mrs.
Pronpimol Promsri, Ms
Ratchanee Archevarahuprok, Mrs.
Tanatporn Boonpee, Mrs.
Tasanee Tantipisankul, Asst. Prof.
Viriya Tripanyasat, Lecturer
Wiwat Sakonsontiset, Lecturer

## *Sponsoring Committee*

Chumchom Pongchavalit, Ph.D., Asst. Prof.
Daow Saguanransirikul, Lecturer
Pirun Dilokpatpongsa, Mrs.
Warisa Yomsatieankul, Ph.D.
Wibulsak Wattayu, Ph.D.
Wiwat Sakonsontiset, Lecturer

## *Committee in Venue and Event*

Angkool Wangwongchai, Ph.D.
Anuwat Sae-tang, Ph.D.
Boobphachard Chansawang, Lecturer
Charoen Sinaphiromsaran, Lecturer
Chatchawan Watchararuangwit, Ph.D.
Dawud Thongtha, Ph.D.
Mareena Swannaphoom, Ms
Ornrudee Suttisri, Assoc. Prof.
Parinya Sa-ngiamsunthorn, Ph.D.
Pariwate Varnakovida, Ph.D.
Pinitnan Ruttananurak, Mr.
Pronpimol Promsri, Ms.
Puchong Praekhaow, Assoc. Prof.
Teerapol Saleewong, Ph.D.
Thanet Chitsuphaphan, Lecturer
Usa Humphries, Ph.D., Assoc. Prof.
Warisa Yomsatieankul, Ph.D.
Yingluk Autsapanich, Ms.

## *Committee in Public Relations and Equipment*

Achirapol Morachart, Mr.
Aphichat Kulpaibuttra, Mr.
Apirada Warashoon, Mrs.
Jintana Sukpetch, Ms.
Kanyapa Wongsawattana, Ms.
Khajonpong Akkarajitsakul, Ph.D.
Panattaporn Meechai, Ms.
Parinya Sa-ngiamsunthorn, Ph.D.
Pawanya Prachumphu, Ms.
Phond Phunchongharn, Ph.D.
Pitaya Tungsiriwatanakul, Mr.
Potmetee Khotsupo, Mr.
Rapeeporn Kuirat, Ms.
Settapat Chinviriyasit, Ph.D., Assoc. Prof.
Supoth Meechart, Mr.
Yingluk Autsapanich, Ms.
Yuttasak Labudin, Mr.

## *Secretary*

Anuwat Sae-tang, Ph.D.
Dawud Thongtha, Ph.D.
Pirun Dilokpatpongsa, Mrs.
Settapat Chinviriyasit, Ph.D., Assoc. Prof.
Sukuman Sarikavanij, Ph.D., Asst. Prof.
Waraporn Chatanin, Ph.D.
Warisa Yomsatieankul, Ph.D.

# Reviewers

1. Andrei I. Volodin, Ph.D., Professor of Statistics, Department of Mathematics and Statistics, University of Regina, Regina, Saskatchewan, Canada
2. Bikas K. Sinha, Ph.D., Professor of Statistics, Indian Statistical Institute, India
3. Bimal K. Sinha, Ph.D., Professor of Statistics, Department of Mathematics and Statistics, University of Maryland, Baltimore County, USA
4. Dankmar BÖhning, Ph.D., Professor in Medical Statistics, Southampton Statistical Sciences Research Institute & School of Mathematics, University of Southampton, UK
5. Guido Knapp, Ph.D., Associate Professor, Clausthal University of Technology, Germany
6. Hung T. Nguyen, Ph.D., Professor of Mathematics and Statistics, New Mexico State University, USA and Chiang Mai University, Thailand
7. Jeong-Soo Park, Ph.D., Professor of Statistics, Department of Statistics, Chonnam National University, South Korea
8. John J. Borkowski, Ph.D., Professor of Statistics, Department of Mathematical Sciences, Montana State University, Bozeman, Montana, USA
9. Marisa Eisenberg, Ph.D., Assistant Professor in Epidemiology and Mathematics, Departments of Mathematics and Epidemiology, University of Michigan, Ann Arbor, USA
10. Sreenivasa Rao Jammalamadaka, Ph.D., Professor of Statistics, Department of Statistics and Applied Probability, University of California, Santa Barbara, CA, USA
11. Tonghui Wang, Ph.D., Professor in Statistics, Department of Mathematical Science, New Mexico State University USA
12. Vasudevan Mangalam, Ph.D., Senior Lecturer in Mathematics, Department of Mathematics, University Brunei Darussalam, Brunei
13. Vladik Kreinovich, Ph.D., Professor of Computer Science, University of Texas at El Paso, USA
14. Adisak Pongpullponsak
    Professor, Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi
15. Akara Prayote, Ph.D.
    Lecturer, Department of Computer and Information Science, Faculty of applied science,
    King Mongkut 's University of Technology North Bangkok
16. Ampai Thongteeraparp, Ph.D.
    Lecturer, Department of Statistics, Faculty of Science, Kasetsart University
17. Anamai Na-udom, Ph.D.
    Assistant Professor, Department of Mathematics, Faculty of Science, Naresuan University
18. Anawat Pinisakul, Ph.D.
    Assistant Professor, Department of Chemistry, Faculty of Science, King Mongkut's University of Technology Thonburi
19. Anirut Luadsong, Ph.D.
    Assistant Professor, Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi
20. Anuwat Sae-Tang, Ph.D.
    Lecturer, Department of Mathematics, Faculty of Science, King Mongkut's University of
    Technology Thonburi
21. Autcha Araweeporn, Ph.D.
    Assistant Professor, Department of Statistics, King Mongkut's Institute of Technology Ladkrabang.
22. Bungon Kumphon, Ph.D.
    Assistant Professor, Department of Mathematics, Faculty of Science, Mahasarakham University.
23. Chatchawan Watchararuangwit, Ph.D.
    Lecturer, Department of Mathematics, Faculty Science, King Mongkut's University of Technology Thonburi
24. Chukiat Worasucheep
    Assistant Professor, Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi
25. Chaiya Dumkum, Ph.D.
    Assistant Professor, Department of Production Engineering, Faculty of Engineering, King Mongkut's University of Technology Thonburi
26. Chunchom Pongchavalit, Ph.D.
    Assistant Professor, Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi
27. Dawud Thongtha, Ph.D.
    Lecturer, Department of Mathematics , Faculty Science ,
    King Monkut's University of Technology Thonburi
28. Dusadee Sukawat, Ph.D.
    Senior Lecturer, Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi
29. Kamon Budsaba, Ph.D.
    Associate Professor, Department of Mathematics and Statistics, Faculty Science and Technology, Thammasat University
30. Katechan Jampachaisri, Ph.D.
    Assistant Professor, Department of Mathematics, Faculty of Science, Naresuan University, Phitsanulok.
31. Krisana Lanumteang, Ph.D.
    Lecturer, Department of Statistics, Faculty of Science, Maejo University
32. Kunlaya Pattanakul, Ph.D.
    Lecturer, Department of Statistics, Faculty of Science, Khon Kaen University
33. Kusaya Plungpongpun, Ph.D.
    Assistant Professor, Department of Statistics, Faculty of Science, Silpakorn University
34. Manad Khamkong, Ph.D.
    Assistant Professor, Department of Statistics, Faculty of Science, Chiang Mai University.
35. Monchaya Chiangpradit, Ph.D.
    Lecturer, Department of Mathematics, Faculty of Science, Mahasarakham University.
36. Montip Tiensuwan, Ph.D.
    Assoc.Prof., Department of Mathematics, Faculty of Science, Mahidol University
37. Montree Piriyakul, Ph.D.
    Associate Professor, Departmet of Statistics, Faculty of Sciences, Ramkhamhaeng University.
38. Naowarut Meejun, Ph.D.
    Lecturer, Department of Statistics, Faculty of Science, Silpakorn University
39. Naratip Jansakul, Ph.D.
    Associate Professor,Department of Mathematics,Faculty of Science, Prince of Songkla University
40. Ngamphol Soonthronworasiri, Ph.D.
    Lecturer, Department of Tropical Hygiene, Faculty of Tropical Medicine, Mahidol University
41. Nipaporn Chutiman, Ph.D.
    Assistant Professor, Department of Mathematics, Faculty of Science, Mahasarakham University.
42. Nitima Aschariyaphotha, Ph.D.
    Assistant Professor, Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi
43. Nuengruithai Tharawatcharasart, Ph.D.
    Lecturer, Mahidol University Kanchanaburi Campus
44. Orathai Chuacharoen, Ph.D.
    Lecturer, Department of Statistics, Faculty of Sciences, Ramkhamhaeng University.

45. Parinya Sa Ngiamsunthorn, Ph.D.
Lecturer, Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi
46. Patrinee Traisathit, Ph.D.
Assistant Professor, Department of Statistics, Faculty of Science, Chiang Mai University.
47. Piyapatr Busababodhin, Ph.D.
Lecturer, Department of Mathematics, Faculty of Science, Mahasarakham University.
48. Pongsa Pornchaiwiseskul, Ph.D.
Associate Professor, Faculty of Economics, Chulalongkorn University
49. Prasit Payakkapong
Associate Professor, Department of Statistics, Faculty of Science, Kasetsart University.
50. Puchong Praekhaow
Associate Professor, Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi
51. Rachanee Phuwapathanapun, Ph.D.
Assistant Professor, Department of Statistics, Faculty of Science, Ramkhamhaeng University.
52. Rawee Suwandechochai, Ph.D.
Lecturer, Department of Mathematics, Faculty of Science, Mahidol University
53. Ronnason Chinram, Ph.D.
Associate Professor, Department of Mathematics and Statistics, Faculty of Science, Prince of Songkla University
54. Sa-aat Niwitpong, Ph.D.
Associate Professor, Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok
55. Saowanit Sukparungsee, Ph.D.
Associate Professor, Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok
56. Settapat Chinviriyasit, Ph.D.
Associate Professor, Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi
57. Siriluck Jermjitpornchai
Assistant Professor, Department of Mathematics, Faculty of Science, Mahasarakham University.
58. Sirima Suwan, Ph.D.
Lecturer, Department of Statistics, Faculty of Science, Chiang Mai University.
59. Sujitta Suraphee, Ph.D.
Assistant Professor, Department of Mathematics, Faculty of Science, Mahasarakham University.
60. Suksan Prombanpong, Ph.D.
Assistant Professor, Department of Production Engineering, Faculty of Engineering, King Mongkut's University of Technology Thonburi
61. Sukuman Sarikavanij, Ph.D.
Assistant Professor, Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi.
62. Sumittra Reungpeerakul
Associate Professor, Department of Statistics, Faculty of Science, Ramkhamhaeng University.
63. Supunnee Ungpansattawong, Ph.D.
Associate Professor, Department of Statistics, Faculty of Science, Khon Kaen University.
64. Tasanee Tantipisalkul
Assistant Professor, Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi
65. Teerapol Saleewong, Ph.D.
Lecturer, Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi
66. Thidaporn Supapakorn, Ph.D.
Lecturer, Department of Statistics, Faculty of Science, Kasetsart University
67. Thipwan Kate-intra, Ph.D.
Lecturer, Department of Mathematics, Faculty of Science, Naresuan University, Phitsanulok.
68. Thiradet Jiarasuksakun, Ph.D.
Assistant Professor, Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi
69. Tosaporn Talangtam, Ph.D.
Department of Mathematics, Faculty of Science, Khon Kaen University
70. Usa Humphries, Ph.D.
Associate Professor, Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi
71. Vanida Pongsakchat, Ph.D.
Lecturer, Department of Mathematics, Faculty of Science, Burapha University.
72. Veeranan Pongsapakdee, Ph.D.
Associate Professor, Department of Statistics, Faculty of Science, Silpakorn University
73. Waraporn Chatanin, Ph.D.
Lecturer, Department of Mathematics, Faculty of Science, King Monkut's University of Technology Thonburi
74. Wararit Panichkitkosolkul, Ph.D.
Assistant Professor, Department of Mathematics and Statistics, Faculty Science and Technology, Thammasat University
75. Watcharin Klongdee, Ph.D.
Risk and Insurance Lab, Department of Mathematics, Faculty of Science, Khon Kaen University
76. Wattanavadee Sriwattanapongse, Ph.D.
Associate Professor, Department of Statistics, Faculty of Science, Chiang Mai University
77. Wibulsak Wattayu, Ph.D.
Lecturer, Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi
78. Winai Bodhisuwan, Ph.D.
Assistant Professor, Department of Statistics, Faculty of Science, Kasetsart University
79. Wirawan Chinviriyasit, Ph.D.
Associate Professor, Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi
80. Woranut Koetsinchai, Ph.D.
Assistant Professor, Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi
81. Wuttichai Srisodaphol, Ph.D.
Lecturer, Department of Statistics, Faculty of Science, Khon Kaen University
82. Yupaporn Areepong, Ph.D.
Associate Professor, Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok

# Editors of Book of Proceedings

Chatchawan Watchararuangwit, Ph.D.
Chukiat Worasucheep, Assistant Professor
Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi

# Oral Presentation Program

| Time | Room A: Park View I | Room B: Park View II | Room C: Park Place |
|---|---|---|---|
| **July 15, 2015** | Chairman: Prof. Bikas K Sinha | Chairman: Prof. Cathy W.S. Chen | Chairman: Prof. N. Rao Chaganty |
| | Co-chairman: Prem Jansawang | Co-chairman: Dr. Suvil Chomchaiya | Co-chairman: Dr. Pariwate Varnakovida |
| 13.40 - 14.00 | Computation of Quantile Function of the Quasi Lindley Distribution and Janardan Distribution by Using Lambert W Function<br>**Panpharisa Khongthip** | A Numerical Solution of Fractional Black-Scholes Equation<br>**Prasert Phaochoo** | The Classification of Severity Level in Exertional Heatstroke Patients by Applying a Decision Tree Technique<br>**Uthumporn Domthong** |
| 14.00 - 14.20 | The Negative Binomial-Sushila (NB-S) Distribution with Application to Count Data<br>**Darika Yamrubboon** | The Negative Hypergeometric Distribution Revisited with Applications to Gaming<br>**Rohan J. Dalpatadu** | Measurement Errors and Misclassifications in Covariates in Logistic Regression: Bayesian Adjustment of Main and Interaction Effects and the Sample Size Implications<br>**Shahadut Hossain** |
| 14.20 - 14.40 | Corrected Score Estimators in Linear Multivariate Regression Model with Measurement Errors of Heteroscedasticity Groups<br>**Wannaporn Junthopas** | Modeling and Forecasting Volatility Series: With Reference to Gold Price<br>**Sandya N. Kumari** | A New Local Likelihood Estimation Method for Semi-parametric Random Effects Models<br>**Lei Liu** |
| 14.40 - 15.00 | Mittag - Leffler Function Distribution - A New Generalization of Hyper-Poisson Distribution<br>**Subrata Chakraborty** | Adaptive Interest Rate Models and Their Empirical Performance: A Case Study of Thailand’s Policy Interest Rate<br>**Krisada Muangsri** | |
| 15.00 - 15.20 | Coffee break | | |
| | Chairman: Assoc. Prof. Kamon Budsaba | Chairman: Prof. Nagaraj K. Neerchal | Chairman: Assoc. Prof. Montip Tiensuwan |
| | Co-chairman: Asst. Prof. Taweesak Siripornpibul | Co-chairman: Dr. Palakorn Seenoi | Co-chairman: Asst. Prof. Katechan Jampachaisri |
| 15.20 - 15.40 | Factor Affecting Operational Performance in Logistics Service Providers in Thailand<br>**Chanikarn Jeenanupan** | Development of Attention, Cognitive and Affective Skills Through Mindfulness Training Among Young Children in a Classroom<br>**Yau Yan Wong** | Tracer Study of the BS Statistic Program AT Eastern VISAYAS State University: 1997 – 2014<br>**Thea C. Galos** |
| 15.40 - 16.00 | Orthogonal Space Filling Design for Computer Simulated Experiments<br>**Jaratsri Rungrattanaubol** | The Role of Manager’s Knowledge and Experience on Hotel Innovation and Performance<br>**Narongsuk Pongsathomwiwat** | Mathematical Competence of Pre-Service Teacher Education Students of State Universities and Colleges in Region 8<br>**Thelma A. Saladaga** |
| 16.00 - 16.20 | Response Surface Methodology using Constrained Nelder-Mead Algorithm<br>**Chantha Wongoutong** | Crew Rostering in Thai Airways Case by Using Greedy Algorithm<br>**Thanaphat Limgitnuwat** | Teachers’ Perceptions and Decision on the Inclusion of History of Mathematics to Instruction and their Students’ Mathematical Competence and Attitude<br>**Grace D. Severo** |
| 16.20 - 16.40 | Factorwise Variance Dispersion Graphs<br>**John Borkowski** | Modified Differential Evolution for Permutation Flow Shop Scheduling<br>**Tawun Remsungnen** | |

| July 16, 2015 | **Room A: Park View I** | **Room B: Park View II** | **Room C: Park Place** |
|---|---|---|---|
|  | Chairman: Prof. Hung T. Nguyen | Chairman: Prof. Sangyeol Lee | Chairman: Assist. Prof. Winai Bodhisuwan |
|  | Co-chairman: Dr.Monthira Duangsaphon | Co-chairman: Dr. Suvil  Chomchaiya | Co-chairman: Assoc. Prof. Supunnee Ungpansattawong |
| 13.00 - 13.20 | The Interval Estimation of Poisson Distribution by Using Maximum Likelihood, Markov Chain Monte Carlo, and Bayes Methods<br>**Autcha Araveeporn** | Using R Statistical Software for Visualizing the Epidemiological Data of Dengue Disease in Chiang Mai, Thailand<br>**Veerasak Punyapornwithaya** | Two-Echelon Inventory Model with Return Items and Lead Time<br>**Jarumas Sangsavang** |
| 13.20 - 13.40 | Discriminating Between the Weibull and Power Lindley Distributions Utilizing Fisher Information<br>**Yuwadee Sangsanit** | An Alternative Attribute Selection to Construct Decision Trees<br>**Weerinrada Wongrin** | A Comparison of Efficiency Between Multivariate Shewhart and Multivariate CUSUM Control Chart for Bivariate Copula<br>**Sasigarn Kuvattana** |
| 13.40 - 14.00 | Statistical Properties and Parameter Estimation of the Two-Sided Length Biased Inverse Gaussian<br>**Teerawat Simmachan** | Statistical Analysis for the Effectiveness of Tuberculosis Tests<br>**Shishen Xie** | Economic Design for  Fuzzy Variable Parameters $\overline{X}$ Control Charts  using $\alpha$ - cut under Non-normality<br>**Chaowalit Panthong** |
| 14.00 - 14.20 | Maximum likelihood estimation of the negative binomial-Erlang distribution<br>**Siriporn Samutwachirawong** | Classroom Voting Approach in Teaching and Learning Discrete Mathematical Structure<br>**Hernando jr.C. Abalos** | Minimum ATI (Average Total Inspection) Chain Sampling Plan - ChSP-4A $(c_1, c_2)$ r<br>**Raju Chockalingam** |
| 14.20 - 14.40 | A Method for Formulating Fuzzy Linear Regression Model and Estimating the Model Parameters<br>**Atchanut Rattanalertnusorn** |  | Economic Design of the New Variable Parameters Hodges-Lehmann Estimator Control Chart<br>**Vadhana Jayathavaj** |
| 14.40 - 15.00 | Approximations of the Coverage Probability by a Confidence Set Centered at the Positive-part James-Stein Estimator<br>**Sujitta Suraphee** |  | The Role of Driver's Attitude to Trucking Firm Performance<br>**Le Thi Ngoc Lan** |
| 15.00 - 15.20 | Coffee break |  |  |
|  | Chairman: Prof. Dankmar Böhning | Chairman: Asst. Prof. Bungon Kumphon | Chairman: Assist. Prof. Winai Bodhisuwan |
|  | Co-chairman: Dr.Krisana  Lanumteang | Co-chairman: Dr. Suvil  Chomchaiya | Co-chairman: Dr. Anuwat Sae-Tang |
| 15.20 - 15.40 | Heuristics for Two-Dimensional Rectangular Guillotine Cutting Stock<br>**Kimseng Tieng** | Principal Components Analysis of Socio- Economic Factors of Renewable Energy Users of Countries Such as Nepal<br>**Jyoti U. Devkota** | Mathematical Analysis of SARS Epidemic Model<br>**Siwaphorn Kanchanarat** |
| 15.40 - 16.00 | Forecasting the Level of Rainfall in Chiang Mai Basin Using Grey Systems Theory<br>**Chalermchai Puripat** | Brand Identity and Brand Loyalty in Thai Social Marketing<br>**Kwanchanok Chumnumporn** | The Meshless Local Petrov-Galerkin Method for Solving the Time Fractional Burgers' Equation<br>**Nuntapon Thamareerat** |
| 16.00 - 16.20 |  | The Pareto ArcTan Distribution: A Simple Way to Model Urban Agglomerations in the Asia Pacific Region<br>**Enrique Calderin** | Enhancing Productivity Through Capacity Modeling in Financial Services Industry<br>**Anurag Seksaria** |
| 16.20 - 16.40 |  |  | Process Capability Index $(\tilde{C}_{pk})$ for a Mixed  Bivariate Data as Applied to Stratified Product Groups<br>**Dennis C. de Paz** |
| 16.40 - 17.00 |  |  | Forecasting Early Warning in Thai Real Estate Crisis<br>**Pichanun Boonpromgul** |

| **July 17, 2015** | **Room A: Park View I** | **Room B: Park View II** | **Room C: Park Place** |
|---|---|---|---|
| | Chairman: Dr.Thidaporn Supapakorn | Chairman: Prof. Cathy W.S. Chen | |
| | Co-chairman: Dr. Saowapa Chaipitak | Co-chairman: Dr. Anuwat Sae-Tang | |
| 09.40 - 10.00 | Road Safety Audit Associated with Analytic Hierarchy Process in Thailand<br>**Napat Munsil** | Analyzing the Dual Long Memory in Thailand Stock Market<br>**Panupong Oonsupwilai** | |
| 10.00 - 10.20 | Measuring and Benchmarking the Efficiency of Solid Waste Removal in Khung Bang Kachao<br>**Anyamanee Palagavong Na Ayudhya** | Modelling Impatient Length of Stay Through Mixture Distributions. The Case of HIV/AIDS Length of Stay in Portuguese Hospitals<br>**Martins, Maria O.** | |
| 10.20 - 10.40 | Imbalanced Data Classification by Incremental Learning with Versatile Elliptic Basis (VEB) Function Neural Network<br>**Prem Junsawang** | | |
| 10.40 – 11.00 | Coffee break | | |
| | Chairman: Prof. S. Rao Jammalamadaka | Chairman: Prof. John P. Morgan | |
| | Co-chairman: Dr. Uthumporn Domthong | Co-chairman: Dr. Anuwat Sae-Tang | |
| 11.00 - 11.20 | Bayesian Approach for Generalized Lindley Distribution and Its Application by Different Loss Function<br>**Pimwarat Nanthaprut** | Evaluating Beaches for Tourism Business Opportunity<br>**Panpawee Chavandig** | |
| 11.20 - 11.40 | Discriminating Between the Generalized Exponential and Gamma Distributions<br>**Orawan Supapueng** | Probability Plot of the Length-biased Exponentiated Inverted Weibull Distribution<br>**Palakorn Seenoi** | |
| 11.40 - 12.00 | Crack Distribution Parameters Estimation by Maximum Likelihood Estimation, Moment Method and Bayesian Approach<br>**Thanasate Akkanphudit** | Assessing and Benchmarking the Efficiency of Local and City Governments in Building Resilience to Natural Disasters in Pathum Thani Province<br>**Morrakot Raweewan** | |
| 12.00 - 12.20 | Asymptotic Properties and Parameter Estimation Based on Two-Sided Crack Distribution<br>**Wikanda Phaphan** | Assessing Greening Activities for Small and Medium Enterprises in the Thai Automotive Sector by DEA Approach<br>**Morrakot Raweewan** | |

# Poster Presentation Program

| Time | Wednesday, July 15, 2015 |
|---|---|
| 13.30 – 16.30 | Stability in the probability of stochastic systems<br>**Manlika Rajchakit** |
| | Optimal guaranteed cost control for stochastic neural networks<br>**Grienggrai Rajchakit** |
| | Model-based Prediction of Length of Stay for Diabetes Patients: A Case Study of Nakornping Hospital, Chiang Mai, Thailand<br>**Taweesak Channgam** |
| | Prevalence and risk factors of Tuberculosis (TB) among HIV-Infected patients: A case study of Doitao hospital<br>**Krisana Lanumteang** |
| | Factors Affecting the Savings of the Immigration Police in Chiang Mai Province<br>**Rungkarn Jaiwongya** |
| **Time** | **Thursday, July 16, 2015** |
| 13.00 – 16.00 | Population Size Estimation Based Upon Zero-Modified Poisson Models<br>**Nupian Thepmong** |
| | Confidence Interval for population size using the Generalized Turing estimator<br>**Parawan Pijitrattana** |
| | A Robustness Study of Fed-Batch Fermentation for the Plasmid DNA Production Using Definitive-Screening Design<br>**Sarawuth Noppiboon** |
| | The Development of Mathematics Learning Achievement on the Topic of Applications of Quadratic Equations for Mathayomsuksa 4 Students by Technique of Mathematical Model<br>**Jaruwan Suwannalai** |
| | The Effect of Organizing Mathematics Learning Activity on the Topic of Applications Using the BSCS 5E Model of Teaching on Mathematics Achievement for Prathomsuksa 5 Student at School District 2 is Wat Om Yai<br>**Sirinapa Prakoobdee** |
| | Principal component mixture regression<br>**Geunyoung Moon** |
| | Motivational and Constructivist Teaching Methods in Mathematics on the Topic of Function for Mathayom 4 Students<br>**Wirakorn Boonchawalit** |
| | Soft prediction strength of clustering with mixture models<br>**Yeyoung Kim** |
| | Using strong and weak heredity for generating weighted A-optimal exact designs in mixture experiments<br>**Wanida Limmun** |

# List of Full Papers

***KEYNOTE SPEAKERS***

***INVITED SPEAKERS***

***ORAL PRESENTATION***

***Room: Park View I***

### *Room: Park View II*

### *Room: Park Place*

***POSTER PRESENTATION***

# KEYNOTE AND INVITED SPEAKERS

# Estimating Risk from Incidence Data

Nagaraj K. Neerchal
*Department of Mathematics and Statistics University of Maryland, Baltimore County, USA*
*e-mail: nagaraj@math.umbc.edu, nagaraj@umbc.edu*

## Abstract

Estimating risk usually requires the knowledge of both total number of individuals exposed and the number injured. Injury data are sometimes collected as new incidences occur, and in such cases the total number exposed is unknown. For example, a national database maintained by an ambulance service company consists of records associated with injuries suffered by emergency service providers such as emergency medical technicians, paramedics, and firefighters, along with adjunct workers such as secretaries, mechanics, and administrators. Repeated injuries to the same person are reported along with demographic data. However, the database consists of only individuals who have suffered at least one injury and the total number of individuals exposed is unavailable. Thus, estimating risk of injury becomes a challenging problem. In this talk we will review the various statistical issues involved in obtaining defensible estimates of injury risk from such a database. The ideas are applicable to a variety of other studies where truncated data are observed.

# Capture-Recapture without Tears – the Size of the Homeless Population in Utrecht (NL)

Dankmar Böhning
*Mathematical Sciences, University of Southampton, UK*
*e-mail: D.A.Bohning@soton.ac.uk*

### Abstract

Capture–recapture methods are frequently used to estimate the size of a population of interest which can be only partially observed. Each member of the population carries a count of identifications by some identifying mechanism – the number of times it has been identified during the observational period. Only positive counts are observed and inference needs to be based on this observed, truncated, count distribution.

One interesting and important application is estimation of the size of a homeless population. In this application we have data from 3 different observational periods of the target population of homeless people in the city of Utrecht (NL). The city of Utrecht runs a shelter where homeless people are allowed to stay overnight. Each observational period consists of 14 nights and for each person entering the shelter at least once it is observed how many nights he/she has been using the shelter. Clearly, only homeless people with shelter contact are entering into the system and the question remains how many have been missed out as not all homeless people are using the shelter, at least not every night. This can be viewed as a capture-recapture problem.

For valid prediction of the number of unobserved units, it is crucial to use an appropriate count data model. Building substantially on recent work by Rocchetti, Alf`o and Holling (2014) we consider ratios of neighboring count probabilities; these can be estimated by ratios of corresponding observed frequencies, independent of whether they arise from zero-truncated or untruncated count distributions. The guiding principle here is that it is often easier to find an appropriate regression model than working directly with modelling the distributional form of the count distribution.

We show how these ratios can be modeled by means of a regression approach, with a suitable link function. It is shown that this process leads not only to a valid count distribution, but also to a wide class of models.

We hope to demonstrate convincingly that a fairly simple, in fact straight line, regression model is fitting the homeless data satisfactorily, for all three periods, and that a simple capture-recapture estimator of the size of the homeless population of Utrecht can be gained from this ratio regression estimator. It will be also shown that this regression model uniquely relates to an old count distribution which will be revisited at this occasion.

Ultimately, the talk establishes links of the ratio regression approach to the mixing of binomial distributions, or in more generality, to the mixing of discrete exponential family distributions, although this will have largely digestive character.

# A New Test for the Mean Vector in High-Dimensional Data

Knavoot Jiamwattanapong and Samruam Chongcharoen
*Graduate School Applied statistics, National Institute of Development Administration, Thailand*
*e-mail: samruam@as.nida.ac.th*

## Abstract

For the testing of the mean vector where the data are drawn from a multivariate normal population, the renowned Hotelling's   test is no longer valid when the dimension of the data equals or exceeds the sample size. In this study, we consider the problem of testing the hypothesis   and propose a new test based on the idea of keeping more information from the sample covariance matrix. The development of the statistic is based on Hotelling's distribution. The asymptotic distribution is derived under the null hypothesis. The simulation results show that the proposed test performs well and is more powerful when the data dimension increases for a given sample size. An analysis of DNA microarray data with the new test is demonstrated.

*Keywords*: high-dimensional data, hypothesis testing, mean vector, block diagonal structure

# SuDoKu Puzzles: A Combinatorial Marvel

Bikas K Sinha
*Retired Professor of Statistics, Indian Statistical Institute, Kolkata, INDIA*
*e-mail: bikas_sinha@hotmail.com*

## Abstract

Sudoku is a popular combinatorial puzzle. It was first interpreted as a statistical experimental design by Subramani and Ponnuswamy (2009), and correctly analyzed by Saba and Sinha (2014). The latter authors took into account the lack of orthogonality between some features, which was missed out by the former authors. Here, we focus on the mathematical properties of a special subclass of Sudoku squares. Details of proofs and discussions of statistical issues are given in Sarkar and Sinha (2014).

*Keywords*: Latin Square, Orthogonal Latin Squares, Mutually Orthogonal Latin Squares, Internal Block, Sudoku Square, Magic Square, Diagonal Sudoku Square, Orthogonal Sudoku Squares, Cylindrical-shift Sudoku Square, Orthogonal Cylindrical-shift Sudoku Squares, Connected Design

# Statistical Consulting—the Need for Theory

S. Rao Jammalamadaka
*University of California, Santa Barbara, USA.*
*e-mail: rchagant@odu.edu*

## Abstract

Although most routine statistical consulting problems involve fairly straightforward, text-bookish solutions, there are often situations where one needs to develop new methodologies that are appropriate for the context. Such solutions break new ground, leading to new research and publications, justifying the view that good statistical research is always motivated by practical situations and examples. A few such examples from the perspective of the speaker, will be presented.

# Why and When We Should Use Bayesian Statistics?

Hung. T. Nguyen
*New Mexico State University (USA) & Chiang Mai University (Thailand)*
*e-mail: hunguyen@nmsu.edu*

## Abstract

We mention here some "critical thinking" in the art of using statistics at the service of economics, with special emphasis on epistemic uncertainty and Bayesian approach to decision-making.

Although there is a striking analogy between physics and economics finance, as they are uncertain dynamical systems, we should realize that there is no invariant laws in economics since they are phenomena involving decisons made by minds. The econophysics view of, say, dynamics of financial markets, should be an excellent example for applied statisticians to think about how to use conventional statistical tools to assist empirical works.

Although we realize that models are what we need to investigate, say,financial markets, e.g., using diffusion models, aspects of uncertainties involved are much more complex than physical systems. Modeling of epistemic uncertainty is delicate in statistical applications. We discuss here the traditional approach to epistemic uncertainty modeling leading to the know well-known Bayesian statistics. A good understanding of what a statistical theory stands for is essential for applied statisticians in their apploed works.

# A Statistical Method for Variable Elimination in Nested DEA Models

Jirawan Jitthavech and Vichit Lorchirachoonkul
*School of Applied Statistics, National Institute of Development Administration Bangkok, Thailand.*
*e-mail: jirawan@as.nidaa.ac.th*

## Abstract

A new definition of a relevant variable in a DEA model is proposed based on the number of efficient DMUs for variable selection by an iterative backward elimination procedure. The DEA results of the full model and a reduced model can be summarized in a correlated contingency table with a structural zero. A test statistic is derived for testing hypotheses by using the McNemar method with Bonferroni adjusted significance levels instead of conventional multiple comparisons. Two simulation populations, one with moderately correlated input variables and one with low correlated input variables, and a data set from a large-scale social experiment in U.S. public school education are used to illustrate the performance of the proposed procedure. From the simulation results, the proposed procedure can identify the relevant variables with high accuracy and eliminate the irrelevant variables effectively. In the data set from a large scale experiment in U.S. public school education, the reduced model selected by the proposed procedure is shown to be the better approximation of the full model than the reduced models selected by the Pastor et al. method.

*Keywords*: Data Envelopment Analysis, Efficiency Analysis, Variable Selection, Hypothesis Testing, Model Selection

# Row-Column Designs for Factorial Experiments

John P. Morgan
*Department of Statistics, Virginia Tech Blacksburg, VA 24061-0439, USA*
*e-mail: jpmorgan@vt.edu*

## Abstract

Mutually Orthogonal Latin Squares (MOLS) provide a large class of row-column designs for use with factorial treatments. Designs based on MOLS have each row and column block complete with respect to the levels of each factor. Alternative strategies based on Youden designs are explored here for situations where blocks cannot be complete. A Youden rectangle is a k by v row-column design on v treatments, where row blocks are complete and column blocks comprise a symmetric BIBD. This design also travels by the name Youden square, owing to the fact that if one interchanges the roles of rows and treatments, the result is a v by v square on k symbols with v (v-k) empty cells. This square, termed a (v,k)-Youden array, has the properties that (i) each of the k symbols appears once in each row and once in each column, and (ii) if each of the k symbols is replaced by a one, and each empty cell is filled with a zero, the result is a BIBD incidence matrix. Consider now two (v,k)-Youden arrays based on the same incidence matrix, and take them as a single array with an ordered pair of integers in each non-empty cell. This is a double Youden array if the ordered pair p,q occurs n1 times if p = q, and n2 times otherwise, where n1 and n2 are constants differing by one. Extending this notion, if any two of m Youden arrays form a double Youden array, one has a (v,k,m)-Youden array, corresponding to a Youden rectangle for m factors.

Here (v,k,m)-Youden arrays are constructed and their properties as statistical designs are explored. The construction is the last statistical work in the long and distinguished career of Professor Donald A. Preece, who died January 6, 2014. The talk will include an overview of his career, and will place (v,k,m)-Youden arrays in the context of other statistically useful combinatorial constructs that he explored.

# Models and Estimation Methods for Repeated Ordinal Data

N. Rao Chaganty
*Department of Mathematics and Statistics, Old Dominion University,Norfolk, VA 23529, USA*
*e-mail: rchagant@odu.edu*

### Abstract

Repeated or clustered ordinal data are common in longitudinal studies arising in biomedical and several other research areas. Statistical analyses of such data are complicated due the presence of dependence among the repeated measurements and the lack of a unique model that encompasses all types of dependence. In this talk we will discuss latent variable likelihood models for repeated or clustered ordinal data using Gaussian copula with probit and logit link functions. These models are flexible because they inherit the properties of the multivariate normal distribution. We will discuss maximum likelihood (ml) and estimating equations approaches for parameter estimation. We will derive the score functions and simplified expressions for the Hessian matrices, which allow easy computation of the standard errors for the marginal regression parameter estimates as well as the dependence parameters. Through asymptotic relative efficiency calculations, we demonstrate that ml estimators are superior as compared to estimators arising from previously established estimating equation approaches. We apply this likelihood-based methodology in an analysis of two real-life data examples using an R package developed specifically for the likelihood estimation. This is joint research work with Drs. Raghavendra Kurada and Roy Sabo.

# Dynamic Quantile Forecasting: Application to Pair Trading

Cathy W. S. Chen and Zona Wang
*Department of Statistics, Feng Chia University, Taiwan*
*e-mail: chenws@mail.fcu.edu.tw*

## Abstract

Pair trading is a statistical arbitrage strategy used to trade in the relative performance of stocks and profit by predicting which stock within a given stock pair will perform better. In this study, we propose a second-order logistic smooth transition generalized autoregressive conditional heteroscedastic (ST-GARCH) models to design trading entry and exit signals. We consider a Bayesian framework, designing an adaptive Markov chain Monte Carlo method for estimation, inference, and forecasting. This model allows us to use the upper and lower threshold values in the model as trading entry and exit signals. The second method is one-step-ahead quantile forecasting for the ST-GARCH model with the second-order logistic function, calculating the upper and lower quantile forecasts as entry and exit signals. The spread of fifteen daily return stock pairs found in U.S. markets are used as an illustration. We consider pair trading profits for two out-of-sample periods within a six-month time frame.

# Change Point Test for Time Series Models

Sangyeol Lee
*Department of Statistics, Seoul National University, Seoul, Korea*
*e-mail: sylee@stats.snu.ac.kr*

## Abstract

Since Page (1955), testing for a parameter change has played an important role in economics, engineering and medicine, and a vast number of articles exist in various research areas. In this talk, we consider the change point test for time series models based on the parameter estimates and residual-based cusum tests. It is shown that their limiting null distributions are a function of independent Brownian bridges. Some empirical results are provided for illustration.

# ORAL PRESENTATIONS

# Computation of Quantile Function of the Quasi Lindley Distribution and Janardan Distribution by Using Lambert W Function

P. Khongthip[1]*, P. Nanthaprut[1], W. Bodhisuwan[1] and C. Pudprommarat[2]

[1] *Department of Statistics, Faculty of Science, Kasetsart University, Bangkok, 10900, Thailand, panpharisa@gmail.com*

[2] *Department of Science, Faculty of Science and Technology, Suan Sunantha Rajabath University,*
*Bangkok, 10300, Thailand, chookait.pu@ssru.ac.th*

## Abstract

In this paper, we propose the quantile functions of a Quasi Lindley distribution and Janardan distribution. This method is based on non-principal branch of Lambert W function. Two theorems are introduced to solve explicit forms of the quantile functions for the Quasi Lindley distribution and Janardan distribution. Furthermore, we apply those theorems to generate a random variable of the Quasi Lindley and the Janardan distributions. In summary, Lambert W function is beneficial to construct the quantile function in explicit form.

*Keywords:* Generating, Janardan distribution, Lambert W function, Quantile function, Quasi Lindley distribution

*Corresponding Author
E-mail address: panpharisa@gmail.com

## 1 Introduction

The quantile function is function that return the value of random variable of which probability greater than or equal to specified probability. On the other word, the quantile function will return the minimum value of a random variable from among all random variables of which distribution functions exeed than specified probability. The quantile function is one way to define distribution function or probability function. It is also called percent point function or inverse cumulative distribution function. Moreover, the quantile function has relation with random variables, thus it is important to generate random variables.

There are two general methods to generate random variable that involve transforming a uniform random generator. The first method, called inverse transform method is related to the quantile function. It is used when one can be able to invert the distribution function of random variable, called the quantile function. [1--3]. The second method is acceptance-rejection method, which is related to probability function. This method is employed when the quantile function can not express in explicit form or it is difficult in inverting the corresponding distribution function [1, 4]. In addition to this basic technique, there are many other technique for generating random variable such as mixtures form of probability, ad hoc method, etc. [2, 5--7]. Obviously, the quantile function plays an important role for choosing the method of generation. Furthermore, if the quantile function can not express in explicit form, the way to solve such problem is use the Lambert W function.

The Lambert W function is a set of function including the branches of the inverse relation associated with exponential function and complex number. It is significantly useful function for solving the equation that involve exponential function. Lambert's transcendental equation was considered by Lambert in 1758 [8]. Consequently, Euler discussed and wrote the paper on its special case in 1783 [9]. However, it is important in applications in physics areas when it was reported and developed for Maple library 1990 [10]. Nowadays, The Lambert W function has various applications such as in biochemistry, chemistry, physics, engineering and statistics field [11--18]. In statistics, we use the Lambert W function to obtain the quantile function in explicit form which will be applied to the inverse transform method. This technique is easy requiring for application and simple way to solve the problem without more complicated method.

In this paper, we concern about the quantile function for the Quasi Lindley and Janardan distributions which are continuous probability distribution. In section 2, we present the Lambert W function then we give the quantile function of these distributions in section 3. The generation algorithms of method not using Lambert W function and the inverse transform method using Lambert W function were shown in section 4. We used R language version 3.1.2 [19]. Finally we give the also provide some conclusions.

## 2 Lambert W function

The Lambert W function, also called the omega function or the product log function that is originally proposed by Lambert in 1758 [8]. Lambert considered the solution to $x^a - x^b = (a-b)\nu x^{a+b}$, which is Lambert's transcendental equation [20]. In 1783, Euler transformed Lambert's transcendental equation into a special case with $a = b$, then it becomes $\ln x = \nu x^b$ [9]. The following three definitions are useful in determining the Lambert W function.

**Definition 1.** Let $z$ be a complex number, then the Lambert W function is a multivalued complex function follows:

$$W(z)exp(W(z)) = z \tag{1}$$

**Definition 2.** If $-1/e \leq z$ is a real number and $W(z) \geq -1$ then Eq. (1) is called principal branch of the Lambert W function and denoted by $W_0(z)$ or $W(z)$.

**Definition 3.** If $-1/e \leq z < 0$ is a real number and $W(z) \leq -1$ Eq. (1) is called non-principal branch of the Lambert W function and denoted by $W_{-1}(z)$
Fig. 1 shows the function along real branches of W from definition 2 and 3. According to Fig. 1, there is two possible real solutions when $z \in (-1/e, 0)$ and a unique solution when $z \geq 0$.

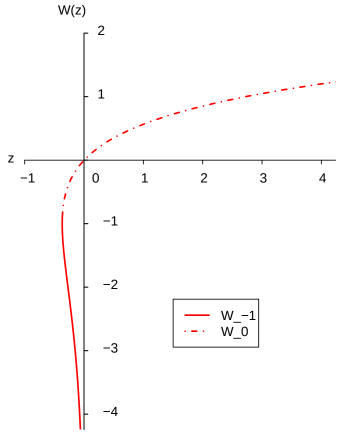


**Figure 1:** The two real branch of Lambert W function

The Lambert W function can be express as the Taylor series expansion which is shown in Corless and et al. [21] as follow

$$W(z) = \sum_{n=1}^{\infty} \frac{(-1)^{n-1} n^{n-2}}{(n-1)!} x^n \tag{2}$$

which the equivalent series expansion of the Eq. (2) is the Lagrange inversion theorem [10],

$$W(z) = \sum_{n=1}^{\infty} \frac{(-n)^{n-1}}{n!} z^n$$

The Lambert W function has various applications. Here we shows in mathmatics and statistics fields. For example Corless et al. are studied real values of W(z) by using Maple [22]. In 2002, Chapeau-Blondeau and Monir used the Lambert W function for address the problem of synthesizing a generalized Gaussian noise with exponent [23]. Wimmer and Duby purposed a class of logarithmic Lambert W random variables for a specific family of distributions [18].

## 3 Quantile function

The quantile function is one of the way to prescribe the distribution function, probability function or characteristic function. In addition, it can be express in infimum function,

$$Q_T(u) = \inf\{t \in R : u \leq F_T(t)\} \qquad ; 0 < u < 1. \tag{3}$$

where $F_T(t)$ is distribution function of random variable T and $t \in R$ [24]. From Eq. (3),

**Definition 4.** Let $T$ be non-negative continuous random variable with distribution function $F_T(t) := P(T \leq t), t \in R$. It has unique inversion when $F_T$ is continuous and nondecreasing and its quantile function is

$$Q_T(u) = F_T^{-1}(u), \qquad ; 0 < u < 1.$$

*3.1 The Quasi Lindley Distribution*

The Quasi Lindley (QL) distribution is continuous probability distribution, it was introduced by Shanker and Mishra in 2013 [25]. This is a mixture of the exponential($\theta$) and gamma($2, \alpha$) distribution.

**Definition 5.** Let $X$ be a random variable from the Quasi Lindley (QL) distribution with parameter $\theta$ and $\alpha$. This distribution function is given by

$$\begin{aligned} F_X(x) =& 1 - \frac{1 + \alpha + \theta x}{\alpha + 1} \exp(-\theta x), \\ & x \in (0, \infty), \theta > 0, \alpha > -1 \end{aligned} \tag{4}$$

From Definition.4 so that the quantile function of X is $F_X^{-1}(u)$, $u \in (0,1)$. Consequently, we obtain the $Q_X$ in term the Lambert W function in Theorem 1.

**Theorem 1.** For any $\theta > 0$ and $\alpha > -1$ the quantile function of a random variable X from the Quasi Lindley distribution is

$$\begin{aligned} Q_X(u) =& \left(-\frac{1}{\theta}\right)(W_{-1}((u-1)(\alpha+1)\exp(-\alpha-1)) \\ & + \alpha + 1), u \in (0,1), \end{aligned}$$

where $W_{-1}$ refer to non-principle branch of the Lambert W function.

**Proof.** For any fixed $\theta > 0$, $\alpha > -1$ and $u \in (0,1)$.

We will solve the equation $F_X(x) = u$ with $x \in (0, \infty)$. From Eq. (4), we get

$$(u-1)(\alpha+1) = -(1+\alpha+\theta x)\exp(-\theta x). \tag{5}$$

Multiplying by $\exp(-1-\alpha)$ for both sides of Eq. (5)

$$\begin{aligned}(u-1)(\alpha+1)\exp(-1-\alpha) = &-(1+\alpha+\theta x)\\ &\times\exp(-1-\alpha-\theta x).\end{aligned} \tag{6}$$

From Eq. (6), we obtain the real argument $(u-1)(\alpha+1)\exp(-1-\alpha)$ which have the Lambert W function $-(1+\alpha+\theta x)$. From Definition 1, we have

$$W((u-1)(\alpha+1)\exp(-1-\alpha)) = -(1+\alpha+\theta x). \tag{7}$$

We consider the Eq. (7) for any $\theta > 0$, $\alpha > -1$ and $x \in (0, \infty)$. By Definition 2 and 3 we obtain $(u-1)(\alpha+1)\exp(-1-\alpha) \in \left(-\frac{1}{e}, 0\right)$ where $u \in (0, 1)$ and $(1+\alpha+\theta x) > 1$. Thus, we will use the non-principal branch of the Lambert W function denoted by

$$W_{-1}((u-1)(\alpha+1)\exp(-\alpha-1)) = -(1+\alpha+\theta x)$$

Some quantile function of QL distribution is shown in Table 1 and Fig. 3.

### 3.2 The Janardan Distribution

The Janardan (JD) distribution, which is continuous probability distribution was introduced by Rama Shanker et al. in 2013 [26]. It can be shown as a mixture of the exponential $\left(\frac{\theta}{\alpha}\right)$ and gamma $\left(2, \frac{\theta}{\alpha}\right)$ distrubutions. The following definition shows the cumulative distribution function of the JD distribution.

**Definition 6.** Let $Y$ be a continuous random variable from the JD distribution with parameter $\theta > 0$ and $\alpha > 0$. This cumulative distribution function is given by

$$F_Y(y) = 1 - \frac{\alpha(\theta+\alpha^2) + \theta\alpha^2 y}{\alpha(\theta+\alpha^2)}\exp\left(-\frac{\theta y}{\alpha}\right), \tag{8}$$

for $y > 0,\ \theta > 0,\ \alpha > 0$.

Let $Q_Y$ be the quantile function of $Y$ by using Lambert W function where $F_Y$ is continuous and absolutely increasing according to Definition 4 and 6. The following Theorem is useful in finding the quantile function of JD distribution.

**Theorem 2.** For any $\theta > 0$ and $\alpha > 0$, the quantile function of a random variable from the JD distribution is

$$Q_Y(u) = -\frac{\alpha}{\theta}\left(\frac{\theta+\alpha^2}{\alpha^2}\right) - \frac{\alpha}{\theta}W_{-1}\left(\frac{(\theta+\alpha^2)(u-1)}{\alpha^2\exp\left(\frac{\theta+\alpha^2}{\alpha^2}\right)}\right)$$

Where $W_{-1}$ is the non-principle branch of the Lambert W function.

**Proof.** For any fixed $\theta > 0,\ \alpha > 0$ and $y \in (0, \infty)$. From Eq. (8), we have to solve the equation $F_Y(y) = u$ as follows

$$\frac{(\theta+\alpha^2)(u-1)}{\alpha^2} = -A\exp\left(-\frac{\theta y}{\alpha}\right), \tag{9}$$

where $A = \frac{(\theta+\alpha^2)}{\alpha^2} + \frac{\theta y}{\alpha}$

Multiplying by $\exp\left(-\frac{\theta+\alpha^2}{\alpha^2}\right)$ both sides of Eq. (9), we obtain

$$\frac{(\theta+\alpha^2)(u-1)}{\alpha^2}\exp\left(-\frac{\theta+\alpha^2}{\alpha^2}\right) = -A\exp(-A) \tag{10}$$

From Eq. (10) and Definition 1, we get the real argument:

$$W\left(\frac{(\theta+\alpha^2)(u-1)}{\alpha^2}\exp\left(-\frac{\theta+\alpha^2}{\alpha^2}\right)\right) = -A \tag{11}$$

By Definition 2 and 3, we consider the Eq. (11) for any $\theta > 0$, $\alpha > 0$ and $y \in (0, \infty)$. We take $\frac{(\theta+\alpha^2)(u-1)}{\alpha^2}\exp\left(-\frac{\theta+\alpha^2}{\alpha^2}\right) \in \left(-\frac{1}{e}, 0\right)$ where $u \in (0, 1)$ and $A > 1$. Therefore, we will use the non-principal branch of the Lambert W function denoted by

$$W_{-1}\left(\frac{(\theta+\alpha^2)(u-1)}{\alpha^2}\exp\left(-\frac{\theta+\alpha^2}{\alpha^2}\right)\right) = -A.$$

Some quantile of the QL distribution is shown in Table 2 and Fig. 3.

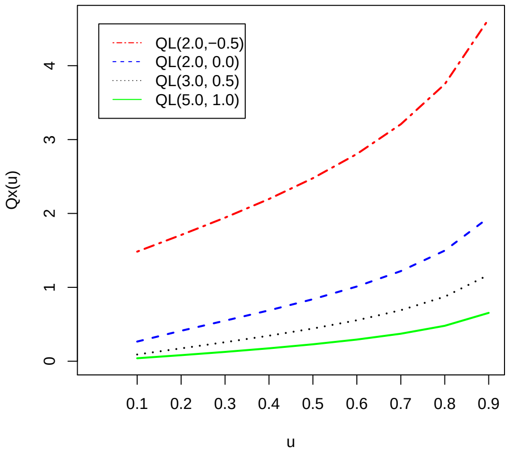


**Figure 2:** $Q_X(u)$ of QL distribution for different values of $\theta$ and $\alpha$

**Table 1:** $Q_X(u)$of the QL distribution for different values of $\theta$ and $\alpha$

| u | $\theta$ | $\alpha$ | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | -0.5 | 0 | 0.5 | 1 | 2 | 5 | 10 |
| 0.1 | 1 | 1.4832391 | 0.5318116 | 0.2720098 | 0.2012293 | 0.1560785 | 0.1261710 | 0.1158360 |
| | 2 | 0.7416196 | 0.2659058 | 0.1360049 | 0.1006147 | 0.0780393 | 0.0630855 | 0.0579180 |
| | 3 | 0.4944130 | 0.1772705 | 0.0906699 | 0.0670764 | 0.0520262 | 0.0420570 | 0.0386120 |
| | 5 | 0.2966478 | 0.1063623 | 0.0544020 | 0.0402459 | 0.0312157 | 0.0252342 | 0.0231672 |
| | 10 | 0.1483239 | 0.0531812 | 0.0272010 | 0.0201229 | 0.0156079 | 0.0126171 | 0.0115836 |
| 0.2 | 1 | 1.7086906 | 0.8243883 | 0.5215361 | 0.4093558 | 0.3264313 | 0.2666214 | 0.2451886 |
| | 2 | 0.8543453 | 0.4121942 | 0.2607681 | 0.2046779 | 0.1632157 | 0.1333107 | 0.1225943 |
| | 3 | 0.5695635 | 0.2747961 | 0.1738454 | 0.1364519 | 0.1088104 | 0.0888738 | 0.0817296 |
| | 5 | 0.3417381 | 0.1648777 | 0.1043072 | 0.0818712 | 0.0652863 | 0.0533243 | 0.0490377 |
| | 10 | 0.1708691 | 0.0824388 | 0.0521536 | 0.0409356 | 0.0326431 | 0.0266621 | 0.0245189 |
| 0.3 | 1 | 1.9430828 | 1.0973492 | 0.7717684 | 0.6308254 | 0.5151428 | 0.4251328 | 0.3916613 |
| | 2 | 0.9715414 | 0.5486746 | 0.3858842 | 0.3154127 | 0.2575714 | 0.2125664 | 0.1958306 |
| | 3 | 0.6476943 | 0.3657831 | 0.2572561 | 0.2102751 | 0.1717143 | 0.1417109 | 0.1305538 |
| | 5 | 0.3886166 | 0.2194698 | 0.1543537 | 0.1261651 | 0.1030286 | 0.0850266 | 0.0783323 |
| | 10 | 0.1943083 | 0.1097349 | 0.0771768 | 0.0630825 | 0.0515143 | 0.0425133 | 0.0391661 |
| 0.4 | 1 | 2.1955899 | 1.3764213 | 1.0359151 | 0.8730540 | 0.7281165 | 0.6072308 | 0.5605268 |
| | 2 | 1.0977949 | 0.6882107 | 0.5179575 | 0.4365270 | 0.3640582 | 0.3036154 | 0.2802634 |
| | 3 | 0.7318633 | 0.4588071 | 0.3453050 | 0.2910180 | 0.2427055 | 0.2024103 | 0.1868423 |
| | 5 | 0.4391180 | 0.2752843 | 0.2071830 | 0.1746108 | 0.1456233 | 0.1214462 | 0.1121054 |
| | 10 | 0.2195590 | 0.1376421 | 0.1035915 | 0.0873054 | 0.0728117 | 0.0607231 | 0.0560527 |
| 0.5 | 1 | 2.4773169 | 1.6783470 | 1.3268424 | 1.1461932 | 0.9744116 | 0.8214614 | 0.7599517 |
| | 2 | 1.2386584 | 0.8391735 | 0.6634212 | 0.5730966 | 0.4872058 | 0.4107307 | 0.3799759 |
| | 3 | 0.8257723 | 0.5594490 | 0.4422808 | 0.3820644 | 0.3248039 | 0.2738205 | 0.2533172 |
| | 5 | 0.4954634 | 0.3356694 | 0.2653685 | 0.2292386 | 0.1948823 | 0.1642923 | 0.1519903 |
| | 10 | 0.2477317 | 0.1678347 | 0.1326842 | 0.1146193 | 0.0974412 | 0.0821461 | 0.0759952 |
| 0.6 | 1 | 2.8048200 | 2.0223130 | 1.6620440 | 1.4662030 | 1.2690760 | 1.0821020 | 1.0036020 |
| | 2 | 1.4024099 | 1.0111566 | 0.8310222 | 0.7331017 | 0.6345379 | 0.5410510 | 0.5018011 |
| | 3 | 0.9349400 | 0.6741044 | 0.5540148 | 0.4887345 | 0.4230253 | 0.3607007 | 0.3345341 |
| | 5 | 0.5609640 | 0.4044626 | 0.3324089 | 0.2932407 | 0.2538152 | 0.2164204 | 0.2007205 |
| | 10 | 0.2804820 | 0.2022313 | 0.1662044 | 0.1466203 | 0.1269076 | 0.1082102 | 0.1003602 |
| 0.7 | 1 | 3.2074700 | 2.4392160 | 2.0714910 | 1.8620150 | 1.6400950 | 1.4158300 | 1.3170630 |
| | 2 | 1.6037348 | 1.2196082 | 1.0357454 | 0.9310073 | 0.8200477 | 0.7079151 | 0.6585315 |
| | 3 | 1.0691565 | 0.8130722 | 0.6904969 | 0.6206715 | 0.5466985 | 0.4719434 | 0.4390210 |
| | 5 | 0.6414939 | 0.4878433 | 0.4142982 | 0.3724029 | 0.3280191 | 0.2831661 | 0.2634126 |
| | 10 | 0.3207470 | 0.2439216 | 0.2071491 | 0.1862015 | 0.1640095 | 0.1415830 | 0.1317063 |
| 0.8 | 1 | 3.7493510 | 2.9943080 | 2.6197700 | 2.3972760 | 2.1497800 | 1.8822980 | 1.7576760 |
| | 2 | 1.8746757 | 1.4971542 | 1.3098851 | 1.1986380 | 1.0748898 | 0.9411490 | 0.8788379 |
| | 3 | 1.2497838 | 0.9981028 | 0.8732567 | 0.7990920 | 0.7165932 | 0.6274326 | 0.5858919 |
| | 5 | 0.7498703 | 0.5988617 | 0.5239540 | 0.4794552 | 0.4299559 | 0.3764596 | 0.3515351 |
| | 10 | 0.3749351 | 0.2994308 | 0.2619770 | 0.2397276 | 0.2149780 | 0.1882298 | 0.1757676 |
| 0.9 | 1 | 4.6310410 | 3.8897200 | 3.5081960 | 3.2718120 | 2.9948780 | 2.6707850 | 2.5079700 |
| | 2 | 2.3155200 | 1.9448600 | 1.7540980 | 1.6359060 | 1.4974390 | 1.3353920 | 1.2539850 |
| | 3 | 1.5436803 | 1.2965734 | 1.1693986 | 1.0906040 | 0.9982928 | 0.8902616 | 0.8359899 |
| | 5 | 0.9262082 | 0.7779440 | 0.7016391 | 0.6543624 | 0.5989757 | 0.5341570 | 0.5015939 |
| | 10 | 0.4631041 | 0.3889720 | 0.3508196 | 0.3271812 | 0.2994878 | 0.2670785 | 0.2507970 |

**Table 2:** $Q_Y(u)$of the JD distribution for different values of $\theta$ and $\alpha$

| u | $\theta$ | $\alpha$ | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 0.5 | 0.7 | 1 | 1.5 | 2 | 5 | 10 |
| 0.1 | 1 | 0.065639 | 0.108579 | 0.201229 | 0.429050 | 0.711807 | 2.470266 | 5.219547 |
| | 2 | 0.029609 | 0.045769 | 0.078039 | 0.158617 | 0.272010 | 1.151296 | 2.561900 |
| | 3 | 0.019017 | 0.028560 | 0.046562 | 0.089728 | 0.151772 | 0.718013 | 1.676938 |
| | 5 | 0.011061 | 0.016188 | 0.025234 | 0.045367 | 0.073577 | 0.381263 | 0.970578 |
| | 10 | 0.005400 | 0.007736 | 0.011584 | 0.019309 | 0.029263 | 0.149413 | 0.445210 |
| 0.2 | 1 | 0.138539 | 0.227189 | 0.409356 | 0.812290 | 1.265430 | 3.930560 | 8.144982 |
| | 2 | 0.062652 | 0.096614 | 0.163216 | 0.320182 | 0.521536 | 1.877747 | 4.024125 |
| | 3 | 0.040259 | 0.060396 | 0.098030 | 0.185263 | 0.302425 | 1.198468 | 2.651215 |
| | 5 | 0.023424 | 0.034265 | 0.053324 | 0.095091 | 0.151482 | 0.663022 | 1.554127 |
| | 10 | 0.011435 | 0.016381 | 0.024519 | 0.040781 | 0.061460 | 0.279608 | 0.734732 |
| 0.3 | 1 | 0.220623 | 0.358682 | 0.630825 | 1.192982 | 1.796257 | 5.294191 | 10.874441 |
| | 2 | 0.100045 | 0.153879 | 0.257571 | 0.490484 | 0.771768 | 2.557876 | 5.388632 |
| | 3 | 0.064322 | 0.096378 | 0.155709 | 0.288991 | 0.458935 | 1.650108 | 3.560641 |
| | 5 | 0.037434 | 0.054737 | 0.085027 | 0.150379 | 0.235721 | 0.930896 | 2.099321 |
| | 10 | 0.018278 | 0.026180 | 0.039166 | 0.064986 | 0.097387 | 0.407915 | 1.006295 |
| 0.4 | 1 | 0.314694 | 0.507180 | 0.873054 | 1.592915 | 2.344939 | 6.688833 | 13.665071 |
| | 2 | 0.143124 | 0.219523 | 0.364058 | 0.675561 | 1.035915 | 3.254160 | 6.783811 |
| | 3 | 0.092074 | 0.137777 | 0.221505 | 0.404128 | 0.627464 | 2.113199 | 4.490610 |
| | 5 | 0.053603 | 0.078339 | 0.121446 | 0.212984 | 0.328933 | 1.206804 | 2.657019 |
| | 10 | 0.026176 | 0.037487 | 0.056053 | 0.092762 | 0.138201 | 0.542187 | 1.284508 |
| 0.5 | 1 | 0.425091 | 0.679015 | 1.146193 | 2.032135 | 2.942010 | 8.197952 | 16.684263 |
| | 2 | 0.193959 | 0.296580 | 0.487206 | 0.883302 | 1.326842 | 4.007985 | 8.293310 |
| | 3 | 0.124863 | 0.186559 | 0.298361 | 0.535428 | 0.815437 | 2.614968 | 5.496837 |
| | 5 | 0.072719 | 0.106213 | 0.164292 | 0.285602 | 0.434965 | 1.506471 | 3.260555 |
| | 10 | 0.035516 | 0.050856 | 0.075995 | 0.125399 | 0.185686 | 0.689317 | 1.585825 |
| 0.6 | 1 | 0.559053 | 0.884708 | 1.466203 | 2.537242 | 3.624640 | 9.917389 | 20.123875 |
| | 2 | 0.256013 | 0.390110 | 0.634538 | 1.125855 | 1.662044 | 4.867133 | 10.013041 |
| | 3 | 0.164943 | 0.246012 | 0.391186 | 0.690631 | 1.033905 | 3.187126 | 6.643242 |
| | 5 | 0.096104 | 0.140272 | 0.216420 | 0.372722 | 0.560048 | 1.848677 | 3.948242 |
| | 10 | 0.046947 | 0.067212 | 0.100360 | 0.165055 | 0.242816 | 0.858271 | 1.929318 |
| 0.7 | 1 | 0.730097 | 1.143833 | 1.862015 | 3.153404 | 4.453998 | 12.001576 | 24.292865 |
| | 2 | 0.335766 | 0.509565 | 0.820048 | 1.425058 | 2.071491 | 5.908751 | 12.097474 |
| | 3 | 0.216538 | 0.322284 | 0.509132 | 0.883991 | 1.302457 | 3.881029 | 8.032795 |
| | 5 | 0.126235 | 0.184092 | 0.283166 | 0.482700 | 0.715621 | 2.264106 | 4.781844 |
| | 10 | 0.061681 | 0.088285 | 0.131706 | 0.215762 | 0.315134 | 1.064162 | 2.345826 |
| 0.8 | 1 | 0.968424 | 1.499973 | 2.397276 | 3.977668 | 5.560162 | 14.776739 | 29.843746 |
| | 2 | 0.447730 | 0.676062 | 1.074890 | 1.828791 | 2.619770 | 7.295904 | 14.872858 |
| | 3 | 0.289116 | 0.429131 | 0.672626 | 1.147077 | 1.663825 | 4.805341 | 9.882990 |
| | 5 | 0.168669 | 0.245693 | 0.376460 | 0.634157 | 0.926987 | 2.817868 | 5.891844 |
| | 10 | 0.082443 | 0.117965 | 0.175768 | 0.286526 | 0.414992 | 1.339379 | 2.900563 |
| 0.9 | 1 | 1.369741 | 2.090769 | 3.271812 | 5.312218 | 7.346903 | 19.253499 | 38.797826 |
| | 2 | 0.638060 | 0.956530 | 1.497439 | 2.487181 | 3.508196 | 9.533850 | 19.349841 |
| | 3 | 0.412832 | 0.610248 | 0.946307 | 1.579280 | 2.251739 | 6.296841 | 12.867583 |
| | 5 | 0.241128 | 0.350604 | 0.534157 | 0.885921 | 1.273762 | 3.711931 | 7.682482 |
| | 10 | 0.117926 | 0.168646 | 0.250797 | 0.405911 | 0.581451 | 1.784708 | 3.795615 |

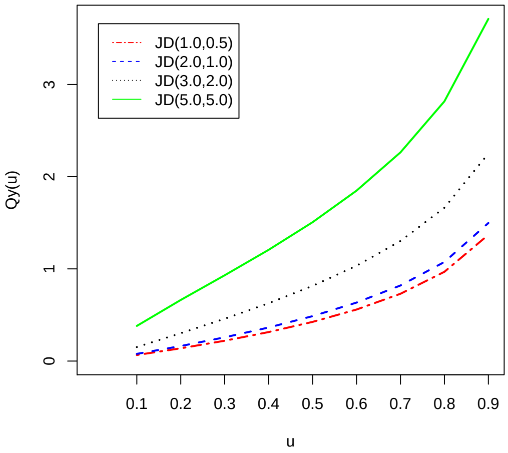


**Figure 3:** $Q_Y(u)$ of JD distribution for different values of $\theta$ and $\alpha$

## 4 Generation algorithms

In this section we will show the algorithms for random variable generation from two methods : 1)method which not use the Lambert W function and 2)the inverse transform method which use the Lambert W function.

*4.1 The Quasi Lindley distribution*

4.1.1 The algorithm of method which not use the Lambert W function of the QL distribution is presented as follows

(1) Generate $U_i \sim Uniform(0,1),\ i = 1,2,\ldots,n;$

(2) Generate $V_i \sim Exponential(\theta),\ i = 1,2,\ldots,n;$

(3) Generate $W_i \sim Gamma(2,\alpha),\ i = 1,2,\ldots,n;$

(4) If $U_i \le \dfrac{\alpha}{\alpha+1}$ then set $X_i = V_i$, otherwise set $X_i$

$= W_i,\ i = 1,2,\ldots,n;$

*The code R of random variable generation of QL distribution for not using Lambert W function*

```
> rquasi.lindley=function(n,theta,alpha){
>  x = rep(0,n)
>  for(i in 1:n){
>    U = runif(1)
>    V = rexp(1,theta)
>    W = rgamma(1,shape=2,rate=theta)
>    p = alpha/(alpha+1)
>    if (U<=p){x[i]=V} else {x[i]=W}
>  }
>  x
> }
> x = rquasi.lindley(15,2,5)
```

4.1.2 The algorithm of the inverse transform method with using Lambert W function is

(1) Generate $U_i \sim Uniform(0,1),\ i = 1,2,\ldots,n;$

(2) Set $X_i = \left(-\dfrac{1}{\theta}\right)(W_{-1}(\exp(-\alpha-1)(u-1)(\alpha$

$+1)) + \alpha + 1), u \in (0,1),$

where $W_{-1}(\cdot)$ is non-principal branch of the Lambert W function.

*The code R of random variable generation of QL distribution by using inverse transform method with Lambert W function*

```
> library("LambertW")
> rquasi.lindley.W=function(n,theta,alpha){
>  x = rep(0,n)
>  for(i in 1:n){
>   U = runif(1)
>   x[i]=(-1/theta)*(W_1(exp(-alpha-1)*(u-1)
>        *(alpha+1))+alpha+1)
>  }
>  x
> }
> x = rquasi.lindley.W(15,2,5)
```

*4.2 The Janardan distribution*

4.2.1 The algorithm which not use Lambert method of the JD distribution is shown as follows

(1) Generate $U_i \sim Uniform(0,1),\ i = 1,2,\ldots,n;$

(2) Generate $V_i \sim Exponential\left(\dfrac{\theta}{\alpha}\right),\ i = 1,2,\ldots,n;$

(3) Generate $W_i \sim Gamma\left(2,\dfrac{\theta}{\alpha}\right),\ i = 1,2,\ldots,n;$

(4) If $U_i \le \dfrac{\theta}{\theta+\alpha^2}$ then set $Y_i = V_i$, otherwise set $Y_i$

$= W_i,\ i = 1,2,\ldots,n;$

*The code R of random variable generation of JD distribution for not using Lambert W function*

```
> rJanardan= function(n,theta,alpha){
>  y<-numeric(n)
>  for (i in 1:n){
>   U = runif (1)
>   V = rexp(1,theta/alpha)
>   W = rgamma(1,2,rate=theta/alpha)
>   p = theta/(theta+alpha^2)
>   if (U<= p){y[i]=V} else {y[i]=W}
>  }
>  y
> }
> y = rJanardan(15,1,0.2)
```

4.2.2 The algorithms of the inverse transform method with using Lambert W function is

(1) Generate $U_i \sim Uniform(0,1),\ i = 1,2,\ldots,n;$

(2) Set $Y_i = -\dfrac{\alpha}{\theta}\dfrac{\theta+\alpha^2}{\alpha^2} - \dfrac{\alpha}{\theta}W_{-1}\left(\dfrac{(\theta+\alpha^2)(u-1)}{\alpha^2\exp\left(\frac{\theta+\alpha^2}{\alpha^2}\right)}\right),$

where $W_{-1}(\cdot)$ is non-principal branch of the Lambert

W function.

*The code R of random variable generation of JD distribution by using inverse transform method with Lambert W function*

```
> library("LambertW")
> rJanardan.W=function(n,theta,alpha){
>  y=rep(0,n)
>  for(i in 1:n){
>   U = runif(1)
>   y[i]=-(alpha/theta)*((theta+alpha^2)
>        /alpha^2)-alpha/theta
>        *W_1(((theta+alpha^2)*(u-1))
>        /(alpha^2*exp((theta+alpha^2)
>        /alpha^2)))
>   }
>   y
> }
> y = rJanardan.W(20,1,0.2)
```

## 5 Conclusion

Some distributions of random variable which the quantile function can not expressed in explicit form such as the Quasi Lindley and Janardan distribution, we used the Lambert W function for obtained its explicit form. Moreover, we used R language version 3.1.2 to obtained the random variables using algorithms from the inverse transform method which use Lambert W function and method which not use Lambert W function.

## Acknowledgements

The authors would like to thank the Science Achievement Scholarship of Thailand (SAST) and Department of Statistics, Kasetsart University.

## References

[1] Ross RM. Simulation. 4th ed. Amsterdam: Elsevier Academic Press; 2006.

[2] Brandimarte P. Handbook in Monte carlo simulation : applications in financial engineering, risk management, and economics. New Jersey: John Wiley & Sons; 2014.

[3] Santos-Neto M, Bourguignon M, Zea LM, Nascimento AD, Cordeiro GM. The Marshall-Olkin extended weibull family of distributions. Journal of Statistical Distributions and Applications. 2014; 1(9).

[4] Kundu D, D Gupta R. A convenient way of generating gamma random variables using generalized exponential distribution. Computational Statistics & Data Analysis. 2007; 51(6): 2796--2802.

[5] Gui W, Zhang S, Lu X. The Lindley-Poisson distribution in lifetime analysis and its properties. Hacettepe Journal of Mathematics and Statistics. 2014; 43(6): 1063--1077.

[6] L Lio Y. A note on Bayesian estimation for the negative-binomial model. Pliska Studia Mathematica Bulgarica. 2009;19: 207--216.

[7] Cakmakyapan S, Ozel Kadilar G. A new customer lifetime duration distribution:The Kumaraswamy Lindley distribution. International Journal of Trade, Economics and Finance. 2014; 5(5): 441--444.

[8] Lambert JH. Observations variae in Mathesin Puram. Acta Helveticae,physico-mathematico-anatomico-botanico-medica. 1758; 3: 128--168.

[9] Euler L. De serie Lambertina Plurimisque eius insignibus proprietatibus. Acta Acad Scient Petropol. 1783; 2: 29--51.

[10] Corless RM, Gonnet GH, Hare DEG, Jeffrey DJ. Lambert's W Function in Maple. The maple technical Newsletter. 1993; 9: 12--22.

[11] Cranmer SR. New views of the solar wind with the Lambert W function. American Journal of Physics. 2004; 72(11): 1397--1403.

[12] Ghitany ME, Al-Mutairi DK, Balakrishnan N, Al-Enezi LJ. Power Lindley distribution and associated inference. Computational Statistics and Data Analysis. 2013; 64(C): 20--33.

[13] Jarlebring E, Damm T. The Lambert W function and the spectrum of some multidimensional time-delay systems. Automatica. 2007; 43(12): 2124--2128.

[14] Kakorin S, Neumann E. Ionic conductivity of electroporated lipid bilayer membranes. Bioelectrochemistry. 2002; 56(1-2): 163--166.

[15] Mathias SA, Roberts AW. A Lambert W function solution for estimating sustainable injection rates for storage of {CO2} in brine aquifers. International Journal of Greenhouse Gas Control. 2013; 17(0): 546--548.

[16] Shinozaki H, Mori T. Robust stability analysis of linear time-delay systems by Lambert function: Some extreme point results. Automatica. 2006; 42(10): 1791 -- 1799.

[17] Tzafriri AR. Michaelis-Menten kinetics at high enzyme concentrations. Bulletin of Mathematical Biology. 2003; 65(6): 1111--1129.

[18] Witkovsky V, Wimmer G, Duby T. Logarithmic Lambert random variables for the family of chi-squared distributions and their applications. Statistics & Probability Letters. 2015; 96: 223--231.

[19] R Core Team. R: A language and environment for statistical computing.; 2015.

[20] Weisstein EW. Lambert's transcendental equation [Internet]; 2015 [cited 2015 March 16]. Available from: http://mathworld.wolfram.com/Lamberts-TranscendentalEquation.html.

[21] Corless RM, Jeffrey DJ, Knuth DE. A Sequence of Series for the Lambert W Function. In: Proceedings of the 1997 International Symposium on Symbolic and Algebraic Computation. ISSAC '97; 1997. .

[22] Corless RM, Gonnet GH, Hare DEG, Jeffrey DJ, Knuth DE. On the Lambert W function. Advances in Computational Mathematics. 1996; 5(1): 329--359.

[23] Chapeau-Blondeau F, Monir A. Numerical evaluation of the Lambert W function and application to generation of generalized Gaussian noise with exponent 1/2. Signal Processing, IEEE Transactions on. 2002 Sep; 50(9): 2160--2165.

[24] Parzen E. Quantile probability and statistical data modeling. Stat, Sci. 2004; 19: 652--662.

[25] Shanker R, Mishra A. A quasi Lindley distribution. African Journal of Mathematixs and Computer Science Research. 2013; 6(4): 64--71.

[26] Shanker R, Sharma S, Shanker U, Shanker R. Janardan distribution and its application to waiting times data. Indian Journal of Applied Reseserch. 2013; 3(8): 500--502.

# Factor Affecting Operational Performance in Logistics Service Providers in Thailand

Chanikarn Jeenanupan, Sakpol Chongprawatsakul, Matawee Waiyatip,
Thanaset Kongtorain, Nuchjarin Intalar, Chawalit Jeenanunta*
*School of Management Technology, Sirindhorn International Institute of Technology, Thammasat University*
*131 M.5 Tiwanont Rd. Bangkadi Muang Pathumthani 12000, Thailand*

email: c.chryztal@gmail, sakpol.ch@gmail.com, minghummer@gmail.com,

eakthanaset@gmail.com, i.nuchjarin@gmail.com, chawalit@siit.tu.ac.th

## Abstract

Logistics Service Providers (LSPs) have more concern to improve its operational system to provide efficient services for their customers and gain competitive advantages. The purpose of this paper is to identify the major factors influencing the operational performance in LSP. This paper fosters a better understanding of the services sector through a firm survey on a sub-service sector and to develop a research base for a future full-fledged research on the sector. The paper collects firm-level data on the trucking industry through a questionnaire survey then makes empirical analysis. Even though the trucking industry is a crucial to develop competitive economies for developing countries, such heterogeneous natures of the trucking industry impede a better understanding of the industry's upgrading mechanism and necessitate firm-level data for detailed analysis. The result shows interesting internal factors that influencing the LSP operational performance which has great contribution to the practitioners and researchers. The research find that the communication between manager and employee also the adopting of ISO 9001 improve company's net profit. The training of the transportation related skills to the driver and the adoption of GPS could reduce the accident.

*Keywords*: *Training truck*: *Regression*: *One-way ANOVA*: *T-test*

Corresponding Author
E-mail Address: chawalit@siit.tu.ac.th

## 1. Introduction

Logistics function plays an important role in supply chain system. In order to improve supply chain effectiveness in the globalization era, firms should improve the logistics function as well as service and innovation to be more efficient. In order to satisfy high demand of logistics services, the using of logistics service providers (LSPs) is increased [1]. However, offering simple transportation and warehousing service is no longer sufficient to satisfy and retain customers [2]. Their performance and competitive advantages depend on the ability to develop their service, technologies, and innovations that add value to the system and deliver to customers.

The demand for logistics services has been growing significantly. According to statistics, the total logistics value has grown by 15.2% (1.64 billion baht). There are a lot of new LSPs started logistics service business in Thailand to satisfy customer's needs. To gain competitive advantages among competitors, the logistics service provider should improve its information and logistics technologies capability. Many researches proposed that logistics companies could increase their performance by employing new technologies [3], new information technologies [4], and pay more attention to innovation [5]. Innovation in logistics can be implemented through technology, knowledge and relationship networks.

Now LSPs in Thailand is in the initial stages of development, and cannot reach the certain logistics ability like developed countries. Moreover, the development of logistics industry is still far from the economy requirement in the structure, scale, quality and efficiency. Therefore, the development of Thailand's logistic industry have not only include innovation but also a sense of service, development strategies and logistics management. Therefore, we will conduct the relevant investigation to verify the model of LSPs operational performance improvement. The main purpose of this paper is to explore the factors affecting the operation performance in LSPs.

## 2. Theory Background and Hypothesis

Operational performance and logistics technology for LSPs will be influenced by the internal factor including organizational encouragement from top manager. The following hypotheses are consequently proposed:

*$H_1$: Communication and information sharing among managers and employees have positive impact on net profit*

Communication and information sharing among management level and employees are important factor which enhance firm performance. Internal communication plays an important role in influence of employee engagement include increased productivity, decreased attrition, improvement of an organization's image, reputation, and increased financial returns [6].

*$H_2$: Training frequency of transportation-related skills has a positive impact on number of accident reduction*

The firms which always train drivers before driving lead to increase driving skill and fewer crashes on the road. The required training in firm usually produces large reductions in driver crash rates. As training methods became more engaging (i.e., requiring trainees' active participation), workers demonstrated greater knowledge acquisition, reductions were seen in accidents, illnesses, and injuries. All methods of training produced meaningful behavioural performance improvements [7].

Carstensen's studies shows that the number of accidents decreased after training. The systematic and structured driver education can be an effective way of eliminating at least some of the new drivers' accidents.

A driver training should improve both safety and skill level; that is, it should teach drivers how to avoid dangerous situations, but should also equip them with adequate skills to deal successfully with unavoidable hazards.

*$H_3$: Adoption of Transportation technologies has a positive impact on number of accident reduction*

Transportation plays an important role of product movement in logistics operations. The transportation must meet customer's demand regarding delivery performance and safety. The transportation technologies are widely used by logistic firms for the delivery monitoring and tracking. There are several technologies that commonly used in the logistic firms include global positioning system (GPS), radio-frequency communication system, and transportation data recorder [8].

*$H_4$: Adoption of Thai/International standards (ISO, Q-mark, 5s) has positive impact on net profit*

International standard helps to bring more scientific analysis into a decision-making process. The study of Wenlong He et al., [9] showed that the adoption of ISO 14001 increases sales and significant net effects on the financial performance. Therefore, we hypothesize that Thai/International standards might have positive effect on firm's net profit.

## 3. Research Methodology

### *3.1 Questionnaire Explanation*

In the basis of the conception and objective of the study, we designed questionnaire about operational performance of logistics service providers in term of innovation driven factors and measurement indicators of LSPs. The former items include logistics technologies, organization cooperation, communication and information sharing, adoption of transportation technologies, training frequency of transportation-related skills.

The questionnaire survey was composed of six sections; respondents'profile; operational performance, and the accident reduction; the communication and information sharing among top manager, employees, and drivers; the adopting of international standard; and the training frequency of transportation-related skills.

For the first section, respondents' profile is about basic information of the company. We use multiple choices ask for single answer.

The second section is for the operational performance in logistics service in fiscal year 2013-2014and the accident reduction, a five-point Likert scale [10] is used, where 1=significantly decreased 2=decreased, 3=almost same, 4=increased and 5 = significantly increased, is used in the net profit, the reduction of accidents without injuries per truck, and the reduction of fatal and injury accidents per truck.

The third section is for the communication and information sharing among top manager, employees, and drivers. A five-point Likert scale [10] is used, where 1=significantly decreased 2=decreased, 3=almost same, 4=increased and 5=significantly increased. The communication and information sharing among top manager, employees, and drivers including sharing the target level of key performance indicators to employees, top manager has established good communication with employees, top manager listen to employees' complaints and discontents.

The fourth section is the adopting of transportation technologies including Driving recorder, Back-eye camera, and GPS where dichotomous question (Yes = 1, No = 0) is used [11].

The fifth section is the adopting of international standard including ISO9001, Q-mark and 5S where dichotomous question (Yes = 1, No = 0) is used [11].

The last section is the training frequency of transportation-related skills. There are four types of transportation-related skills training; safety-driving skill, fuel-efficient driving skill, truck maintenance, and laws or regulations compliance. There are three types of the frequency of training; training new drivers; periodically training; and no training.

### *3.2 Sample*

The trucking company list is from Siam List Database Marketing Company which contains 13,418 logistics companies and the additional 9,607 logistics companies are from Department of Land Transport, the ministry of transportation, Thailand. Total of 1,200 questionnaires were distributed randomly to logistic company in Bangkok and surrounding, by direct interviewing, mailing, calling. There are 115 responses which account for 10% of response rate, where 94 from mailing (81.7%), 15 from calling (13.0%) and 6 from direct interviewing (5.2%).

### *3.3 Statistical method*

The descriptive statistics frequencies, reliability, regression, T-test and One-way ANOVA are used to analyse the data. The reliability is usually expressed on the basis of the Cronbach's alpha coefficient (reliability coefficients). Levels of 0.70 or more are generally accepted as representing good reliability [12].

For *$H_1$: Communication and information sharing among managers and employees have positive impact on net profit,* we use regression to analyse the data.

Significant value (p value) must be lower than 0.05 for data accepted [13]. R-squared is a statistical measure of how close the data are to the fitted regression line.

For $H_2$: *Training frequency of transportation-related skills has a positive impact on number of accident reduction*, we use One-way ANOVA to analyse the data. Significant value (p value) must be lower than 0.05 [14].

For $H_3$: *Adoption of Transportation technologies has a positive impact on number of accident reduction*, we use independent-samples t-test to analyse the data. Significant value (p value) must be lower than 0.05 and the value of *t* must be greater than zero [13].

For $H_4$: *Adoption of Thai/International standards (ISO, Q-mark, 5s) has positive impact on net profit*, we use independent-samples t-test to analyse the data. Significant value (p value) must be lower than 0.05 and the value of *t* must be greater than zero [13].

## 4. Research Results and Discussion

### *4.1 Data description*

Description statistical analysis of each factor that improve LSPs performance is calculate the number of respondents and percentage of respondent of each factor as shown in Table 1. For the form of legal organization, 62.6% of the respondents is individual proprietors and 37.4% is company limited, public company limited. There are 92.2% of the firms is locally owned while 7.0% is joint venture firms, and 0.9% is foreign-owned firm. Most of the response firms are small size at 75.9% while there are 21.4% of the medium size and 2.7% is large size of the response firm. For the form of sales, 64.3% of the respondents are 10 million or less, 11-50 million is 21.7%, 51-100 million is 8.7% and 101-999 million is 5.2%, respectively.

Table 1: Respondents' profile

| | **Number of respondents** | **Percentage of respondents** |
|---|---|---|
| **Form of legal organization** | | |
| Individual proprietor | 72 | 62.6 |
| Company limited, Public company limited | 43 | 37.4 |
| **Proportion of foreign investment** | | |
| 100% locally owned | 106 | 92.2 |
| Joint Venture (JV) | 8 | 7.0 |
| 100% foreign owned(MNC) | 1 | 0.9 |
| **Size of Company** | | |
| Small | 85 | 75.9 |
| Medium | 24 | 21.4 |
| Large | 3 | 2.7 |
| **Sales (THB)** | | |
| 10 million or less | 74 | 64.3 |
| 11-50 million | 25 | 21.7 |
| 51-100 million | 10 | 8.7 |
| 101-999 million | 6 | 5.2 |
| **Thai/International Standards** | | |
| ISO series adoption | 62 | 54.4 |
| Q-mark adoption | 7 | 6.1 |
| 5S adoption | 23 | 20.2 |
| **Quality control and training for drivers** | | |
| **Safe-driving skill** | | |
| Periodically | 74 | 64.3 |
| Only new drivers | 28 | 24.3 |
| No training | 13 | 11.3 |
| **Fuel-efficient driving skill** | | |
| Periodically | 60 | 52.2 |
| Only new drivers | 42 | 36.5 |
| No training | 13 | 11.3 |
| **Truck maintenance** | | |
| Periodically | 61 | 53.0 |
| Only new drivers | 40 | 34.8 |
| No training | 13 | 11.3 |
| **Laws or regulations** | | |
| Periodically | 58 | 50.4 |
| Only new drivers | 44 | 38.3 |
| No training | 12 | 10.4 |
| **Transportation Technologies and equipment** | | |
| Driving recorder | 9 | 7.9 |
| Back eye camera | 3 | 2.6 |
| GPS | 99 | 86.8 |

Table 2 shows the descriptive statistics of communication and information sharing and logistic function perceived status by top manager.

In the view of the respondents, they are agreed that top manager who listen to employees' complaints and discontents is the driving shill of driver (Mean=4.16) the result comes from the survey which the responder consider to employee complaints. The second rank is establishing a good communication with employees (means scores is 4.04. From the results in Table 2, most of top managers agree that training and enhancing skills of drivers is the most important activity which should be carried out to improve logistics function and performance. Followed by compliance with Thai laws and regulations (Mean=4.38), Safety (Mean=4.13), environmental protection (Mean=4.10), and delivery on time (Mean=3.90), respectively.

Table 2: The important of communication and information sharing and Logistic function perceived by top manager

| | Mean | SD |
|---|---|---|
| **Communication and information sharing** | | |
| Target level of key performance indicators is shared to employees | 3.82 | 0.823 |
| Top manager has established good communication with employees | 4.04 | 0.754 |
| Top manager listen to employees' complaints and discontents | 4.16 | 0.744 |
| **Logistic function perceived by top manager** | | |
| Delivery on time | 3.90 | 0.882 |
| Safety | 4.13 | 0.800 |
| Environmental protection. | 4.10 | 0.783 |
| Compliance with Thai laws and regulations | 4.38 | 0.732 |
| Training and enhancing skills of drivers | 4.40 | 0.747 |

Note: 1 = unimportant, 5 = very important and SD = standard deviation

Table 3 shows the operational performance that has significantly increased is repairing and maintenance expense per truck (Mean=4.34). The second performances that has increased are delivery in full on time (DIFOT) and number of accidents without injuries per truck (Mean=4.30). The other performances also increased in the fiscal year 2013-2014 as shown in Table 3.

Table 3: The operational performance in logistics service in fiscal year 2013-2014

| | Mean | SD |
|---|---|---|
| Net profit | 3.83 | 0.939 |
| Load efficiency | 4.05 | 0.815 |
| Laden miles | 4.10 | 0.872 |
| Usage efficiency | 4.21 | 0.893 |
| Delivery in Full on Time | 4.30 | 0.870 |
| Number of damages or losses of cargos | 4.31 | 0.968 |
| Number of accidents without injuries per truck | 4.30 | 0.982 |
| Number of fatal and injury accidents per truck | 4.27 | 1.012 |
| Fuel expense | 4.29 | 0.915 |
| Repair and maintenance expense per truck | 4.34 | 0.887 |

Note: 1 = significantly decreased 5 = significantly increased and SD = standard deviation

*4.2 Reliability test*

The instrument from communication and information sharing group in logistic function perceived status by top manager group have high level of reliability with Cronbach's alpha as 0.887 and 0.885, respectively. Hence, the instrument for these two groups are appropriate for analyzing.

*4.3. Test of hypothesis*

*4.3.1 Hypothesis 1: Communication and information sharing among top manager and employees have positive impact on net profit*

In order to test Hypothesis 1, we use the linear regression analysis. There are three independent variables, which are target level of key performance indicators is shared to employees, top manager has established good communication with employees, and top manager listen to employees' complaints and discontents. The result in Table 4 show that each independent variable has positive impact on net profit.

Moreover, we group three observed variables as Communication and information sharing among manager and employees by using the factor analysis. Communication and information sharing among manager and employees significantly predicted net profit, $\beta$=1.244, $t(115)=2.741$, $p<0.001$. Communication and information sharing among manager and employees also explained a significant proportion of variance in net profit, $R^2 = 0.236$, $F(1, 115) = 12.711$, $p<0.001$.The result shows that communication and information sharing among manager and employees has positive impact on net profit in LSPs.

Table 4: Regression analysis of communication and information sharing among manager and employees as a predictor of net profit

| | $R^2$ | F-test | Sig | Conclude |
|---|---|---|---|---|
| Target level of key performance indicators is shared to employees | 0.210 | 30.015 | 0.000 | Accepted |
| Top manager has established good communication with employees | 0.232 | 34.137 | 0.000 | Accepted |
| Top manager listen to employees' complaints and discontents | 0.163 | 21.933 | 0.000 | Accepted |
| H1 | 0.236 | 12.711 | 0.000 | Accepted |

*4.3.2 Hypothesis 2: The frequency of training on transportation-relatedskills has a positive impact on number of accident reduction*

We use one-way ANOVA to test and analyse the relationship between the frequency of training and the number of accidents reduction. Table 5 and 6 show the results from ANOVA Analysis of the number of accidents reduction related to the training. In table 5 ANOVA was conducted to compare the effect of the reduction of accidents without injuries in safety-driving skill, fuel-efficient driving skill, truck maintenance, laws or regulations.

Table 5: ANOVA result of comparing the effect of the frequency of training on transportation-related skills on the reduction of accidents without injuries per truck

| Training Type | Mean | F-test | Sig. | Conclude |
|---|---|---|---|---|
| H2.1a<br>Safety-driving skill | 9.691 | 11.984 | 0.000 | Accepted |
| H2.2a<br>Fuel-efficient driving skill | 9.559 | 11.787 | 0.000 | Accepted |
| H2.3a<br>Truck maintenance | 8.805 | 11.700 | 0.000 | Accepted |
| H2.4a<br>Laws or regulations | 7.822 | 10.040 | 0.000 | Accepted |

A one-way between subjects ANOVA was conducted to compare the effect of the frequency of training on transportation-related skills on the reduction of accidents without injuries per truck in training new drivers, training periodically, and no training conditions.

There was a significant effect of frequency of the training Safety-driving skill on reduction of accidents without injuries per truck at the p less than 0.05 level for the three conditions [F(2,112) = 11.984 p= 0.00]. Post hoc comparisons using the TukeyHSD test indicated that the mean score for the training periodically (Mean = 4.47, SD = 0.954) was not significantly different than the new drivers training (Mean = 4.36, SD = 0.84).But the firms without training employees (Mean = 3.15, SD = 0.89) was significantly different from the training periodically and the new drivers training conditions.

The fuel-efficient driving skill on reduction of accident without injuries per truck. For the three conditions [F (2,112) = 11.787 .p= 0.00]. The mean score for the training periodically (Mean = 4.45, SD = 1.015) was not significantly different than the new drivers training (Mean = 4.43, SD = 0.67). But the firms without training employees (Mean = 3.15, SD = 0.987) was significantly different from the training periodically and the new drivers training conditions.

The training truck maintenance on reduction of accidents without injuries per truck. For the three conditions [F (3,111) = 11.700 p= 0.00]. Training periodically (Mean = 4.64, SD = 0.708), new drivers training (Mean = 4.18, SD = 1.035). But the firms without training employees (Mean = 3.15, SD = 0.89) was significantly different from the training periodically and the new drivers training conditions.

The training laws or regulations on reduction of accidents without injuries per truck where [F (3,111) = 10.040 p= 0.00]. Post hoc comparisons using the Tukey HSD test indicated that the mean score for the training periodically (Mean = 4.62, SD = 0.745) was not significantly different than the new drivers training (Mean = 4.20, SD = 1.002). But the firms without training employees (Mean = 3.17, SD = 1.030) was significantly different from the training periodically and only new drivers training conditions.

A one-way between subjects ANOVA was conducted to compare the effect of the frequency of training on transportation-related skills on the reduction of fatal and injury accidents per truck in training new drivers, training periodically and no training conditions.

There was a significant effect of frequency of the training on Safety-driving skill on reduction of accidents of fatal and injury per truck at the p less than 0.05 level for the three conditions [F(2,112) = 12.698 p= 0.00]. For the training periodically (Mean = 4.47, SD = 0.940) was not significantly different than the new drivers training (Mean = 4.29, SD = 0.810). But the firms without training employees (Mean = 3.08, SD = 1.038) was significantly different from the training periodically and the new drivers training conditions.

The training Fuel-efficient driving skill on reduction of accidents of fatal and injury per truck where [F(2,112) = 12.190 p= 0.00] ], (Mean = 4.43, SD = 1.013) was not significantly different than the new drivers training (Mean = 4.42, SD = 0.737).But the firms without training employees (Mean = 3.08, SD = 1.038) was significantly different from the training periodically and the new drivers conditions.

The training truck maintenance on reduction of accidents of fatal and injury per truck at the p less than 0.05 level for the three conditions [F (3,111) = 12.030 p= 0.00]. For the training periodically (Mean = 4.61, SD = 0.759) was not significantly different than the new drivers training (Mean = 4.18, SD = 1.035). But the firms without training employees (Mean = 3.15, SD = 0.987)

The training laws or regulations on reduction of accidents of fatal and injury per truck where [F (3,111) = 11.379 p= 0.00]. For the training periodically (Mean = 4.62, SD = 0.745) was not significantly different than the new drivers training (Mean = 4.16, SD = 1.033). But the firms without training employees (Mean = 3.17, SD = 1.030)

The results show that the periodical training and training only to the new drivers on transportation-related skills has significant impact on reducing the number of accident in both without injuries and fatal and injury accidents.

Table 6: ANOVA result of comparing the effect of the frequency of training on transportation-related skills on the reduction of fatal and injury accidents per truck

| Training Type | Mean | F-test | Sig. | Conclude |
|---|---|---|---|---|
| H2.1b Safety-driving skill | 10.780 | 12.698 | 0.000 | Accepted |
| H2.2b Fuel-efficient driving skill | 10.426 | 12.190 | 0.000 | Accepted |
| H2.3b Truck maintenance | 9.540 | 12.030 | 0.000 | Accepted |
| H2.4b Laws or regulations | 9.145 | 11.379 | 0.000 | Accepted |

*4.3.3 Hypothesis 3: Adoption of Transportation technologies has a positive impact on number of accident reduction*

Hypothesis 3, an independent-samples t-test was conducted to compare the number of accident reduction without injuries per truck in adopting transportation technologies and not adopting transportation technologies conditions. Table 7 shows the result from t-test analysis between each transportation technology and the reduction of accidents without injuries per truck.

There was not a significant different in the score for adopting driving record (Mean =3.78, SD=1.302) and not adopting driving record (Mean =4.37, SD=0.891) condition; t (112) = -1.845, p = 0.068.

There was not a significant different in the score for adopting back-eye camera (Mean =4.33, SD=1.155) and not adopting back-eye camera (M=4.32, SD=0.936) condition; t (112) = 0.016, p = 0.987.

There was a significant different in the score for adopting GPS (Mean =4.46, SD=0.837) and not adopting GPS (Mean =3.4, SD=1.056) condition; t (112) = 4.431, p = 0.00.These results suggest that adopting GPS does have an effect on reduction of accident without injury per truck.

Table 7: An independent-samples t-test result of comparing the number of accident reduction without injuries per truck in adopting transportation technologies

| Technologies | T-test | F-test | Sig. | Conclude |
|---|---|---|---|---|
| Driving recorder | -1.845 | 1.584 | 0.068 | Rejected |
| Back-eye camera | 0.016 | 0.176 | 0.987 | Rejected |
| GPS | 4.431 | 2.398 | 0.000 | Accepted |

Moreover, an independent-samples t-test was conducted to compare the number of reduction of fatal and injury accidents per truck in adopting transportation technologies and not adopting transportation technologies conditions.

There was not a significant different in the score for adopting driving record (Mean =3.78, SD=1.302) and not adopting driving record (Mean =4.34, SD=0.928) condition; t (112) = -1.695, p = 0.093.

There was not a significant different in the score for adopting back-eye camera (Mean =4.33, SD=1.155) and not adopting back-eye camera (Mean =4.30, SD=0.969) condition; t (112) = 0.062, p = 0.950.

There was a significant different in the score for adopting GPS (Mean =4.43, SD=0.859) and not adopting GPS (Mean =3.40, SD=1.183) condition; t (112) = 4.121, p = 0.00.These results suggest that adopting GPS does have an effect on reduction of fatal and injury accidents per truck.

There is not enough evident to support that installing driving record and back-eye camera could decrease the number of accidents without injury per truck. However, the results suggest that adopting GPS does have an effect on reduction of accident without injury and the fatal and injury accidents per truck.

Table 8: An independent-samples t-test result of comparing the number of accident reduction of fatal and injury accidents per truck in adopting transportation technologies

| Technologies | T-test | F-test | Sig. | Conclude |
|---|---|---|---|---|
| Driving recorder | -0.1695 | 1.097 | 0.093 | Rejected |
| Back-eye camera | 0.063 | 0.101 | 0.950 | Rejected |
| GPS | 4.121 | 5.226 | 0.000 | Accepted |

*4.3.4 Hypothesis 4: Adoption of Thai/International standards (ISO, Q-mark, and 5s) have positive impact on net profit*

An independent-samples t-test is conducted to compare the changes of net profit in adopting Thai/International standard and not adopting Thai/International standard.

There was a significant different in the score for adopting ISO9001 (Mean =4.32, SD=0.696) and not adopting ISO9001 (Mean =3.21, SD=0.825) condition; t (112) = 7.803, p = 0.00. These results suggest that adopting ISO9001 does have an effect on net profit. Specifically, our results suggest that when adopting ISO9001, their net profit increase.

There was not a significant different in the score for adopting Q-mark (Mean =3.86, SD=0.69) and not adopting Q-mark (Mean =3.81, SD=0.953) condition; t(112) = 0.12, p = 0.905.

There was not a significant different in the score for adopting 5S (Mean =3.74, SD=1.010) and not adopting 5S (Mean =3.84, SD=0.922) condition; t (112) = -0.438, p = 0.662.

From testing by T-test shows that firms which adopted ISO9001 increase their net profit in firm because ISO9001 is a standard which based on a number of quality management principles including a strong customer focus, the motivation and implication of top manager, the process approach and continual improvement.

Table 9: An independent-samples t-test result in comparing the changes of net profit in adopting Thai/International standard

| Thai/ International standard | T-test | F-test | Sig. | Conclude |
|---|---|---|---|---|
| ISO 9001 | 7.803 | 0.286 | 0.000 | Accepted |
| Q-mark | 0.120 | 1.908 | 0.905 | Rejected |
| 5S | -0.414 | 0.320 | 0.681 | Rejected |

## 4. Conclusion

This research aims to study the major factors affecting the operational performance for logistics service providers. The firm-level data on the trucking industry was collected in a questionnaire survey.

The research finding shows that information sharing between manager and employee has a positive impact on net profit. In addition, the results also confirm that the companies that provide periodical training and training to the new driver could reduce their both fatal and injuries accident, and without injuries accident than the company which has no driver training. Moreover, the adoption of GPS also help to reduce the number of accident during working hour. Lastly, adopting of ISO9001 can help to improve company’s net profit.

This research has several limitations that suggest promising avenues for future research. First, data set is limited to only logistics service providers in Thailand. It could be extend to other industry or other countries. Second, we could extend this research by applying more robust statistical analysis to gain insight understanding. In addition to this internal factors inside the firms, the investigation on the external factors such as the requirement from the customers should be conducted in the future research.

## Acknowledgements

This paper was partially supported by Sirindhorn International Institute of Technology (SIIT), Thammasat University.

## References

[1] Lieb R., Miller J., The use of third-party logistics services by large US manufacturers: The 2000 survey. International Journal of Logistics: Research and Applications. 2002; 5(1): 1-12.

[2] Wagner M.S., et al., A qualitative investigation of innovation between third-party logistics providers and customers. International Journal of Logistics: Research and Applications, 2012; 140: 944-958.

[3] Speakman J.P. Innovation leads to new efficiencies. Logistics Management, 2002; 41: 71.

[4] Nixon M., Innovations in logistics technology: Generating top-line value and bottom-line ROI. World Trade. 2001; 14: 62-64.

[5] Chapman R. L., Soosay C., Kandampully J. Innovation in logistic services and the new business model: A conceptual framework. International Journal of Physical Distribution & Logistics Management. 2003; 33(7): 630-650.

[6] Ronald F. The Arrangement of Field Experiments. Journal of the Ministry of Agriculture of Great Britain. 1926; 33:503-513.

[7] Burke M.J., Sarpy S.A., Smith-Crowe K., Chan-Serafin S., Salvador R.O., Islam G., Relative effectiveness of worker safety and health training methods.American Journal of Public Health. 2006; 96(2): 315–324.

[8] Lin C.Y., Ho Yi-hui, Technological Innovation for China’s Logistics Industry. Journal of Technology Management & Innovation. 2007; 2(4): 1-19.

[9] He W., Liu C., Lu J. , Cao J. Impacts of ISO 14001 adoption on firm performance: Evidence from China. 2015; 32: 43-56.

[10] Bernstein I.H. Likert Scale Analysis & Encyclopedia of Social Measurement. 2005; 497–504.

[11] William M.K. Trochim, Types of Questions. 2006 [updated 2006 Oct 20]. Available from: http://www.socialresearchmethods.net/kb/questype.php.

[12] Hair, J.F.,. Multivariate Data Analysis with Readings..4th ed. Pearson Prentice Hall; 1995.

[13] Craparo R.t M. Significance level. In Salkind, Neil J. Encyclopedia of Measurement and Statistics 3. Thousand Oaks, CA: SAGE Publications; 2007: 889–891.

[14] Karanges E., Johnston K., Lings I. The influence Of internal communication on employee engagement: A pilot study. Public Relations Review. 2015; 41(1): 129-131.

# Factorwise Variance Dispersion Graphs

John Borkowski[1]* and Boonorm Chomtee[2]

[1]*Department of Mathematical Sciences, Montana State University, Bozeman, MT 59715, USA*

[2]*Department of Statistics, Kasetsart University, Bangkok, 10900, Thailand*

**Abstract**

In 2006, Borkowski introduced the componentwise variance dispersion graph (CVDG) which is a graphical tool for studying the prediction variance properties of experimental designs for mixtures. In this paper, the CVDG will be generalized to response surface designs in hypercube and hypersphere design spaces. The new plots called factorwise variance dispersion graphs (FVDGs) display predication variances throughout the design space. FVDGs supplement the use of variance dispersion graphs (VDGs) and fraction of design space (FDS) plots when assessing the prediction variance properties of a response surface design. Examples of FVDGs and other graphical methods will be presented.

*Keywords:* componentwise variance dispersion graphs, fraction of design space plots, scaled prediction variance, variance dispersion graphs

*Corresponding author.
E-mail address: jobo@math.montana.edu

## 1 Introduction

Consider the common experimental situation of having to select a response surface (RS) design having $k$-factors. The problem of choosing a ``best'' RS design for fitting the parameters of a linear model depends on the criterion of choice. For example, the experimenter could want a design that will produce RS model coefficient estimates with smallest variance or instead could want an orthogonal design. For a second or higher-order polynomial RS models, there is not a unique class of ``best'' designs, and it is recommended that coefficient estimates should be studied simultaneously [4]. Therefore, one desirable RS design property is to produce predicted values with small variance.

To address this issue, graphical techniques for evaluating prediction variance properties throughout the experimental region have been developed to supplement the use of single-value design optimality criteria, such as $D$ and $G$ optimality [1]. These include variance dispersion graphs (VDGs) [5,7], fraction of design space (FDS) plots [11], and quantile dispersion graphs (QDGs) [6]. Nguyen and Borkowski [8] introduced an enhancement to VDGs, called prediction variance volatility (PVV) plots.

For experiments with mixtures having constraints on the component proportions, the mixture design space becomes an irregularly-shaped polyhedron contained in a simplex. Piepel and Anderson [9] developed plots that display properties of the prediction variance on shrunken polyhedral spaces while Vining et al. [10] developed the prediction variance trace in which the prediction variance is plotted in each Cox-effect direction. Borkowski [3] introduced componentwise variance dispersion graphs (CVDGs) for mixture experiments in which the the minimum, maximum, and average scaled predictions are plotted across the set of feasible component levels for each mixture component.

The graphical concepts behind CVDGs for mixture experiments will be generalized to response surface designs in hypercube and hypersphere design spaces. In addition, PVV plotting will be incorporated to create new plots called factorwise variance dispersion graphs (FVDGs). FVDGs are a set of plots of the scaled prediction variance throughout the entire design space. FVDGs can be used to supplement the use of variance dispersion graphs (VDGs) and fraction of design space (FDS) plots when assessing the prediction variance properties of a response surface design throughout the design space.

In this paper, a brief review of VDGs, QDGs, PVV plots, and FDS plots will be given. Then FVDGs will be defined with design examples in hypercube and hypersphere design spaces.

## 2 Research Methodology

**Variance Dispersion Graphs (VDGs)**

Consider a $k$-factor response surface (RS) experiment having $N$ experimental runs. The *design matrix* $\mathbf{D}$ is the $N \times k$ matrix whose rows correspond to factor settings for the $N$ experimental runs. For any $p$-parameter polynomial RS model, $\mathbf{D}$ is expanded to form the $N \times p$ *model matrix* $\mathbf{X}$ with columns corresponding to the $p$ terms in the model.

Assuming IID errors, the prediction variance at point associated with $\mathbf{X}$ is $\mathbf{x}$ is $\mathrm{var}(\widehat{Y}(\mathbf{x})) = \sigma^2 V_1(\mathbf{x})$ where $V_1(\mathbf{x}) = \mathbf{x_m}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_m}$ and row vector $\mathbf{x_m}$ is the RS model expansion of $\mathbf{x}$. The *scaled prediction variance* (SPV) function $V(\mathbf{x})$ is defined as

$$V(\mathbf{x}) \;=\; \frac{N}{\sigma^2}\mathrm{var}(\widehat{Y}(\mathbf{x})) \;=\; NV_1(\mathbf{x}) \qquad (1)$$

For VDGs, three properties of interest involve the scaled prediction variance (SPV) function $V(\mathbf{x})$ in (1). The first property is the *average spherical prediction variance* $V_\rho$ which is the expected value of $V(\mathbf{x})$ on the surface of the sphere $S_\rho$ of radius $\rho$ from the origin:

$$V_\rho \;=\; \frac{1}{\omega_\rho}\int_{S_\rho} N\mathbf{x_m}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_m}\,d\mathbf{x} \qquad (2)$$

where $\omega_\rho$ is the surface area of $S_\rho$. The other two properties are the *minimum* and *maximum spherical prediction variances* given radius $\rho$:

$$VMIN_\rho \;=\; \min_{\mathbf{x}\in S_\rho} V(\mathbf{x}) \quad \text{and} \quad VMAX_\rho \;=\; \max_{\mathbf{x}\in S_\rho} V(\mathbf{x}) \qquad (3)$$

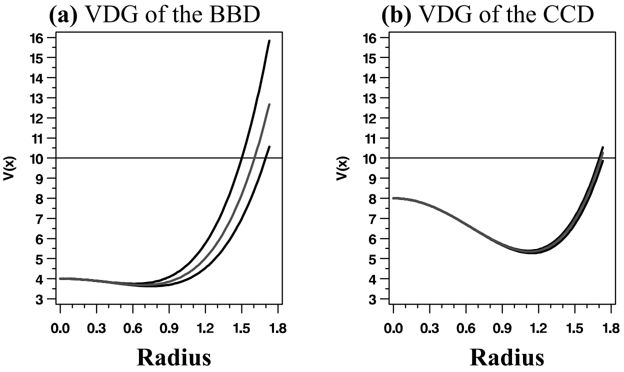
To compare prediction variance properties of RS designs, $V_\rho$, $VMIN_\rho$ and $VMAX_\rho$ are plotted against $\rho$ generating **variance dispersion graphs** or **VDG**s [2,5,7]. The $VMIN_\rho$ and $VMAX_\rho$ plots provide information regarding the relative stability of $V(\mathbf{x})$ throughout a spherical design space. A horizontal line at $V(\mathbf{x}) = p$ (the number of model parameters) is included because $p$ is the optimal value of $VMAX_\rho$ in the design space. Figure 1 contains VDGs of $V_\rho$, $VMIN_\rho$, and $VMAX_\rho$ for two 3-factor 16-point designs : the Box-Behnken design (BBD) having 4 center points and the central composite design (CCD) having 2 center points (see Table 1).

## Quantile Dispersion Graphs (QDGs)

One method for displaying properties of the distribution of $V(\mathbf{x})$ across various values of $\rho$ is through the use of boxplots of the values of $V(\mathbf{x}|\rho)$ across a sequence of $\rho$ values. The maximum and minimum points on a boxplots correspond to $VMIN_\rho$ and $VMAX_\rho$ for a specific $\rho$ value. The median provides a measure of central tendency as $\rho$ increases while the range and interquartile range provide information regarding the stability of $V(\mathbf{x}|\rho)$. This is an example of a **quantile dispersion graph** or QDG which is a plot of the quantiles of the distribution of $V(\mathbf{x}|\rho)$ [6].

**Table 1:** 16-Point Response Surface Designs

| BBD | | | CCD | | |
|---|---|---|---|---|---|
| $x_1$ | $x_2$ | $x_3$ | $x_1$ | $x_2$ | $x_3$ |
| $-1$ | $-1$ | 0 | $-1$ | $-1$ | $-1$ |
| $-1$ | 1 | 0 | $-1$ | $-1$ | 1 |
| 1 | $-1$ | 0 | $-1$ | 1 | $-1$ |
| 1 | 1 | 0 | $-1$ | 1 | 1 |
| $-1$ | 0 | $-1$ | 1 | $-1$ | $-1$ |
| $-1$ | 0 | 1 | 1 | $-1$ | 1 |
| 1 | 0 | $-1$ | 1 | 1 | $-1$ |
| 1 | 0 | 1 | 1 | 1 | 1 |
| 0 | $-1$ | $-1$ | 0 | 0 | $-\sqrt{3}$ |
| 0 | $-1$ | 1 | 0 | 0 | $\sqrt{3}$ |
| 0 | 1 | $-1$ | 0 | $-\sqrt{3}$ | 0 |
| 0 | 1 | 1 | 0 | $\sqrt{3}$ | 0 |
| 0 | 0 | 0 | $-\sqrt{3}$ | 0 | 0 |
| 0 | 0 | 0 | $\sqrt{3}$ | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |


**(a)** VDG of the BBD

**(b)** VDG of the CCD

**Figure 1:** VDGs for 3-factor 16-point designs: BBD (Figure 1a) and CCD (Figure 1b)

Figure 2 summarizes the distributions across $\rho$ of the 3-factor, 13-point small composite design (SCD). It clearly shows the increasing trend for the center of the distribution of $V(\mathbf{x}|\rho)$ as $\rho \to \sqrt{3}$. Also, $VMAX_\rho$ and the third quartile increase much more rapidly with increasing $\rho$ than either $VMIN_\rho$, the first quartile, or the median. Hence, as $\rho$ increases, these QDG boxplots show that the distribution of $V(\mathbf{x}|\rho)$ becomes increasingly right-skewed and with increasing variability.

## Prediction Variance Volatility (PVV) Plots

A **prediction variance volatility (PVV)** plot is a modification of a VDG [8]. For a PVV plot, $V(\mathbf{x})$ is calculated for a large number of randomly-generated points in the design space $\mathcal{R}$ and are then superimposed on VDGs.

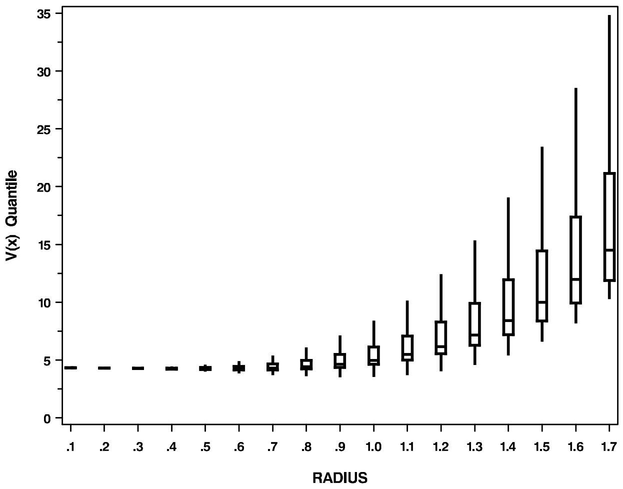
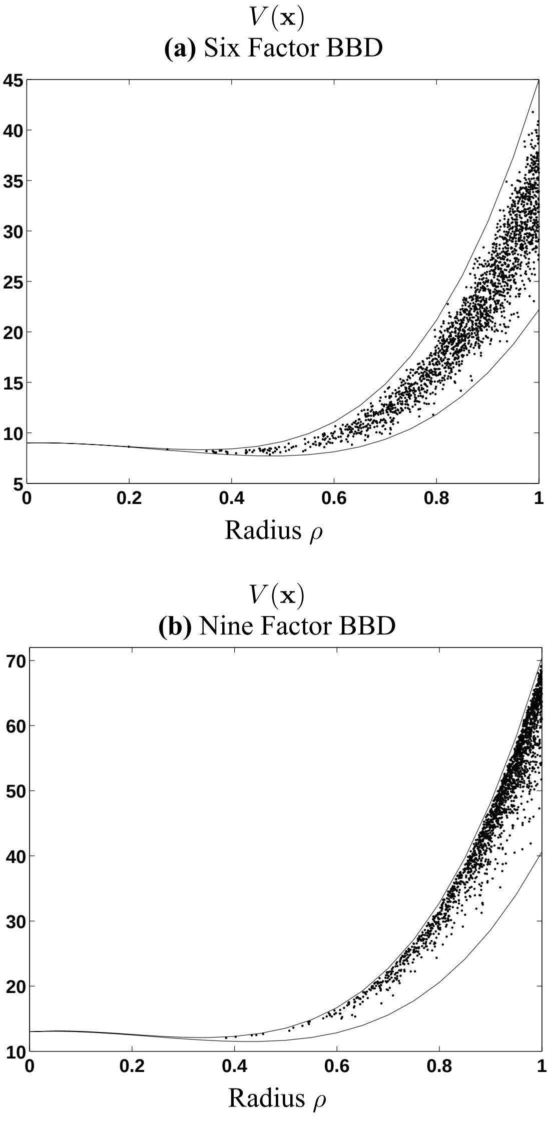


**Figure 2:** QDG Boxplots for a 13-Point SCD


$V(\mathbf{x})$

**(a)** Six Factor BBD

$V(\mathbf{x})$

**(b)** Nine Factor BBD

**Figure 3:** PVV Plots for 6-factor and 9-Factor BBDs

Figure 3 contains PVV plots for 6-factor and 9-factor Box-Behnknen designs (BBDs) in the hypercube. In each PVV plot, the SPVs for 2000 randomly-generated points are plotted against the associated $\rho$ values. For the 6-factor BBD, the points are concentrated midway between $V_{min}$ and $V_{max}$ for all $\rho$ with very low concentrations near the extremes. However, for the 9-factor BBD, the points are concentrated near $V_{max}$ for all $\rho$ with very low concentrations everywhere else. These PVV plots show the ``volatility" of the scaled prediction variances.

Nguyen and Borkowski [8] also showed that PVV plots are useful for studying rotatability properties of a design. That is, for designs approaching near-rotatable, the PVV plot becomes less volatile (with the PPV plot being a single curve for a rotatable design).

### Fraction of Design Space (FDS) Plots

For a response surface design, it is common to have regions with small $V(\mathbf{x})$ values while other regions having large $V(\mathbf{x})$ values. Therefore, it would be useful to assess the prediction variance properties of a design throughout the entire design space $\mathcal{R}$. Although VDGs, QDGs, and PVV plots show the maximum and minimum values of $V(\mathbf{x})$ for a given $\rho$, the information provided in these plots visually assign the same weight at each $\rho$. That is, they do not provide any information relating $\rho$ to the volume of the design region. It was recommended that the prediction variances should be weighted by the proportion of the design region [11].

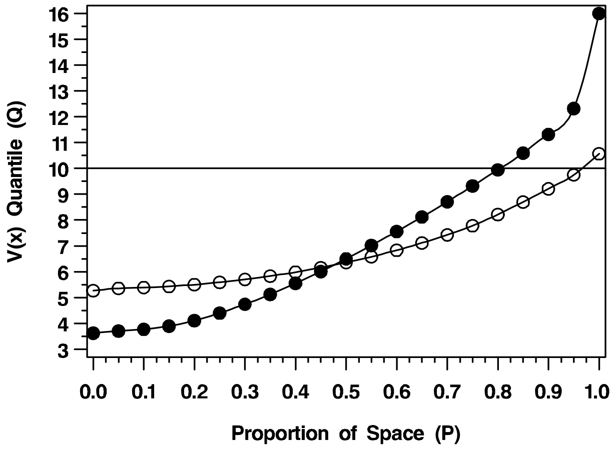


**Figure 4:** FDS Plots for 16-Point BBD (● ● ●) and CCD with $\alpha = \sqrt{3}$ (○ ○ ○)

The proportion of a spherical design region accounted for by all points within radius $\rho$ of the design center is an increasing function of $\rho$. As the number of design factors increases, the proportion of the design region becomes negligible for $\rho$ close to zero but increases rapidly as $\rho$ increases. Zahran et al. [11] proposed the **fraction of design space** or **FDS** plot. An FDS plot involves plotting the quantiles $Q$ of $V(\mathbf{x})$ against $P$, the proportion of the volume of the design region $R$ for which $V(\mathbf{x}) \leq Q$ for any specified value $Q$. Figure 4 contains the FDS plots for the two 16-point BBD and CCD plots given in Table 1.

## 3 Research Results and Discussion

### Factorwise Variance Dispersion Graphs (FVDGs)

A new graphical tool called the **factorwise variance dispersion graph (FVDG)** will be presented. A FVDG is a generalization of CVDGs and is a modification of VDGs and PVV plots [3,8].

To generate a FVDG for the $i^{th}$ design factor ($i = 1, 2, \ldots, k$), fix the factor level $x_i = a$ subject to $L_i \leq a \leq U_i$ where $L_i$ and $U_i$ are the minimum and maximum levels considered for $x_i$ in the design space $\mathcal{R}$. For example, if the design space is a hypercube, then $L_i = -1$ and $U_i = +1$, or if the design space is a hypersphere of radius $\sqrt{k}$, then $L_i = -\sqrt{k}$ and $U_i = +\sqrt{k}$.

Let $V(\mathbf{x}|x_i = a)$, $VMIN(\mathbf{x}|x_i = a)$, and $VMAX(\mathbf{x}|x_i = a)$ be the average, minimum, and maximum of $V(\mathbf{x})$ conditioned on $x_i = a$ for all design points $x \in \mathcal{R}$. The following steps outline the procedure for generating a FVDG for design factor $x_i$.

1. For the $i^{th}$ factor plot $V(\mathbf{x}|x_i = a)$, $VMIN(\mathbf{x}|x_i = a)$, and $VMAX(\mathbf{x}|x_i = a)$ across the set of $x_i = a$ values such that $L_i \leq a \leq U_i$.

2. Calculate $V(\mathbf{x})$ for a large number of randomly-generated points in $\mathcal{R}$. Add these points to the plot from Step 1.

3. A FVDG is created for the remaining factors using the points from Step 2.

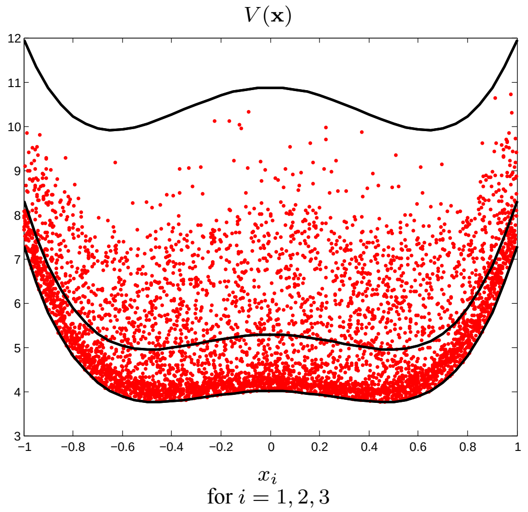
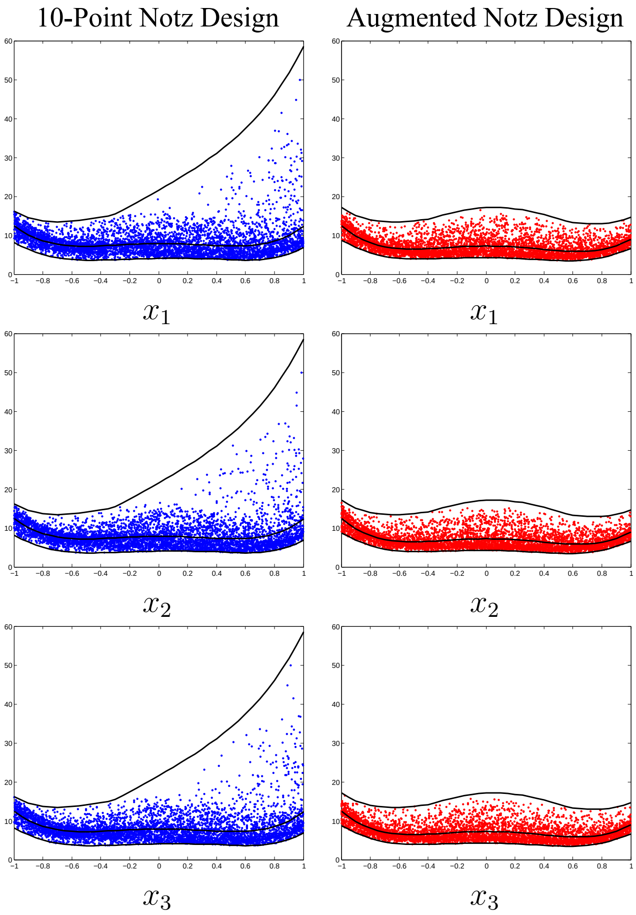

$V(\mathbf{x})$

**Figure 5:** The FVDG for the Face-Centered CCD

Figure 5 contains the FVDG for the 15-point face-centered CCD in Table 2a. Because of design symmetry (i.e., permuting $x_1$, $x_2$, $x_3$ yields the same design), it is not necessary to also include the FVDGs for $x_2$ and $x_3$. Now we can study the distribution of $V(\mathbf{x})$ throughout the cube $\mathcal{R}$. Figure 5 shows that the high density of SPV values near the minimum. Thus, for any $x_1$ values, the distribution of SPV values is strongly right-skewed.

For the second example, consider the 10-point 3-factor symmetric Notz design (see Table 2b). Note that the point (1,1,1) is not in this design. Thus, it is likely that the prediction variance at that point will be large. The FVDGs in column 1 of Figure 6 support that claim. Suppose we augment the Notz design by adding design point (1,1,1). Note the impact of adding the vertex $(1, 1, 1)$ on the FVDGs (given in column 2 of Figure 6). Note the stability (low volatility) in the prediction variances across all values of $x_i$ for $i = 1, 2, 3$.


10-Point Notz Design

Augmented Notz Design

**Figure 6:** FVDGs for 10-point and augmented 11-point Notz Designs

**Table 2:** Three Response Surface Designs

| **(a)** 15-Point CCD | | | **(b)** 10-Point Notz Design | | | **(c)** 15-point *SAS* Design | | |
|---|---|---|---|---|---|---|---|---|
| $x_1$ | $x_2$ | $x_3$ | $x_1$ | $x_2$ | $x_3$ | $x_1$ | $x_2$ | $x_3$ |
| $-1$ | $-1$ | $-1$ | $-1$ | $-1$ | $-1$ | $-1$ | $-1$ | $-1$ |
| $-1$ | $-1$ | 1 | 1 | $-1$ | $-1$ | $-1$ | $-1$ | 1 |
| $-1$ | 1 | $-1$ | $-1$ | 1 | $-1$ | $-1$ | $-1$ | 0 |
| $-1$ | 1 | 1 | $-1$ | $-1$ | 1 | 1 | 0 | 0 |
| 1 | $-1$ | $-1$ | 1 | 1 | $-1$ | 1 | $-1$ | $-1$ |
| 1 | $-1$ | 1 | 1 | $-1$ | 1 | 1 | $-1$ | 1 |
| 1 | 1 | $-1$ | $-1$ | 1 | 1 | 1 | $-1$ | 0 |
| 1 | 1 | 1 | 1 | 0 | 0 | $-1$ | 1 | $-1$ |
| 0 | 0 | $-1$ | 0 | 1 | 0 | $-1$ | 1 | 1 |
| 0 | 0 | 1 | 0 | 0 | 1 | $-1$ | 1 | 0 |
| 0 | $-1$ | 0 | | | | 1 | $-1$ | 1 |
| 0 | 1 | 0 | | | | 1 | 0 | 0 |
| $-1$ | 0 | 0 | | | | $-1$ | 0 | 0 |
| 1 | 0 | 0 | | | | 0 | 0 | 0 |
| 0 | 0 | 0 | | | | 1 | 1 | 0 |

In the third example, consider the 15-point 3-factor nonsymmetric $D$-optimal design generated by *SAS* Proc Optex (Table 3c) assuming the second-order model:

$$\begin{aligned} f(x) \quad &= \quad \beta_0 + \sum_{i=1}^{3} \beta_i x_i + \sum_{i=1}^{3} \beta_{ii} x_i^2 \\ &\quad + \beta_{12}x_1x_2 \ + \ \beta_{13}x_1x_3 \ + \ \beta_{23}x_2x_3 \end{aligned}$$

Suppose, however, the **interaction model** is the true model:

$$f(x) \ = \ \beta_0 + \sum_{i=1}^{3} \beta_i x_i + \beta_{12}x_1x_2 + \beta_{13}x_1x_3 + \beta_{23}x_2x_3$$

For the second-order model, the FVDGs are given in column 1 of Figure 7. Because the design is not symmetric, the FVDGs for $x_1$, $x_2$ and $x_3$ are very different. For example, the prediction variances are dense near the minimum for all values of $x_1$, while they are much more dispersed for $x_2$ and $x_3$. The shapes of the minimum, maximum and average curves also vary greatly across the factors.

For the interaction model, the FVDGs are given in column 2 of Figure 7. These FVDGs highlight the dramatic changes that can occur when different models are considered. Now the FVDGs for $x_1$ and $x_2$ are very similar, but were very different for the second-order model. For the interaction model, the prediction variances are now concentrated near the minimum for all three factors.

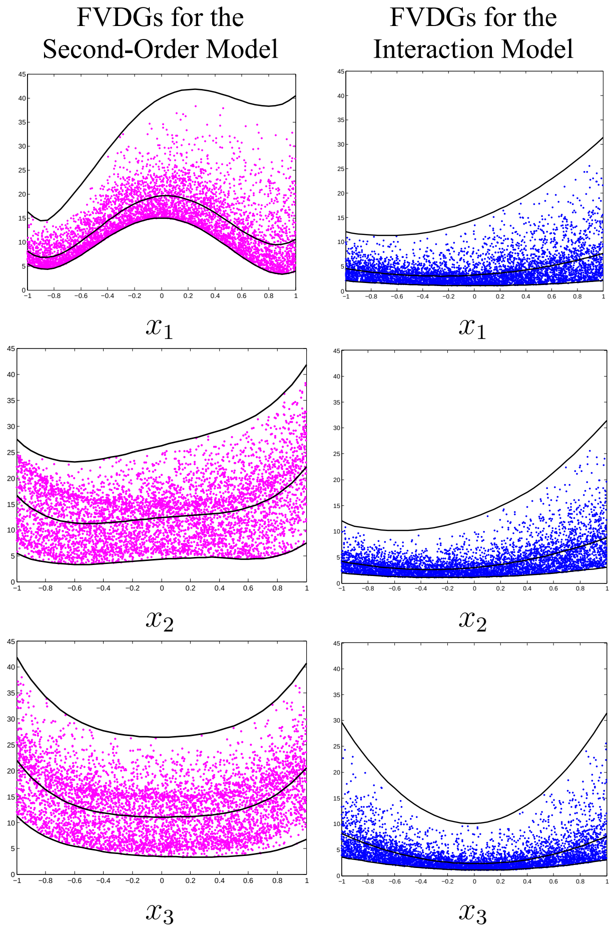


**Figure 7:** FVDGs for a 15-Point Design Generated by Proc Optex in *SAS*

Now suppose the design space $\mathcal{R}$ is spherical:

$$\mathcal{R} \ = \ \left\{ x = (x_1, x_2, \ldots, x_k) : \left( \sum_{i=1}^{k} x_i^2 \right) \ \leq \sqrt{k} \right\}$$

Consider the symmetric 15-point and 16-point Central Composite designs (CCDs) having 1 and 2 center points, respectively. Therefore, $-\sqrt{3} \leq x_i \leq \sqrt{3}$ for $i = 1, 2, 3$. Note the effect of the additional centerpoint in the FVDGs for $x_i$ near 0 in Figure 8.

Finally, lets return the the BBD and CCD given in Table 1 with VDGs in Figure 1 and FDS plots in Figure 4. These can be supplemented with their FVDGs shown in Figure 9. Because of design symmetry, only one FVDG is presented for each design. Note that the prediction variance is less stable for the BBD for most values of each $x_i$.

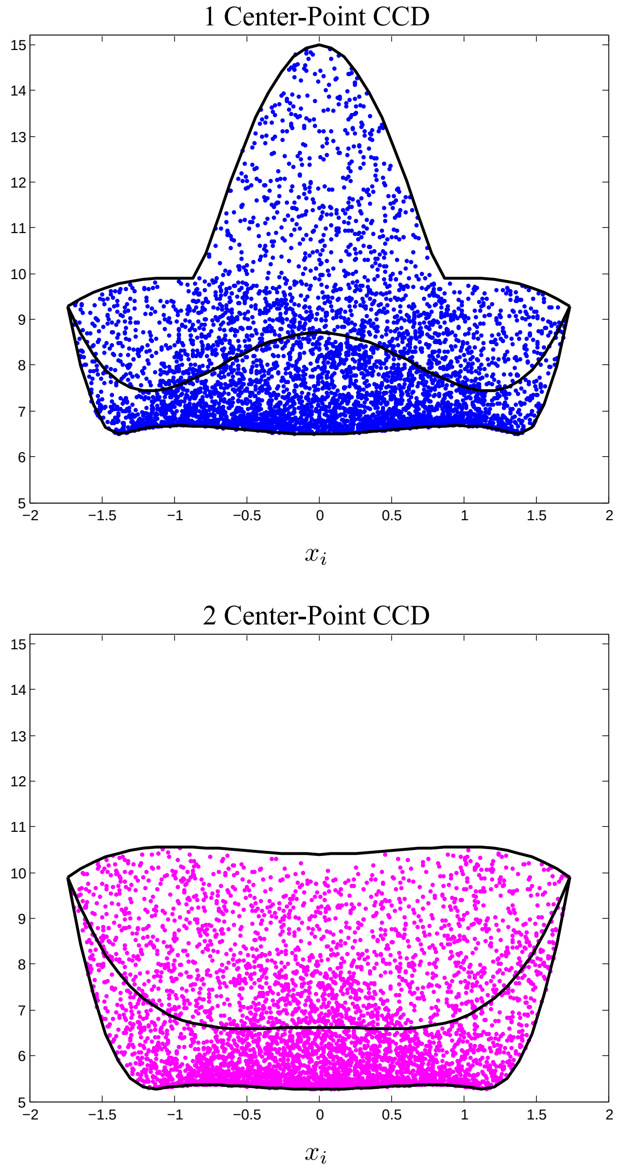


**Figure 8:** FVDGs for 15-Point and 16-Point CCDs

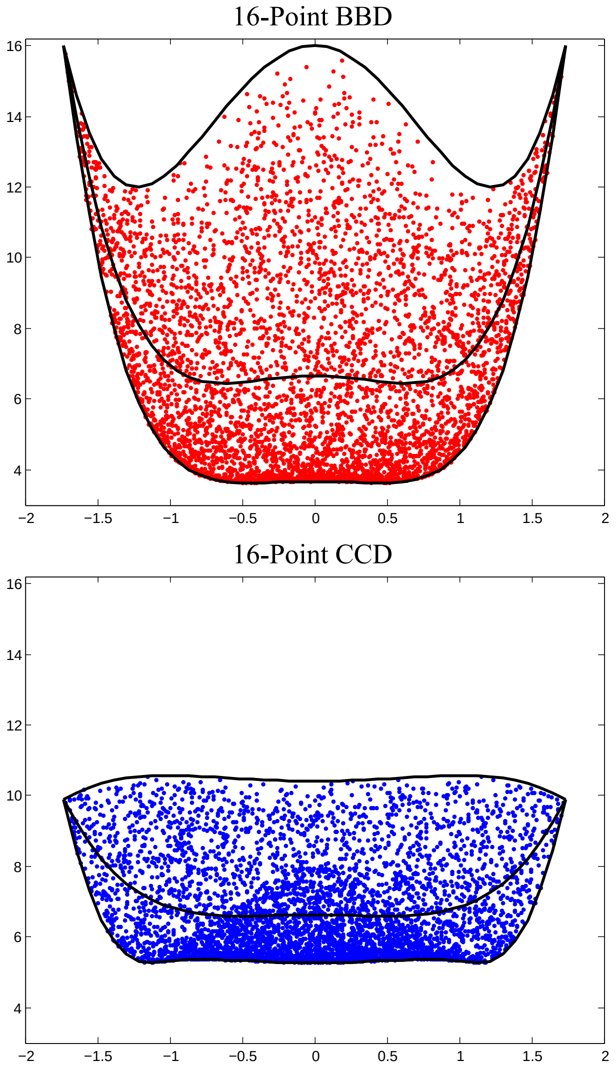


**Figure 9:** FVDGs for the BBD and CCD in Table 1

## 4 Conclusions

To determine which response surface design to run, the experimenter should study the prediction variance properties of each design. For designs with $\geq 3$ factors, we can compare designs using design optimality criteria (e.g., $D$ or $G$). These criteria, however, provide no information about the distribution of the prediction variance function over $\mathcal{R}$. To address this weakness, graphical methods were developed (VDGs, QDGs, PVV and FDS plots, and now FVDGs), which provide information about the prediction variance properties of a design throughout the design space.

Before running an experiment it is recommended to look at multiple graphical tools for studying the prediction variance properties of potential designs to aid in the design selection process. All of the graphical methods help the experimenter to understand $V(\mathbf{x})$ throughout the design space. Thus, the use of these graphical methods is a practical way to compare the prediction variance properties of competing designs.

## References

[1] Atkinson AC, Donev AN, Tobias R. Optimum experimental designs, with SAS. Clarendon: Oxford Press; 2007.

[2] Borkowski JJ. Spherical prediction-variance properties of central composite and Box-Behnken designs. Technometrics. 1995; 7: 399-410.

[3] Borkowski JJ. Chapter 14: Graphical methods for assessing the prediction capability of response

surface designs. Response surface methodology and related topics. Hackensack NJ: World Scientific; 2006.

[4] Box GEP, Hunter JS. Multi-factor experimental designs for exploring response surfaces. The Annals of Mathematical Statistics. 1957; 72: 195-241.

[5] Giovannitti-Jensen A, Myers RH. Graphical assessment of the prediction capability of response surface designs. Technometrics. 1989; 31: 159-171.

[6] Khuri AI, Kim HJ, Um Y. Quantile plots of the prediction variance for response surface designs. Computational Statistics & Data Analysis. 1996; 22: 395-407.

[7] Myers RH, Vining GG, Giovannitti-Jensen A, Myers SL (1992). Variance dispersion properties of second-order response surface designs. Journal of Quality Technology. 1992; 24: 1-11.

[8] Nguyen N-K, Borkowski JJ. New 3-level response surface designs constructed from incomplete block designs. Journal of Statistical Planning and Inference. 2008; 138: 294-305.

[9] Piepel GF, Anderson CM. Variance dispersion graphs for designs on polyhedral regions, 1992 Proceedings of the Section on Physical and Engineering Sciences; Joint Statistical Meetings; 1992 August 9-13; Boston, USA. 1993. p.111-117.

[10] Vining GG, Cornell JA, Myers RH. A graphical approach for evaluating mixture designs. Applied Statistics. 1993; 42: 127-138.

[11] Zahran A, Anderson-Cook CM, Myers, RH. Fraction of design space plots to assess prediction capability of response surface designs. Journal of Quality Technology. 2003; 35: 377-386.

# The Interval Estimation of Poisson Distribution by Using Maximum Likelihood, Markov Chain Monte Carlo, and Bayes methods

Autcha Araveeporn*
*Department of Statistics, Faculty of Science,*
*King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand*

## Abstract

The goal of this research is to compare the interval estimation based on Poisson distribution. The interval estimation is evaluated by Maximum Likelihood (ML) method, Markov Chain Monte Carlo (MCMC) method, and Bayes method from a point estimation to estimate confidence interval. The confidence coefficients are approximated by considering the proportion when the upper and lower of confidence interval are covered the true parameters. If the Confidence Coefficients (CC) are greater than the fixed confidence interval, the Average Width (AW) of the confidence interval will focus the performance of these methods. In this case, the data is generated by Monte Carlo process depended on Poisson distribution with true parameters 0.5, 2, 5, 10, and 20, sample sizes 30, 50, 100, 300, and 500, and the 90%, 95%, and 99% confidence interval. The performance of three methods is compared by the CC and the AW values. The output is showed that ML method outperforms the other methods when true parameter is small values (0.5) for all sample sizes. For the large sample sizes (100,300,500), MCMC method is reasonably performed when true parameter is 5. Furthermore, the Bayes method is a good performance in most cases.

*Keywords*: Bayes, Confidence Coefficient, Markov Chain Monte Carlo, Maximum Likelihood, Poisson distribution

*Corresponding Author
E-mail Address: kaautcha@hotmail.com, kaautcha@kmitl.ac.th

## 1. Introduction

The estimation is a part of statistical inference for studying data from sample which is described the characteristic population. The estimation contains point estimation and interval estimation. The point estimation is to estimate an estimator from sample that referred to a population parameter. The estimator presents a single values or point, but the interval estimation presents a range which is constructed along with point estimation to show the reliability of the estimation.

Confidence interval is a type of interval estimation to approximate the range of a population parameter. Interval estimation consists of lower confidence interval and upper confidence interval which depended on the level of confidence and standard deviation. The level of confidence interval is a range of probability that captured this population parameter. When the population parameter hold on the narrow confidence intervals, it can be concluded that the estimator is a high accuracy.

In this case, we interested the discrete random variable in form of Poisson distribution. The population parameter defined an average number of successes that occurred in a specified region of time and space. A Poisson random variable is a number of successes in a length, an area, a period of time such as the number of accident occurring at the express way per day.

The interval estimation depends on the point estimation which is referred to the population parameter by the estimator. The Maximum Likelihood (ML) method is widely used a popular method to estimate the population parameter because the estimator is to be a class of uniformly minimum variance unbiased estimator [1]. Moreover, the Markov Chain Monte Carlo (MCMC) [2] method can produce the estimator of prior distribution and posterior distribution by generating sample from these functions. But the posterior distribution is provided by Bayesian model, the prior distribution can be used to evaluate this estimator. The Bayes method is also to interest since the distribution of the population parameter or called prior distribution and posterior distribution are used to approximate estimator[3]. The Bayes method is so complicate for computing estimator, then the MCMC method is used to help for estimating the population parameter.

For this reason, we propose the MLE, MCMC, and Bayes methods to estimate confidence interval with Poisson distribution using simulation studies.

## 2. Scope

2.1 Let random variable X be independent and identically distributed (iid) random variables following a Poisson distribution with parameter $\lambda$, and the probability density function is

$$f(x_i \mid \lambda) \;=\; \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \quad , \quad x_i = 0,1,2,\ldots \quad .$$

2.2 Let the prior distribution of $\lambda$ be a gamma distribution with parameters $a,b$, and the probability density function is

$$g(\lambda \mid a,b) \;=\; \frac{\lambda^{a-1}e^{-\frac{\lambda}{b}}}{\Gamma(a)b^{a}} \quad , \quad \lambda > 0.$$

The gamma distribution shows the parameter format as the Poisson distribution or called conjugate distribution.

2.3 The sample sizes are considered at n = 30, 50, 100, 300, and 500.

2.4 The true parameter of Poisson distribution is defined as 0.5, 2, 5, 10, and 20.

2.5 Let the significance confidence level $(\alpha)$ be 3 levels at 0.1, 0.05, and 0.01 following the confidence interval (1-α)100% is to be 90%, 95%, and 99%.

2.6 The R program [4] is used to generate data at 500 replicates for each cases.

## 3. Methodology

3.1 The random variable X is generated in a class of Poisson distribution following the sample sizes, the true parameter, and the significance confidence level.

3.2 The methods for computing the confidence interval consist of following 3 methods:

*3.2.1 Maximum Likelihood (ML) Method*

The $\hat{\lambda}_{ML}$ is ML estimator of $\overline{X}$ [3] or written as $\hat{\lambda}_{ML} = \overline{X}$, and the variance of ML estimator is given by

$$Var(\hat{\lambda}_{ML}) = Var(\overline{X}) = \frac{1}{n}Var(X) \because Var(X) = \lambda$$
$$= \frac{\lambda}{n}.$$

The population parameter $(\lambda)$ is estimated by ML then $\lambda \Rightarrow \hat{\lambda}_{ML} = \overline{X}$.

The confidence interval (1-α)100% of $\lambda$ is approximated by

$$\lambda = \hat{\lambda}_{ML} \pm z_{\alpha/2}\sqrt{Var(\hat{\lambda}_{ML})}$$
$$= \overline{X} \pm z_{\alpha/2}\sqrt{\frac{\overline{X}}{n}}.$$

*3.3.2 Markov Chain Monte Carlo (MCMC) Method*

The Markov Chain Monte Carlo (MCMC) method is operated by sequentially sampling parameter values from a Markov Chain whose stationary distribution which is desired from posterior distribution. The Gibbs sampling ([6], [7]) is a algorithm for MCMC computing. The algorithm of Gibbs sampling can be seen at [5]. We carry out the WinBUGS Program [3] to obtain the estimating estimator from the posterior distribution function. The MCMC estimators can be computed by

$$\hat{\lambda}_{MCMC} \;=\; \frac{1}{T}\sum_{t=1}^{T}\lambda^{(t)},$$

$$Var\left(\hat{\lambda}_{MCMC}\right) \;=\; \frac{1}{T-1}\sum_{t=1}^{T}\left(\lambda^{(t)} - \bar{\hat{\lambda}}\right)^{2},$$

where $\lambda^{(t)}$ is generated from the posterior distribution based on the gamma distribution at parameter $a^{(t)}$ and $b^{(t)}$, $\bar{\hat{\lambda}}$ is a sample mean of independent observation, and T is a iteration of posterior distribution function .

The confidence interval (1-α)100% of $\lambda$ is written by

$$\lambda = \hat{\lambda}_{MCMC} \pm z_{\alpha/2}\sqrt{Var\left(\hat{\lambda}_{MCMC}\right)}.$$

For the MCMC method, the parameter $a$, and $b$ are approximated by

$$\hat{a}_{MCMC} = \frac{1}{T}\sum_{t=1}^{T}a^{(t)} \quad ,$$

$$\hat{b}_{MCMC} = \frac{1}{T}\sum_{t=1}^{T}b^{(t)} \quad ,$$

where $a^{(t)}$ is generated from exponential distribution , and $b^{(t)}$ is generated from gamma distribution.

*3.2.3 Bayes Method*

Let $X_1,\ldots,X_n$ is the random variable of Poisson distribution with parameter $\lambda$ while exponential and gamma distributions are considered the prior distribution with parameter $a$, and $b$. The resulting posterior distribution is equal

$$\begin{aligned} h(\lambda|x_i) &\propto f(x_i|\lambda)g(\lambda|a,b) \\ &\propto \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n}X_i}}{\prod_{i=1}^{n}X_i!}\times\frac{\lambda^{a-1}e^{-\frac{\lambda}{b}}}{\Gamma(a)b^a} \\ &\propto e^{-(n+\frac{1}{b})\lambda}\lambda^{n\bar{X}+a-1}. \end{aligned}$$

We reach to conclusion that

$$\lambda|x_i \sim gamma\left(n\overline{X}+a, n+\frac{1}{b}\right)$$ [8].

Therefore the gamma distribution is conjugate to the Poisson distribution. The posterior mean or called Bayes estimator is given by

$$E(\lambda|x_i)=\frac{n\overline{X}+a}{n+\frac{1}{b}},$$

while the posterior variance is given by

$$Var(\lambda|x_i)=\frac{n\overline{X}+a}{\left(n+\frac{1}{b}\right)^2}.$$

The confidence interval (1-α)100% of $\lambda$ is written by

$$\lambda=\frac{n\overline{X}+a}{n+\frac{1}{b}}\pm z_{\alpha/2}\sqrt{\frac{n\overline{X}+a}{\left(n+\frac{1}{b}\right)^2}}.$$

Hence, the parameter $a$, and $b$ are obtained from MCMC method. Recall that

$$\lambda=\frac{n\overline{X}+\hat{a}_{MCMC}}{n+\frac{1}{\hat{b}_{MCMC}}}\pm z_{\alpha/2}\sqrt{\frac{n\overline{X}+\hat{a}_{MCMC}}{\left(n+\frac{1}{\hat{b}_{MCMC}}\right)^2}}.$$

*3.3 The estimating confidence coefficient*

The confidence interval is approximated by ML, Bayes, MCMC methods at significance level 0.1, 0.05, and 0.01. If the confidence intervals cover the true parameters, we will count the number and compute the proportion denoted the confidence coefficient $(1-\hat{\alpha})$.

*3.4 The comparison of the confidence coefficient and the fixed confidence interval*

The confidence coefficient $(1-\hat{\alpha})$ is to compare with the fixed confidence interval $(1-\alpha_0)$ that we define the significance level at 0.05. If the confidence coefficient is more than the fixed confidence interval , we will perform these methods. The comparison is given by

$$1-\hat{\alpha}\geq 1-\alpha_0 = P_0 - z_{\alpha/2}\sqrt{\frac{P_0(1-P_0)}{M}},$$

where $P_0$ is the fixed probability given by 0.9, 0.95, and 0.99, M is the number of replications.

The fixed confidence intervals are computed by: $P_0=0.9$,

$$1-\alpha_0=0.9-1.96\sqrt{\frac{0.9(1-0.9)}{500}}=0.8737,$$

$P_0=0.95$,

$$1-\alpha_0=0.95-1.96\sqrt{\frac{0.95(1-0.95)}{500}}=0.9308,$$

$P_0=0.99$,

$$1-\alpha_0=0.99-1.96\sqrt{\frac{0.99(1-0.99)}{500}}=0.9812.$$

*3.5 The average width of confidence interval*

There are several the confidence coefficients are more than the fixed confidence interval, then the average width of confidence interval will be considered instead. The average width of confidence interval is evaluated by computing the average of difference values between upper limit and lower limit or written as $\sum_{j=1}^{500}\frac{\left(U_j-L_j\right)}{500}$, where $U_j$ is the upper confidence interval, and $L_j$ is the lower confidence interval.

## 4. Results

The estimating confidence interval of population parameter with Poisson distribution is presented by the Confidence Coefficient (CC) and the Average Width (AW) at Table 1-3. The first column and the second columns of these tables are shown the sample sizes and the true parameters. The Confidence Coefficient (CC) and the Average Width (AW) are presented in the next six columns for three methods. The minimizing AW values are illustrated the performance of these methods ,but some AW values are in the blank because the confidence coefficient is less than the fixed confidence interval. By observing the CC and AW, the results appear as follow:

*4.1 A 90% confidence interval*

From Table 1, the AW of ML method is a minimum values when $\lambda = 0.5$ for all sample sizes. For MCMC method, the AW outperforms at $\lambda = 5$ with $n$ =100, 300, and 500 , and $\lambda$ = 10 ,20 with $n$ = 500. The AW of Bayes method appears the minimum values at most cases especially when $n$ = 30 and 50.

*4.2 A 95% confidence interval*

From Table 2, the CC and AW of ML method are a minimum values when $\lambda = 0.5$ for all sample sizes, when $\lambda = 2$ for $n$ =100 and 500. The AW of the MCMC method shows the minimum values of the same 90% confidence interval when $\lambda = 5$ with $n$ = 100, 300, and 500 , and $\lambda$ = 10 ,20 with $n$ = 500. The AW of Bayes method appears the minimum values at most cases especially when $n$ =30 and 50. For $\lambda = 0.5$ and $n = 50, 100$ , the CC are the same values for these methods.

*4.3 A 99% confidence interval*

From Table 3, the CC and AW of ML method are a minimum values when $\lambda = 0.5$ for all sample sizes, when $\lambda = 2$ for $n = 100$ . The AW of the MCMC method shows the minimum values of the same 90% and 95% confidence interval when $\lambda = 5$ with $n$ =100, 300, and 500 , and $\lambda$ = 10 ,20 with $n$ = 500. The AW of Bayes method appears the minimum values at most cases especially when $n$ =30 and 50. For $\lambda = 0.5$ and $n = 30, 50$ , the CC values are equal on three methods.

Table1 : The Confidence Coefficient (CC) and Average Width (AW) obtained via 90% confidence interval.

| n | $\lambda$ | Methods | | | | | |
|---|---|---|---|---|---|---|---|
| | | ML | | MCMC | | Bayes | |
| | | CC | AW | CC | AW | CC | AW |
| 30 | 0.5 | 0.974 | **0.4201** | 0.874 | 0.4282 | 0.874 | 0.4221 |
| | 2 | 0.912 | 0.8454 | 0.912 | 0.8577 | 0.906 | **0.8426** |
| | 5 | 0.900 | 1.3415 | 0.900 | 1.3383 | 0.900 | **1.3375** |
| | 10 | 0.890 | 1.8983 | 0.894 | 1.9115 | 0.890 | **1.8949** |
| | 20 | 0.904 | 2.6842 | 0.910 | 2.6840 | 0.904 | **2.6814** |
| 50 | 0.5 | 0.874 | **0.3265** | 0.906 | 0.3341 | 0.906 | 0.3273 |
| | 2 | 0.912 | 0.6568 | 0.916 | 0.6634 | 0.912 | **0.6554** |
| | 5 | 0.894 | 1.0412 | 0.894 | 1.0418 | 0.894 | **1.0394** |
| | 10 | 0.916 | 1.4722 | 0.916 | 1.4888 | 0.916 | **1.4707** |
| | 20 | 0.898 | 2.0793 | 0.898 | 2.0845 | 0.898 | **2.0780** |
| 100 | 0.5 | 0.894 | **0.2317** | 0.894 | 0.2362 | 0.890 | 0.2320 |
| | 2 | 0.888 | 0.4661 | 0.896 | 0.4670 | 0.896 | **0.4657** |
| | 5 | 0.896 | 0.7353 | 0.900 | **0.7320** | 0.896 | 0.7346 |
| | 10 | **0.872** | - | 0.880 | 1.0601 | **0.872** | - |
| | 20 | 0.896 | 1.4706 | 0.900 | 1.4884 | 0.896 | **1.4701** |
| 300 | 0.5 | 0.898 | **0.1346** | 0.908 | 0.1371 | 0.908 | 0.1347 |
| | 2 | 0.914 | 0.2684 | 0.920 | 0.2684 | 0.918 | **0.2683** |
| | 5 | 0.912 | 0.4250 | 0.904 | **0.4280** | 0.908 | 0.4249 |
| | 10 | 0.898 | 0.6004 | 0.902 | 0.6058 | 0.898 | **0.6003** |
| | 20 | 0.932 | 0.8494 | 0.932 | 0.8496 | 0.932 | 0.8493 |
| 500 | 0.5 | 0.886 | **0.1039** | 0.890 | 0.1055 | 0.890 | **0.1039** |
| | 2 | 0.886 | 0.2080 | 0.892 | 0.2103 | 0.886 | **0.2079** |
| | 5 | 0.902 | 0.3289 | 0.902 | **0.3264** | 0.902 | 0.3288 |
| | 10 | 0.894 | 0.4653 | 0.898 | **0.4642** | 0.894 | 0.4652 |
| | 20 | 0.884 | 0.6578 | 0.884 | **0.6572** | 0.884 | 0.6578 |

Table 2: The Confidence Coefficient (CC) and Average Width (AW) obtained via 95% confidence interval.

| n | $\lambda$ | Methods | | | | | |
|---|---|---|---|---|---|---|---|
| | | ML | | MCMC | | Bayes | |
| | | CC | AW | CC | AW | CC | AW |
| 30 | 0.5 | **0.910** | - | 0.952 | 0.5103 | 0.952 | **0.5037** |
| | 2 | 0.954 | 1.0087 | 0.954 | 1.0227 | 0.950 | **1.0053** |
| | 5 | 0.956 | 1.6026 | 0.954 | 1.5982 | 0.954 | **1.5978** |
| | 10 | 0.946 | 2.2584 | 0.946 | 2.2745 | 0.946 | **2.2543** |
| | 20 | 0.958 | 3.1982 | 0.958 | 3.1981 | 0.958 | **3.1949** |
| 50 | 0.5 | **0.916** | - | **0.916** | - | **0.916** | - |
| | 2 | 0.954 | 0.7843 | 0.956 | 0.7922 | 0.954 | **0.7826** |
| | 5 | 0.934 | 1.2387 | 0.936 | 1.2395 | 0.934 | **1.2366** |
| | 10 | 0.954 | 1.7512 | 0.958 | 1.7711 | 0.954 | **1.7493** |
| | 20 | 0.960 | 2.4780 | 0.958 | 2.4842 | 0.960 | **2.4765** |
| 100 | 0.5 | 0.942 | **0.2752** | 0.942 | 0.2804 | 0.942 | 0.2756 |
| | 2 | **0.926** | - | 0.934 | 0.5526 | 0.934 | **0.5505** |
| | 5 | 0.954 | 0.8756 | 0.954 | **0.8414** | 0.960 | 0.8749 |
| | 10 | 0.946 | 1.2400 | 0.950 | 1.2656 | 0.946 | **1.2393** |
| | 20 | 0.964 | 1.7535 | 0.966 | 1.7751 | 0.964 | **1.7530** |
| 300 | 0.5 | 0.962 | **0.1600** | 0.970 | 0.1630 | 0.962 | **0.1600** |
| | 2 | 0.970 | 0.3201 | 0.970 | 0.3200 | 0.970 | **0.3199** |
| | 5 | 0.940 | 0.5059 | 0.938 | **0.5008** | 0.940 | 0.5057 |
| | 10 | 0.942 | 0.7153 | 0.944 | 0.7218 | 0.942 | **0.7152** |
| | 20 | 0.942 | **1.0120** | 0.942 | **1.0120** | 0.942 | 1.0124 |
| 500 | 0.5 | 0.942 | **0.1239** | 0.942 | 0.1240 | 0.942 | 0.1258 |
| | 2 | 0.942 | **0.2480** | 0.942 | 0.2508 | 0.942 | **0.2480** |
| | 5 | 0.952 | 0.3914 | 0.950 | **0.3885** | 0.952 | 0.3913 |
| | 10 | 0.958 | 0.5546 | 0.960 | **0.5533** | 0.958 | 0.5545 |
| | 20 | 0.943 | 0.7840 | 0.946 | **0.7833** | 0.946 | 0.7840 |

Table 3: The Confidence Coefficient (CC) and Average Width (AW) obtained via 99% confidence interval.

| n | $\lambda$ | Methods | | | | | |
|---|---|---|---|---|---|---|---|
| | | ML | | MCMC | | Bayes | |
| | | CC | AW | CC | AW | CC | AW |
| 30 | 0.5 | **0.978** | - | **0.978** | - | **0.978** | - |
| | 2 | 0.984 | 1.3292 | 0.988 | 1.3489 | 0.988 | **1.3248** |
| | 5 | 0.99 | 2.1069 | 0.99 | 2.1020 | 0.99 | **2.1006** |
| | 10 | 0.986 | 2.9736 | 0.986 | 2.9940 | 0.986 | **2.9681** |
| | 20 | 0.982 | 4.2022 | 0.984 | 4.2021 | 0.982 | **4.1977** |
| 50 | 0.5 | **0.968** | - | **0.968** | - | **0.968** | - |
| | 2 | 0.998 | 1.0258 | 0.998 | 1.0358 | 0.998 | **1.0236** |
| | 5 | 0.984 | 1.6278 | 0.984 | 1.6284 | 0.984 | **1.6250** |
| | 10 | 0.992 | 2.2982 | 0.992 | 2.3238 | 0.992 | **2.2958** |
| | 20 | 0.992 | 3.2558 | 0.994 | 3.2642 | 0.992 | **3.2538** |
| 100 | 0.5 | 0.99 | **0.3652** | 0.994 | 0.3720 | 0.99 | 0.3656 |
| | 2 | 0.992 | **0.7270** | 0.992 | 0.7289 | 0.992 | 0.7271 |
| | 5 | 0.992 | 1.1505 | 0.992 | **1.1451** | 0.992 | 1.1495 |
| | 10 | 0.996 | 1.6286 | 0.996 | 0.6621 | 0.996 | **1.6277** |
| | 20 | 0.988 | 2.3037 | 0.988 | 2.3321 | 0.988 | **2.3320** |
| 300 | 0.5 | 0.994 | **0.2100** | 0.996 | 0.2140 | 0.992 | 0.2101 |
| | 2 | 0.990 | 0.4209 | 0.996 | **0.4208** | 0.994 | **0.4208** |
| | 5 | 0.994 | 0.6652 | 0.994 | **0.6586** | 0.994 | 0.6650 |
| | 10 | 0.992 | 0.9398 | 0.992 | 0.9484 | 0.992 | **0.9397** |
| | 20 | 0.988 | 1.3296 | 0.988 | 1.3299 | 0.988 | **1.3294** |
| 500 | 0.5 | 0.990 | **0.1631** | 0.990 | 0.1655 | 0.988 | **0.1631** |
| | 2 | 0.984 | 0.3263 | 0.986 | 0.3300 | 0.984 | **0.3262** |
| | 5 | 0.996 | 0.5150 | 0.996 | **0.5111** | 0.996 | 0.5150 |
| | 10 | 0.996 | 0.7281 | 0.996 | **0.7260** | 0.996 | 0.7280 |
| | 20 | 0.988 | 1.0307 | 0.988 | **1.0298** | 0.988 | 1.0307 |

## 5. Conclusions

In this research, we generated data from a Poisson distribution and estimated the confidence interval which is obtained confidence coefficient and average width to perform ML, MCMC, and Bayes Methods. The MCMC method is proposed to estimate population parameter of prior distribution function that can be used for estimating Bayes estimator. The posterior distribution function is related with the MCMC method to evaluate MCMC estimator. Through a simulation study, the ML method is a good performance when true parameter is small values ($\lambda = 0.5$) for all sample sizes. For the large sample sizes (n=100,300,500), MCMC method outperforms ML and Bayes when true parameter is 5. Moreover, the Bayes method is a good performance in most cases. However, the confidence coefficient and average width are equal in some case, so we would recommend user to use MCMC method because the Bayes method depends on the MCMC method.

## Acknowledgements

This Research was financially supported by the Faculty of Science of King Mongkut’s Institute of Technology Ladkrabang.

## References

[1] Rohatgi V.K., and Saleh Ehasnes., An Introduction to Probability and Statistics, Canada: A Wieley-Interscience Publication; 2001.

[2] Ntzoufras I, Bayesian Modeling Using WinBUGS, New Jersey: John Wiley & Sons; 2009.

[3] Christensen R, Johnson W, Branscum A, Hanson TE, Bayesian ideas and data analysis : an introduction for scientists and statisticians. Florida: Taylor & Francis Group; 2011.

[4] R Development Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria; 2004.

[5] Gemam S, Geman D, Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration of Images. IEEE Transactions on Pattern analysis and Machine Intelligence. 1984; 6: 721-741.

[6] Gelfand A, Hills S, Racine-Poon A., Smith A, Illustration of Bayesian inference in normal data models using Gibbs sampling. Journal of the American Statistical Association, 1990; 85: 972-985.

[7] Araveeporn A. Parameter Estimation of Poisson Distribution by Using Maximum Likelihood, Markov Chain Monte Carlo, and Bayes method. Thammasat International Journal of Science and Technology. 2014; 19(3): 1-14.

[8] Carlin B P, and Louis T A, Bayesian Methods for Data analysis. Florida: CRC Press Taylor & Francis Group; 2009.

# Maximum likelihood estimation of the negative binomial-Erlang distribution

Siriporn Samutwachirawong[1]*, Winai Bodhisuwan[2] and Prasit Payakkapong [3]

[1]*Department of Statistics, Kasetsart University, Chatuchak, Bangkok, 10900, Thailand, g5517400141@ku.ac.th*

[2]*Department of Statistics, Kasetsart University, Chatuchak, Bangkok, 10900, Thailand, fsciwnb@ku.ac.th*

[3]*Department of Statistics, Kasetsart University, Chatuchak, Bangkok, 10900, Thailand, fscipsp@ku.ac.th*

**Abstract**

A Poisson distribution and a negative binomial distribution are the most important topics in model for overdispersed count data. The objective of this paper is to provide a new negative binomial mixed distribution for count data. We propose a negative binomial-Erlang (NB-EL) distribution which is obtained by mixing the negative binomial distribution with an Erlang distribution. This distribution can be used to describe count data distribution with a large number of zeros. This distribution contains the negative binomial distribution-exponential distribution as a special case. In addition, we present some properties of the new distribution such as the factorial moments, mean, variance, skewness and kurtosis. The simulation study for comparison of parameter estimation is also implemented using the maximum likelihood method. A simulation study of the negative binomial-Erlang distribution is illustrated that this distribution could give the smaller value of the root-mean-square-error (RMSE) when the sample size are large.

*Keywords:* mixed distribution, count data, negative-binomial distribution, negative binomial-Erlang distribution, overdispersed.

*Siriporn Samutwachirawong.
E-mail address: g5517400141@ku.ac.th

## 1 Introduction

The paper is introduced a new negative binomial mixed distribution and more flexible alternative to the Poisson distribution. A negative binomial-Erlang (NB-EL) distribution which is a mixed negative Binomial (NB) distribution obtained by mixing the distribution of NB$(r, p)$ where, $p = exp(-\lambda)$ with distribution of a Erlang distribution $(k, c)$. This distribution can be used to describe count data distribution with the overdispersion.

The Poisson distribution provides a standard framework for the analysis of count data. In practice, in recent years there have been considerable study models for counts data that allow for overdispersion. The NB distribution is employed as a function from that which relaxes the overdispersion restriction of the Poisson distribution.

The NB distribution can be obtained as a mixture of Poisson and Gamma distributions [1--3] and this distribution has been traditionally used in many fields, such as medicine, public health, epidemiology, psychology, and agriculture, among others. The NB distribution has become increasingly popular as a more flexible alternative to Poisson distribution. In some cases, it is proven that mixed distributions, in particular, mixed NB model provided a better fit to count data compared to other distributions. For examples, of mixed NB distributions are the NB-Pareto [4,5], the NB-Beta Exponential [6], the NB-Generalized Exponential [7] and the NB-Crack distribution [8].

The NB-exponential distribution, which has two-parameter [9], is a special case of the NB-EL distribution. Moreover, the aim of this work is to describe our proposed NB-EL distribution and show that it includes many well-known distributions which are the closed form and the factorial moment, e.g., mean, variance, skewness, and kurtosis. The simulation study is applied in order to compare, for some situations, numerical results under estimating parameters of the NB-EL distribution by maximum likelihood (MLE) method.

## 2 Research Methodology

In this research, we have derived estimated parameters of NB-EL distribution by MLE method and simulation study is applied in order to compare, for some situations, numerical results under estimating parameters of the NB-EL distribution by MLE method, R Program is used for simulation study.

# 3 Research Results and Discussion

### *3.1 The Proposed distribution*

In this section, the closed form and the factorial moment of the NB-EL distribution are given. A classical NB distribution is denoted with probability mass function (pmf)

$$f(x;r,p) = \binom{r+x-1}{x} p^r(1-p)^x;\ x = 0,1,2,\ldots. \tag{1}$$

where, $r > 0$ and $0 < p < 1$. Note that $p$ is a probability of success, the experiment is repeated as many times as required to obtain $r$ successes.

The moments about zero and the factorial moment of the order of the NB distribution are given respectively by Gomez [10]

$$\begin{aligned} E(X) &= \frac{r(1-p)}{p}, \\ E(X^2) &= \frac{r(1-p)[1+r(1-p)]}{p^2} \quad \text{and} \\ \mu_{[m]}(X) &= E(X(X-1)\ldots(X-m+1)) \\ &= \frac{\Gamma(r+m)]}{\Gamma(r)} - \frac{(1-p)^m}{p^m}, m = 1,2,\ldots \end{aligned} \tag{2}$$

The Erlang distribution function which is specified by the probability density function (pdf)

$$g(x) = \frac{c^k x^{k-1} e^{-cx}}{(k-1)!}, \quad x > 0,\ k,c > 0, \tag{3}$$

was introduced by Erlang [11]. The moment generating function of the Erlang distribution is given by

$$M_X(t) = \left(1 - \frac{t}{c}\right)^{-k} \quad \text{for} \quad t > 0. \tag{4}$$

**Definition 1:** Let $X$ be a random variable of the NB-EL $(r,k,c)$ distribution. The NB distribution has parameters $r > 0$ and $p = \exp(-\lambda)$, where $\lambda$ is distributed as the Erlang distribution with positive parameters $k$ and $c$, i.e.,

$X|\lambda \sim \text{NB}(r, p = \exp(-\lambda))$ and $\lambda \sim \text{EL}(k,c)$.

**theorem 1:** Let $X \sim$ NB-EL$(r,k,c)$. The pmf of $X$ is given by

$$\begin{aligned} f(x;r,k,c) &= \binom{r+x-1}{x} \sum_{j=0}^{x} \binom{x}{j} (-1)^j \\ &\times \left(\frac{c}{c+(r+j)}\right)^k, \end{aligned} \tag{5}$$

for $x > 0$ and $k,c > 0$.

**Proof**. If $X|\lambda \sim \text{NB}(r, p = \exp(-\lambda))$ in Eq.(1) and $\lambda \sim$ Erlang$(k,c)$ in Eq.(3), then the pmf of $X$ can be obtained by

$$h(x) = \int_0^\infty f(x|\lambda) g(\lambda;k,c) \mathrm{d}\lambda \tag{6}$$

where, $f(x|\lambda)$ is defined by

$$\begin{aligned} f(x|\lambda) &= \binom{r+x-1}{x} e^{-\lambda r}(1-e^{-\lambda})^x, \\ &= \binom{r+x-1}{x} \sum_{j=0}^{x} \binom{x}{j} (-1)^j e^{-\lambda(r+j)} \end{aligned} \tag{7}$$

By substituting Eq.(7) into Eq.(6), we obtain

$$\begin{aligned} h(x) &= \binom{r+x-1}{x} \sum_{j=0}^{x} \binom{x}{j} (-1)^j \\ &\times \left(\int_0^\infty e^{-\lambda(r+j)} f_2(\lambda;k,c) \mathrm{d}\lambda\right), \\ &= \binom{r+x-1}{x} \sum_{j=0}^{x} \binom{x}{j} (-1)^j \\ &\times \left(M_\lambda(-(r+j))\right). \end{aligned} \tag{8}$$

Substituting the moment generating function of Erlang distribution in Eq.(4) into Eq.(8), the pmf of NB-EL $(r,k,c)$ is finally given as

$$f(x;r,k,c) = \binom{r+x-1}{x} \sum_{j=0}^{x} \binom{x}{j} (-1)^j \left(\frac{c}{c+(r+j)}\right)^k.$$

**Theorem 2:** If $X \sim$ NB-EL$(r,k,c)$, then the factorial moment of order $m$ of $X$ is given by

$$\mu_{[m]}(X) = \frac{\Gamma(r+m)}{\Gamma(r)} \sum_{j=0}^{m} \binom{m}{j} (-1)^j \left(\frac{c}{c-(m-j)}\right)^k,$$

for $x = 0,1,2,\ldots$ and $r,k,c > 0$.

**Proof**. The factorial moment of order $m$ of the NB distribution can be expressed in the terms of elementary function Eq.(2) by

$$\begin{aligned} \mu_{[m]}(X) &= E_\lambda\left(\frac{\Gamma(r+m)}{\Gamma(r)} \frac{(1-e^{-\lambda})^m}{e^{-\lambda m}}\right), \\ &= \frac{\Gamma(r+m)}{\Gamma(r)} E_\lambda(e^\lambda - 1)^m. \end{aligned}$$

A binomial expansion of $(e^\lambda - 1)^m$ shows that $\mu_{[m]}(X)$ can be written as

$$\begin{aligned} \mu_{[m]}(X) &= \frac{\Gamma(r+m)}{\Gamma(r)} \sum_{j=0}^{m} \binom{m}{j} (-1)^j E_\lambda(e^{\lambda(m-j)}), \\ &= \frac{\Gamma(r+m)}{\Gamma(r)} \sum_{j=0}^{m} \binom{m}{j} (-1)^j M_\lambda(m-j). \end{aligned}$$

From the moment generating function of the Erlang random variable in Eq. (4) with $t = m - j$, we have finally that $\mu_{[m]}(X)$ can be written as

$$\mu_{[m]}(X) \;=\; \frac{\Gamma(r+m)}{\Gamma(r)} \sum_{j=0}^{m} \binom{m}{j} (-1)^j \left(\frac{c}{c-(m-j)}\right)^k .$$

From the factorial moments of the NB-EL distribution, it is straightforward to derive the first four moments by

$$E(X) \;=\; r\left(\varphi_1 - 1\right), \qquad (9)$$

$$E(X^2) \;=\; (r^2+r)\varphi_2 - (2r^2+r)\varphi_1 + r^2, \qquad (10)$$

$$\begin{aligned} E(X^3) \;=\; & (r^3+3r^2+2r)\varphi_3 \\ & -(3r^3+6r^2+3r)\varphi_2 \\ & +(3r^3+3r^2+r)\varphi_1 - r^3, \qquad (11) \end{aligned}$$

$$\begin{aligned} E(X^4) \;=\; & (r^4+6r^3+11r^2+6r)\varphi_4 \\ & -(4r^4+18r^3+26r^2+12r)\varphi_3 \\ & +(6r^4+18r^3+19r^2+7r)\varphi_2 \\ & -(4r^4+6r^3+4r^2+r)\varphi_1 + r^4, \qquad (12) \end{aligned}$$

$$\begin{aligned} Var(X) \;&=\; E(X^2) - \left(E(X)\right)^2 \\ &=\; (r^2+r)\varphi_2 - r\varphi_1\left(1 + r\varphi_1\right), \qquad (13) \end{aligned}$$

$$\begin{aligned} \text{Skewness}(X) \;&=\; \big[E(X^3) - 3E(X^2)E(X) \\ &\quad +2[E(X)]^3/\big]\,\sigma^3 \\ &=\; \big[(r^3+3r^2+2r)\varphi_3 - (3r^2+3r)\varphi_2 \\ &\quad +r\varphi_1 + 3r^2\left(\varphi_2\right)^2 - (3r^3+3r^2)\varphi_1\varphi_2 \\ &\quad +2r^3\left(\varphi_2\right)^3\big]/\sigma^3, \qquad (14) \end{aligned}$$

$$\begin{aligned} \text{Kurtosis}(X) \;&=\; \big[E(X^4) - 4E(X^3)E(X) \\ &\quad +6E(X^2)\left(E(X)\right)^2 - 3[E(X)]^4/\big]\,\sigma^4 \\ &=\; \big[(r^4+6r^3+11r^2+6r)\varphi_4 \\ &\quad -(6r^3+18r^2+12r)\varphi_3 - 3r^4\left(\varphi_1\right)^4 \\ &\quad +(7r^2+7r)\varphi_2 - 4r\varphi_1 - 6r^3\left(\varphi_1\right)^3 \\ &\quad +(12r^3+12r^2)\varphi_1\varphi_2 \\ &\quad -(4r^4+12r^3+8r^2)\varphi_1\varphi_3 \\ &\quad +(6r^4+6r^3)\left(\varphi_1\right)^2\varphi_2\big]/\sigma^4, \qquad (15) \end{aligned}$$

where

$$\varphi_i \;=\; \left(\frac{c-i}{c}\right)^{-k}.$$

**Corollary 1:** If $k = 1$ then the NB-EL distribution reduces to the NB-exponential distribution with pmf given by

$$\begin{aligned} f(x_1) \;=\; & \binom{r+x_1-1}{x_1} \sum_{j=0}^{x_1} \binom{x_1}{j} (-1)^j \\ & \times \left(\frac{c}{c+(r+j)}\right), \qquad (16) \end{aligned}$$

for $x_1 = 0, 1, 2, \ldots$ for $r$ and $c > 0$.

**Proof.** If $X|\lambda \sim \text{NB}(r, p = (\exp - \lambda))$ in Eq. (1) and $\lambda \sim \text{Erlang}(k = 1)$, then the pmf of $X$ is

$$\begin{aligned} f(x_1; r, c) \;=\; & \binom{r+x_1-1}{x_1} \sum_{j=0}^{x_1} \binom{x_1}{j} (-1)^j \\ & \times \left(\frac{c}{c+(r+j)}\right). \end{aligned}$$

Throughout this research, we will use the notation the NB-EL $(r, k, c)$ as a reference for the negative binomial-Erlang distribution. Figs 1-3 show that the pmf of the NB-EL $(r, k, c)$has the highest mass at zero for several values of $r, k$ and $c$.

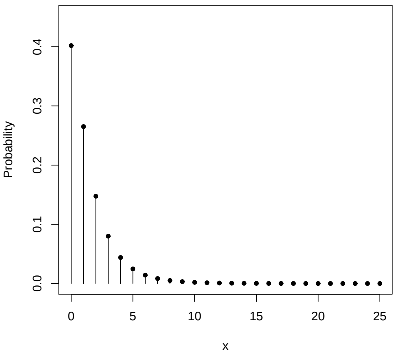


**Figure 1:** Plot of pmf for NB-EL $(r = 2, k = 5, c = 10)$

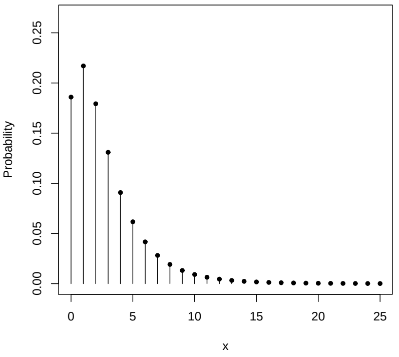


**Figure 2:** Plot of pmf for NB-EL $(r = 4, k = 5, c = 10)$

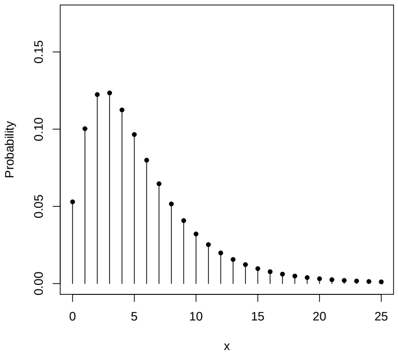


**Figure 3:** Plot of pmf for NB-EL $(r = 8, k = 5, c = 10)$

*3.2 Random variate generation of the NB-EL distribution*

To generate a random variable $X$ from the NB-EL $(r, k, c)$, one can use the following algorithm:
1) Generate $U$ from the uniform distribution, $U(0, 1)$.
2) Set $\lambda = -\frac{1}{\beta}\log(1 - U^{1/\alpha})$ from the Erlang distribution
3) Generate $X$ from the $NB(r, p = exp(-\lambda))$ distribution.

*3.3 Parameters estimation*

In this section, the estimation of parameters for the NB-EL $(r, k, c)$ distribution via the MLE method is provided.

The likelihood function of the NB-EL distribution is

$$L(r,k,c) = \prod_{i=1}^{n} \binom{r + x_i - 1}{x_i} \sum_{j=0}^{x_i} \times \binom{x_i}{j} (-1)^j \left( \frac{c}{c + (r+j)} \right)^k . \quad (17)$$

The log likelihood function of the NB-EL is given by

$$\mathcal{L} = \log L(r,k,c) = \sum_{i=1}^{n} \log \binom{r + x_i - 1}{x_i} + \sum_{i=1}^{n} \left( \log \sum_{j=0}^{x_i} \binom{x_i}{j} (-1)^j \left( \frac{c}{c + (r+j)} \right)^k \right) . \quad (18)$$

The first order conditions for finding the optimal values of the parameters obtained by differentiating Eq.(18) with respect to $r, k$ and $c$ give rise to the following differential equations

$$\frac{\partial \mathcal{L}}{\partial r} = \sum_{i=1}^{n} \sum_{k=0}^{x_i - 1} \log(r + k) + \sum_{i=1}^{n} \left\{ \frac{-k \sum_{j=0}^{x_i} \binom{x_i}{j} (-1)^j \frac{c^k}{(c+(r+j))^{k+1}}}{\sum_{j=0}^{x_i} \binom{x_i}{j} (-1)^j \left[ \frac{c}{c+(r+j)} \right]^k} \right\} , \quad (19)$$

$$\frac{\partial \mathcal{L}}{\partial k} = \sum_{i=1}^{n} \left\{ \frac{\sum_{j=0}^{x_i} \binom{x_i}{j} (-1)^j \left[ \frac{c}{c+(r+j)} \right]^k \log \left[ \frac{c}{c+(r+j)} \right]}{\sum_{j=0}^{x_i} \binom{x_i}{j} (-1)^j \left[ \frac{c}{c+(r+j)} \right]^k} \right\} , \quad (20)$$

and

$$\frac{\partial \mathcal{L}}{\partial c} = \sum_{i=1}^{n} \left\{ \frac{\sum_{j=0}^{x_i} \binom{x_i}{j} (-1)^j \left[ \frac{kc^{k-1}(r+j)}{(c+r+j)^{k+1}} \right]}{\sum_{j=0}^{x_i} \binom{x_i}{j} (-1)^j \left[ \frac{c}{c+(r+j)} \right]^k} \right\} (21)$$

The Eqs.(19)-(21) can be obtained by solving the resulting equations simultaneously using a numerical procedure such as the Newton-Raphson method.

*3.4 Generating simulated the NB-EL data sets*

We present three types of simulated data sets were examined. Our objective here is to compare the true values of the parameters of the NB-EL distribution and their estimates using MLE method. The true parameter values are $k = 5, c = 10, r = 2, 4, 8$. All cases of simulation study were then generated from the NB-EL distribution with sample of sizes=20, 50, 100, 200 and 500 respectively. We used R program to generate each sample size and repeated this for 500 trials. The bias is defined as usual as the difference between the estimated and true parameter values. Parameters, expected value, variance, root-mean-square-error (RMSE) are shown in Tables 1-3.

**Table 1:** Simulation study of NB-EL distribution (True parameters : $r = 2, k = 5, c = 10$)

| Sample Size | Parameter | Estimate | Var | RMSE |
|---|---|---|---|---|
| 20 | r | 13.542 | 2.906 | 2.299 |
| | c | 14.517 | 83.156 | 10.176 |
| | k | 5.193 | 29.243 | 5.411 |
| 50 | r | 3.145 | 2.551 | 1.965 |
| | c | 12.867 | 69.565 | 8.819 |
| | k | 5.052 | 23.851 | 4.884 |
| 100 | r | 2.970 | 2.245 | 1.785 |
| | c | 12.121 | 44.622 | 7.009 |
| | k | 5.031 | 15.684 | 3.960 |
| 200 | r | 2.816 | 2.238 | 1.704 |
| | c | 11.552 | 32.456 | 5.905 |
| | k | 5.008 | 13.141 | 3.625 |
| 500 | r | 2.439 | 0.535 | 0.853 |
| | c | 10.502 | 10.095 | 3.217 |
| | k | 5.007 | 7.178 | 2.679 |

**Table 2:** Simulation study of NB-EL distribution (True parameters : $r = 4, k = 5, c = 10$)

| Sample Size | Parameter | Estimate | Var | RMSE |
|---|---|---|---|---|
| 20 | r | 8.361 | 90.244 | 10.453 |
| | c | 15.982 | 115.631 | 12.305 |
| | k | 6.074 | 31.616 | 5.724 |
| 50 | r | 7.804 | 77.874 | 9.609 |
| | c | 15.618 | 88.531 | 10.959 |
| | k | 6.060 | 23.467 | 4.959 |
| 100 | r | 6.947 | 45.762 | 7.379 |
| | c | 14.268 | 61.751 | 8.942 |
| | k | 6.022 | 23.432 | 4.947 |
| 200 | r | 6.328 | 35.559 | 6.4024 |
| | c | 13.267 | 32.505 | 6.571 |
| | k | 5.996 | 15.688 | 4.084 |
| 500 | r | 5.007 | 5.776 | 2.606 |
| | c | 11.686 | 8.879 | 3.424 |
| | k | 5.903 | 11.094 | 3.451 |

**Table 3:** Simulation study of NB-EL distribution (True parameters : $r = 8, k = 5, c = 10$)

| Sample Size | Parameter | Estimate | Var | RMSE |
|---|---|---|---|---|
| 20 | r | 14.204 | 79.225 | 10.850 |
| | c | 15.623 | 89.365 | 10.999 |
| | k | 5.336 | 8.944 | 3.009 |
| 50 | r | 11.545 | 51.754 | 8.020 |
| | c | 13.768 | 53.037 | 8.199 |
| | k | 5.322 | 3.705 | 1.951 |
| 100 | r | 9.635 | 5.040 | 2.777 |
| | c | 12.263 | 10.897 | 4.002 |
| | k | 5.314 | 2.590 | 1.640 |
| 200 | r | 9.097 | 2.617 | 1.954 |
| | c | 11.704 | 7.467 | 3.220 |
| | k | 5.230 | 1.171 | 1.106 |
| 500 | r | 8.694 | 0.505 | 0.993 |
| | c | 11.364 | 0.970 | 1.682 |
| | k | 5.192 | 0.413 | 0.671 |

### *3.5 Discussion*

We present a simulation study illustrating some of the theory covered. Our objective here is to compare the true values of the parameters of the NB-BE distribution and their estimates using MLE method. All cases of simulation study were then generated from this NB-EL distribution with sample of sizes n = 20, 50, 100, 200 and 500 respectively. We used R program to generate each sample of a fixed size and repeated this for 500 trials. A simulation study of the NB-EL distribution is illustrated that this distribution could give the smaller value of the RMSE when the sample size are large.

## 4 Conclusion

As mentioned earlier, we introduce a new three parameter negative binomial-Erlang distribution, NB-EL$(r, k, c)$. This distribution obtained by mixing the NB with the Erlang distribution (when the NB distribution have parameters $r > 0$ and $p = \exp(-\lambda)$, where $\lambda$ is distributed as the Erlang distribution with positive parameters $k$ and $c$). In addition, the moments of the NB-EL distribution which include the factorial moments, mean, variance, skewness and kurtosis are derived. Moreover, the parameter estimation of the NB-EL using MLE are developed. We hope that NB-EL distribution may attract wider applications in count data analysis.

## Acknowledgements

We would like to thank Dr. Chookait Pudprommarat, who has given many useful comments and valuable suggestions. This work was supported by research funding from the Research Professional Development

Project under the Science Achievement Scholarship of Thailand (SAST).

## References

[1] Greenwood M, Yule GU. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. Journal of the Royal Statistical Society. 1920;83:255--279.

[2] Lemaire J. How to define a bonus-malus system with an exponential utility function. Astin Bulletin. 1979;10:274--282.

[3] Simon L. Fitting negative binomial distribution by the method of maximum likelihood. In: Proceedings of the Casualty Actuarial Society; 1961. p. 45--53.

[4] Klugman S. Loss models: From Data to Decisions. 3rd ed. Wiley - Inter Science; 2008.

[5] S Meng YW, Whitmore GA. HAccounting for individual overdispersion in a bonus-malus system. Astin Bulletin. 1999;.

[6] Pudprommarat C, Bodhisuwan W, Zeephongsekul P. A new mixed negative binomial distribution. Applied Sciences. 2012;12(17): 1853--1858.

[7] Aryuyuen S, Bodhisuwan W. The negative binomial-generalized exponential (NB-GE) distribution. Applied Mathematical Sciences. 2013;7(22):1093--1105.

[8] Saengthong P, Bodhisuwan W. negative binomial-Crack (NB-CR) distribution. International Journal of Pure and Applied Mathematics. 2013;84(3):213--230.

[9] Klugman S, Panjer H, Willmot G. Loss Models : From Data to Decisions. New York: Wiley; 1998.

[10] Gómez-Déniz E, Sarabia JM, Calderin-Ojeda E. Univariate and multivariate versions of the negative binomial-inverse Gaussian distributions with applications. Insurance: Mathematics and Economics. 2008;42(1):39--49.

[11] Erlang AK. Solution of some problems in the theory of probabilities of significance in automatic telephone exchange. Elektrotkeknikeren. 1917;13:138--155.

# A method for formulating fuzzy linear regression model and estimating the model parameters

Atchanut Rattanalertnusorn[1*], Sathit Phannarook[2] and Uraiwan Jaroengeratikun[3]

[1]*Department of Mathematics and Computer Science, Rajamangala University of Technology Thanyaburi, Pathumtani, 12110, Thailand, atchanut_r@rmutt.ac.th*

[2]*Department of Mathematics and Computer Science, Rajamangala University of Technology Thanyaburi, Pathumtani, 12110, Thailand, ph.sthit@rmutt.ac.th*

[3]*Department of Applied Statistics, King Mongkut's University of Technology North Bangkok, Bangkok, 10800, Thailand, urj@kmutnb.ac.th*

## Abstract

The objectives of this research are to formulate a fuzzy linear regression model using LR-fuzzy number and estimate the model parameters. A distance between two LR-fuzzy numbers is defined as an error of fuzzy model. Consequently, the total error is obtained as the model fuzziness. The methodology of estimating model parameters is presented, which is based on the least squares method. Furthermore, the coefficient of determination is measured as the goodness of fit of the fuzzy model. Finally, a numerical example for demonstration the proposed method is shown.

*Keywords*: fuzzy linear regression model, fuzzy numbers, LR-fuzzy numbers

*Corresponding Author
E-mail Address: atchanut_r@rmutt.ac.th

## 1. Introduction

The concept of fuzzy regression has been first presented by Tanaka, Uejima and Asai[14]. They developed the fuzzy linear regression model (FLRM) with the fuzzy linear function as:

$$\tilde{y}_i = \tilde{A}_0 + \tilde{A}_1 x_{i1} + \cdots + \tilde{A}_k x_{ik} \quad \text{for } i = 1, \ldots, n, \tag{1}$$

where $\tilde{y}_i$ is the *i*th fuzzy output, $\tilde{A}_j$ is the *j*th fuzzy parameters which are corresponding to $x_{ij}$ except $\tilde{A}_0$ is a fuzzy intercept and $x_{ij}$ is the *j*th input variable for the *i*th sample (observation). However, the model has some drawbacks. For example, when using more observations to establish the model leads to the fuzzier estimation of parameters, thus making the spread of the estimated fuzzy output wider as stated by [15-18]. In addition, Redden and Woodall[12] pointed out the Tanaka's method sensitive to outliers.

To prevent the above problems, Diamond[4] and Kao and Chyu[5-6] adopted numeric parameters to describe the fuzzy relationship between fuzzy output variable and fuzzy input variables. Diamond[4] also defined a new distance metric to measure error of two fuzzy numbers. Kao and Chyu[6] also proposed the two-stages method to formulate the FLRM. In the first stage, fuzzy output data or fuzzy input data or both, are defuzzified into crisp data. Then they used the least squares method to estimate the model parameters. In the second stage, they calculated the estimated error in the model which is based on the criterion of Kim and Bishu[7]. This criterion minimizes the difference of the membership values between the observed and estimated fuzzy output data.

In the other approaches, Buckley [1] proposed a new method to estimate fuzzy parameters by using a set of confidence intervals, associated with the $\alpha$ level of fuzzy numbers. The set of confidence intervals can estimate the triangular fuzzy number. He also employed fuzzy prediction and fuzzy hypothesis testing. Kim, Kim and Choi [8] proposed the fuzzy least absolute deviation method and constructed the FLRM with fuzzy input. They also provided some numerical examples and evaluated the effectiveness of proposed method compared with the fuzzy least squares method. Chen and Hsueh[2] proposed a mathematical programming to build FLRM which is based on the concept of distance. They defined the total error in the model as a sum of the distance between the observed and estimated fuzzy output data. Chen and Hsueh[3] also proposed the new fuzzy linear regression model which is added the adjustment term in the model. They adapted the least squares method to estimate the model parameters.

Recently, Rattanalertnusorn, Thongteeraparp and Bodhisuwan[11] presented the extension of Chen and Hsueh method. They extended fuzzy components in FLRM from triangular fuzzy numbers to trapezoidal fuzzy numbers. They also derived the proposed method. Unfortunately, the computational method of the proposed method is not convenient for the beginners. It is quite complex and not suitable for practice. Therefore, this paper has adjusted some steps out and added some steps that should to do with the fuzzy model. For example, we added an index of goodness of fit for FLRM. We also interpreted the estimated fuzzy output or fuzzy prediction. However, we cut off the adjustment term in FLRM because it is quite complex for computation. Thus, we proposed the simple method to build the FLRM and estimate the parameters.

## 2. **Research Methodology**

In this section, the research method can be divided into three parts according to the following:

### *2.1 Preliminaries*

Zadeh[21] stated that let $X$ be a universal set, then a fuzzy set $\tilde{A}$ of $X$ is defined by its membership function $\mu_{\tilde{A}} : X \to [0,1]$. In general, the universal set $X$ is assumed to be the set of real number; i.e., $X \in \Re$.

*Definition 1.* The fuzzy set $\tilde{A}$ of $\Re$ is called a fuzzy number if the following conditions are satisfied:

1) $\tilde{A}$ is a normal and convex fuzzy set.
2) Its membership function $\mu_{\tilde{A}}$ is upper semi-continuous.
3) The $\alpha$ level of $\tilde{A}$ is closed set for $\alpha \in [0,1]$.

Let $\tilde{A}$ be a fuzzy set of $\Re$. $\tilde{A}$ is called a normal fuzzy set if there exists some $x \in \tilde{A}$ such that $\mu_{\tilde{A}} = 1$. $\tilde{A}$ is called a convex fuzzy set if

$\mu_{\tilde{A}}(\lambda x + (1-\lambda)y) \geq \min\{\mu_{\tilde{A}}(x), \mu_{\tilde{A}}(y)\}$ for $\lambda \in [0,1]$.

The membership function $\mu_{\tilde{A}}$ is said to be upper semi-continuous if the $\alpha$ level of $\tilde{A}$, denoted by $A_{\alpha} = \{x : \mu_{\tilde{A}}(x) \geq \alpha\}$ is a closed set for $\alpha \in [0,1]$.

*Definition 2.* The fuzzy set $\tilde{A}$ of $\Re$ is called nonnegative fuzzy number if its membership function $\mu_{\tilde{A}}(x) = 0$ for all $x < 0$ [20].

*Definition 3.* The fuzzy number $\tilde{A}$ is called LR –fuzzy number if it has the membership function as follows:

$$\mu_{\tilde{A}}(x) = \begin{cases} L(x), & if\ x \in [a,b) \\ 1, & if\ x \in [b,c] \\ R(x), & if\ x \in (c,d] \\ 0, & otherwise \end{cases}$$

for all $x \in \Re$ and $a < b < c < d$ ,
where $L, R : [0,1] \to [0,1]$ are two shape functions such that $R[0] = L[0] = 0$ and $R[1] = L[1] = 1$. The support and the core of LR-fuzzy number Ã are closed intervals; i.e., $[a,d]$ and $[b,c]$, respectively.

If $L[x] = \dfrac{x-b}{b-a}$ , $R[x] = \dfrac{d-x}{d-c}$ and $b = c$, then LR-fuzzy number Ã is called triangular fuzzy number (TFN). In general, TFN $\tilde{A}$ can be denoted by $\tilde{A} = (a,b,d)_T$ with the membership function as:

$$\mu_{\tilde{A}}(x) = \begin{cases} \dfrac{x-b}{b-a}, & if\ x \in [a,b) \\ \dfrac{d-x}{d-b}, & if\ x \in [b,d] \\ 0, & otherwise \end{cases} .$$

*Definition 4.* Let $\tilde{A}$ and $\tilde{B}$ be triangular fuzzy numbers, denoted by $\tilde{A} = (a_1, b_1, d_1)_T$ and $\tilde{B} = (a_2, b_2, d_2)_T$.
Then the following fuzzy operations are satisfied:

1) Addition

$\tilde{A} \oplus \tilde{B} = (a_1 + a_2, b_1 + b_2, d_1 + d_2)_T$.

2) Multiplication by *k*, *k* is a real number

$k \otimes \tilde{A} = (ka_1, kb_1, kd_1)_T$ if $k > 0$ or

$k \otimes \tilde{A} = (kd_1, kb_1, ka_1)_T$ if $k < 0$ or

$k \otimes \tilde{A} = (0,0,0)_T$ if $k = 0$.

Noted that: $(0,0,0)_T$ is crisp number 0.

*Definition 5.* Let $\tilde{A}$ be TFN, denoted by $\tilde{A} = (a,b,d)_T$.
If *a*=*b*=*d*, then TFN $\tilde{A}$ is called crisp number *A*=*a* .

### *2.2 Fuzzy linear regression models*

In this paper, the fuzzy linear regression model can be defined as:

$$\tilde{y}_i = \beta_0 + \beta_1 \tilde{x}_{i1} + \cdots + \beta_k \tilde{x}_{ik} \quad \text{for } i = 1, \ldots, n, \qquad (2)$$

where $\tilde{y}_i$ is fuzzy output data for the *i*th sample, $\{\tilde{x}_{i1}, ..., \tilde{x}_{ik}\}$ is a set of *k*-fuzzy input data for the *i*th sample and $\{\beta_0, \beta_1, ..., \beta_k\}$ is a set of *k*-regression coefficients except $\beta_0$ is an intercept.

### *2.3 The proposed method to formulate the FLRM*

In this part, the method to formulate the FLRM in Eq.(2) can be presented according to the following steps:

*First step,* we defined LR-fuzzy numbers in the fuzzy model.

Let $\tilde{y}_i$ and $\tilde{x}_{ij}$ are TFNs, which are denoted by

$\tilde{y}_i = (y_{il}, y_{im}, y_{ir})_T$ and $\tilde{x}_{ij} = (x_{ijl}, x_{ijm}, x_{ijr})_T$ ,
$j = 1, \ldots, k$ .

Also the membership function defined as:

$$\mu_{\tilde{y}_i}(y) = \begin{cases} \dfrac{y - y_{im}}{y_{im} - y_{il}}, & y_{il} \leq y < y_{im} \\ \dfrac{y_{ir} - y}{y_{ir} - y_{im}}, & y_{im} \leq y \leq y_{ir} \\ 0, & otherwise \end{cases}$$

and

$$\mu_{\tilde{x}_{ij}}(x) = \begin{cases} \dfrac{x - x_{ijm}}{x_{ijm} - x_{ijl}}, & x_{ijl} \le x < x_{ijm} \\ \dfrac{x_{ijr} - x}{x_{ijr} - x_{ijm}}, & x_{ijm} \le x \le x_{ijr} \\ 0, & otherwise \end{cases} .$$

*Second step,* we transformed fuzzy components into crisp number by using the centroid formulae [19] as the following:

$y_{ic} = \dfrac{y_{il} + y_{im} + y_{ir}}{3}$ and $x_{ijc} = \dfrac{x_{ijl} + x_{ijm} + x_{ijr}}{3}$ if fuzzy components are TFNs. That is, we obtained the conventional linear regression model as

$$y_{ic} = \beta_0 + \beta_1 x_{i1c} + \cdots + \beta_k x_{ikc} + \varepsilon_i , \tag{3}$$

where $\varepsilon_i$ is a random error, which is assumed to be independently and identically distributed (i.i.d.) and has mean $E(\varepsilon_i) = 0$ and variance $V(\varepsilon_i) = \sigma^2$ . It is well-known the least squares estimators are

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y} , \tag{4}$$

where

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11c} & \cdots & x_{1kc} \\ 1 & x_{21c} & \cdots & x_{2kc} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1c} & \cdots & x_{nkc} \end{bmatrix}_{n \times p} , \ \mathbf{y} = \begin{bmatrix} y_{1c} \\ y_{2c} \\ \vdots \\ y_{nc} \end{bmatrix}_{n \times 1} ,$$

$\mathbf{X}^T$ is a transpose matrix $\mathbf{X}$ and $p = k + 1$ .

*Third step,* we investigated the regression coefficient what each $\beta_j$ means in the regression model using hypothesis testing method. Then the t-test can be used to test the following hypothesis:

$\mathrm{H}_0 : \beta_j = 0$ versus $\mathrm{H}_1 : \beta_j \ne 0$ , for $j = 0,1,...,k.$

t- test statistic:

$$t_s = \frac{\hat{\beta}_j}{s(\hat{\beta}_j)} , \ df = n - p$$

and the acceptance region is $-t_{1-\alpha/2, n-p} \le t_s \le t_{1-\alpha/2, n-p}$ . If the value of $t_s$ falls in the acceptance region, then we conclude $\mathrm{H}_0$ , otherwise we conclude $\mathrm{H}_1$ at the level of significance $\alpha$ . Further details see in [9].

*Fourth step,* we formulated FLRM in Eq.(2) by using fuzzy output data and fuzzy input data are TFNs. Thus, the Eq.(2) can be rewritten by

$$(y_{il}, y_{im}, y_{ir})_T = \beta_0 + \beta_1 (x_{i1l}, x_{i1m}, x_{i1r})_T + \cdots + \beta_k (x_{ikl}, x_{ikm}, x_{ikr})_T$$

or

$$\tilde{y}_i = (\beta_0 + \sum_{j=1}^{k} \beta_j x_{ijl}, \ \beta_0 + \sum_{j=1}^{k} \beta_j x_{ijm}, \ \beta_0 + \sum_{j=1}^{k} \beta_j x_{ijr}) .$$

That is, the FLRM is built completely. The details of the estimation parameter method are in the next section.

## 3. The estimation parameter method

In this paper, we present the estimated parameters in FLRM which is based on the least squares method as the following:

Let an error for the *i*th sample, defined by using Diamond's distance of two TFNs[4] as:

$$d(\tilde{y}_i, \hat{\tilde{y}}_i) = (y_{il} - \hat{y}_{il})^2 + (y_{im} - \hat{y}_{im})^2 + (y_{ir} - \hat{y}_{ir})^2$$

Thus, total error for *n* samples is given by

$$SS(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} d(\tilde{y}_i, \hat{\tilde{y}}_i) = \sum_{i=1}^{n} \left( (y_{il} - \hat{y}_{il})^2 + (y_{im} - \hat{y}_{im})^2 + (y_{ir} - \hat{y}_{ir})^2 \right)$$

In this paper, the total error can be defined as sum squares error in the fuzzy model. According to the least squares method, the estimator can be obtained by minimizing sum squares error, i.e., minimized $SS(\hat{\boldsymbol{\beta}})$ . In general, $SS(\hat{\boldsymbol{\beta}})$ can be rewritten in the quadratic form as follows:

$$SS(\hat{\boldsymbol{\beta}}) = [\mathbf{y}_l - \hat{\mathbf{y}}_l]^{\mathbf{T}} [\mathbf{y}_l - \hat{\mathbf{y}}_l] + [\mathbf{y}_m - \hat{\mathbf{y}}_m]^{\mathbf{T}} [\mathbf{y}_m - \hat{\mathbf{y}}_m] + [\mathbf{y}_r - \hat{\mathbf{y}}_r]^{\mathbf{T}} [\mathbf{y}_r - \hat{\mathbf{y}}_r] , \tag{5}$$

where

$$\mathbf{y}_l = \begin{bmatrix} y_{1l} \\ y_{2l} \\ \vdots \\ y_{nl} \end{bmatrix}_{n \times 1} , \ \mathbf{y}_m = \begin{bmatrix} y_{1m} \\ y_{2m} \\ \vdots \\ y_{nm} \end{bmatrix}_{n \times 1} , \ \mathbf{y}_r = \begin{bmatrix} y_{1r} \\ y_{2r} \\ \vdots \\ y_{nr} \end{bmatrix}_{n \times 1} ,$$

$$\mathbf{x}_l = \begin{bmatrix} 1 & x_{11l} & \cdots & x_{1kl} \\ 1 & x_{21l} & \cdots & x_{2kl} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1l} & \cdots & x_{nkl} \end{bmatrix}_{n \times p} , \ \mathbf{X}_m = \begin{bmatrix} 1 & x_{11m} & \cdots & x_{1km} \\ 1 & x_{21m} & \cdots & x_{2km} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1m} & \cdots & x_{nkm} \end{bmatrix}_{n \times p} ,$$

$$\mathbf{X}_r = \begin{bmatrix} 1 & x_{11r} & \cdots & x_{1kr} \\ 1 & x_{21r} & \cdots & x_{2kr} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1r} & \cdots & x_{nkr} \end{bmatrix}_{n \times p} , \ \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}_{p \times 1} \text{ and}$$

$p = k + 1$. Since $\hat{\mathbf{y}}_l = \mathbf{X}_l \hat{\boldsymbol{\beta}}$, $\hat{\mathbf{y}}_m = \mathbf{X}_m \hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{y}}_r = \mathbf{X}_r \hat{\boldsymbol{\beta}}$, the Eq.(5) can be rewritten as

$$SS(\hat{\boldsymbol{\beta}}) = [\mathbf{y}_l - \mathbf{X}_l \hat{\boldsymbol{\beta}}]^{\mathbf{T}} [\mathbf{y}_l - \mathbf{X}_l \hat{\boldsymbol{\beta}}] + [\mathbf{y}_m - \mathbf{X}_m \hat{\boldsymbol{\beta}}]^{\mathbf{T}} [\mathbf{y}_m - \mathbf{X}_m \hat{\boldsymbol{\beta}}] + [\mathbf{y}_r - \mathbf{X}_r \hat{\boldsymbol{\beta}}]^{\mathbf{T}} [\mathbf{y}_r - \mathbf{X}_r \hat{\boldsymbol{\beta}}] .$$

By applying the partial derivative with respect to $\hat{\boldsymbol{\beta}}$ , we obtained the fuzzy estimators vector,

$$\hat{\boldsymbol{\beta}}_F = [\mathbf{X}_l^T \mathbf{X}_l + \mathbf{X}_m^T \mathbf{X}_m + \mathbf{X}_m^T \mathbf{X}_m]^{-1} [\mathbf{X}_l^T \mathbf{y}_l + \mathbf{X}_m^T \mathbf{y}_m + \mathbf{X}_r^T \mathbf{y}_r]$$

or

$$\hat{\boldsymbol{\beta}}_F = \begin{bmatrix} \hat{\beta}_{F_0} \\ \hat{\beta}_{F_1} \\ \vdots \\ \hat{\beta}_{F_k} \end{bmatrix}_{p\times 1} \qquad (6)$$

where, $\hat{\boldsymbol{\beta}}_F$ is called a vector of fuzzy least squares estimator with dimension $p\times 1$. Further details see in [10].

## 4. Goodness of fit in fuzzy model

For the evaluation of the fuzzy model, an index of goodness of fit will be presented by:

The coefficient of determination ( $R^2$ ) is denoted as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left((y_{il}-\hat{y}_{il})^2 + (y_{im}-\hat{y}_{im})^2 + (y_{ir}-\hat{y}_{ir})^2\right)}{\sum_{i=1}^{n}\left((y_{il}-\overline{y}_{l})^2 + (y_{im}-\overline{y}_{m})^2 + (y_{ir}-\overline{y}_{r})^2\right)}$$

where $\overline{y}_l = \frac{\sum_{i=1}^{n} y_{il}}{n}$, $\overline{y}_m = \frac{\sum_{i=1}^{n} y_{im}}{n}$ and $\overline{y}_r = \frac{\sum_{i=1}^{n} y_{ir}}{n}$ .

Generally, $0 \le R^2 \le 1$, and if $R^2$ is a high value, the model is the best fit.

## 5. Fuzzy prediction and interpretation

By substituting $\hat{\boldsymbol{\beta}}_F$ to Eq.(2), the estimated fuzzy output can be expressed as

$$\hat{\tilde{y}}_i = (\hat{\beta}_{F0} + \sum_{j=1}^{k} \hat{\beta}_{Fj} x_{ijl},\ \hat{\beta}_{F0} + \sum_{j=1}^{k} \hat{\beta}_{Fj} x_{ijm},\ \hat{\beta}_{F0} + \sum_{j=1}^{k} \hat{\beta}_{Fj} x_{ijr})$$

where $i = 1,\ldots,n$, $\hat{\tilde{y}}_i$ is the estimated fuzzy output for the $i$th sample, it can be denoted by $\hat{\tilde{y}}_i = (\hat{y}_{il}, \hat{y}_{im}, \hat{y}_{ir})_T$ and its membership function defined as

$$\mu_{\hat{\tilde{y}}_i}(\hat{y}) = \begin{cases} \dfrac{\hat{y}-\hat{y}_{im}}{\hat{y}_{im}-\hat{y}_{il}}, & \hat{y}_{il} \le \hat{y} < \hat{y}_{im} \\ \dfrac{\hat{y}_{ir}-\hat{y}}{\hat{y}_{ir}-\hat{y}_{im}}, & \hat{y}_{im} \le \hat{y} \le \hat{y}_{ir} \\ 0, & otherwise \end{cases} .$$

Since $\hat{\tilde{y}}_i$ is fuzzy number (TFN), the estimated fuzzy output (or fuzzy prediction) can be interpreted as the following approaches:

Approach 1: $\hat{\tilde{y}}_i$ means approximately $\hat{y}_{im}$.

Approach 2: $\hat{\tilde{y}}_i$ means a closed interval [ $\hat{y}_{il}$ , $\hat{y}_{ir}$ ].

## 6. Numerical example

In this section, a numerical example is presented to demonstrate the proposed method.

*Example*

The fuzzy data come from Example 5.1 in the Wu's paper[20]. Suppose that we have the following triangular fuzzy data as shown in Table 1.

Table 1: Fuzzy input and fuzzy output.

| $i$ | $\tilde{y}_i$ | $\tilde{x}_{i1}$ | $\tilde{x}_{i2}$ |
|---|---|---|---|
| 1 | (111, 162, 194) | (151, 274, 322) | (1432, 2450, 3461) |
| 2 | (88, 120, 161) | (101, 180, 291) | (2448, 3254, 4463) |
| 3 | (161, 223, 288) | (221, 375, 539) | (2592, 3802, 5116) |
| 4 | (83, 131, 194) | (128, 205, 313) | (1414, 2838, 3252) |
| 5 | (51, 67, 83) | (62, 86, 112) | (1024, 2347, 3766) |
| 6 | (124, 169, 213) | (132, 265, 362) | (2163, 3782, 5091) |
| 7 | (62, 81, 102) | (66, 98, 152) | (1687,3008, 4325) |
| 8 | (138, 192, 241) | (151, 330, 463) | (1524, 2450, 3864) |
| 9 | (82, 116, 159) | (115, 195, 291) | (1216, 2137, 3161) |
| 10 | (41, 55, 71) | (35, 53, 71) | (1432, 2560, 3782) |
| 11 | (168, 252, 367) | (307, 430, 584) | (2592, 4020, 5562) |
| 12 | (178, 232, 346) | (284, 372, 498) | (2792, 4427, 6163) |
| 13 | (111, 144, 198) | (121, 236, 370) | (1734, 2660, 4094) |
| 14 | (78, 103, 148) | (103, 157, 211) | (1426,2088, 3312) |
| 15 | (167, 212, 267) | (216, 370, 516) | (1785, 2605, 4042) |

$\tilde{y}_i$ and $\tilde{x}_{ij}$ are TFNs ,and defined by $\tilde{y}_i = (y_{il}, y_{im}, y_{ir})_T$ and $\tilde{x}_{ij} = (x_{ijl}, x_{ijm}, x_{ijr})_T$ , respectively.

The proposed method can be summarized as follows:

**Step 1:** Fuzzy input and fuzzy output data are defuzzified to crisp data as

$$y_{ic} = \frac{y_{il}+y_{im}+y_{ir}}{3} \quad \text{and} \quad x_{ijc} = \frac{x_{ijl}+x_{ijm}+x_{ijr}}{3} .$$

The results are shown in Table 2.

**Step 2:** In the classical regression model, the parameters can be obtained by the least squares method. Then the least squares estimators are

$$\hat{\beta}_0 = 0.3403,\ \hat{\beta}_1 = 0.5122 \ \text{ and } \ \hat{\beta}_2 = 0.0098 .$$

**Step 3**: Testing hypothesis $\mathrm{H}_1 : \beta_j \ne 0$ for j=0,1,2 by using lm package in R. The following results are

```
> ols.obj<- lm(y.c ~ x1.c + x2.c)
> summary(ols.obj)

Call:
lm(formula = y.c ~ x1.c + x2.c)
Residuals:
    Min      1Q  Median      3Q     Max
-8.3078 -2.8103 -0.7879  2.9215 10.7821

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.340349   6.588405   0.052  0.95965
x1.c        0.512216   0.016740  30.598 9.32e-13 ***
x2.c        0.009829   0.002741   3.586  0.00374 **
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 5.598 on 12 degrees of freedom
Multiple R-squared:  0.9936,   Adjusted R-squared:  0.9925
F-statistic: 927.4 on 2 and 12 DF,  p-value: 7.053e-14
```

As the results, we concluded $\beta_0 = 0$ . That is, $\beta_0$ is not in the model.

Table 2: Defuzzified fuzzy input and fuzzy output data.

| i | y.c | x1.c | x2.c |
|---|---|---|---|
| 1 | 155.67 | 249.00 | 2447.67 |
| 2 | 123.00 | 190.67 | 3388.33 |
| 3 | 224.00 | 378.33 | 3836.67 |
| 4 | 136.00 | 215.33 | 2501.33 |
| 5 | 67.00 | 86.67 | 2379.00 |
| 6 | 168.67 | 253.00 | 3678.67 |
| 7 | 81.67 | 105.33 | 3006.67 |
| 8 | 190.33 | 314.67 | 2612.67 |
| 9 | 119.00 | 200.33 | 2171.33 |
| 10 | 55.67 | 53.00 | 2591.33 |
| 11 | 262.33 | 440.33 | 4058.00 |
| 12 | 252.00 | 384.67 | 4460.67 |
| 13 | 151.00 | 242.33 | 2829.33 |
| 14 | 109.67 | 157.00 | 2275.33 |
| 15 | 215.33 | 367.33 | 2810.67 |

**Step 4**: The FLRM can be formulated as

$$(y_{il}, y_{im}, y_{ir})_T = \beta_0 + \beta_1 (x_{i1l}, x_{i1m}, x_{i1r})_T$$
$$+\beta_2 (x_{i2l}, x_{i2m}, x_{i2r})_T \quad .$$

**Step 5:** The estimation parameter method.
According to Section 3, the fuzzy least squares estimator vector is

$$\hat{\boldsymbol{\beta}}_F = [\mathbf{X}_l^T \mathbf{X}_l + \mathbf{X}_m^T \mathbf{X}_m + \mathbf{X}_m^T \mathbf{X}_m]^{-1} [\mathbf{X}_l^T \mathbf{y}_l + \mathbf{X}_m^T \mathbf{y}_m + \mathbf{X}_r^T \mathbf{y}_r] \quad .$$

We wrote R code to solve the fuzzy estimators. Thus the $\hat{\boldsymbol{\beta}}_F$ can be obtained as:

$$\hat{\beta}_{F1} = 0.4856 \text{ and } \hat{\beta}_{F2} = 0.0110 \quad .$$

**Step 6:** Since $\hat{\tilde{y}}_i$ is triangular fuzzy number and its membership function is

$$\mu_{\hat{\tilde{y}}_i}(\hat{y}) = \begin{cases} \dfrac{\hat{y} - \hat{y}_{im}}{\hat{y}_{im} - \hat{y}_{il}}, & \hat{y}_{il} \le \hat{y} < \hat{y}_{im} \\ \dfrac{\hat{y}_{ir} - \hat{y}}{\hat{y}_{ir} - \hat{y}_{im}}, & \hat{y}_{im} \le \hat{y} \le \hat{y}_{ir} \\ 0, & otherwise \end{cases} \quad ,$$

where $\hat{y}_{il} = 0.4856 x_{i1l} + 0.0110 x_{i2l}$,

$\hat{y}_{im} = 0.4856 x_{i1m} + 0.0110 x_{i2m}$,

and $\hat{y}_{ir} = 0.4856 x_{i1r} + 0.0110 x_{i2r}$. Thus we obtained the estimated fuzzy output as:

$$\hat{\tilde{y}}_i = 0.4856 (x_{i1l}, x_{i1m}, x_{i1r})_T + 0.0110 (x_{i2l}, x_{i2m}, x_{i2r})_T \ .$$

For example, in the first observation, the estimated fuzzy output (or fuzzy prediction) can be calculated by

$$\hat{\tilde{y}}_1 = 0.4856(151,\ 274,\ 322)_T + 0.0110(1432,\ 2450,\ 3461)_T$$

$$\hat{\tilde{y}}_1 = (89.0838,\ 160.0156,\ 194.4479)_T \ .$$

For interpretation, fuzzy prediction $\hat{\tilde{y}}_1$ has the following approaches:

1. The fuzzy prediction $\hat{\tilde{y}}_1$ means approximately $160.0156$.

2. The fuzzy prediction $\hat{\tilde{y}}_1$ means a closed interval $[89.0838, 194.4479]$.

**Step 7:** The coefficient of determination of the fuzzy model equals to

$$R^2 = 0.9176 \quad .$$

It indicates that the model fits the data well.
For the result, the fuzzy model is as

$$\hat{\tilde{y}}_i = 0.4856 (x_{i1l}, x_{i1m}, x_{i1r})_T + 0.0110 (x_{i2l}, x_{i2m}, x_{i2r})_T$$

## 7. Conclusion

The objectives of this research are to formulate a fuzzy linear regression model using LR-fuzzy number and estimate the model parameters. Diamond's distance, a distance between two LR-fuzzy numbers, is defined as an error of fuzzy model. Consequently, the total error is defined as a sum squares error or the model fuzziness. The proposed method based on the least squares method is presented to estimate the model parameters. Furthermore, the fuzzy prediction and interpretation are also presented. Finally, we defined a coefficient of determination ($R^2$) in the fuzzy model as an index of goodness of fit.

*Discussion and Further study*

The advantage of the proposed method to build the FLRM is convenient for the beginners. Also, the proposed estimation parameter method is based on the least squares method in regression analysis. It is a well-known method. However, the proposed estimation parameter method is associated with the formulation of FLRM. For example, if we change the FLRM to

$$\tilde{y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 \tilde{x}_{i1} + \cdots + \tilde{\beta}_k \tilde{x}_{ik} \quad \text{for } i = 1, \ldots, n$$

(fuzzy coefficients and fuzzy data model)
or the other

$$\tilde{y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_{i1} + \cdots + \tilde{\beta}_k x_{ik} \quad \text{for } i = 1, \ldots, n$$

(fuzzy coefficients model).
The important question will be occurring. How can we estimate the model parameters? Is the least squares method continuing to work? What is the goodness of fit in fuzzy model? These questions are interesting to find out the solution for the further study.

**Acknowledgements**

We gratefully thank Assistant Professor Dr. Winai Bodhisuwan, Department of Statistics, Faculty of Science, Kasetsart University. We are also grateful to Rajamangala University of Technology Thanyaburi (RMUTT), Pathum Thani, Thailand, for financial support. Finally, we gratefully thank the anonymous reviewers for their valuable suggestions.

## References

[1] Buckley JJ. Fuzzy statistics: regression and prediction. Soft Comput. 2005; 9: 769–775.

[2] Chen LH, Hsueh CC. A mathematical programming method for formulating a fuzzy regression model based on distance criterion. IEEE Transactions on Systems, Man, and Cybernetics. 2007; l37: 705–712.

[3] Chen LH, Hsueh CC. Fuzzy regression models using the least-squares method based on the concept of distance. IEEE Transactions on fuzzy system. 2009; 17: 1259–1272.

[4] Diamond P. Fuzzy least squares. Information Sciences. 1988; 46: 141–157.

[5] Kao C, Chyu CL. A fuzzy linear regression model with better explanatory power. Fuzzy Sets and Systems. 2002; 126: 401–409.

[6] Kao C, Chyu CL. Least-squares estimates in fuzzy regression analysis. European Journal of Operations Re--search. 2003; 148: 426–435.

[7] Kim B, Bishu RR. Evaluation of fuzzy linear regression models by comparing membership functions. Fuzzy Sets and Systems. 1998; 100: 343–352.

[8] Kim KJ, Kim DH, Choi SH. Least absolute deviation estimator in fuzzy regression. J. Appl. Math. & Computing. 2005; 18: 649–656.

[9] Myers RH, Milton JS. A first course in the theory of linear statistical models. 2nd ed. Boston: PWS-KENT Publishing Company; 1991.

[10] Rattanalertnusorn A. Parameter estimation of fuzzy linear regression models [Thesis]. Bangkok: Kasetsart University; 2014.

[11] Rattanalertnusorn A, Thongteeraparp A, Bodhisuwan W. Parameter estimation of fuzzy linear regression model: The extension of Chen and Hsueh. Journal of Applied Sciences. 2014; 14(7): 631–640.

[12] Redden DT, Woodall WH. Properties of certain fuzzy linear regression methods. Fuzzy Sets and Systems. 1994; 64: 361–375.

[13] Sakawa M, Yano H. Multiobjective fuzzy linear regression analysis for fuzzy input–output data. Fuzzy Sets and Systems. 1992; 47: 173–181.

[14] Tanaka H, Uejima S, Asai K. Linear regression analysis with fuzzy model. IEEE Transactions on Systems, Man, and Cybernetics. 1982; 12: 903–907.

[15] Tanaka H. Fuzzy data analysis by possibilistic linear models. Fuzzy Sets and Systems. 1987; 24: 363–375.

[16] Tanaka H, Watada J. Possibilistic linear systems and their application to the linear regression model. Fuzzy Sets and Systems. 1988; 27: 275–289.

[17] Tanaka H, Hayashi I, Watada J. Possibilistic linear regression analysis for fuzzy data. European Journal of Operations Research. 1989; 40: 389–396.

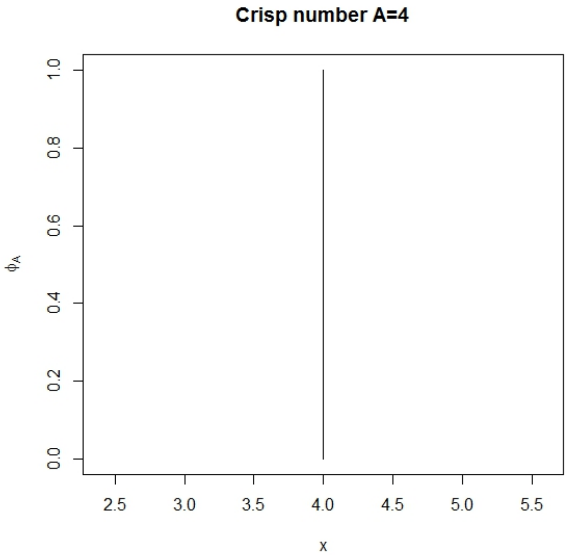
[18] Tanaka H, Lee H. Interval regression analysis by quadratic programming approach. IEEE Transactions on fuzzy system. 1998; 6: 473–481.

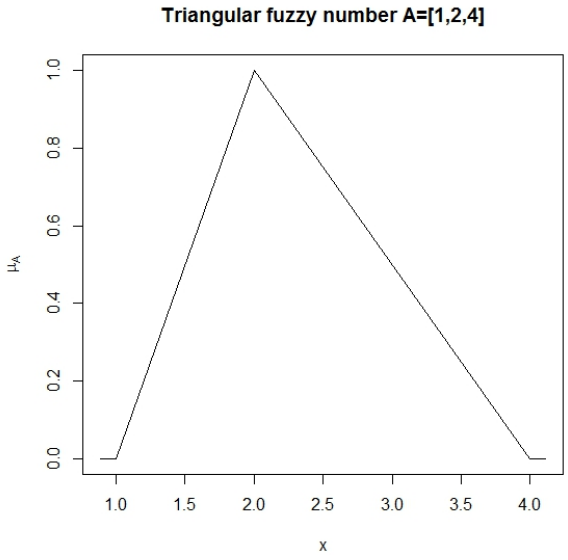
[19] Wang YM, Yang JB, Xu DL, Chin KS. On the centroids of fuzzy numbers. Fuzzy Sets and Systems. 2006; 157: 919–926.

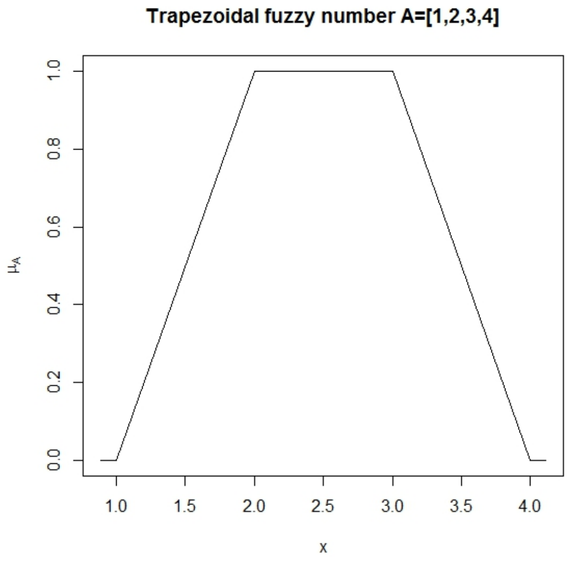
[20] Wu HC. Linear regression analysis for fuzzy input and output data using extension principle. Computers and mathematics with applications. 2003; 45: 1849–1859.

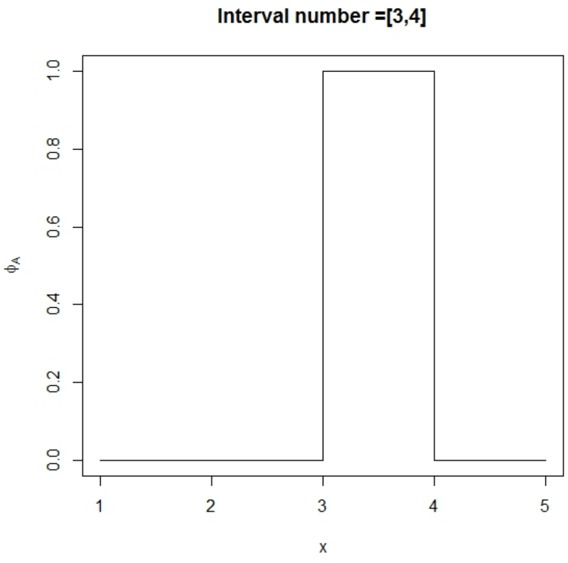
[21] Zadeh LA. Fuzzy sets. Information and Control. 1965; 8(3): 338–353.

**Appendix A. Figures of crisp number, fuzzy numbers and interval number.**


Crisp number A=4


Triangular fuzzy number A=[1,2,4]


Trapezoidal fuzzy number A=[1,2,3,4]


Interval number =[3,4]

## Appendix B. R script for numerical example.

```
#########################################
#      paper 126: numerical example
#      written by: Rattanalertnusorn A.
#########################################
# To storage fuzzy output and fuzzy input
y.l=c(111,88,161,83,51,124,62,138,82,41
      ,168,178,111,78,167)
y.m=c(162,120,223,131,67,169,81,192,116
      ,55,252,232,144,103,212)
y.r=c(194,161,288,194,83,213,102,241,
      159,71,367,346,198,148,267)
x1.l=c(151,101,221,128,62,132,66,151,
       115,35,307,284,121,103,216)
x1.m=c(274,180,375,205,86,265,98,330,
       195,53,430,372,236,157,370)
x1.r=c(322,291,539,313,112,362,152,463
       ,291,71,584,498,370,211,516)
x2.l=c(1432,2448,2592,1414,1024,2163,
       1687,1524,1216,1432,2592,2792,
       1734,1426,1785)
x2.m=c(2450,3254,3802,2838,2347,3782,
       3008,2450,2137,2560,4020,4427,
       2660,2088,2605)
x2.r=c(3461,4463,5116,3252,3766,5091,
       4325,3864,3161,3782,5562,
       6163,4094,3312,4042)
df=data.frame(y.l,y.m,y.r,x1.l,x1.m,x1.r
       ,x2.l,x2.m,x2.r)
df
###########################
#  Defuzzification
###########################
y.c=(y.l+y.m+y.r)/3
x1.c=(x1.l+x1.m+x1.r)/3
x2.c=(x2.l+x2.m+x2.r)/3
df2=data.frame(y.c,x1.c,x2.c)
df2
#########################################
# Least squares method and testing
# hypothesis
#   H0: beta[j] equal to 0 versus
#   H1: beta[j] unequal to 0
#########################################
ols.obj<- lm(y.c ~ x1.c + x2.c)
summary(ols.obj)
b0<-ols.obj$coefficient[1]
b1<-ols.obj$coefficient[2]
b2<-ols.obj$coefficient[3]
c(b0,b1,b2)
###################################
#  fuzzy least squares estimator
###################################
n=length(df$y.m)
YL<-matrix(nrow=n,ncol=1)
YM<-matrix(nrow=n,ncol=1)
YR<-matrix(nrow=n,ncol=1)
XL<-matrix(nrow=n,ncol=2)
XM<-matrix(nrow=n,ncol=2)
XR<-matrix(nrow=n,ncol=2)
XL[,1]=df$x1.l;XL[,2]=df$x2.l
XM[,1]=df$x1.m;XM[,2]=df$x2.m
XR[,1]=df$x1.r;XR[,2]=df$x2.r
YL=df$y.l
YM=df$y.m
YR=df$y.r
XTXL=t(XL)%*%XL
XTXM=t(XM)%*%XM
XTXR=t(XR)%*%XR
B=solve(XTXL+XTXM+XTXR)%*%(t(XL)%*%YL+t(X
M)%*%YM+t(XR)%*%YR)
B1=as.numeric(B[1,1])
B2=as.numeric(B[2,1])
c(B1,B2)
################################
# To calculate MSE and R-square
################################
ypred.l=df$x1.l*B1 +df$x2.l*B2
ypred.m=df$x1.m*B1 +df$x2.m*B2
ypred.r=df$x1.r*B1 +df$x2.r*B2
df3=data.frame(ypred.l,ypred.m,ypred.r)
df3
SSE=sum((y.l-ypred.l)^2)+sum((y.m-
ypred.m)^2)+sum((y.r-ypred.r)^2)
SST=sum((y.l-mean(y.l))^2)+sum((y.m-
mean(y.m))^2)+sum((y.r-mean(y.r))^2)
MSE=SSE/(n-2)
Rsq=1-(SSE/SST)
Rsq
###############################
```

# Road Safety Audit Associatedwith Analytic HierarchyProcessin THAILAND

NapatMunsil*, HatsakornWanna,AttapornPohkaew,ChoorachSrikhumdeeand, and VeerisAmmarapala
*School of Management Technology,Sirindhorn Institute of Technology, Thammasat University,*
*PathumThani, 12000, THAILAND*
*+662-501-3505, Email: tap43344@hotmail.com*

## Abstract

Road accidents is a major problem that causes a life and injuries, occur mostly in a developing country such as Thailand, to prevent the loss from the traffic accidents, the use of Road Safety Audit (RSA) is a good alternative. This paper rely on a Ranker which is one of the RSA methods, started by look at the factors that might lead to a road accidents and applied the factors appropriate to the country for instance, Road Alignment, Pavement, Intersection, Roadside, and Signal & Sign and Lighting. Furthermore, on the ranker, the additional of another factor from road audit manual of South Africa which is the factor of Traffic operation. After the factors are specified, the data and information will collected by audit team, the result of evaluated the road is call as Road Safety Index (RSI) and will calculated by the auditor in order to know a level of safety of each particular road by the weight of each factor is equally. In addition, to make the result from RSA more accurately and more efficiency, Analytic Hierarchy Process (AHP) is the decision making tools that can help the user acquired the weight of each criteria by ask the experts base on group about the important of each factors and calculated base on pair-wise method. The result from AHP will help the auditor know which factor is the most important and improve the accurate of the result by calculated the RSA again, but change the weight of each factor to the weight that acquired from AHP. The recommendation from the auditor will help the Ministry of Transport (MOT) in term of road safety and can reduce the number of road accidents.

*Keywords*: Road Safety Audit, Road Safety Index, Analytic Hierarchy Process, Ministry of Transport

*Corresponding Author
E-mail Address: *tap43344@hotmail.com*

## 1. Introduction

The Road Safety Audit (RSA) was started by UK in 1980s, after that Australia and New Zealand start establish RSA in 1990s. The requirement of a Road Safety Audit is contained in the Design Manual for Roads and Bridges. To do the Road Safety Audit the auditor mostly be an expert of road safety and have various experiences, and can include road safety engineering, to be more effective team the auditor can invite to the local experienced people, together the Road Safety Audit will have an independent team of expert. According to the Journal of Society for Transportation and Traffic Studies (JSTS) Vol.1 No.3, World Health Organization (WHO, 2004) indicated that without increased the effort and novel in Thailand the numbers of accident occurred by crashed cars have peaked in 1994 and 1995, the total number of road traffic deaths around the world and injuries is forecast to rise by 65% between the year 2000 and 2020, and for the developing countries and develop countries traffic deaths are expected to be 80%. This information gave us the real situation of the road that is tend to be more significantly for worldwide countries because the cost of death from the accident could damage the countries so much, so it is very important to promote the RSA worldwide

in order to reduce the number of traffic deaths. In order to use RSA the auditors are required to have RSA manual which have several kind of manual, some countries have their own manual and some countries did not have their own RSA manual usually accommodated other manual and make the new one. Furthermore, sometime the result of RSA might have more than two alternatives, to choose the best ways the owner of the RSA project can utilized of the decision making tools

The Analytic Hierarchy process is one of the most useful decision making tools to deal with the decision problems such as Which alternative is the best and how can we selected them if those alternative have a different weight, priority have played the big role in order to make a decision. We considered that each alternative have different relative importance (score) or weight, to find the optimal solution we can use the AHP to supported our project associated with the RSA and RSI. The used of AHP can improve the effective outcome of RSA and RSI such as which road we will maintain first if we have several roads that slightly the same problems and need to be fixed in time, by understand the important of the weight of each criteria the auditor could tell which road needed a maintenances before the other road. In order to start the AHP, many phase of process were involved such as selected five or six experts the reason was about the time, cost and the consistency of the result. In additional, have expert more than 6 person it can be the time consuming when the expert needed to reach a consensus on each judgments, and the consistency of result might over the standard level. In this case, the selection process of expert was essential. The other phase of AHP was number of comparison in pair-wise comparison; the number of comparison was a combination of number of factors or criterion and it will much more depended on the number of thing to be compared (Table 1).

Table 1: Number of comparison

| Number of Factor/ Criteria | 1 | 2 | 3 | 4 | 5 | 6 | 7 | n |
|---|---|---|---|---|---|---|---|---|
| Number of comparisons | 0 | 1 | 3 | 6 | 10 | 15 | 21 | $\frac{n(n-1)}{2}$ |

According to the table above, the more factor or criteria to compare the more the number of comparisons is. So it difficult to use when then number of criteria or alternative is high or more than 7.

## 2. Research Methodology

*2.1 Road Infrastructure*

The road infrastructure measurement has been categorized in 6 main factors for this paper. The scope or the items that involved with each factors are shown in figure below.

Table 2: Scope of Road Infrastructure

| | |
|---|---|
| Road Alignment | Road Geometry, Lanes and Shoulder Width, Overtaking, Traffic Marking and Delineator |
| Intersection & Junction | Junction, Private Access, U-turn, and U-turn lane |
| Pavement | Cracks, Roughness, Skid Resistance, Wheel rut, Pothole, and Road Surface |
| Roadside | Guard Rail, Obstacles, Safety Equipment, Pedestrians, Footway, and Roadside Barriers |
| Signal, Sign, and Lighting | Traffic Signal, Traffic Sign, Lighting |
| Traffic Operation | Special Traffic Sign, Special Traffic Management |

*2.1.1 Prepare theroad safety manual*

The manual is accommodated from the senior project that based on Rankers [7, 11], and combination of road audit manual of South Africa [15]. The manual needed to be extensive, so the other auditor can utilize it with any other road type such as rural road.

*2.1.2 Road Section Length*

Based on Rankers the length of application of the road must be long enough to be cost-effective but it must also be short enough so any compensation effect within too long sections can be avoided. In addition, the length used in the evaluation process is 2 kilometers.

*2.2.3 Collected Information*

After selected the road to audited, and the manual is readiness, the audit team will gathered the information by walk along the roadside in total of 10 kilometers and filled the data in the manual and the information will calculated by the same audit team.

*2.2 Road Safety Index (RSI)*

The Rankers believe that the best way to define the evaluation scale is to base on the urgency actuation started from 1 until 4 (1 is the worst, and 4 is the best). But the senior think that in reality even the road is scored as 4 it is willing to have routine maintenance properly, so the table below shown the score associated with the action. And the rearrange of the number started from 0 to 3 (0 is the worst, and 3 is the best)

Table 3: RSI score and meaning

| | |
|---|---|
| 0 | Need urgent remedial measurement and improvement |
| 1 | There are deficiencies and need improvement |
| 2 | Need some maintenance |
| 3 | Only need routine and periodic maintenance |

Moreover, the score from the previous section will calculated and find the total score and average of each factor, and will express as a percentages in order to define each score clearly. To make it clearly to see the level of safety of each road section and easier to plot the result on the draft map, the use of color associated with the RSI score is in use as below.

Table 4: Color associated with RSI scores

| Color | Percentage | Meaning |
|---|---|---|
| Red | 0%-25% | Need urgent remedial measurement and improvement |
| Orange | 26%-51% | There are deficiencies and need improvement |
| Yellow | 52%- 75% | Need some maintenance |
| Green | 76%-100 | Only need routine and periodic maintenance |

After the score is obtained from the calculation it will be translated to the above color scale, and also the overall average of the road section can be calculated and translated to the color scale (Table 5).

Table 5: Road section evaluation result

| | | |
|---|---|---|
| 🟩 Road Alignment | 🟨 Intersection & Junction | 🟩 Pavement |
| 🟨 Roadside | 🟥 Signal, Sign, and Lighting | 🟨 Traffic Operation |
| 🟨 Overall | | |

At this step the score of RSI can be now utilized in reality, and also can be used depend on the objective of each users. This paper also used the other tools to help classify the level of safety of the road along with RSI which needed to be done before calculated the information above which call as Analytic Hierarchy process (AHP) the information and the way to calculated will be shown in the next topic.

### *2.3 Analytic Hierarchy Process*

The Analytic Hierarchy Process (AHP) is a decision making process which was established by Thomas Saaty. AHP is a utilize tool for selecting the best alternative of complex situation between the alternatives. AHP model can be divided into at least two levels include of goal, objectives or criteria. In level 3, criteria have sub-criteria as an additional division which is related to those criteria. Before being AHP, many measurements are needed, for example, pair-wise comparisons, eigenvector calculation, and consistency computation. This paper, however, does not choose the best alternative. Its objective is to find appropriate weight of each criterion in order to apply in Road Safety Audit (RSA) and Road Safety Index(RSI) performance. Furthermore, Calculating of AHP can be used in Microsoft Excel or a utilized program called Expert Choice, which was developed by Thomas Saaty.

The AHP model was shown in figure below (Figure 1).

Table 6: Fundamentals scale

| **Intensity of Importance On an Absolute Scale** | **Definition** | **Explanation** |
|---|---|---|
| 1 | Equal importance | Two activities contribute equally to the objective |
| 3 | Moderate importance of one over another | Experience and judgment strongly favor one activity over another |
| 5 | Essential or strong importance | Experience and judgment strongly favor one activity over another |
| 7 | Very strong importance | An activity is strongly favored and its dominance demonstrated in practice |
| 9 | Extreme importance | The evidence favoring one activity over another is of the highest possible order of affirmation |
| 2,4,6,8 | Intermediate values between the two adjacent judgments | When compromise is needed |
| Reciprocals of Above non zero | If activity i has one of the above numbers assigned to it when compared with activity j, then j has the reciprocals value when compared with i | |

Figure 1: AHP Model

For our AHP model was placed in appendices section (Appendix A).

### *2.3.1 Measuring of AHP*

The experts of the related field with the project were invited to attend a personal interview. Since the experts were familiar with the goal of AHP, thus they are willing to determine the relative weight against a list of factors and sub-factor that involved in the project. Moreover, the scales of measuring the weight of each criterion are provided by Saaty call as Fundamental scale [3] or nine-scale [23], the detailed and the meaning are shown in the table below (Table 6).

### *2.3.2 Pair-wise comparisons*

Pair-wise comparison method is a comparison between two variables in order to acquire a new variable which one is more importance than another in matrix table. An importance levels depend on what experiences and skills of decision maker had. In addition, the variable is determined by multiplied number between one to nine (Table 6). For example, road alignment is more importance or safe than pavement five times for the fact that the accident could be occurred, it mean that the expert selected number 5 in Fundamental. This report compares between two variables in each levels until every blank space in the metric box is completed.

Some examples of pair-wise comparisons in matrix are shown below.

Table 7: Pair-wise matrix

| **Roadside** | **A** | **B** | **C** |
|---|---|---|---|
| **A** | 1 | 4 | 5 |
| **B** | 1/4 | 1 | 7 |
| **C** | 1/5 | 1/7 | 1 |

### *2.3.3 Geometric mean*

Geometric mean is the most common method used by group to set priorities. To calculate geometric mean, each level was multiplied and takes nth root of criteria in a formula below.

$$a_{ij} = \sqrt[1/n]{a_{ij}^{1} \times a_{ij}^{2} \times \ldots \times a_{ij}^{n}}$$

Where a is variable of each criteria
i is element
j is decision maker or expert

For example, adjust the weight from the pair-wise comparisons assume that we have 5 expert involved and each score of the first factors is as follow; expert1 (0.14), expert2 (0.04), expert3 (0.03), expert4 (0.36), and expert5 (0.12), and n = 5 (number of expert) the formula will be.

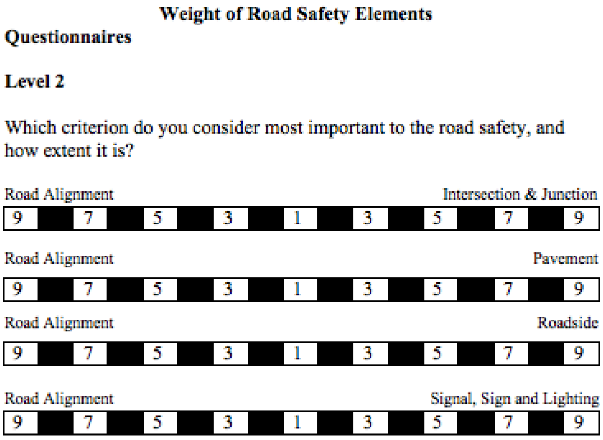
$$= \sqrt[1/5]{0.14} \times 0.04 \times 0.03 \times 0.36 \times 0.12$$

## 3. Research Results and Discussion

The aim of the paper was to show the differencs on result between the one that utilize the ability of AHP to set the priorities of each factors by weight them difference, in the other hand the other one that normally set the weight of every factor equaly and calculated the reuslt directly on RSI. This stage will shown the way to evaluated the raw data and translated it into the information by EC 2000, and some simple calculation.

### *3.1 Weighting of the criteria*

The first important thing before the calculation stage was to classify weight of each factors by using AHP. By asked the opinion of 5 experts about which factors did they think that can be the cause of traffic accident the most or which factors have the most important in term of road safety. In this stage, each expert will reveived a questionnaires and have time to done it around 1 week, after that the audit team will collected it and expert might have their advise and recommendation about the questionnaries and road safety project, some example of questionnaries shown in the figure below.



**Weight of Road Safety Elements**

**Questionnaires**

**Level 2**

Which criterion do you consider most important to the road safety, and how extent it is?

Road Alignment — Intersection & Junction
9 ■ 7 ■ 5 ■ 3 ■ 1 ■ 3 ■ 5 ■ 7 ■ 9

Road Alignment — Pavement
9 ■ 7 ■ 5 ■ 3 ■ 1 ■ 3 ■ 5 ■ 7 ■ 9

Road Alignment — Roadside
9 ■ 7 ■ 5 ■ 3 ■ 1 ■ 3 ■ 5 ■ 7 ■ 9

Road Alignment — Signal, Sign and Lighting
9 ■ 7 ■ 5 ■ 3 ■ 1 ■ 3 ■ 5 ■ 7 ■ 9

Figure 2: AHP Questionnaires

### *3.2 Obtained AHP results*

By using of EC2000 to help in term of solving a huge matrix, and can save a lot of time in calculation. The number of matrix are depended on the number of expert, so the total matrix is equal to 5 and the size of matrix is 6x6 from the factor above for instance Road Alignment, Intersection & Junction, Pavement, Roadside, Signal, Sign, and Lighting, and Traffic Operation. So first thing was to set a goal of the project which is to know the most important factor that have an effect to road safety, and setup the AHP level 2 which refer to the 6 factors above, and last but not least was to set the sub-criteria of each factors which call as AHP level 3, and let the EC 2000 solving the matrix of pair-wise comparisons. After put in the data from the road safety questionnaires to the expert choice program, the result are shown in term of weight of each factors both in Level 2 and Sub-criteria. Some of the result was shown in the table below (Table 7).

The weight that show below the factor name in the figure above is not the real weight yet, to finding the real one that can be used calculated the road safety score, the calculation of geometric mean was involved by using the formula in previous section or used the MS excel program and type in the formula =Geomean(number1, number2, …., number n), after that sum up all of the score and put that number divided by the prior number, continued until get all of the real weight of each factors, and do the same process with the sub-criteria to find the real weight of each sub-criteria. The results of inspection of geometric mean are summarized into a table below (Table 8).

Table 8: Result from the Expert choice

| Name | Goal Road Safety | Road Alignment (W:0.1480) | Intersection (W: 0.3730) | Pavement (W:0.0640) | Roadside (W:0.0290) | Signal, Sign and Lighting (W:0.1510) | Traffic Operation (W:0.2340) |
|---|---|---|---|---|---|---|---|
| Expert1 | 0.0948 | 0.0000 | 0.0595 | 0.0739 | 0.0164 | 0.0000 | 0.0000 |

Table 9: Calculation of real weight of each factor

| Name | Road Align ment | Inter sectio n | Paveme nt | Roadsi de | Signal, Sign, and Lighting | Traffic Operati on | Sum |
|---|---|---|---|---|---|---|---|
| **Expert1** | 0.148 | 0.373 | 0.064 | 0.029 | 0.151 | 0.235 | **1** |
| **Expert2** | 0.044 | 0.468 | 0.157 | 0.058 | 0.251 | 0.022 | **1** |
| **Expert3** | 0.03 | 0.166 | 0.108 | 0.068 | 0.408 | 0.22 | **1** |
| **Expert4** | 0.36 | 0.113 | 0.078 | 0.048 | 0.237 | 0.164 | **1** |
| **Expert5** | 0.126 | 0.29 | 0.033 | 0.03 | 0.068 | 0.453 | **1** |
| **Geometri c mean** | 0.097 | 0.248 | 0.077 | 0.043 | 0.190 | 0.1532 | **0.811** |
| **Real Weight** | **0.120** | **0.306** | **0.095** | **0.054** | **0.234** | **0.188** | **1** |

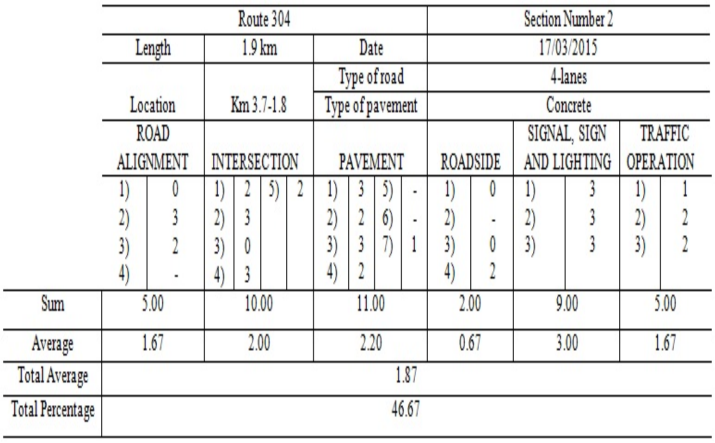
The weights in table above are ready to utilize in RSI to find the result of road safety which detailed in next sub-section, and the result of calculated the geometric mean for the sub-criteria are in appendices section ( Appendix B).

*3.3 Combination of RSI and AHP*

This paper selected Changwattana road in route 304 and we segmented the road into 6 sections, each section length is around 1.8-1.9 Km. inthe total of11 Km.

In order to get the result from the RSI, the users need the sample model base on the Rankers or the users can develop their own model instead of using everything the same as the Rankers model, the example of model are shown in table below.

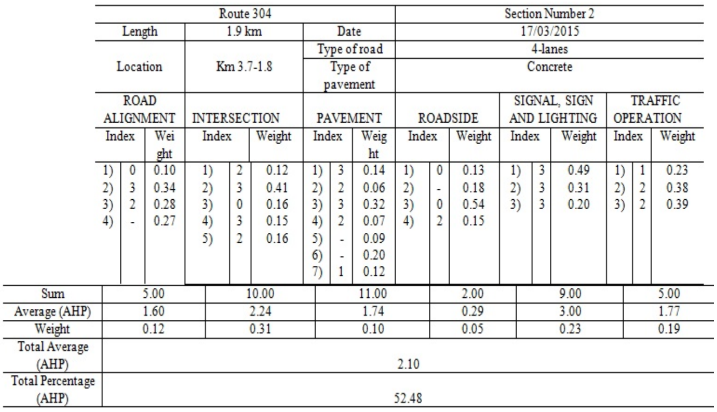
Table 10: RSI result from road section 2



| | Route 304 | | | Section Number 2 | | |
|---|---|---|---|---|---|---|
| | Length | 1.9 km | Date | 17/03/2015 | | |
| | Location | Km 3.7-1.8 | Type of road | 4-lanes | | |
| | | | Type of pavement | Concrete | | |
| | ROAD ALIGNMENT | INTERSECTION | PAVEMENT | ROADSIDE | SIGNAL, SIGN AND LIGHTING | TRAFFIC OPERATION |
| | 1) 0<br>2) 3<br>3) 2<br>4) - | 1) 2<br>2) 3<br>3) 0<br>4) 3<br>5) 2 | 1) 3<br>2) 2<br>3) 3<br>4) 2<br>5) -<br>6) -<br>7) 1 | 1) 0<br>2) -<br>3) 0<br>4) 2 | 1) 3<br>2) 3<br>3) 3 | 1) 1<br>2) 2<br>3) 2 |
| Sum | 5.00 | 10.00 | 11.00 | 2.00 | 9.00 | 5.00 |
| Average | 1.67 | 2.00 | 2.20 | 0.67 | 3.00 | 1.67 |
| Total Average | 1.87 | | | | | |
| Total Percentage | 46.67 | | | | | |

The result from the table shown that the total percentage or the level of safety of this road section is only 46.67% or it in range of orange color meaning some deficiencies are exist and need an improvement. But as mentioned before, the model without AHP it mean each factors have the same weight, in this example the weight of each one equal to 0.16666667 from divided 1 by 6. Therefore, in reality some factors might important than the other, so the combination of AHP and RSI can supported this truth by changed the weight of each factors.By calculation AHP in previous section, the auditor will know the new weight of each criterion and sub-criteria, the next table we shown the combination of AHP and RSI together by developed a new model (Table 10).

Table 11: RSI & AHP result from road section 2



| | Route 304 | | | | | | Section Number 2 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Length | | 1.9 km | | Date | | 17/03/2015 | | | | | |
| | Location | | Km 3.7-1.8 | | Type of road | | 4-lanes | | | | | |
| | | | | | Type of pavement | | Concrete | | | | | |
| | ROAD ALIGNMENT | | INTERSECTION | | PAVEMENT | | ROADSIDE | | SIGNAL, SIGN AND LIGHTING | | TRAFFIC OPERATION | |
| | Index | Wei ght | Index | Weight | Index | Weig ht | Index | Weight | Index | Weight | Index | Weight |
| | 1) 0<br>2) 3<br>3) 2<br>4) - | 0.10<br>0.34<br>0.28<br>0.27 | 1) 2<br>2) 3<br>3) 0<br>4) 3<br>5) 2 | 0.12<br>0.41<br>0.16<br>0.15<br>0.16 | 1) 3<br>2) 2<br>3) 3<br>4) 2<br>5) -<br>6) -<br>7) 1 | 0.14<br>0.06<br>0.32<br>0.07<br>0.09<br>0.20<br>0.12 | 1) 0<br>2) -<br>3) 0<br>4) 2 | 0.13<br>0.18<br>0.54<br>0.15 | 1) 3<br>2) 3<br>3) 3 | 0.49<br>0.31<br>0.20 | 1) 1<br>2) 2<br>3) 2 | 0.23<br>0.38<br>0.39 |
| Sum | 5.00 | | 10.00 | | 11.00 | | 2.00 | | 9.00 | | 5.00 | |
| Average (AHP) | 1.60 | | 2.24 | | 1.74 | | 0.29 | | 3.00 | | 1.77 | |
| Weight | 0.12 | | 0.31 | | 0.10 | | 0.05 | | 0.23 | | 0.19 | |
| Total Average (AHP) | 2.10 | | | | | | | | | | | |
| Total Percentage (AHP) | 52.48 | | | | | | | | | | | |

From the result above, the significant thing from this combination is that in change the overall score of the RSI because of changed the weight, and the total percentage of this road section is increase in a good way from 46.67% to 52.48% along with the range of the score is changed from orange to yellow too, even though it is a slightly increase, but that enough to supported the objective of this paper about the important of the weight.

*3.4 Discussion the RSI & AHP result*

From overall of the result above it shown that the weight of each factor are as follow; Road Alignment (0.120333724), Intersection & Junction (0.306474), Pavement (0.095523), Roadside (0.054228), Signal, Sign, and Lighting (0.23453), and Traffic Operation (0.188911). Therefore the most important factor was Intersection & Junction with the score of 0.306474 and the factor that has least important was Roadside with the score of 0.054228. But not all road sections have differences in total percentage as the example, the table below will shown the differences in percentage of each factors above (Table 11).

Table 12: The percent different

| Section | Non-AHP | AHP | %Different |
|---|---|---|---|
| **Number 1** | 39.44 | 39.00 | 1.14 |
| **Number 2** | 46.67 | 52.48 | 12.45 |
| **Number 3** | 46.67 | 46.70 | 0.06 |
| **Number 4** | 44.44 | 42.29 | 4.85 |
| **Number 5** | 45.83 | 43.43 | 5.25 |
| **Number 6** | 37.00 | 23.25 | 37.15 |

As you can see, there are only two road sections that have a big in percentage different which is road section number 2, and road section number 6. Therefore the highest safety score of each road section is road section number 2 and road section number 3 as a score of 46.67% and the lowest safety was road section number 6 the score of 37.00%, after applied the AHP there are some changed in the best score of road section from road section number 2 and number 3 to only road section number 2, any other is still the same placed. Thus, the AHP wasrearrange the result from almost similar score to a more range of score, so AHP help in term of making decision and what will be the best alternative of the objective. The other result sheets of other road section are available in appendices section (Appendix B).

## 4. Conclusion

Mostly the traffic accidents have a high rate in developing countries like stated at the beginning the main reason of the accident almost came from the driving behavior of driver or human factor, and followed by a road infrastructure or road assessments factor [12]. Even though the road infrastructure was not the number one of factors causes a traffic accidents, but in reality human factor was hard to predict and hard to control, therefore aiming to make a better road rather than have set up a law and regulation to control the driving behavior was a must way to do, to reduce the number of serious injuries and death on a road way. Furthermore about making the better road, it has many methods and tools to take care of this case such as iRAP (Star Rating), RSI (Proactive), Black spot (Reactive). By selected the RSI and used the method based-on Rankers, the overall evaluation process and the standard is quite good because the manual used to audit the road was extensive and can easily adapt with other RSA manuals, so it be the standalone manual that normally can used right away. But there were some case to aware for instance some type of road was not recommended to use a manual from Rankers such as single road section or highway. But overall performance of RSI was quite high and the auditor can get the RSI score faster than the other method because it is a proactive tools, the requirement of RSI was only a manual for auditing. Moreover about AHP, the method itself was good and were recommended by many experts, and the result after applied AHP to the RSI was interested because the range of the result are extended so it easily to select which one is the best among the other. In this case we know which road infrastructure needed to be focuses the most if Changwattana road route 304 needed maintenance and improvement, the answer was a intersection factor. If we don't apply an AHP we might not sure which factor needed to be focuses the most and hardly to selected which road section need a maintenance firstbecause road section number 2 and road section number 3 have the same score. In short, combination of other software or other method can made the result of RSI more realistic, and more accurately. The used of AHP can make the data more valuable by changed of weight. For the future works, adding another tools of AHP such as the cost of maintenance, can make the result changed and the optimal solution of this paper might changed, so there are several tools to add in to this project, but the objective of the goal didn't change it is to reduce the number of accident occurred on the road way.

## 5. Acknowledgements

This project was supported by Sirindhorn Institute of Technology, Thammasat University. With valuable guidance and advice from Dr. VeerisAmmarapala towards the completion of this project. Special thanks goes to Mr.SujinMungnimit and Mr.BhanitizAursudkij from Bureau of Highway Safety, Department of Highways also for Mr.PonlathepLertworawanich, Mr.SophonSangpan, and Mr.JoompojPhumpradub from Nonthaburi Highway Maintenance District Office for Recommendations and informative answers to the questionnaires. Additionally, thank for whose aren't mentions in here but inspired and encouraged with this project until this project reach the goals.

## 6. References

[1]Alessio Ishizaka and Ashraf Labbib. Analytic Hierarchy Process and Expert Choice: Benefits andLimitations.ORInsight. 2009; 22(4): 201-220.

[2] Arun B, Dheeraj K. Road safety audit. Journal of Mechanical and Civil Engineering (IOSRJMCE). 2012; 1(6): 1-56.

[3]Cesar P, Michael L. Weighting sustainable development indicators (SDIs) for surface mining operations using the analytical hierarchy process (AHP). International Journal of the Analytic Hierarchy Process. 2013; 5(2): 200-222.

[4] Colin B, Ian C, et al. Road safety audit procedures for projects. New Zealand; 2013.

[5] Evangelos T, Stuart M. Using the analytic hierarchy process for decision making in engineering applications: some challenges. Inter'l Journal of industrial. 1995; 2(1): 35-44.

[6] Jain S, Singh K, and Parida M. Road safety audit forfour lane national highways [Dissertation].India; 2011.

[7] Jose P, Guiller R. Road safety index, Ranking for European road safety. 2008: 1-58.

[8] Karen S. Quantifying Safety in the Road Safety Audit Process [Internet]. 2014 Available from: http://safety.fhwa.dot.gov/tools/crf/resources/cmfs/

[9] Kuo-Piao Chung, et al. Application of the analytic hierarchy process in the performance measurement of colorectal cancer care for the design of a pay-for-performance program in Taiwan. International Journal for Quality in Health Care 2013; 25(1), 81 – 91.

[10] Luz V, LAGUNZAD, and MEngSci. Introducing road safety audit in the Philippines.Journal of the Eastern Asia Sociely for Trarsportation Studies. 1999: 197-208.

[11] Monica GM, Pablo, AB, and M. Carmen GC. An AHP-based evaluation procedure for innovative educational project: A face-to-face vs. computer-mediated case study. Omega.2008; 36: 754-765.

[12] Morakot K, Thananthorn P, et al. The Analysis of Bangkok Metropolitan Road Traffic Safety.Songklanakarin journal of science and technology. 2014; 36(1): 97-105.

[13] Mujgan S, Ozdemir, Thomas L. Satty. The unknow in decision making what to do about it. European Journal of Operation Research. 2006; 174(1): 349-359

[14]Pichai T, Weeradej C, and Kittipol A. Development of road safety audit in Thailand. Journal of the Eastern Asia Society for Transportation Studies. 1999; Vol3: 175-186.

[15] Pichai T, Yordphol T, et al. Implementing road safety audit in Thailand. Journal of the Eastern Asia Society for Transportation Studies. 2003; Vol5: 2650-2663.

[16] Road Traffic Management Corporation. South African Road Safety Audit Manual. 2012: 1-109.

[17] Thomas L. Saaty. Decision-making with the AHP Why is the principal eigenvector necessary. European Journal of Operational Research, 2003; 145(1), 85–91.

[18] Thomas L, Saaty,Mujgan S. Ozdemir. Negative Priorities in the Analytic Hierarchy Process.Mathematical and Computer Modelling, 2003; vol.37 1063-1075.

[19] Saaty, T. L, et al. The Analytic Hirarchy Process and human resource allocation: Half the story. Mathematics and Computer modeling. 2007: 1041-1053.

[20] Saaty, T. L. Decision making with the analytic hierarchy process Int. J. Services Sciences. 2008; Vol1:83-98.

[21] Thomas L. Saatty. Rank from comparisons and from ratings in the analytic hierarchy/network processes. European Journal of Operational Research 168, 2006, 227-570.

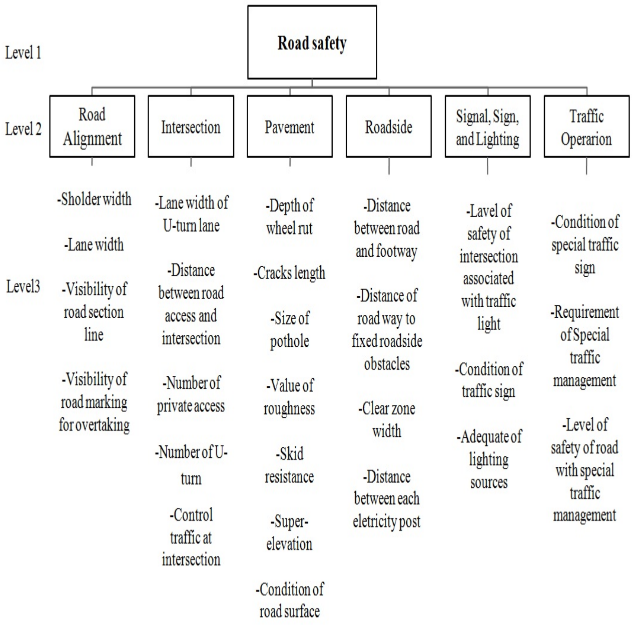
[22] Thomas L. Saaty. The analytic hierarchy process without the theory of oskarperron. International Journal of the Analytic Hierarchy Process.2013; 5(2): 268-283.

[23] Veeris A, Thanwadee C, et al. Selection of weigh station locations in Thailand using the analytic hierarchy process [Dissertation].Bangkok; Sirindhorn International Institute of Technology, Thammasat Univ.

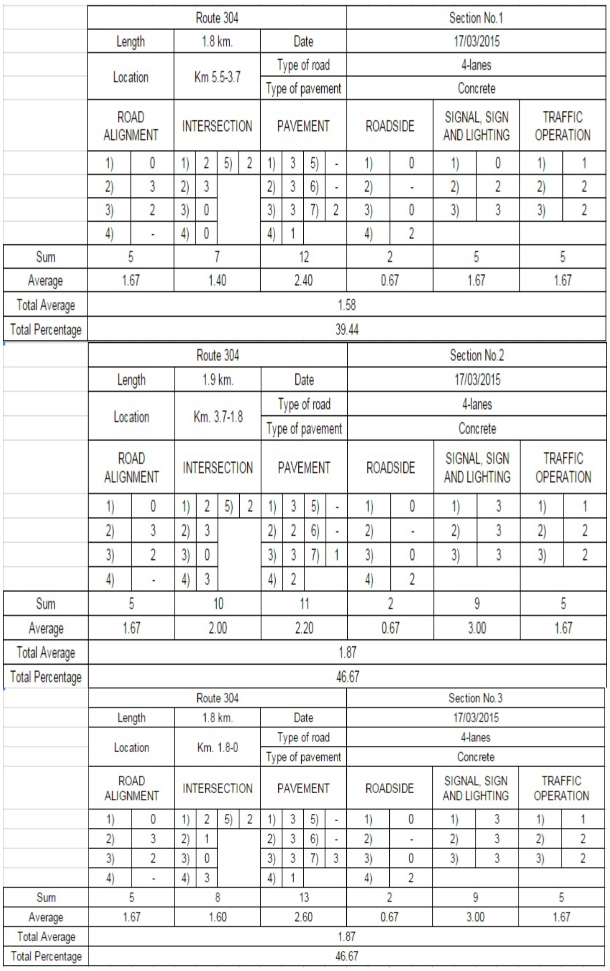
[24] Vincent SL, Bo KW, et al. Group decision making in a multiple criteria environment: A case using the AHP in software selection. European Journal of Operational Research. 2002: 134-144.

## 7. Appendices

### Appendix A. Analytic Hierarchy Process Model


Road safety
Level 1
Level 2
Road Alignment
Intersection
Pavement
Roadside
Signal, Sign, and Lighting
Traffic Operarion
Level3
-Sholder width
-Lane width
-Visibility of road section line
-Visibility of road marking for overtaking
-Lane width of U-turn lane
-Distance between road access and intersection
-Number of private access
-Number of U-turn
-Control traffic at intersection
-Depth of wheel rut
-Cracks length
-Size of pothole
-Value of roughness
-Skid resistance
-Super-elevation
-Condition of road surface
-Distance between road and footway
-Distance of road way to fixed roadside obstacles
-Clear zone width
-Distance between each eletricity post
-Level of safety of intersection associated with traffic light
-Condition of traffic sign
-Adequate of lighting sources
-Condition of special traffic sign
-Requirement of Special traffic management
-Level of safety of road with special traffic management

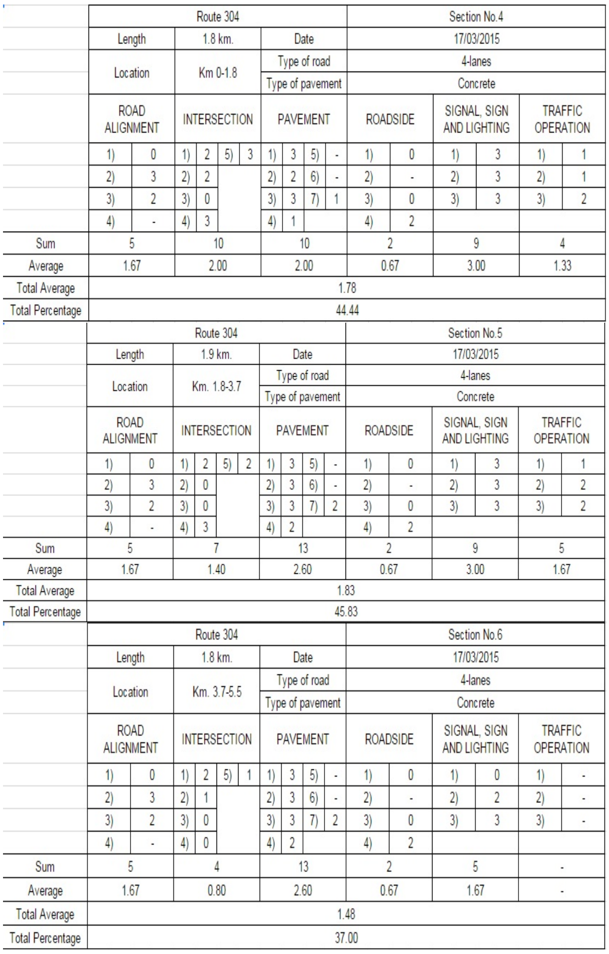
### Appendix B. Result sheet of Road Safety Index



| | Route 304 | | | | | | | | | | Section No.1 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Length | | 1.8 km. | | | | Date | | | | 17/03/2015 | | | | | |
| | Location | | Km 5.5-3.7 | | | | Type of road | | | | 4-lanes | | | | | |
| | | | | | | | Type of pavement | | | | Concrete | | | | | |
| | ROAD ALIGNMENT | | INTERSECTION | | | | PAVEMENT | | | | ROADSIDE | | SIGNAL, SIGN AND LIGHTING | | TRAFFIC OPERATION | |
| | 1) | 0 | 1) | 2 | 5) | 2 | 1) | 3 | 5) | - | 1) | 0 | 1) | 0 | 1) | 1 |
| | 2) | 3 | 2) | 3 | | | 2) | 3 | 6) | - | 2) | - | 2) | 2 | 2) | 2 |
| | 3) | 2 | 3) | 0 | | | 3) | 3 | 7) | 2 | 3) | 0 | 3) | 3 | 3) | 2 |
| | 4) | - | 4) | 0 | | | 4) | 1 | | | 4) | 2 | | | | |
| Sum | 5 | | 7 | | | | 12 | | | | 2 | | 5 | | 5 | |
| Average | 1.67 | | 1.40 | | | | 2.40 | | | | 0.67 | | 1.67 | | 1.67 | |
| Total Average | 1.58 | | | | | | | | | | | | | | | |
| Total Percentage | 39.44 | | | | | | | | | | | | | | | |

| | Route 304 | | | | | | | | | | Section No.2 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Length | | 1.9 km. | | | | Date | | | | 17/03/2015 | | | | | |
| | Location | | Km. 3.7-1.8 | | | | Type of road | | | | 4-lanes | | | | | |
| | | | | | | | Type of pavement | | | | Concrete | | | | | |
| | ROAD ALIGNMENT | | INTERSECTION | | | | PAVEMENT | | | | ROADSIDE | | SIGNAL, SIGN AND LIGHTING | | TRAFFIC OPERATION | |
| | 1) | 0 | 1) | 2 | 5) | 2 | 1) | 3 | 5) | - | 1) | 0 | 1) | 3 | 1) | 1 |
| | 2) | 3 | 2) | 3 | | | 2) | 2 | 6) | - | 2) | - | 2) | 3 | 2) | 2 |
| | 3) | 2 | 3) | 0 | | | 3) | 3 | 7) | 1 | 3) | 0 | 3) | 3 | 3) | 2 |
| | 4) | - | 4) | 3 | | | 4) | 2 | | | 4) | 2 | | | | |
| Sum | 5 | | 10 | | | | 11 | | | | 2 | | 9 | | 5 | |
| Average | 1.67 | | 2.00 | | | | 2.20 | | | | 0.67 | | 3.00 | | 1.67 | |
| Total Average | 1.87 | | | | | | | | | | | | | | | |
| Total Percentage | 46.67 | | | | | | | | | | | | | | | |

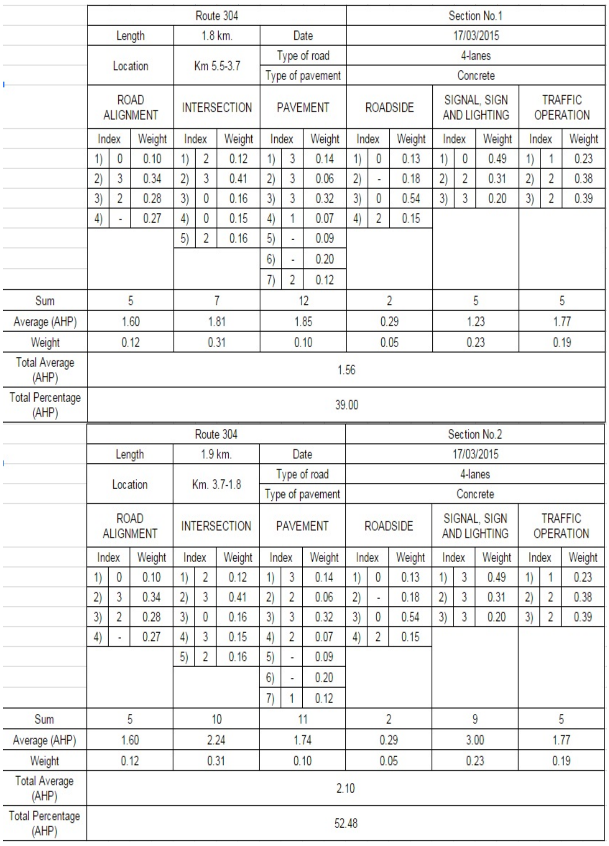
| | Route 304 | | | | | | | | | | Section No.3 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Length | | 1.8 km. | | | | Date | | | | 17/03/2015 | | | | | |
| | Location | | Km. 1.8-0 | | | | Type of road | | | | 4-lanes | | | | | |
| | | | | | | | Type of pavement | | | | Concrete | | | | | |
| | ROAD ALIGNMENT | | INTERSECTION | | | | PAVEMENT | | | | ROADSIDE | | SIGNAL, SIGN AND LIGHTING | | TRAFFIC OPERATION | |
| | 1) | 0 | 1) | 2 | 5) | 2 | 1) | 3 | 5) | - | 1) | 0 | 1) | 3 | 1) | 1 |
| | 2) | 3 | 2) | 1 | | | 2) | 3 | 6) | - | 2) | - | 2) | 3 | 2) | 2 |
| | 3) | 2 | 3) | 0 | | | 3) | 3 | 7) | 3 | 3) | 0 | 3) | 3 | 3) | 2 |
| | 4) | - | 4) | 3 | | | 4) | 1 | | | 4) | 2 | | | | |
| Sum | 5 | | 8 | | | | 13 | | | | 2 | | 9 | | 5 | |
| Average | 1.67 | | 1.60 | | | | 2.60 | | | | 0.67 | | 3.00 | | 1.67 | |
| Total Average | 1.87 | | | | | | | | | | | | | | | |
| Total Percentage | 46.67 | | | | | | | | | | | | | | | |



| | Route 304 | | | | | | | | | | Section No.4 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Length | | 1.8 km. | | | | Date | | | | 17/03/2015 | | | | | |
| | Location | | Km 0-1.8 | | | | Type of road | | | | 4-lanes | | | | | |
| | | | | | | | Type of pavement | | | | Concrete | | | | | |
| | ROAD ALIGNMENT | | INTERSECTION | | | | PAVEMENT | | | | ROADSIDE | | SIGNAL, SIGN AND LIGHTING | | TRAFFIC OPERATION | |
| | 1) | 0 | 1) | 2 | 5) | 3 | 1) | 3 | 5) | - | 1) | 0 | 1) | 3 | 1) | 1 |
| | 2) | 3 | 2) | 2 | | | 2) | 2 | 6) | - | 2) | - | 2) | 3 | 2) | 1 |
| | 3) | 2 | 3) | 0 | | | 3) | 3 | 7) | 1 | 3) | 0 | 3) | 3 | 3) | 2 |
| | 4) | - | 4) | 3 | | | 4) | 1 | | | 4) | 2 | | | | |
| Sum | 5 | | 10 | | | | 10 | | | | 2 | | 9 | | 4 | |
| Average | 1.67 | | 2.00 | | | | 2.00 | | | | 0.67 | | 3.00 | | 1.33 | |
| Total Average | 1.78 | | | | | | | | | | | | | | | |
| Total Percentage | 44.44 | | | | | | | | | | | | | | | |

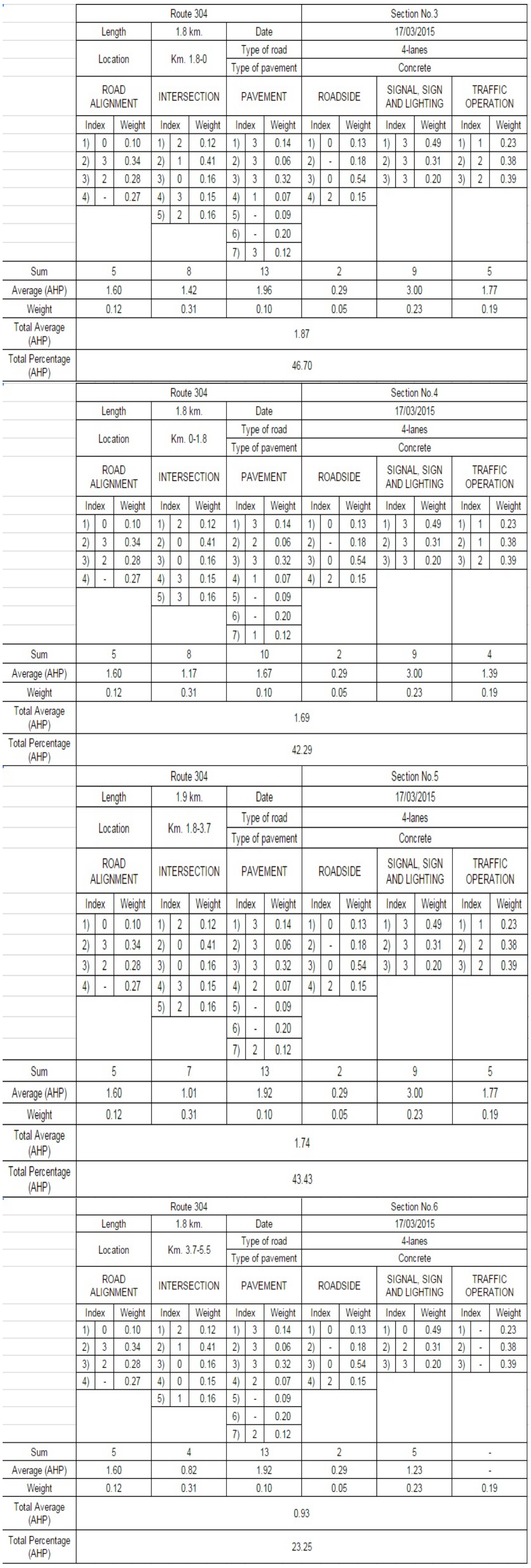
| | Route 304 | | | | | | | | | | Section No.5 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Length | | 1.9 km. | | | | Date | | | | 17/03/2015 | | | | | |
| | Location | | Km. 1.8-3.7 | | | | Type of road | | | | 4-lanes | | | | | |
| | | | | | | | Type of pavement | | | | Concrete | | | | | |
| | ROAD ALIGNMENT | | INTERSECTION | | | | PAVEMENT | | | | ROADSIDE | | SIGNAL, SIGN AND LIGHTING | | TRAFFIC OPERATION | |
| | 1) | 0 | 1) | 2 | 5) | 2 | 1) | 3 | 5) | - | 1) | 0 | 1) | 3 | 1) | 1 |
| | 2) | 3 | 2) | 0 | | | 2) | 3 | 6) | - | 2) | - | 2) | 3 | 2) | 2 |
| | 3) | 2 | 3) | 0 | | | 3) | 3 | 7) | 2 | 3) | 0 | 3) | 3 | 3) | 2 |
| | 4) | - | 4) | 3 | | | 4) | 2 | | | 4) | 2 | | | | |
| Sum | 5 | | 7 | | | | 13 | | | | 2 | | 9 | | 5 | |
| Average | 1.67 | | 1.40 | | | | 2.60 | | | | 0.67 | | 3.00 | | 1.67 | |
| Total Average | 1.83 | | | | | | | | | | | | | | | |
| Total Percentage | 45.83 | | | | | | | | | | | | | | | |

| | Route 304 | | | | | | | | | | Section No.6 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Length | | 1.8 km. | | | | Date | | | | 17/03/2015 | | | | | |
| | Location | | Km. 3.7-5.5 | | | | Type of road | | | | 4-lanes | | | | | |
| | | | | | | | Type of pavement | | | | Concrete | | | | | |
| | ROAD ALIGNMENT | | INTERSECTION | | | | PAVEMENT | | | | ROADSIDE | | SIGNAL, SIGN AND LIGHTING | | TRAFFIC OPERATION | |
| | 1) | 0 | 1) | 2 | 5) | 1 | 1) | 3 | 5) | - | 1) | 0 | 1) | 0 | 1) | - |
| | 2) | 3 | 2) | 1 | | | 2) | 3 | 6) | - | 2) | - | 2) | 2 | 2) | - |
| | 3) | 2 | 3) | 0 | | | 3) | 3 | 7) | 2 | 3) | 0 | 3) | 3 | 3) | - |
| | 4) | - | 4) | 0 | | | 4) | 2 | | | 4) | 2 | | | | |
| Sum | 5 | | 4 | | | | 13 | | | | 2 | | 5 | | - | |
| Average | 1.67 | | 0.80 | | | | 2.60 | | | | 0.67 | | 1.67 | | - | |
| Total Average | 1.48 | | | | | | | | | | | | | | | |
| Total Percentage | 37.00 | | | | | | | | | | | | | | | |



| | Route 304 | | | | | | | | | Section No.1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Length | | | 1.8 km. | | | Date | | | 17/03/2015 | | | | | | | | |
| | Location | | | Km 5.5-3.7 | | | Type of road | | | 4-lanes | | | | | | | | |
| | | | | | | | Type of pavement | | | Concrete | | | | | | | | |
| | ROAD ALIGNMENT | | | INTERSECTION | | | PAVEMENT | | | ROADSIDE | | | SIGNAL, SIGN AND LIGHTING | | | TRAFFIC OPERATION | | |
| | Index | | Weight | Index | | Weight | Index | | Weight | Index | | Weight | Index | | Weight | Index | | Weight |
| | 1) | 0 | 0.10 | 1) | 2 | 0.12 | 1) | 3 | 0.14 | 1) | 0 | 0.13 | 1) | 0 | 0.49 | 1) | 1 | 0.23 |
| | 2) | 3 | 0.34 | 2) | 3 | 0.41 | 2) | 3 | 0.06 | 2) | - | 0.18 | 2) | 2 | 0.31 | 2) | 2 | 0.38 |
| | 3) | 2 | 0.28 | 3) | 0 | 0.16 | 3) | 3 | 0.32 | 3) | 0 | 0.54 | 3) | 3 | 0.20 | 3) | 2 | 0.39 |
| | 4) | - | 0.27 | 4) | 0 | 0.15 | 4) | 1 | 0.07 | 4) | 2 | 0.15 | | | | | | |
| | | | | 5) | 2 | 0.16 | 5) | - | 0.09 | | | | | | | | | |
| | | | | | | | 6) | - | 0.20 | | | | | | | | | |
| | | | | | | | 7) | 2 | 0.12 | | | | | | | | | |
| Sum | 5 | | | 7 | | | 12 | | | 2 | | | 5 | | | 5 | | |
| Average (AHP) | 1.60 | | | 1.81 | | | 1.85 | | | 0.29 | | | 1.23 | | | 1.77 | | |
| Weight | 0.12 | | | 0.31 | | | 0.10 | | | 0.05 | | | 0.23 | | | 0.19 | | |
| Total Average (AHP) | 1.56 | | | | | | | | | | | | | | | | | |
| Total Percentage (AHP) | 39.00 | | | | | | | | | | | | | | | | | |

| | Route 304 | | | | | | | | | Section No.2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Length | | | 1.9 km. | | | Date | | | 17/03/2015 | | | | | | | | |
| | Location | | | Km. 3.7-1.8 | | | Type of road | | | 4-lanes | | | | | | | | |
| | | | | | | | Type of pavement | | | Concrete | | | | | | | | |
| | ROAD ALIGNMENT | | | INTERSECTION | | | PAVEMENT | | | ROADSIDE | | | SIGNAL, SIGN AND LIGHTING | | | TRAFFIC OPERATION | | |
| | Index | | Weight | Index | | Weight | Index | | Weight | Index | | Weight | Index | | Weight | Index | | Weight |
| | 1) | 0 | 0.10 | 1) | 2 | 0.12 | 1) | 3 | 0.14 | 1) | 0 | 0.13 | 1) | 3 | 0.49 | 1) | 1 | 0.23 |
| | 2) | 3 | 0.34 | 2) | 3 | 0.41 | 2) | 2 | 0.06 | 2) | - | 0.18 | 2) | 3 | 0.31 | 2) | 2 | 0.38 |
| | 3) | 2 | 0.28 | 3) | 0 | 0.16 | 3) | 3 | 0.32 | 3) | 0 | 0.54 | 3) | 3 | 0.20 | 3) | 2 | 0.39 |
| | 4) | - | 0.27 | 4) | 3 | 0.15 | 4) | 2 | 0.07 | 4) | 2 | 0.15 | | | | | | |
| | | | | 5) | 2 | 0.16 | 5) | - | 0.09 | | | | | | | | | |
| | | | | | | | 6) | - | 0.20 | | | | | | | | | |
| | | | | | | | 7) | 1 | 0.12 | | | | | | | | | |
| Sum | 5 | | | 10 | | | 11 | | | 2 | | | 9 | | | 5 | | |
| Average (AHP) | 1.60 | | | 2.24 | | | 1.74 | | | 0.29 | | | 3.00 | | | 1.77 | | |
| Weight | 0.12 | | | 0.31 | | | 0.10 | | | 0.05 | | | 0.23 | | | 0.19 | | |
| Total Average (AHP) | 2.10 | | | | | | | | | | | | | | | | | |
| Total Percentage (AHP) | 52.48 | | | | | | | | | | | | | | | | | |



| | Route 304 | | | | | | | | | Section No.3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Length | | | 1.8 km. | | | Date | | | 17/03/2015 | | | | | | | | |
| | Location | | | Km. 1.8-0 | | | Type of road | | | 4-lanes | | | | | | | | |
| | | | | | | | Type of pavement | | | Concrete | | | | | | | | |
| | ROAD ALIGNMENT | | | INTERSECTION | | | PAVEMENT | | | ROADSIDE | | | SIGNAL, SIGN AND LIGHTING | | | TRAFFIC OPERATION | | |
| | Index | | Weight | Index | | Weight | Index | | Weight | Index | | Weight | Index | | Weight | Index | | Weight |
| | 1) | 0 | 0.10 | 1) | 2 | 0.12 | 1) | 3 | 0.14 | 1) | 0 | 0.13 | 1) | 3 | 0.49 | 1) | 1 | 0.23 |
| | 2) | 3 | 0.34 | 2) | 1 | 0.41 | 2) | 3 | 0.06 | 2) | - | 0.18 | 2) | 3 | 0.31 | 2) | 2 | 0.38 |
| | 3) | 2 | 0.28 | 3) | 0 | 0.16 | 3) | 3 | 0.32 | 3) | 0 | 0.54 | 3) | 3 | 0.20 | 3) | 2 | 0.39 |
| | 4) | - | 0.27 | 4) | 3 | 0.15 | 4) | 1 | 0.07 | 4) | 2 | 0.15 | | | | | | |
| | | | | 5) | 2 | 0.16 | 5) | - | 0.09 | | | | | | | | | |
| | | | | | | | 6) | - | 0.20 | | | | | | | | | |
| | | | | | | | 7) | 3 | 0.12 | | | | | | | | | |
| Sum | 5 | | | 8 | | | 13 | | | 2 | | | 9 | | | 5 | | |
| Average (AHP) | 1.60 | | | 1.42 | | | 1.96 | | | 0.29 | | | 3.00 | | | 1.77 | | |
| Weight | 0.12 | | | 0.31 | | | 0.10 | | | 0.05 | | | 0.23 | | | 0.19 | | |
| Total Average (AHP) | 1.87 | | | | | | | | | | | | | | | | | |
| Total Percentage (AHP) | 46.70 | | | | | | | | | | | | | | | | | |

| | Route 304 | | | | | | | | | Section No.4 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Length | | | 1.8 km. | | | Date | | | 17/03/2015 | | | | | | | | |
| | Location | | | Km. 0-1.8 | | | Type of road | | | 4-lanes | | | | | | | | |
| | | | | | | | Type of pavement | | | Concrete | | | | | | | | |
| | ROAD ALIGNMENT | | | INTERSECTION | | | PAVEMENT | | | ROADSIDE | | | SIGNAL, SIGN AND LIGHTING | | | TRAFFIC OPERATION | | |
| | Index | | Weight | Index | | Weight | Index | | Weight | Index | | Weight | Index | | Weight | Index | | Weight |
| | 1) | 0 | 0.10 | 1) | 2 | 0.12 | 1) | 3 | 0.14 | 1) | 0 | 0.13 | 1) | 3 | 0.49 | 1) | 1 | 0.23 |
| | 2) | 3 | 0.34 | 2) | 0 | 0.41 | 2) | 2 | 0.06 | 2) | - | 0.18 | 2) | 3 | 0.31 | 2) | 1 | 0.38 |
| | 3) | 2 | 0.28 | 3) | 0 | 0.16 | 3) | 3 | 0.32 | 3) | 0 | 0.54 | 3) | 3 | 0.20 | 3) | 2 | 0.39 |
| | 4) | - | 0.27 | 4) | 3 | 0.15 | 4) | 1 | 0.07 | 4) | 2 | 0.15 | | | | | | |
| | | | | 5) | 3 | 0.16 | 5) | - | 0.09 | | | | | | | | | |
| | | | | | | | 6) | - | 0.20 | | | | | | | | | |
| | | | | | | | 7) | 1 | 0.12 | | | | | | | | | |
| Sum | 5 | | | 8 | | | 10 | | | 2 | | | 9 | | | 4 | | |
| Average (AHP) | 1.60 | | | 1.17 | | | 1.67 | | | 0.29 | | | 3.00 | | | 1.39 | | |
| Weight | 0.12 | | | 0.31 | | | 0.10 | | | 0.05 | | | 0.23 | | | 0.19 | | |
| Total Average (AHP) | 1.69 | | | | | | | | | | | | | | | | | |
| Total Percentage (AHP) | 42.29 | | | | | | | | | | | | | | | | | |

| | Route 304 | | | | | | | | | Section No.5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Length | | | 1.9 km. | | | Date | | | 17/03/2015 | | | | | | | | |
| | Location | | | Km. 1.8-3.7 | | | Type of road | | | 4-lanes | | | | | | | | |
| | | | | | | | Type of pavement | | | Concrete | | | | | | | | |
| | ROAD ALIGNMENT | | | INTERSECTION | | | PAVEMENT | | | ROADSIDE | | | SIGNAL, SIGN AND LIGHTING | | | TRAFFIC OPERATION | | |
| | Index | | Weight | Index | | Weight | Index | | Weight | Index | | Weight | Index | | Weight | Index | | Weight |
| | 1) | 0 | 0.10 | 1) | 2 | 0.12 | 1) | 3 | 0.14 | 1) | 0 | 0.13 | 1) | 3 | 0.49 | 1) | 1 | 0.23 |
| | 2) | 3 | 0.34 | 2) | 0 | 0.41 | 2) | 3 | 0.06 | 2) | - | 0.18 | 2) | 3 | 0.31 | 2) | 2 | 0.38 |
| | 3) | 2 | 0.28 | 3) | 0 | 0.16 | 3) | 3 | 0.32 | 3) | 0 | 0.54 | 3) | 3 | 0.20 | 3) | 2 | 0.39 |
| | 4) | - | 0.27 | 4) | 3 | 0.15 | 4) | 2 | 0.07 | 4) | 2 | 0.15 | | | | | | |
| | | | | 5) | 2 | 0.16 | 5) | - | 0.09 | | | | | | | | | |
| | | | | | | | 6) | - | 0.20 | | | | | | | | | |
| | | | | | | | 7) | 2 | 0.12 | | | | | | | | | |
| Sum | 5 | | | 7 | | | 13 | | | 2 | | | 9 | | | 5 | | |
| Average (AHP) | 1.60 | | | 1.01 | | | 1.92 | | | 0.29 | | | 3.00 | | | 1.77 | | |
| Weight | 0.12 | | | 0.31 | | | 0.10 | | | 0.05 | | | 0.23 | | | 0.19 | | |
| Total Average (AHP) | 1.74 | | | | | | | | | | | | | | | | | |
| Total Percentage (AHP) | 43.43 | | | | | | | | | | | | | | | | | |

| | Route 304 | | | | | | | | | Section No.6 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Length | | | 1.8 km. | | | Date | | | 17/03/2015 | | | | | | | | |
| | Location | | | Km. 3.7-5.5 | | | Type of road | | | 4-lanes | | | | | | | | |
| | | | | | | | Type of pavement | | | Concrete | | | | | | | | |
| | ROAD ALIGNMENT | | | INTERSECTION | | | PAVEMENT | | | ROADSIDE | | | SIGNAL, SIGN AND LIGHTING | | | TRAFFIC OPERATION | | |
| | Index | | Weight | Index | | Weight | Index | | Weight | Index | | Weight | Index | | Weight | Index | | Weight |
| | 1) | 0 | 0.10 | 1) | 2 | 0.12 | 1) | 3 | 0.14 | 1) | 0 | 0.13 | 1) | 0 | 0.49 | 1) | - | 0.23 |
| | 2) | 3 | 0.34 | 2) | 1 | 0.41 | 2) | 3 | 0.06 | 2) | - | 0.18 | 2) | 2 | 0.31 | 2) | - | 0.38 |
| | 3) | 2 | 0.28 | 3) | 0 | 0.16 | 3) | 3 | 0.32 | 3) | 0 | 0.54 | 3) | 3 | 0.20 | 3) | - | 0.39 |
| | 4) | - | 0.27 | 4) | 0 | 0.15 | 4) | 2 | 0.07 | 4) | 2 | 0.15 | | | | | | |
| | | | | 5) | 1 | 0.16 | 5) | - | 0.09 | | | | | | | | | |
| | | | | | | | 6) | - | 0.20 | | | | | | | | | |
| | | | | | | | 7) | 2 | 0.12 | | | | | | | | | |
| Sum | 5 | | | 4 | | | 13 | | | 2 | | | 5 | | | - | | |
| Average (AHP) | 1.60 | | | 0.82 | | | 1.92 | | | 0.29 | | | 1.23 | | | - | | |
| Weight | 0.12 | | | 0.31 | | | 0.10 | | | 0.05 | | | 0.23 | | | 0.19 | | |
| Total Average (AHP) | 0.93 | | | | | | | | | | | | | | | | | |
| Total Percentage (AHP) | 23.25 | | | | | | | | | | | | | | | | | |

# Measuring and Benchmarking the Efficiency of Solid Waste Removal in Khung Bang Krachao

TheerawanChansang, PhanitaYansivamoke*, AnyamaneePalagavong Na Ayudhya, Aunchaleeporn Manovachirasun, and Morrakot Raweewan
*School of Management Technology, Sirindhorn International Institute of Technology,*
*Thammasat University, PathumThani, 12000, THAILAND*
*Email Address: nannamenam@gmail.com*

## Abstract

With its unique culture and natural landscape, Khung Bang Krachao has become a popular tourist destination near Bangkok in SamutPrakan Province, central Thailand. The popularity has brought people to the region and activated the use of resources; as a result, municipal solid waste has become an environmental concern for local residents and environmentalists. The purpose of this study is to assess and benchmark municipal solid waste removal in Khung Bang Krachao in all six subdistricts – Bang Krachao, Bang Kasorb, Bang Korbua, Bang Nampeang, Bang Yor, and Song Kanong. Data including an amount of solid waste, a size of population, a number of households, budget and other resources for solid waste removal program such as a number of garbage collectors, a number of trash pickup trucks, and a number of trash wheeled bins have been collected for the analysis. The Data Envelopment Analysis (DEA) is employed to assess and benchmark undesirable outputs (an amount of solid waste) given multiple inputs. The most efficient Data Making Unit (DMU) (i.e. Subdistrict Administrative Organization or SAO) is identified as the best practice. Slacks of the models suggest a target forinefficient DMU to improve a solid waste removal program.

*Keywords*: Performance Measurement, DEA, Khung Bang Krachao, Subdistrict Administrative Organization

*Corresponding Author
E-mail Address: *nannamenam@gmail.com*

## 1. Introduction

The solid waste management is one of the most important problem of a Subdistrict Administrative Organization (SAO).It is a key utility service for the public health and the external "image" of SAO. On one hand, uncollected solid waste is a serious public health issue in many developing countries [13]. On the other hand, a clean city is an attractive to tourists, business, population and investors. It has been suggested that the effectiveness of a SAO solid waste management system can be used as a proxy indicator of good governance [11] and hence of a local government who can be trusted and with whom can do business. Appropriate benchmark indicators allow a city to judge its own performance regarding delivery of solid waste management services; provide information for decision-making on priorities for the limited budget available for service improvements.

Nowadays, wastes came from many directions and sources such as increasing population, agriculture, industrial and manufacturing plants, office buildings, households, and nature. Either as solid or liquid, these materials follow a variety of routes toward specific disposal sites [2]. Thus, solid waste management has become a considerable issue, in addition to other environmental problems, especially for densely populated cities in developing countries [10]. Solid waste management is one of the most difficult environmental problems in every country. Rapid urban growth, accompanied by the increasing density of population, traffic jam, air and water population, increasing per capita generation of solid waste and the lack of land conveniently situated for waste disposal are difficulties of solid waste management in urban centers [12]. SAO solid waste management refers to collection, transfer, and disposal of solid waste [4].

Khung Bang Krachao has become a developing and popular tourist destination near Bangkok in SamutPrakan Province, central Thailand. The popularity has brought people to the region and activated the use of resources; as a result, SAO solid waste has become an environmental concern for local residents and environmentalists. The cumulative solid waste at Khung Bang Kachao is 18 tons per day [9]. This is a big problem at Khung Bang Krachao. The area now has focused on solid waste collection and environmental sustainability.

Shown in Figure 2, Khung Bang Krachao consists of six subdistricts – Bang Yor, Song Kanong, Bang Nampeang, Bang Krachao, Bang Korbua, and Bang Kasorb. SAOs have tried to manage solid waste system by using three activities: reduce, reuse, and recycle (3R). The objective of conducting this paper is to benchmark efficiency of solid waste removal of six SAOs in Khung Bang Krachao by using Data Envelopment Analysis (DEA). Data in year 2012 to 2014 is collected for an analysis.

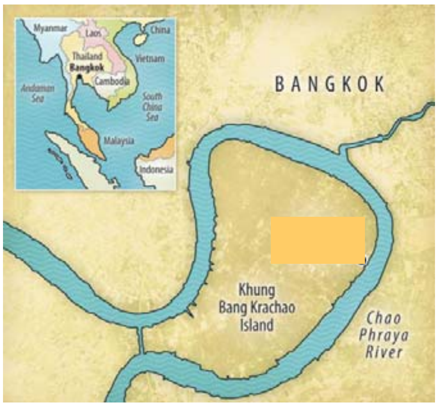


Figure 1: Map of Khung Bang Krachao [3]

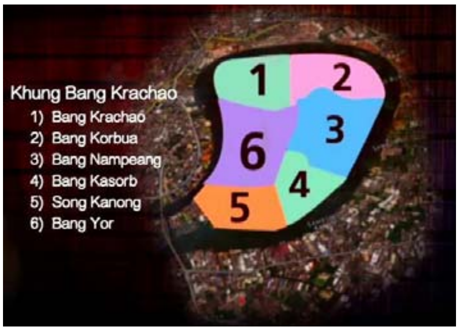


Figure 2: Khung Bang Krachao has six subdistricts [9]

*1.1. Subdistrict Administrative Organization waste collection*

Solid waste collection is started from collecting waste from household, business (retail store and grocery shop, supermarket), then loading, unloading and transferring to a final destination such as a disposal site [7]. Dual collection systems separating waste in reduce, re-use and recycling resource for non-recyclable waste can entail multi-compartmentalized vehicles or separate collection rounds on the same or different days [6]. Transportation costs are very important in SAO waste management collection. Many SAOs try to meet their legal mandate of providing a weekly waste collection service to all households due to limited budgets [6]. Solid waste collections in most developing countries do not satisfy the full demand in urban areas. In the poorest countries, the service sometimes reaches only 10 percent to 40 percent of the urban population; in the better-organized middle-income countries, the services reach 50 percent to 85 percent of the urban population [8].

*1.2. Assessing and benchmarking SAO solid waste collection*

Data Envelopment Analysis (DEA) is often used to benchmark an efficiency of the peer group to improve performance. DEA has used in benchmarking waste collection in many countries, see [14] for example. In an output-oriented model, DEA maximizes output while inputs are fixed. But an amount of solid waste, that is an output, should not be maximized. Thus, it is treated as undesirable output in data transition before using DEA. Here, our analysis uses variable returns to scale (VRS) models because the ratio between input and output is not constant.

## 2. Research Methodology

*2.1 The basic DEA model and data envelopment analysis model*

Data Envelopment Analysis (DEA) model developed by Charnes et. al [1] to measure the relative efficiency where market price are not available and the performance of Decision Making Units (DMUs) that convert model multiple-input and multiple-output relationships without a priori underlying functional form assumption. DEA has also been widely applied, for example, bank performance, agricultural economics, production planning, hospital, and web security. In particular, DEA is an excellent tool for improving the productivity of service businesses [15]. We adopted DEA model to measure the efficiency of the solid waste removal of SAO at Khung Bang Krachao. The following DEA model is an input-oriented model where the inputs are minimized and the outputs are kept at their current levels [15], has been used for this analysis. The model can be stated as:

$$\min \theta - \varepsilon(\sum_{i=1}^{m} S_i^- + \sum_{r=1}^{s} S_r^+)$$

Subject to

$$\sum_{j=1}^{n} \lambda_j x_{ij} + S_i^- = \theta x_{io} i = 1,2,\dots,m;$$

$$\sum_{j=1}^{n} \lambda_{rj} y_{rj} - S_r^+ = y_{ro} r = 1,2,\dots,s; \qquad (1)$$

$$\sum_{j=1}^{n} \lambda_j = 1$$

$$\lambda_j \geq 0 \qquad j = 1,2,\dots,n.$$

Where:
$\theta$= input technical efficiency measure,
$\varepsilon$ = a non-Archimedean number,
$\lambda_j$=efficiency measure to be calculated for each DMUj,
$y_{rj}$= quantityof output r produced by subdistrict j,
$x_{ij}$ = quantity of input i produced by subdistrictj.

Thus, Formula (1) is using the minimum inputs resource to receive a given resource of outputs. Efficiency is measured by the maximum reduction in inputs which will still allow a given output resource which allows the minimization over $\theta$ to preempt the optimization involving the slack $S_i^-$and $S_r^+$ [15].

For input oriented, the DMUs that shows the result that if $\theta^* = 1$ and$S_i^{-^*} = S_r^{+^*} = 0$ for all $i$ and $r$ that means efficient and if $\theta^* = 1$ and $S_i^{-^*} \neq 0$ and $S_r^{+^*} \neq 0$ for some *i* and *r* that means weakly efficient.

$$S_r^- = \theta^* X_{io} - \sum_{j=1}^{n} \lambda_j x_{ij} i = 1,2,\dots,m;$$

$$S_r^+ = \sum_{j=1}^{n} \lambda_j y_{ij} - y_{ro} r = 1,2,\dots,s;$$

Next, the output oriented can be computed by formula (2), the model can be stated as:

$$\max \emptyset - \varepsilon(\sum_{i=1}^{m} S_i^- + \sum_{r=1}^{s} S_r^+)$$

Subject to (2)

$$\sum_{j=1}^{n} \lambda_j x_{ij} + S_i^- = x_{io} i = 1,2,\dots,m;$$

$$\sum_{j=1}^{n} \lambda_{rj} y_{rj} - S_r^+ = \phi y_{ro} r = 1,2,\dots,s;$$

$$\sum_{j=1}^{n} \lambda_j = 1$$

$$\lambda_j \geq 0 \qquad j = 1,2,\dots,n.$$

Where:
$\emptyset$= output technical efficiency measure,
$\varepsilon$ = a non-Archimedean number,
$\lambda_j$= efficiency measure to be calculated for each DMUj,
$y_{rj}$= quantityof output r produced by subdistrict j,
$x_{ij}$ = quantity of input i produced by subdistrict j.

The optimization involving the slack $S_i^-$and $S_r^+$ which is DMUs is efficient if $\emptyset = 1$ and$S_i^{-^*} = S_r^{+^*} = 0$ for all $i$ and $r$ that means efficient and if $\emptyset^* = 1$ and $S_i^{-^*} \neq 0$ and $S_r^{+^*} \neq 0$ for some $i$ and $r$ that means weakly efficient can be computed.

$$S_r^- = X_{io} - \sum_{j=1}^{n} \lambda_j x_{ij} i = 1,2,\dots,m;$$

$$S_r^+ = \sum_{j=1}^{n} \lambda_j y_{ij} - \emptyset^* y_{ro} r = 1,2,\dots,s;$$

*2.2 Modeling undesirable measures*

Reference [15] was the first introduce undesirable measures DEA model. In the case, the management of solid waste at Khung Bang Krachao in order to improve the performance of the amount of waste to be treated should be decreased rather than increased as assumed in the conventional DEA models [15].

The model undesirable output can separate in two ways, either directly or indirectly. Directly way is considered variables as inputs aim at its minimization and indirectly way with a preprocessing on data before putting into the model that can be completed by the use of the inverse additive or multiplicative inverse. When treated by the additive inverse, because the values become negative, which makes the linear programming model impracticable, the data set can be transformed, adding it to a previously chosen translation vector, turning all the elements into positive [5].

In this case waste management that mean we have to use undesirable output to solve this model in DEA because solid waste is bad output. For undesirable outputs $(y_{rj}^b)$ denote a bad output, which was to decrease the undesirable output to increasing improvement the performance. In order to multiply each undesirable output by -1 and then find a proper value $\mathrm{v_r}$ to let all negative undesirable outputs be positive. That is,$\bar{y}_{rj}^b = -y_{rj}^b + v_r > 0$. This can be achieved by $v_r = \max\{y_{rj}^b\} + 1$ [15], for example.

Max h
subject to

$$\sum_{j=1}^{n} \lambda_j y_{rj}^g \geq h y_{r0}^g$$

$$\sum_{j=1}^{n} \lambda_j \bar{y}_{rj}^b \geq h \bar{y}_{r0}^b \qquad (3)$$

$$\sum_{j=1}^{n} \lambda_j x_{ij} \geq x_{i0}$$

$$\sum_{j=1}^{n} \lambda_j = 1$$

$$\lambda_j \geq 0, \quad j = 1,\dots,n$$

## 3. Research Results and Discussion

*3.1 Data and variables*

The model of evaluating efficiency of solid waste removal in Khung Bang Krachao is based on DEA model. The DEA model introduces performance evaluation and benchmarking to positively encourage any unit to be improved. There are three models. Model 1 aims to benchmark the ratio between the amount of solid waste in each subdistrict and population and household in the areas by applying the undesirable output in an output-orientation model. Model 2.1 and 2.2 analyze how each subdistrict efficiently uses resources in solid waste removal. Models 2.1 and 2.2 use an undesirable output in input-orientation model. Table 1 shows the list of inputs and outputs. Data in year 2012 to 2014 is collected for this study.

Note that Khung Bang Krachao has six subdistricts. But Bang Krachao and Bang Kasorb work together, thus, there are only five DMUs in this study. For confidential reason, they are called DMU1 – DMU5.

Table 1: Inputs and outputs in DEA models

| DEA Model | | Variable | Units |
|---|---|---|---|
| Model 1 | Input | Population Number of household | Person Family |
| | Output | Amount of SAO solid waste | Tons per year |
| Model 2.1 | Input | Salary | Bath per year |
| | | Cost of fuel | Bath per year |
| Model 2.2 | Input | Garbage collectors | Persons per month |
| | | Number of trucks | Trucks per month |
| | | Number of wheeled bins | Number per month |
| Model 2.1 & 2.2 | Output | Amount of SAO solid waste | Tons per year |

### *3.1.1 Result of Model 1: output orientation*

Table 2: Efficiency in six subdistricts
1 = efficiency, more than 1 = inefficiency

| 2012 | | | 2013 | | | 2014 | | |
|---|---|---|---|---|---|---|---|---|
| | Efficiency | Measuring the Effectiveness | | Efficiency | Measuring the Effectiveness | | Efficiency | Measuring the Effectiveness |
| DMU 1 | 824.93 | Inefficiency | DMU 1 | 813.25 | Inefficiency | DMU 1 | 779.59 | Inefficiency |
| DMU 2 | 4.84 | Inefficiency | DMU 2 | 5.60 | Inefficiency | DMU 2 | 7.48 | Inefficiency |
| DMU 3 | 1.55 | Inefficiency | DMU 3 | 1.60 | Inefficiency | DMU 3 | 1.63 | Inefficiency |
| DMU 4 | 1.00 | Efficiency | DMU 4 | 1.00 | Efficiency | DMU 4 | 1.00 | Efficiency |
| DMU 5 | 1.20 | Inefficiency | DMU 5 | 1.29 | Inefficiency | DMU 5 | 1.23 | Inefficiency |

Model 1 If efficiency value equal to 1, DMU is efficientin the peer group. Efficiency value more than 1 means DMU is inefficient. Efficiency scores are ranked and presented as follows.

*DMU4*:It is the best practice of the group. From on-site visit and survey, we found that the subdistrict is suitable for agriculture and also attractive for tourism such as floating market and homestay. Thus, it has good program to manage solid waste.

*DMU5*:Even though it has the biggest area, the size of population and the number of households are the smallest. It has greenland and a park to attract tourists.

*DMU3*:Most of the area is for agriculture. Solid waste is created by local people.

*DMU2*: The subdistrict is the most crowded area. It has a lot of factories and industrial warehouses. The solid waste is from factories and local people.

*DMU1*: This subdistrict has both agricultural area and factory. Thus, solid waste is from both sectors and local people.

### *3.1.2 Result of Model 2.1 and 2.2: input orientation*

Model 2.1 and 2.2 are input orientation with undesireble output. If efficiency score equals to 1, DMU is efficient; less than 1 meansinefficient. However, if the efficiency score is 1 but it has slacks, this means DMU can adjust resouces to improve the management. This situation is called weakly efficient.

Table 3: Result of Model 2.1
1 = efficiency, 1* = weakly efficiency, less than 1 = inefficiency

| 2012 | | | Input Slack | | Output Slack |
|---|---|---|---|---|---|
| | Efficiency | Measuring the Effectiveness | Salary (baht per year) | Cost of fuel (baht per year) | Amount of SAO solid waste (tons per year) |
| DMU 1 | 0.61 | Inefficiency | 0.00 | 274160.67 | 682.20 |
| DMU 2 | 0.38 | Inefficiency | 0.00 | 5080.83 | 513.00 |
| DMU 3 | 0.88 | Inefficiency | 0.00 | 104659.81 | 151.30 |
| DMU 4 | 1.00 | Efficiency | 0.00 | 0.00 | 0.00 |
| DMU 5 | 1.00 | Efficiency | 0.00 | 0.00 | 0.00 |
| 2013 | | | Input Slack | | Output Slack |
| | Efficiency | Measuring the Effectiveness | Salary (baht per year) | Cost of fuel (baht per year) | Amount of SAO solid waste (tons per year) |
| DMU 1 | 0.61 | Inefficiency | 0.00 | 274160.67 | 629.30 |
| DMU 2 | 0.38 | Inefficiency | 0.00 | 5080.83 | 485.30 |
| DMU 3 | 0.88 | Inefficiency | 0.00 | 104659.81 | 123.60 |
| DMU 4 | 1.00 | Efficiency | 0.00 | 0.00 | 0.00 |
| DMU 5 | 1.00 | Efficiency | 0.00 | 0.00 | 0.00 |

| 2014 | | | Input Slack | | Output Slack |
|---|---|---|---|---|---|
| | Efficiency | Measuring the Effectiveness | Salary (baht per year) | Cost of Fuel (baht per year) | Amount of SAO solid waste (tons per year) |
| DMU 1 | 0.61 | Inefficiency | 0.00 | 274160.67 | 629.62 |
| DMU 2 | 0.38 | Inefficiency | 0.00 | 5080.83 | 526.42 |
| DMU 3 | 0.88 | Inefficiency | 0.00 | 104659.81 | 152.78 |
| DMU 4 | 1.00 | Efficiency | 0.00 | 0.00 | 0.00 |
| DMU 5 | 1.00 | Efficiency | 0.00 | 0.00 | 0.00 |

Model 2.1considered the annual costs in management. Input slacks are salary and cost of fuel;the output slack is the amount of municiple solid waste (tons per year). DMU4 and DMU5 are efficient in the peer group. DMU1, DMU2, and DMU3 are inefficient. They can improve by using suggested slacks. The slack values show how much input and undesirable output measures needed to be reduced.Data in year 2012 is used to illustrated. For example, in Table 3, DMU1 has the efficiency score equal to 0.61, the input slack of the fuel cost is 274,160.67 baht per year and the output slack of amount of solid waste is 682.20 tons per year. DMU1's fuel cost is 708,000.00 baht per year and the amount of solid waste is 1,882.80 tons per year (see Table 7 in the appendix). Therefore, Cost of Fuel – Input Slack of Fuel Cost = 708,000.00 – 274,160.67 = 433,839.33 baht per year and the Undesirable Output – Output Slack = 1,882.80 – 682.20 = 1,200.60 tons per year. This means DMU1 shoud reduce fuel cost 433,839.33 baht per year and also reduce amount of solid waste 1,200.60 tons per year. Similar calculation can be done for DMU2 and DMU3.

Table 4: Result of Model 2.2
1 = efficiency, 1* = weakly efficiency, less than 1 = inefficiency

| 2012 | | | Input Slack | | | Output Slack |
|---|---|---|---|---|---|---|
| | Efficiency | Measuring the Effectiveness | Number of garbage collectors per month | Number of trucks per month | Number of bins per month | Amount of SAO solid waste (tons per year) |
| DMU 1 | 1.00* | Weakly Efficiency | 2.00 | 0.00 | 135.00 | 823.93 |
| DMU 2 | 0.67 | Inefficiency | 2.67 | 0.33 | 0.00 | 654.73 |
| DMU 3 | 1.00* | Weakly Efficiency | 0.00 | 0.00 | 70.00 | 293.05 |
| DMU 4 | 1.00 | Efficiency | 0.00 | 0.00 | 0.00 | 0.00 |
| DMU 5 | 1.00* | Weakly Efficiency | 0.00 | 0.00 | 145.00 | 141.73 |
| 2013 | | | Input Slack | | | Output Slack |
| | Efficiency | Measuring the Effectiveness | Number of garbage collectors per month | Number of trucks per month | Number of bins per month | Amount of SAO solid waste (tons per year) |
| DMU 1 | 1.00* | Weakly Efficiency | 2.00 | 0.00 | 177.00 | 812.25 |
| DMU 2 | 0.55 | Inefficiency | 1.50 | 0.10 | 0.00 | 668.25 |
| DMU 3 | 1.00* | Weakly Efficiency | 0.00 | 0.00 | 125.00 | 306.57 |
| DMU 4 | 1.00 | Efficiency | 0.00 | 0.00 | 0.00 | 0.00 |
| DMU 5 | 1.00* | Weakly Efficiency | 0.00 | 0.00 | 189.00 | 182.98 |
| 2014 | | | Input Slack | | | Output Slack |
| | Efficiency | Measuring the Effectiveness | Number of garbage collectors per month | Number of trucks per month | Number of bins per month | Amount of SAO solid waste (tons per year) |
| DMU 1 | 1.00* | Weakly Efficiency | 2.00 | 0.00 | 181.00 | 778.59 |
| DMU 2 | 0.52 | Inefficiency | 1.19 | 0.04 | 0.00 | 675.39 |
| DMU 3 | 1.00* | Weakly Efficiency | 0.00 | 0.00 | 180.00 | 301.75 |
| DMU 4 | 1.00 | Efficiency | 0.00 | 0.00 | 0.00 | 0.00 |
| DMU 5 | 1.00* | Weakly Efficiency | 0.00 | 0.00 | 185.00 | 148.97 |

Model 2.2 investigates other resources in solid waste program – number of garbage collectors per month, number of trucks per month, and number of bins per month. The results found as follows. DMU4 is the only DMU that is efficient in the peer group. DMU2 is inefficient. The slacks suggest how much input resources and output should be decreased to reach efficiency. For example, in 2012 in Table 4, the efficiency score of DMU2 is 0.67 with slack of number of garbage collectors (2.67≈3), slack of number of trucks (0.33≈1) and the slack of output (654.73 tons per year). From Table 8 in appendix, these slacks suggest that DMU2 should have 7 (10 - 3) garbage collectors, 1 truck (2-1), and control solid waste at 1,058.87 (1,713.60 – 654.73) tons per year. DMU1, DMU3 and DMU5 are weakly efficiency, that is, the efficient score is 1 but there are slacks. Similar calculation can be done to achieve the target of improvement.

Table 5: Summary results of Model 1, Model 2.1, and Model 2.2
1 = efficiency, 1* = weakly efficiency, less than 1 = inefficiency

| | Output Oriented | | | Input Oriented | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Year | Model 1 | | | Model 2.1 | | | Model 2.2 | | |
| Efficiency | 2012 | 2013 | 2014 | 2012 | 2013 | 2014 | 2012 | 2013 | 2014 |
| DMU 1 | 824.93 | 813.25 | 779.59 | 0.61 | 0.61 | 0.61 | 1.00* | 1.00* | 1.00* |
| DMU 2 | 4.84 | 5.60 | 7.48 | 0.38 | 0.38 | 0.38 | 0.67 | 0.55 | 0.52 |
| DMU 3 | 1.55 | 1.60 | 1.63 | 0.88 | 0.88 | 0.88 | 1.00* | 1.00* | 1.00* |
| DMU 4 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| DMU 5 | 1.20 | 1.29 | 1.23 | 1.00 | 1.00 | 1.00 | 1.00* | 1.00* | 1.00* |

In summary, results from Models 1, 2.1, and 2.2 are considered together. DMU4 has the best performance among subdistricts in its peer group. All inefficient DMUs should decrease the amount of solid waste. Further to fuel cost, DMU1, DMU2, and DUM3 should reduceit.For the number of garbage collectors, DMU1 and DMU2 should decrease it. Only DMU2 should reduce the number of truck. For the number of bins, DMU1, DMU3, and DMU5 should take out some bins from its area.

## 4. Conclusion

We explore Khung Bang Krachao and use DEA to access and benchmark the effiency of solid waste removel of five DMUs (six subdistricts but two of them work together). The analysis consists of three models. Model 1 finds the ratio between amount of solid waste produced and number of population and households to see which area has good proportion of producing solid waste. Model 2.1 and 2.2 find efficiency of using resources in subdistricts to remove solid waste. Slacks of inputs and outputs are identified in order to adjust resource to improve efficiency.

The important resultstates that each SAO except the best one should reduce the amount of solid waste. So the problem of cummalative waste (18 tons per day) can be taken care. One way to reduce solid waste is to use 3R policy (reduce, reuse, and recycle). DEA idenifies DMU4 as the best practice; however, it does not mean DMU4 is the best in the country. But it is the best in the peer group. There are still room for improvement for the best practice.

## Acknowledgements

We would like to thank the staff of all Subdistrict Administrative OrganizationinKhung Bang Krachaofor their useful information.

## References

[1] A. Charnes, W.W. Cooper and E. Rhodes, "Measuring the efficiency of decision making units, "European Journal of Operational Research, vol. 2, no. 6, pp. 429-444, 1978.

[2] A. Mazculak, Waste Treatment Reducing Global Waste, Facts On File, USA, 2010.

[3] Adams J. An Ecofriendly tree house in Bangkok [Internet]. 2014 [updated 2014 Nov/Dec]. Available from: http:// www.sierraclub.org

[4] Boyd GA, Pang JX. Estimating the linkage between energy efficiency and productivity. Energy Policy 2000;28:289–96.

[5] Claudia A, Mariana R and Djalma R. Efficiency in Brazilian Refineries Under Different DEA Technologies.Engineering Business Journal Management. 2012.

[6] CSIR, Municipal Waste Management-Good Practices, CSIR, Pretoria, 2011.

[7] M.Coffey, A. Coad, Collection of Municipal Solid Waste in Developing Countries", Unit ed Nations Human Settlements Programme, Kenya,2010.

[8] S.C. Levine, Private Sector Participation in Municipal Solid Waste Management, Part II Guidance Note, SKAT, Switzerland, 2000.

[9] SpringNews.2015 [Updated 2015 Apr 10]. Avaiable from: http://www.springnews.co.th/program/documentary/kaipomkao/201433

[10] S. S. Nas, A. Bayram "Municipal Solid Waste Characteristics and Management in Gümüşhane,Turkey", Waste Management, Vol:28,No:12,2008.

[11] Whiteman, A., Smith, P. and Wilson, D.C., 2001. Waste management: an indicator of urban governance. In: Presented at UN-Habitat Global Conference on Urban Development, New York. Prepared for: UK Department for International Development (DFID). [Online] available at: <http://www.davidcwilson.com/Waste_Management_An_Indicator_of_Urban_Governance.pdf> (accessed 13.05.13).

[12] Wilson DC, Rodic L, Cowing MJ, Velis CA, Whiteman AD, Scheinberg A, Vilches R, Masterson D, Stretz J, and Oelz B. 'Wasteaware' benchmark indicators for integrated sustainable waste management in cities. Journal Waste Management Journal 2014.

[13] Wilson, D.C., Velis, C.A., Rodic, L., 2013c. Integrated sustainable waste management in developing countries. Proc. Inst. Civil Eng. Waste Res. Manage. 166 (WR2), 52–68.

[14] Yüksel H. Evaluating the Success of Waste Collection Programs of Municipalities with Data Envelopment Analysis. International Journal of Environmental Protection. 2012.

[15] Zhu J. Quantitative models for performance evaluation and benchmarking. 2nd ed. USA: Worcester Polytechnic Institute; 2009.

## Appendix: Data in year 2012 used in Model 1, Model 2.1, and Model 2.2

Table 6: Data used in Model 1

| | Input | | Undesirable output | Translating of undesirable output |
|---|---|---|---|---|
| | Population (person per year) | Number of household (Families per year) | Amount of SAO solid waste (ton per year) | Amount of SAO solid waste (tons per year) |
| DMU 1 | 8,312 | 2,495 | 1,882.80 | 1.00 |
| DMU 2 | 11,144 | 3,283 | 1713.60 | 170.20 |
| DMU 3 | 7,178 | 2,410 | 1351.92 | 531.88 |
| DMU 4 | 4,928 | 1,411 | 1058.87 | 824.93 |
| DMU 5 | 7,564 | 2,225 | 1200.60 | 683.20 |

Table 7: Data used in Model 2.1

| | Input | | Undesirable output | Translating of undesirable output |
|---|---|---|---|---|
| | Salary (baht per year) | Cost of fuel (baht per year) | Amount of SAO solid waste (tons per year) | Amount of SAO solid waste (tons per year) |
| DMU 1 | 649,800 | 708,000 | 1,882.80 | 1.00 |
| DMU 2 | 1,039,200 | 424,000 | 1713.60 | 170.20 |
| DMU 3 | 449,400 | 302,400 | 1351.92 | 531.88 |
| DMU 4 | 401,400 | 210,000 | 1058.87 | 824.93 |
| DMU 5 | 394,800 | 156,000 | 1200.60 | 683.20 |

Table 8: Data used in Model 2.2

| | Input | | | Undesirable output | Translating of undesirable output |
|---|---|---|---|---|---|
| | Number of garbage collectors per month | Number of trucks per month | Number of bins per month | Amount of SAO solid waste (ton per year) | Amount of SAO solid waste (tons per year) |
| DMU 1 | 6 | 1 | 400 | 1,882.80 | 1.00 |
| DMU 2 | 10 | 2 | 405 | 1713.60 | 170.20 |
| DMU 3 | 4 | 1 | 348 | 1351.92 | 531.88 |
| DMU 4 | 4 | 1 | 223 | 1058.87 | 824.93 |
| DMU 5 | 4 | 1 | 412 | 1200.60 | 683.20 |

# Bayesian approach for generalized Lindley distribution and its application by different loss function

P. Nanthaprut[1*], P. Khongthip[1], W. Bodhisuwan[1] and C. Pudprommarat[2]

[1] *Department of Statistics, Faculty of Science, Kasetsart University, Bangkok, 10900, Thailand, pimwarat.n@gmail.com*

[2] *Department of Science, Faculty of Science and Technology, Suan Sunantha Rajabath University, Bangkok, 10300, Thailand, chookait.pu@ssru.ac.th*

## Abstract

In this paper, we would like to estimate the parameters of the generalize Lindley (GL) distribution by using Bayesian approach with the uniform and gamma distribution. Bayesian approach is derived under the squared error, absolute error and zero-one loss functions by using Markov Chain Monte Carlo technique. Furthermore, applications to real data sets are illustrated by using R language and open Bayesian inference using Gibbs sampling (OpenBUGS) program. Next, we show estimating parameter of GL distribution by using Bayesian approach with different loss error function, log-likelihood, Akaike information criterion, Bayesian information criterion, Kolmogorov-Smirnov test and trace plot of samples for parameters.

*Keywords:* Generalize Lindley distribution, Gamma distribution, Loss functions, OpenBUGS.

[*] Corresponding author
E-mail address: pimwarat.n@gmail.com

## 1 Introduction

Inferential statistics, the distinguish inference to estimation about population is mainly separated into classical and Bayesian approach. According to the classical inference based on maximum likelihood estimator, the unknown parameter is constant. However the Bayesian approach belief in uncertain about the true value of the parameter, which consider the parameter as random variable [1]. So, the Bayesian approach is obtained from the posterior distribution which consists of prior distribution and observed data.

Bayesian approach is called Bayesian inference. The word Bayesian refer to Thomas Bayes, who proposed a special case of Bayes'theorem (Bayes'law or Bayes rule). The Bayes published in his paper 1763 and he estimated the parameter of the Bernoulli distribution and used the uniform prior distribution [2, 3]. In 1774, Laplace shown the Bayes'theorem in general form which led to widely application. In addition, many researchers used it to approach problems in celestial mechanics, medical statistics, reliability and jurisprudence [4, 5]. Although, the Bayesian approach has origin since 18th century, the practical application has limited by difficulty of calculation when the posterior distribution is not expressed in closed form such as sumation and integrate. So, when the sampling method is developed in middle 20th century such as Markov Chain Monte Carlo (MCMC) and the speed and memory capacity of computer bring about calculation of the posterior distribution part easy [6]. For the general method, comparison of estimation performance of Bayesian approach usually use loss function, which is consider difference between estimated and true parameter [7, 8]. The loss function is called cost function, it map one or more variable values to a real number. The important type of the loss function is square error, absolute error, zero-one loss function, etc. [9--11]. Furthermore, different loss functions lead to the various estimators of Bayesian approach [11, 12].

In this paper, we concern estimating of parametors by using Bayesian approach under the squared error, absolute error and zero-one loss functions for the generalized Lindley (GL) distribution. We use the R language and open Bayesian inference using Gibbs sampling (OpenBUGS) program. For the content, we will present the GL distribution and some properties in Section 2 and the estimating parameters of Bayesian approach is presented in Section 3. Then we will show the application for estimating parameters of Bayesian approach in Section 4 . Finally, we give some conclusions.

## 2 Generalize Lindley distribution

In this section, we consider some properties of the GL distribution, which is flexible for lifetime data. The GL distribution is a mixture distribution between gamma$(\alpha, \theta)$ and gamma$(\alpha + 1, \theta)$ [13].

**Definition 1.** Let $X$ be an independent and identical distributed random variable (IID) of the GL distribution with parameters $\alpha$, $\theta$ and $\gamma$, denoted as $X \sim GL(\alpha, \theta, \gamma)$. The probability density function of X is given by

$$f_X(x; \alpha, \theta, \gamma) = \frac{\theta^2(\theta x)^{\alpha-1}(\alpha + \gamma x)\exp\{-\theta x\}}{(\gamma + \theta)\Gamma(\alpha + 1)}, \quad (1)$$

where $\alpha, \theta, \gamma, x > 0$.

The pdf of the GL distribution reduces to the pdf of gamma distribution with parameters $\alpha$ and $\theta$, when $\gamma = 0$. In particular case, $\alpha = \gamma = 1$ then Eq.1 becomes to pdf of the Lindley distribution. Furthermore, $\alpha = 1$ and $\gamma = 0$ reduce to the pdf of the ordinary exponential distribution. Figure 1 shows some pdf plots of the GL distribution with different values.

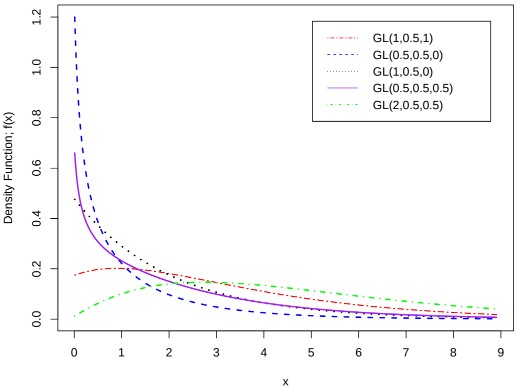


**Figure 1:** The pdf plots of the GL distribution for some values of $\alpha, \theta$ and $\gamma$

The moment generating function (mgf) of $X$ is

$$M_X(t) = \left(\frac{\theta}{\theta - t}\right)^{\alpha+1}\left(\frac{\theta - t + \gamma}{\theta + \gamma}\right).$$

Then, we got mean and variance of a three parameters GL random variable can be expressed as

$$E(X) = \frac{(\alpha+1)(\theta+\gamma) - \theta}{\theta(\theta+\gamma)},$$

and

$$Var(X) = \frac{\alpha + 2 - (\alpha+1)^2}{\theta^2} + \frac{\alpha+1}{\theta(\theta+\gamma)} - \frac{1}{(\theta+\gamma)^2}.$$

The likelihood function of the GL distribution is given by

$$L(\underset{\sim}{x}; \alpha, \theta, \gamma) = \left[\frac{\theta^{\alpha+1}}{(\gamma+\theta)\Gamma(\alpha+1)}\right]^n \exp\left\{-\theta\sum_{i=1}^{n} x_i\right\} \times \prod_{i=1}^{n} x_i^{\alpha-1}(\alpha + \gamma x_i). \quad (2)$$

## 3 Bayesian approach under different loss functions

This section, the existing Bayesian approach includes posterior distribution and loss function. The posterior distribution contains the prior distribution and likelihood function. When the prior distributions are the uniform and gamma distributions. Moreover, we are interested in considering various type of loss function that are the squared error, absolute error and zero-one loss functions.

### *3.1 The Posterior distribution*

**Definition 2.** Let $f(\Theta)$ is prior distribution where $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$ and $L(\underset{\sim}{x}|\Theta)$ is the likelihood function of X then the posterior distribution, denoted by $f(\Theta|\underset{\sim}{x})$, can be computed as

$$f(\Theta|\underset{\sim}{x}) = \frac{L(\underset{\sim}{x}|\Theta)f(\Theta)}{\int_\Theta L(\underset{\sim}{x}|\Theta)f(\Theta)\mathrm{d}\Theta}.$$

#### *3.1.1 The Posterior distribution using the uniform prior distribution*

**Theorem 1.** The posterior distribution of the GL distribution with the uniform prior distribution is

$$f(\alpha, \theta, \gamma|\underset{\sim}{x}) = \frac{w\exp\left\{-\theta\sum\limits_{i=1}^{n} x_i\right\}}{C(x)},$$

where $w = \left[\frac{\theta^{\alpha+1}}{(\gamma+\theta)\Gamma(\alpha+1)}\right]^n \prod\limits_{i=1}^{n} x_i^{\alpha-1}(\alpha + \gamma x_i)$ and $C(x) = \int\limits_0^\infty \int\limits_0^\infty \int\limits_0^\infty w\exp\left\{-\theta\sum\limits_{i=1}^{n} x_i\right\}\mathrm{d}\alpha\mathrm{d}\theta\mathrm{d}\gamma$.

**Proof.** The prior of parameters $\alpha, \theta, \gamma$ are the uniform distribution as

$$\alpha \sim Uniform(a_\alpha, b_\alpha);\ a_\alpha \le \alpha \le b_\alpha$$
$$\theta \sim Uniform(a_\theta, b_\theta);\ a_\theta \le \theta \le b_\theta$$
$$\gamma \sim Uniform(a_\gamma, b_\gamma);\ a_\gamma \le \gamma \le b_\gamma$$

According to Definition 2 and the uniform prior distribution of parameters $\alpha, \theta, \gamma$, the posterior distribution of the GL distribution is obtained by

$$\begin{aligned} f(\alpha, \theta, \gamma|\underset{\sim}{x}) &= \frac{L(\underset{\sim}{x}|\alpha, \theta, \gamma)f(\alpha)f(\theta)f(\gamma)}{\int\limits_0^\infty \int\limits_0^\infty \int\limits_0^\infty L(\underset{\sim}{x}|\alpha, \theta, \gamma)f(\alpha)f(\theta)f(\gamma)\mathrm{d}\alpha\mathrm{d}\theta\mathrm{d}\gamma} \\ &= \frac{\dfrac{w\exp\left\{-\theta\sum\limits_{i=1}^{n} x_i\right\}}{(b_\alpha - a_\alpha)(b_\theta - a_\theta)(b_\gamma - a_\gamma)}}{C(x)/(b_\alpha - a_\alpha)(b_\theta - a_\theta)(b_\gamma - a_\gamma)} \\ &= \frac{w\exp\left\{-\theta\sum\limits_{i=1}^{n} x_i\right\}}{C(x)}. \end{aligned} \quad (3)$$

#### *3.1.2 The Posterior distribution using the gamma prior distribution*

**Theorem 2.** The posterior distribution of the GL ditribution with the gamma prior distribution is

$$g(\alpha,\theta,\gamma|\underset{\sim}{x}) = \frac{wy\exp\left\{-(\theta\sum\limits_{i=1}^{n} x_i + d_\alpha\alpha + d_\theta\theta + d_\gamma\gamma)\right\}}{D(x)},$$

where $y = \alpha^{c_\alpha-1}\theta^{c_\theta-1}\gamma^{c_\gamma-1}$ and $D(x) = \int\limits_0^\infty\int\limits_0^\infty\int\limits_0^\infty wy\exp\Big\{ - \big(\theta\sum\limits_{i=1}^{n} x_i + d_\alpha\alpha + d_\theta\theta + d_\gamma\gamma\big)\Big\}\mathrm{d}\alpha\mathrm{d}\theta\mathrm{d}\gamma$.

**Proof.** Follow Definition 2 and the gamma prior distribution of three parameters are

$$\begin{aligned}&\alpha \sim Gamma(c_\alpha, d_\alpha);\ c_\alpha, d_\alpha > 0\\ &\theta \sim Gamma(c_\theta, d_\theta);\ c_\theta, d_\theta > 0\\ &\gamma \sim Gamma(c_\gamma, d_\gamma);\ c_\gamma, d_\gamma > 0.\end{aligned}$$

The gamma prior distribution of parameters $\alpha,\theta,\gamma$ then the posterior distribution of the GL distribution can be written as

$$\begin{aligned} g(\alpha,\theta,\gamma|\underset{\sim}{x}) &= \frac{L(\underset{\sim}{x}|\alpha,\theta,\gamma)f(\alpha)f(\theta)f(\gamma)}{\int\limits_0^\infty\int\limits_0^\infty\int\limits_0^\infty L(\underset{\sim}{x}|\alpha,\theta,\gamma)f(\alpha)f(\theta)f(\gamma)\mathrm{d}\alpha\mathrm{d}\theta\mathrm{d}\gamma}\\ &= \frac{\dfrac{wy d_\alpha^{c_\alpha} d_\theta^{c_\theta} d_\gamma^{c_\gamma}\exp\left\{-(\theta\sum\limits_{i=1}^{n} x_i + z)\right\}}{\Gamma(c_\alpha)\Gamma(c_\theta)\Gamma(c_\gamma)}}{D(x)\dfrac{d_\alpha^{c_\alpha} d_\theta^{c_\theta} d_\gamma^{c_\gamma}}{\Gamma(c_\alpha)\Gamma(c_\theta)\Gamma(c_\gamma)}}\\ &= \frac{wy\exp\left\{-(\theta\sum\limits_{i=1}^{n} x_i + z)\right\}}{D(x)}. \end{aligned} \tag{4}$$

*3.2 Loss function*

The loss function is used to measure accuracy of the Bayesian approach. In this section we would like to show the Bayesian approach under the three loss functions.

**Definition 3.** Let the estimator $\hat{\theta}$ is real-valued function that used estimate the parameter $\theta$. The squared error, absolute and zero-one loss functions are following

1)$L_S(\hat{\theta},\theta) = (\theta-\hat{\theta})^2$,

2)$L_A(\hat{\theta},\theta) = |\theta-\hat{\theta}|$,

3)$L_Z(\hat{\theta},\theta) = \begin{cases} 0 & ; |\theta-\hat{\theta}| \le c,\\ 1 & ; |\theta-\hat{\theta}| > c.\end{cases}$

The squared error of loss function is sometimes called quadratic loss function. It is widely used for the Bayesian approach [12].

**Theorem 3** Let $X \sim GL(\alpha,\theta,\gamma)$. By using uniform and gamma prior distribution, the Bayesian approach for the parameter $\alpha,\theta$ and $\gamma$ under squared error loss function are $E_{Uni}(\alpha|\underset{\sim}{x})$, $E_{Gam}(\alpha|\underset{\sim}{x})$, $E_{Uni}(\theta|\underset{\sim}{x})$, $E_{Gam}(\theta|\underset{\sim}{x})$, $E_{Uni}(\gamma|\underset{\sim}{x})$ and $E_{Gam}(\gamma|\underset{\sim}{x})$, respectively.

**Proof** The squared error loss function from Definition 3 and uniform prior distribution from Theorem 1 are applied in posterior expectation,

$$E(L_S(\hat{\alpha},\alpha)|\underset{\sim}{x}) = \int_0^\infty (\alpha-\hat{\alpha})^2 f(\alpha,\theta,\gamma|\underset{\sim}{x})d\alpha. \tag{5}$$

For minimizing the expected loss function the derivative Eq. 5 with respect to $\alpha$ is equated to zero

$$\begin{aligned} \frac{\partial}{\partial\hat{\alpha}}\int_0^\infty (\alpha-\hat{\alpha})^2 f(\alpha,\theta,\gamma|\underset{\sim}{x})d\alpha &= 0,\\ 2\int_0^\infty (\alpha-\hat{\alpha}) f(\alpha,\theta,\gamma|\underset{\sim}{x})d\alpha &= 0,\\ 2\Big(\int_0^\infty \alpha f(\alpha,\theta,\gamma|\underset{\sim}{x})d\alpha - \hat{\alpha}\Big) &= 0,\\ 2(E_{Uni}(\alpha|\underset{\sim}{x}) - \hat{\alpha}) &= 0,\\ \hat{\alpha} &= E_{Uni}(\alpha|\underset{\sim}{x}). \end{aligned}$$

Similarly, the $f(\alpha,\theta,\gamma|\underset{\sim}{x})$ in Eq. 5 is replaced by $g(\alpha,\theta,\gamma|\underset{\sim}{x})$ from Theorem 2. We then obtain $E_{Gam}(\alpha|\underset{\sim}{x})$. If we use the parameter $\theta$ instead of parameter $\alpha$ then we obtain $E_{Uni}(\theta|\underset{\sim}{x})$ and $E_{Gam}(\theta|\underset{\sim}{x})$, respectively. Moreover, if we chance the parameter $\alpha$ to $\gamma$, we got $E_{Uni}(\theta|\underset{\sim}{x})$ and $E_{Gam}(\theta|\underset{\sim}{x})$, respectively.

Some situations of the data have outlying observations, the Bayesian approach based on the square error loss function is sensitive to this situation. Since, it is fast-growing loss function [12]. So, the absolute error loss function that is slowly growing loss function is play important role for this situation.

**Theorem 4** Let $X \sim GL(\alpha,\theta,\gamma)$. By using uniform and gamma prior distribution, the Bayesian approach for the parameter $\alpha,\theta$ and $\gamma$ under absolute error loss function are $Med_{Uni}(\alpha|\underset{\sim}{x})$, $Med_{Gam}(\alpha|\underset{\sim}{x})$, $Med_{Uni}(\theta|\underset{\sim}{x})$, $Med_{Gam}(\theta|\underset{\sim}{x})$, $Med_{Uni}(\gamma|\underset{\sim}{x})$ and $Med_{Gam}(\gamma|\underset{\sim}{x})$, respectively.

**Proof** From Definition 3 and Theorem 1, the absolute error loss function and uniform prior distribution are plugged into posterior expectation as follows

$$E(L_A(\hat{\alpha},\alpha)|\underset{\sim}{x}) = \int_0^\infty |\alpha-\hat{\alpha}| f(\alpha,\theta,\gamma|\underset{\sim}{x})d\alpha. \tag{6}$$

we will minimize the Eq. 6 by

$$\begin{aligned} \frac{\partial}{\partial\hat{\alpha}}\int_0^\infty |\alpha-\hat{\alpha}| f(\alpha,\theta,\gamma|\underset{\sim}{x})d\alpha &= 0,\\ \frac{\partial}{\partial\hat{\alpha}}\Big(\int_{\alpha\le\hat{\alpha}} (\hat{\alpha}-\alpha)) f(\alpha|\underset{\sim}{x})d\alpha + &\\ \int_{\alpha>\hat{\alpha}} (\alpha-\hat{\alpha}) f(\alpha,\theta,\gamma|\underset{\sim}{x})d\alpha\Big) &= 0,\\ \frac{\partial}{\partial\hat{\alpha}}\Big(\int_0^{\hat{\alpha}} \hat{\alpha} f(\alpha,\theta,\gamma|\underset{\sim}{x})d\alpha - \int_0^{\hat{\alpha}} \alpha f(\alpha,\theta,\gamma|\underset{\sim}{x})d\alpha + &\\ \int_{\hat{\alpha}}^\infty \alpha f(\alpha,\theta,\gamma|\underset{\sim}{x})d\alpha - \int_{\hat{\alpha}}^\infty \hat{\alpha} f(\alpha,\theta,\gamma|\underset{\sim}{x})d\alpha\Big) &= 0,\\ \int_0^{\hat{\alpha}} f(\alpha,\theta,\gamma|\underset{\sim}{x})d\alpha - \int_{\hat{\alpha}}^\infty f(\alpha,\theta,\gamma|\underset{\sim}{x})d\alpha &= 0. \end{aligned}$$

The last equation is only satisfied by median, $Med_{Uni}(\alpha|\underset{\sim}{x})$. Similarly, if use the Theorem 2 instead of the Theorem 1. We obtain $Med_{Gam}(\alpha|\underset{\sim}{x})$ for gamma prior distribution case. We got the $Med_{Uni}(\theta|\underset{\sim}{x})$ and $Med_{Gam}(\theta|\underset{\sim}{x})$, respectively if we plugged in the parameter $\theta$ instead of parameter $\alpha$. In addition, we obtain $Med_{Uni}(\gamma|\underset{\sim}{x})$ and $Med_{Gam}(\gamma|\underset{\sim}{x})$, respectively.

Furthermore, the zero-one loss function is one of the insensitive loss functions that is interesting [14].

**Theorem 5** Let $X \sim GL(\alpha,\theta,\gamma)$. By using uniform and gamma prior distribution, the Bayesian approach for the parameter $\alpha,\theta$ and $\gamma$ under zero-one loss function are $Mod_{Uni}(\alpha|\underset{\sim}{x})$, $Mod_{Gam}(\alpha|\underset{\sim}{x})$, $Mod_{Uni}(\theta|\underset{\sim}{x})$, $Mod_{Gam}(\theta|\underset{\sim}{x})$, $Mod_{Uni}(\gamma|\underset{\sim}{x})$ and $Mod_{Gam}(\gamma|\underset{\sim}{x})$, respectively.

**Proof** We replace the zero-one loss function in posterior expectation

$$\begin{aligned} E(L_Z(\hat{\alpha},\alpha|\underset{\sim}{x}) &= \int_0^{\infty} L_Z(\hat{\alpha},\alpha) f(\alpha,\theta,\gamma|\underset{\sim}{x})d\alpha \\ &= \int_0^{\hat{\alpha}-c} f(\alpha,\theta,\gamma|\underset{\sim}{x})d\alpha + \\ &\qquad \int_{\hat{\alpha}+c}^{\infty} f(\alpha,\theta,\gamma|\underset{\sim}{x})d\alpha \\ &= 1 - \int_{\hat{\alpha}-c}^{\hat{\alpha}+c} f(\alpha,\theta,\gamma|\underset{\sim}{x})d\alpha \qquad (7) \end{aligned}$$

For this case, we consider Eq. 7. It has minimize value when $\int_{\hat{\alpha}-c}^{\hat{\alpha}+c} f(\alpha,\theta,\gamma|\underset{\sim}{x})d\theta$ close to 1 with small value of c which is maximized by posterior mode, $Mod_{Uni}(\alpha|\underset{\sim}{x})$. For gamma prior distribution, we changed the Theorem 1 is the Theorem 2 and obtain $Mod_{Gam}(\alpha|\underset{\sim}{x})$. If we use the $\theta$ instead of the parameter $\alpha$, we obtain $Mod_{Uni}(\theta|\underset{\sim}{x})$ and $Mod_{Gam}(\theta|\underset{\sim}{x})$, respectively. Moreover, if we replace the parameter $\alpha$ by parameter $\gamma$ the we got $Mod_{Uni}(\gamma|\underset{\sim}{x})$ and $Mod_{Gam}(\gamma|\underset{\sim}{x})$, respectively.

The estimating parameters by Bayesian approach in Theorems 3, 4 and 5 cannot be expressed in closed form, so, we use the MCMC techniques base on Gibbs sampling to solve the problem.

## 4 Application

In this section, we illustrate the applicability of the GL distribution which estimate parameters values by Bayesian approach for two real data sets. The criteria used for comparison the estimator are log-likelihood (LL), Akaike information criterion (AIC), Bayesian information criterion (BIC), and Kolmogorov-Smirnov test.

$$\begin{aligned} &LL = \log L(\underset{\sim}{x};\Theta), \\ &AIC = -2LL + 2k, \\ &BIC = -2LL + k * \log(n), \\ &K - S = \underset{x}{\text{Sup}}|F_n(x) - F(x)|, \end{aligned}$$

when $L(\underset{\sim}{x};\Theta)$ is the likelihood function, k is the number of estimated parameter in model, n is the number of observation, $F_n(x)$ is empirical distribution, and $F(x)$ is distribution function.

The example I is the lifetime data (years) for more complex system and its components and they are 17.2, 10.85, 33.7, 55.55, 52.85, 11.93, 39.5, 9.21, 55.14, 26.68, 52.42, 30.85, 31.27, 14.85, 29.87, 43.51, 46.44, 58.67, 63.11, 28.45 [15]. The example II is the real data set for the lifetime data which given arose in tests on endurance of deep groove ball bearings and they are 17.88, 28.92, 33.00, 41.52, 42.12, 45.60, 48.80, 51.84, 51.96, 54.12, 55.56, 67.80, 68.44, 68.64, 68.88, 84.12, 93.12, 98.64, 105.12, 105.84, 127.92, 128.04, 173.40 [16]. Tables 1 arises from the Bayes approach under different loss functions and prior distributions by using R language and OpenBUGS program with 25,000 iterations after a 5,000 iterations burn-in. Figs. 2 - 5 show trace plot of simulated parameters with different prior distributions for two examples.

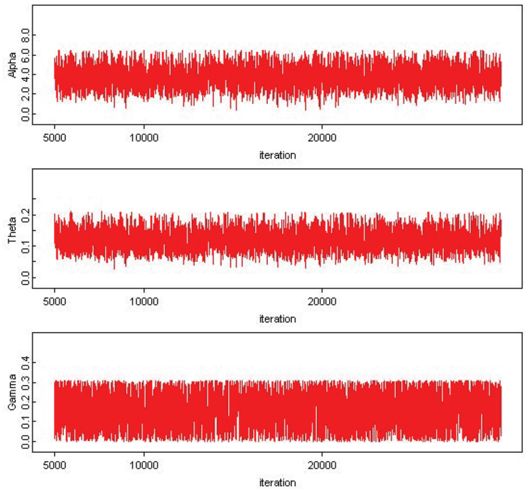


**Figure 2:** Trace plot of samples for parameters $\alpha,\theta,\gamma$ respectively with the uniform prior distribution for example I.

**Table 1:** The value of estimated parameters, log-likelihood, AIC, BIC and KS-statistic

| Example | Parameter | Squared error loss | | Absolute error loss | | Zero-one loss | |
|---|---|---|---|---|---|---|---|
| | | Uniform prior | Gamma prior | Uniform prior | Gamma prior | Uniform prior | Gamma prior |
| I | $\alpha$ | 3.682 | 3.233 | 3.618 | 3.178 | 3.530 | 3.866 |
| | $\theta$ | 0.119 | 0.100 | 0.117 | 0.098 | 0.112 | 0.095 |
| | $\gamma$ | 0.158 | 0.162 | 0.158 | 0.046 | 0.288 | 2.07E-05 |
| | LL | -85.149 | -85.274 | -85.135 | -85.140 | -85.334 | -85.803 |
| | AIC | 176.297 | 176.548 | 176.269 | 176.280 | 176.667 | 177.607 |
| | BIC | 179.284 | 179.535 | 179.257 | 179.267 | 179.654 | 180.594 |
| | KS | 0.136 | 0.125 | 0.131 | 0.129 | 0.140 | 0.152 |
| | $p$-value | 0.807 | 0.876 | 0.840 | 0.854 | 0.777 | 0.686 |
| II | $\alpha$ | 3.429 | 2.934 | 3.358 | 2.865 | 2.778 | 2.946 |
| | $\theta$ | 0.062 | 0.054 | 0.061 | 0.053 | 0.059 | 0.056 |
| | $\gamma$ | 13.689 | 13.679 | 13.600 | 13.530 | 20.270 | 12.210 |
| | LL | -113.085 | -113.030 | -113.066 | -113.039 | -113.784 | -113.043 |
| | AIC | 232.171 | 232.061 | 232.132 | 232.077 | 233.568 | 232.087 |
| | BIC | 235.577 | 235.467 | 235.538 | 235.484 | 236.974 | 235.493 |
| | KS | 0.121 | 0.123 | 0.122 | 0.124 | 0.145 | 0.109 |
| | $p$-value | 0.853 | 0.838 | 0.846 | 0.832 | 0.664 | 0.922 |

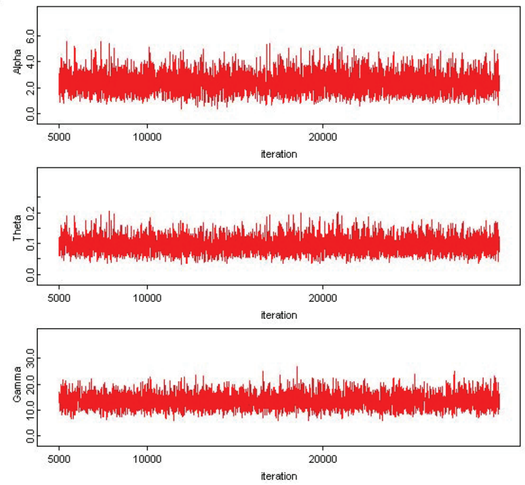


**Figure 3:** Trace plot of samples for parameters $\alpha, \theta, \gamma$ respectively with the gamma prior distribution for example I.

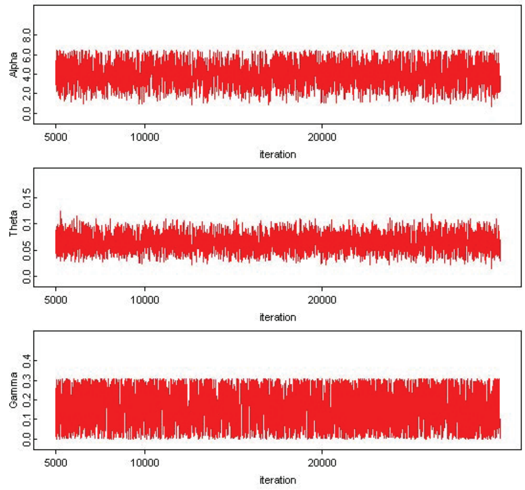


**Figure 4:** Trace plot of samples for parameters $\alpha, \theta, \gamma$ respectively with the uniform prior distribution for example II.

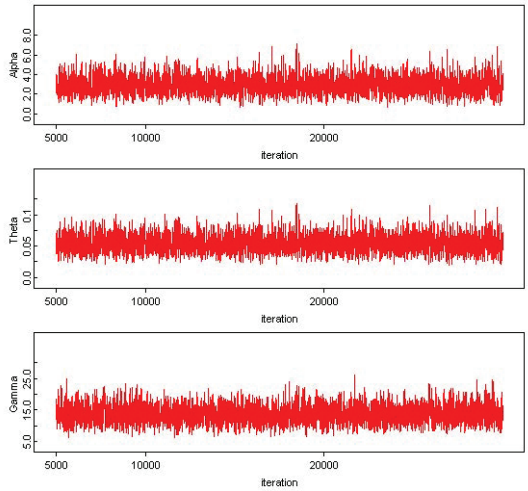


**Figure 5:** Trace plot of samples for parameters $\alpha, \theta, \gamma$ respectively with the gamma prior distribution for example II.

## 5 Conclusion and discussion

In this paper, parameter estimation of GL distribution by Bayesian approach using different loss functions which are the squared error, absolute error and zero-one loss functions are presented. Investigations are done based on the uniform and gamma prior distributions. In case of the uniform prior distribution we found that the absolute error loss function gives the highest $p-$ value in example I, but the squared loss function provides the highest $p-$ value in example II. Next the gamma prior distribution, the squared error loss function show the highest $p-$ value in example I. In the other hand, the zero-one loss function presents the highest $p-$ value in example II.

We found that from real data sets based on fitting with GL distribution and parameter estimation of the distribution using Bayesian approach, the results of parameter estimation showed that there are some different results based on specified criteria. We may conduct some simulation study for comparison purpose.

## Acknowledgements

The authors would like to thank the Science Achievement Scholarship of Thailand (SAST). This work was partially supported by Department of Statistics, Faculty of Science, Kasetsart University and Graduate school Kasetsart University.

## References

[1] Bolstad WM. Introduction to Bayesian statistics. 2nd ed. John Wiley; 2007.

[2] McGrayne SB. The theory that would not die : how Bayes' rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy. Yale University Press; 2011.

[3] Bayes T. An Essay towards Solving a Problem in the Doctrine of Chances. Philosophical transactions of the royal society of London. 1763;53: 370--418.

[4] Fienberg SE. When did Bayesian inference become "Bayesian". 2006;1: 1--40.

[5] Laplace PS. Mémoire sur la Probabilité des Causes par les Événements. Mémoires de Mathématique et de Physique Presentés ál'Académie Royale des Sciences. 1774;6: 621--656.

[6] Bishop CM. Pattern Recognition and Machine Learning. 1st ed. Information science and statistics. Springer; 2006.

[7] Singh PK, Singh SK, Singh U. Bayes Estimator of Inverse Gaussian Parameters Under General Entropy Loss Function Using Lindley's Approximation. Communications in Statistics - Simulation and Computation. 2008;37(9): 1750--1762.

[8] Wald A. Statistical decision functions. Wiley publications in statistics. New York, Wiley; 1950.

[9] Dey S. Bayesian Estimation of the Shape Parameter of the Generalised Exponential Distribution under Different Loss Functions. Pakistan Journal of Statistics and Operation Research. 2010;6(2): 163--174.

[10] Preda V. Bayes estimators of Modified-Weibull distribution parameters using Lindley's approximation. Wseas transactions on mathematics. 2010;9: 539--549.

[11] Ali S, Aslam M, Kazmi SMA. A study of the effect of the loss function on Bayes Estimate, posterior risk and hazard function for Lindley distribution. Applied Mathematical Modelling. 2013;37(8): 6068 -- 6078.

[12] Lehmann EL, Casella G. Theory of Point Estimation. 2nd ed. Springer texts in statistics. Springer; 1998.

[13] Zakerzadeh H, Dolati A. Generalized Lindley Distribution. Journal of Mathematical Extension. 2009;.

[14] Leonhard Held DSBa. Applied Statistical Inference: Likelihood and Bayes. Springer Berlin Heidelberg; 2014.

[15] Guo J. Bayesian methods for system reliability and community detection. 2011;.

[16] Lieblein J, Zelen M. Statistical investigation of the fatigue life of deep groove ball bearings. Journal of Res Nat Bur Std. 1956;p. 273--315.

# Crack Distribution Parameters Estimation by Maximum Likelihood Estimation, Method of moments and Bayesian Approach

T. Akkanphudit[1*], W. Bodhisuwan[2] and C. Pudprommarat[3]

[1]Departmen of statistics, Faculty of Science, Kasetsart University, Bangkok, 10900, Thailand, thanasate_stat@hotmail.com

[2] Departmen of statistics, Faculty of Science, Kasetsart University, Bangkok, 10900, Thailand, fsciwnb@ku.ac.th

[3] Department of statistics, Faculty of Science and Technology, Suan Sunantha Rajabath University, Bangkok, 10300, Thailand, chookait.pu@ssru.ac.th

## Abstract

In this paper, we study estimating parameters of crack distribution by using maximum likelihood estimation, method of moments and Baysian approach are derived under square error loss function. For application, we are compare estimator parameter of maximum likelihood estimation, method of moments and Baysian approach. Based on the results, we have shown that estimating parameters of crack distribution by using Bayesian approach provides a better fit compared to maximum likelihood estimation and method of moments.

*Keywords*: Crack distribution, Maximum likelihood estimation, Method of moments, Bayesian approach

***** Thanasate Akkanphudit.
E-mail address: thanasate_stat@hotmail.com

## 1. Introduction

The crack distribution has three parameters distributions, also known as the inverse Gaussian mixture distribution which are Inverse Gaussian (IG) and Length Biased Inverse Gaussian (LBIG) distribution, was studied by Jorgensen et al. [1], Balakrishnan et al. [2] and Bowonrattanaset and Budsaba. [3]. This distribution is used as a lifetime distribution in the various models of reliability theory. Applications using the Crack distribution can be found in many areas, for instance, physics, engineering, biomedical and economics.

Estimation (or estimating) parameter is the process of finding an estimate, or approximation, which is a value that is usable for some purpose even if input data may be incomplete, uncertain or unstable. The value is nonetheless usable because it is derived from the best information available.Typically, estimation involves "using the value of a statistic derived from a sample to estimate the value of a corresponding population parameter". The sample provides information that can be projected, through various formal or informal processes, to determine a range most likely to describe the missing information. An estimate that turns out to be incorrect will be an overestimate if the estimate exceeded the actual result, and anunderestimate if the estimate fell short of the actual result. Generally, several methods of estimating parameters will be analysed. In order to estimate the parameters, it is necessary to know the sampling theory and statistical inference.

This manual will use one of the general methods most commonly used in the estimation of parameters which are maximum likelihood (MLE), the moments method (MM). In many cases this method uses iterative processes, which require the adoption of initial values. Therefore, particular methods will also be presented, which obtain estimates close to the real values of the parameters. In many situations, these initial estimates also have a practical interest.

In this paper, we considers estimating parameters of crack distribution by using maximum likelihood estimation, method of moments and Bayesian approach are derived under square error loss function. The engineering interpretation of Crack random variable is time after a machine element is started to be forced by a cyclic or non-cyclic loading until the crack achieves the critical value. After a machine element is forced, a slightly crack may happen but the element could still works. When it arrives the critical point, tolerance exceeds and the element does not properly work anymore.

The plan of the paper is as follows. First we show estimating parameters of Crack distribution by using maximum likelihood estimation, method of moments and Baysian approach are derived under square error loss function. Finally, we have shown application of Crack distribution is carried out on one sample of lifetime data.

## 2. Crack distribution

The crack distribution related to two distributions: the Inverse Gaussian distribution and the Length Biased Inverse Gaussian distribution is denoted as CR (P, λ, θ) and has the density function :

$$f(x;\, P,\, \lambda,\, \theta) \;=\; pf_{IG}(x;\, \lambda,\, \theta) \;+\; pf_{LB}(x;\, \lambda,\, \theta),$$

where $x > 0$ ; $\lambda > 0$ , $\theta > 0$ , $0 \le p \le 1$.

That is

$$f_{CR}(x, p, \lambda, \theta) = p\left[\frac{\lambda}{\theta\sqrt{2\pi}}\left(\frac{\theta}{x}\right)^{\frac{3}{2}} \exp\left\{-\frac{1}{2}\left(\sqrt{\frac{x}{\theta}} - \sqrt{\frac{x}{\lambda}}\right)^{2}\right\}\right]$$

$$+(1-p)\left[\frac{1}{\sqrt{\theta 2\pi}}\left(\frac{x}{\theta}\right)^{\frac{3}{2}} \exp\left\{-\frac{1}{2}\left(\sqrt{\frac{x}{\theta}} - \sqrt{\frac{x}{\lambda}}\right)^{2}\right\}\right],$$

where $x > 0$ ; $\lambda > 0$ , $\theta > 0$ , $0 \le p \le 1$.

The relations between classical parameters $p, \lambda$ and $\theta$, which mean and varain follows:

$$\mu = \theta\left(1 - p + \lambda\right),$$

$$\sigma^2 = \theta^2\left[3\left(1-p\right) - \left(1-p\right)^2 + \lambda\right].$$

## 3. Parameters Estimation by Maximum Likelihood

The parameter estimation for the crack distribution via the Maximum Likelihood (MLE) method procedure will be discussed. Let $X_1, X_2, ..., X_n$ be a random sample from $X : CR(p, \lambda, \theta)$ and let $\delta=(p,\lambda,\theta)^T$ be the vector of the model parameters, the likelihood function for $\delta$ is given by:

$$L(\delta) = L(\delta \mid x_1, x_2, x_3, ..., x_n),$$

$$L(\delta) = \prod_{i=1}^{n}\left\{\frac{1}{\theta\sqrt{2\pi}}\left[p\lambda\left(\frac{\theta}{x_i}\right)^{\frac{3}{2}} + \left(1-p\right)\left(\frac{\theta}{x_i}\right)^{\frac{1}{2}}\right] \times \exp\left[-\frac{1}{2}\left(\sqrt{\frac{x_i}{\theta}} - \lambda\sqrt{\frac{\theta}{x_i}}\right)^2\right]\right\}, \quad (1)$$

where as $x_i > 0,\ i = 1,2,3,\ldots,n.$

The log-likelihood function is give by

$$\log \mathrm{L}(\delta) = -\frac{n}{2}\log\theta - \frac{n}{2}\log\left(2\pi\right) + n\lambda - \frac{1}{2\theta}\sum_{i=1}^{n} x_i - \frac{\lambda^2\theta}{2}\sum_{i=1}^{n}\frac{1}{x_i} - \frac{3}{2}\sum_{i=1}^{n}\log x_i + \sum_{i=1}^{n}\log(\mathrm{p}\lambda\theta + (1-\mathrm{p})x_i). \quad (2)$$

The first step for finding the optimal values of the parameters obtained by differentiating in Eg.(2) with respect to $p, \lambda$ and $\theta$ give rise to the following equation:

$$\frac{\partial}{\partial p}\log L(\delta) = \sum_{i=1}^{n}\frac{\lambda\theta - x_i}{p\lambda\theta + (1-p)x_i}, \quad (3)$$

$$\frac{\partial}{\partial \lambda}\log L(\delta) = n - \lambda\theta\sum_{i=1}^{n}\frac{1}{x_i} + \sum_{i=1}^{n}\frac{p\theta}{p\lambda\theta + (1-p)x_i}, \quad (4)$$

$$\frac{\partial}{\partial \theta}\log L(\delta) = -\frac{n}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{n} x_i - \frac{\lambda^2}{2}\sum_{i=1}^{n}\frac{1}{x_i} + \sum_{i=1}^{n}\frac{p\lambda}{p\lambda\theta + (1-p)x_i}. \quad (5)$$

The MLE of the parameter $p, \lambda$ and $\theta$ are $\widehat{p}_{MLE,}$ $\hat{\lambda}_{MLE}$ and $\hat{\theta}_{MLE}$ respectively, which are obtained by solving iteratively Eq.(3) – (5). Since there are not insimple form, a numerical method can be employed to obtain the expectations of the parameter estimation by using R language.

## 4. Parameters Estimation by Method of moments

We can also estimate parameter of crack distribution by the method of moments (MM) following Bowonrattanaset and Budsaba [3], the first second and thrid moment of the Crack distribution X are

$$m_1 = E(X) = (\lambda + 1 - p)\theta, \quad (6)$$

$$m_2 = E(X^2) = \left[\lambda^3 + 3\lambda - 2\lambda p - 3p + 3\right]\theta^2, \quad (7)$$

$$m_3 = E(X^3) = [\lambda^3 + 6\lambda^2 + 15\lambda - 3\lambda^2 p - 12\lambda p - 15p + 15]\theta^3. \quad (8)$$

The MM estimates $\hat{p}_{MM}$, $\hat{\lambda}_{MM}$ and $\hat{\theta}_{MM}$ for the parameters $p$, $\lambda$ and $\theta$ respectively, are obtained by solving iteratively Eq.(6) – (8). We use gmm function in statistical package of R language.

## 5. Parameters Estimation by Bayesian Approach

The crack distribution using Bayes approach under square error loss function will be discussed. The parameters $p, \lambda$ and $\theta$ are assumed to be unknown, the prior distribution for $p$ is take to be a beta distribution and $\lambda$ and $\theta$ are take to be a gamma distribution, which given by

$$p \sim beta(a,b),$$
$$\lambda \sim Gamma(\alpha_\lambda,\beta_\lambda),$$
$$\theta \sim Gamma(\alpha_\theta,\beta_\theta),$$

which pdf of

$$\pi(p) = \frac{1}{B(a,b)} p^{a-1}(1-p)^{b-1}, \quad a,\, b > 0,$$

$$\pi(\lambda) = \frac{1}{\Gamma(\alpha_\lambda)\left(\beta_\lambda\right)^{\alpha_\lambda}} \lambda^{\alpha-1}\exp(-\frac{\lambda}{\beta_\lambda}), \qquad \alpha_\lambda,\, \beta_\lambda > 0,$$

$$\pi(\theta) = \frac{1}{\Gamma(\alpha_\theta)\left(\beta_\theta\right)^{\alpha_\theta}} \theta^{\alpha_\theta-1}\exp(-\frac{\theta}{\beta_\theta}), \qquad \alpha_\theta,\, \beta_\theta > 0.$$

Since $p,\lambda$ and $\theta$ are assumed to be independence, then the joint prior distribution of $p,\lambda$ and $\theta$ is form

$$\pi(p,\lambda,\theta) = \frac{1}{B(a,b)} p^{a-1}(1-p)^{b-1}$$
$$\times \frac{1}{\Gamma(\alpha_\lambda)\left(\beta_\lambda\right)^{\alpha_\lambda}} \lambda^{\alpha-1}\exp(-\frac{\lambda}{\beta_\lambda})$$
$$\times \frac{1}{\Gamma(\alpha_\theta)\left(\beta_\theta\right)^{\alpha_\theta}} \theta^{\alpha_\theta-1}\exp(-\frac{\theta}{\beta_\theta}). \text{(9)}$$

Combining the prior distribution with likelihood function given by Eq.(9) and Eq.(1) respectively, we get the joint distribution has the pdf which is obtain by

$$\pi(x,p,\lambda,\theta) = \prod_{i=1}^{n}\left\{ \frac{1}{\theta\sqrt{2\pi}}\left[ p\lambda\left(\frac{\theta}{x_i}\right)^{\frac{3}{2}} + (1-p)\left(\frac{\theta}{x_i}\right)^{\frac{1}{2}} \right]\right.$$
$$\left.\times \exp\left[ -\frac{1}{2}\left( \sqrt{\frac{x_i}{\theta}} - \lambda\sqrt{\frac{\theta}{x_i}} \right)^2 \right]\right\}$$
$$\times \frac{1}{B(a,b)} p^{a-1}(1-p)^{b-1}$$
$$\times \frac{1}{\Gamma(\alpha_\lambda)\left(\beta_\lambda\right)^{\alpha_\lambda}} \lambda^{\alpha-1}\exp(-\frac{\lambda}{\beta_\lambda})$$
$$\times \frac{1}{\Gamma(\alpha_\theta)\left(\beta_\theta\right)^{\alpha_\theta}} \theta^{\alpha_\theta-1}\exp(-\frac{\theta}{\beta_\theta}).$$

In addition, the marginal distribution can be written as

$$\pi(x) = \int_0^1\int_0^\infty\int_0^\infty \pi(x,p,\lambda,\theta)\, d\theta\, d\lambda\, dp$$

Therefore, we can easily obtain the posterior distribution of $p,\lambda$ and $\theta$ is given by

$$\pi(p,\lambda,\theta \mid x) = \frac{\pi(x,p,\lambda,\theta)}{\pi(x)} \qquad (10)$$

As a result, the Bayes approach of $p,\lambda$ and $\theta$ based on the square error loss function are respectively obtained as

$$\hat{p}_{Bayes} = \int_0^\infty\int_0^\infty\int_0^1 p\,\pi(p,\lambda,\theta \mid x)\, dp\, d\lambda\, d\theta,$$

$$\hat{\lambda}_{Bayes} = \int_0^1\int_0^\infty\int_0^\infty \lambda\pi(p,\lambda,\theta \mid x)\, d\lambda\, d\theta\, dp,$$

$$\hat{\theta}_{Bayes} = \int_0^1\int_0^\infty\int_0^\infty \theta\pi(p,\lambda,\theta \mid x)\, d\theta\, d\lambda\, dp.$$

The posterior distribution, Eq.(10) can not be expressed in an explicit form. Therefore the Bayesian approach is derived by using numerical integration which is Markov Chain Monte Carlo (MCMC) technique to generate samples from the posterior distribution and in turn computing the Bayesian approach estimation such as Metropolis, Metropolis – Hastings and Gilbbs sampling. In this study, the Gibbs sampling is performed for obtaining posterior distribution to estimated the parameters of the Crack distribution. Furthermore, the computational method is illustrated by using R language and Open Bayesian Using Gibbs Sampling (OpenBUGS) program.

## 6. Application

In this section, we applied to real data set which was taken from Hsieh's data [5]. The data provides information on active repair times (in hours) for an airborne communication transceiver which are given in Table 1. The parameter estimates of this distribution for parameters $p,\lambda$ and $\theta$ by using the MLE, MM, and Bayes approach. Bayes approach under square error loss function will be discussed, the prior distribution for $p \sim beta(0.17, 0.20)$, $\lambda \sim Gamma(0.05, 0.11)$ and $\theta \sim Gamma(1.52, 0.30)$. We compare the parameter estimation based on the MLE, the MM, and the Bayes approach by histogramin Figure 4.

Table 1: Repair lifetimes (in hours) of an airborne transceiver

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0.2 | 0.3 | 0.5 | 0.5 | 0.5 | 0.5 | 0.6 | 0.6 |
| 0.7 | 0.7 | 0.7 | 0.8 | 0.8 | 1.0 | 1.0 | 1.0 |
| 1.0 | 1.1 | 1.3 | 1.5 | 1.5 | 1.5 | 1.5 | 2.0 |
| 2.0 | 2.2 | 2.5 | 2.7 | 3.0 | 3.0 | 3.3 | 3.3 |
| 4.0 | 4.0 | 4.5 | 4.7 | 5.0 | 5.0 | 5.4 | 7.0 |
| 7.5 | 8.8 | 9.0 | 10.3 | 22.0 | 24.5 | | |

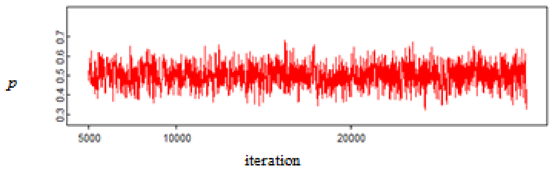


**Figure 1**: Trace plot of samples for parameter $p$ with the beta prior distribution for real data.

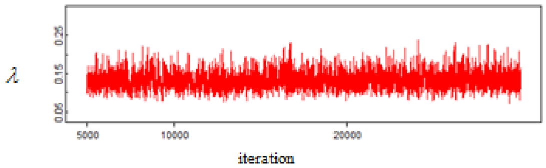


**Figure 2**: Trace plot of samples for parameter $\lambda$ with the gamma prior distribution for real data.

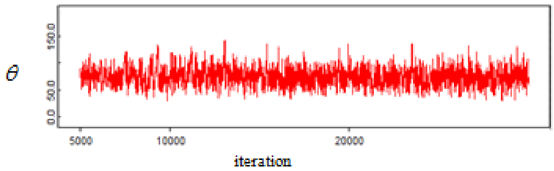


**Figure 3**: Trace plot of sample for parameter $\theta$ with the gamma prior distribution for real data.

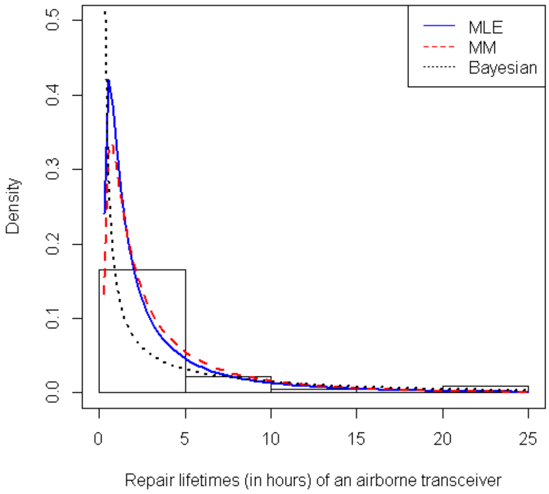


**Figure 4** The density function of the repair lifetimes (in hours) of an airborne transceiver

## 7. Conclusion

In this paper, we consider estimating parameters of crack distribution by using maximum likelihood estimation, method of moments and Baysian approach are derived under square error loss function. For application, we are compare estimator parameter of maximum likelihood estimation, method of moments and Baysian approach. Bayesian approach is the most perfecte parameter eatimation for data set.

**Table 2**: MLE, Moment, Bayesian of the model parameters for the repair lifetime of an airborne transceiver, LL ,AIC, BIC, MSE and KS

| Parameters Estimation | Estimate parameters | | | LL | AIC | BIC | KS | *p*-value |
|---|---|---|---|---|---|---|---|---|
| | $P$ | $\lambda$ | $\theta$ | | | | | |
| Maximum Likelihood | 0.8475 | 0.5652 | 5.0251 | 14.3552 | 34.7104 | 40.1964 | 1.8964 | 0.876 |
| Method of moments | 0.6694 | 0.7163 | 3.7320 | 76.1465 | 98.1601 | 92.6742 | 1.4173 | <0.001 |
| Bayesian Approach | 0.5005 | 0.1325 | 7.5118 | 11.4963 | 37.0271 | 41.9846 | 1.9432 | 0.963 |

## Acknowledgements

The authors would like to thank my friend for helpful this research, reviewer for review our paper and Department of Statistics, Kasetsart University.

## References

[1]Jorgensen, B., V. Seshadri and G.A. Whitmore. On the mixture of the inverse gaussian distribution with its complementary reciprocal. Scand. J. Stat.,1991, 18: 77-89.

[2]Bowonrattanaset, P., Point estimation for the crack lifetime distribution. Ph.D. Thesis, Thammasat University, Thailand, 2011.

[3]Balakrishnan, N., V. Leiva, A. Sanhueza and E. Cabrera,Mixture inverse gaussian distributions and its transformations, moments and applications. Statistics, 2009, 43: 91-104.

[4] Phitchaphat Bowonrattanaset and Kamon Budsaba. Some Properties of the Three-Parameter Crack Distribution. Journal of Thailand Statistician, 2011. 9: 195-203.

[5] Birnbaum, Z. W., & Saunders, S. C. A new family of life distributions, J. Applied Probability, 1969; 6: 319-327.

[6] Birnbaum, Z. W., & Saunders, S. C. Estimation for a family of life distributions with applications to fatigue, J. Applied Probability, 1969; 6: 328-347.

[7] Desmond, A. F. On the relationship between two fatigue-life models, IEEE Reliability, 1986; 35: 167-169.

[8] Shuster, J. On the Inverse Gaussian distribution function, J.American Statistical Association, 1968; 63: 1514-1516.

[9] Chhikara, R. S., & Folks, J. L. The inverse Gaussian distribution-theory, methodology, and applications, Now York: Marcel Dekker, 2008.

[10] Khattree, R. Characterization of inverse-Gaussian and gamma distributions through their length-biased distributions, IEEE Reliability, 1989; 38: 610-611.

[11] Ahmed S. E., Budsaba K., Lisawadi S., & Volodin, A. I. Parametric estimation for the Birnbaum-Saunders lifetime distribution based on a new parametrization, Thailand Statistician, 2008; 6: 213-240.

[12] Gradshteyn, I. S., & Ryzhik, I. M. Table of integrals, series, and products (7ed.), MA, Elsevier Academic Press, 2007.

[13] Jorgensen, B., V. Seshadri and G.A. Whitmore. On the mixture of the inverse gaussian distribution with its complementary reciprocal. Scand. J. Stat., 1991. 18: 77-89.

[14]Balakrishnan, N., V. Leiva, A. Sanhueza and E. Cabrera, 2009. Mixture inverse gaussian distributions and its transformations, moments and applications. Statistics, 43: 91-104.

[15] Bowonrattanaset, P. Point estimation for the crack lifetime distribution. Ph.D. Thesis, Thammasat University, Thailand. 2001.

[16] Pornpop Saengthong and Winai Bodhisuwan. A New Two-parameter Crack Distribution. Journal of Applied Sciences, 2014. 14: 758-766.

# A numerical solution of fractional Black-Scholes equation

P.Phaochoo*, A.Luadsong

*Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi (KMUTT)*
*126 Pracha-utid Road, Bangmod, Toongkru, Bangkok, 10140, Thailand*
*e-mail: prasert.p@rmutk.ac.th:anirut.lua@kmutt.ac.th.*

**Abstract**

The Black-Scholes equation is the famous financial model that relates with an option. Nowadays, a fractional calculus is an excellent tool for solving problems in many vital fields. A fractional differential equation are widely used for many research with continuous and discrete approaches. In this research, the fractional Black-Scholes equation in financial problem is solved by using the numerical method. This equation is a fractional partial differential equation for the option price of a European call or European put under the Black-Scholes model. The implicit finite difference method and MLPG2 are used for discretizing the governing equation in time variable and option price, respectively. The time fractional derivative uses the Caputo partial derivative of fractional order $\alpha$. The numerical examples for varieties of variables are also included.

*Keywords*: European option, fractional Black-Scholes equation, MLPG, moving kriging interpolation.

## 1. Introduction

The first idea of fractional calculus is considered to be the Leibniz's letter to L'Hospital in 1965. Fractional calculus is a name for the theory of derivatives and integrals of arbitrary order. The famous definition of a fractional calculus are the Riemann-Liouville and Grunwald-Letnikov definition [1]. Caputo reformulated the definition of the Riemann-Liouvillein order to use integer order initial conditions to solve fractional differential equation [2]. Fractional differential equation have attracted much attention during the past few decades. This is the fact that fractional calculus supply an competent and excellent tool for the description of many important phenomena such as electromagnetic, physics, chemistry, biology, economy and many more.

Black-Scholes equation, which is proposed by Fisher Black and Myron Scholes [3], is the financial model that concern with option. An option is a contract between the seller and the buyer. It consists of a call option and a put option. Option valuation depends on the underlying asset price and time. The European option can only be exercised at the expiration date, but the American option can be exercised at any time before expiration date. The solution of Black-Scholes equation provides an option pricing formula for European option. The analytic solution is used in general case with basic assumption but it is not satisfied in some conditions. Some restrictions were appeared in the classical Black-Scholes equation that is the weaken of this model. Original assumptions were relieved by other models such as models with transaction cost [4-5], Jump-diffusion model [6], Stochastic volatility model [7] and Fractional Black-Scholes model [8-9].

Fractional Black-Scholes model is derived by many researchers. Some restrictions were appeared in the classical Black-Scholes equation that is the weaken of this model [10]. The Fractional Black-Scholes models are derived by substitute the standard Brownian motion with fractional Brownian motion.

In this paper, we propose a numerical method base on Meshless Local Petrov-Galerkin (MLPG) method to solve a fractional Black-Scholes equation. The MLPG is a truly meshless method, which involves not only a meshless interpolation for the trial functions, but also a meshless integration of the weak-form, [11]. MLPG2 is chosen for this research so the Kronecker delta is the test function. This method will avoid the domain integral in the weak-form.

## 2. Problem Formulation

The Black-Scholes equation is the outstanding financial equation that solve the European option pricing without a transaction cost. Moreover, underlying asset price distributed on the lognormal random walk, risk-free interest rate, no dividend and no arbitrate opportunity are fundamental assumption. The fractional Black-Scholes equation is following

$$\frac{\partial^\alpha u}{\partial \tau^\alpha}+r(\tau)s\frac{\partial u}{\partial s}+\frac{1}{2}\sigma^2(s,\tau)s^2\frac{\partial^2 u}{\partial s^2}-r(\tau)u=0 \tag{2.1}$$

$(s,\tau)\in\mathbb{R}^+\times[0,T]$ with the terminal and boundary condition

$$u(s,T)=\max(s-E,0), s\in\mathbb{R}^+, u(0,\tau)=0, \tau\in[0,T],$$

where *u(s, $\tau$ )* is the value of European call option at underlying asset price $s$ at time $\tau$, $T$ *is* the expiration date, *r* is the risk-free interest rate, $\sigma$ is the volatility of underlying asset price and *E* is the strike price. Basic definition of fractional calculus as following

**Definition 1.** The Riemann-Liouville fractional integral operator of order $\alpha > 0$, of a function $f(t) \in C_\mu,\ \mu \ge -1$ is defined as [12],

$$J^\alpha f(t) = \frac{1}{\Gamma(\alpha)}\int_0^t (t-\tau)^{\alpha-1} f(\tau)\,d\tau, (\alpha > 0)$$

$$J^0 f(t) = f(t).$$

For the The Riemann-Liouville fractional integral, we have:

$$J^\alpha t^\gamma = \frac{\Gamma(\gamma+1)}{\Gamma(\gamma+\alpha+1)} t^{\alpha+\gamma}$$

**Definition 2.** The fractional derivative of *f(t)* in the Caputo sense is defined as [13],

$$D_\tau^\alpha f(t) = J^{m-\alpha} D^m f(t) = \frac{1}{\Gamma(m-\alpha)}\int_0^t (t-\tau)^{m-\alpha-1} f^{(m)}(\tau)\,d\tau,$$

for $m-1 < \alpha \le m, m \in N, t > 0.$

For the Riemann-Liouville fractional integral and the Caputo fractional derivative, we have the following relation

$$J_\tau^\alpha D_\tau^\alpha f(t) = f(t) - \sum_{k=0}^{m-1} f^{(k)}(0_+)\frac{t^k}{k!}.$$

**Definition 3.** The Mittag-Leffler is defined as [14]

$$E_\alpha(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k+1)}, (\alpha \in C, \mathrm{Re}(\alpha) > 0).$$

From Eq.(2.1), when $s$ goes to zero then degenerating will occur in approximation. We transform the Black-Scholes equation into a nondegenerate partial differential equation by using a logarithmic

transformation $x = \ln s, t = \frac{1}{2}\sigma^2(T-\tau),$ and define the

computational domain for convenient in numerical experiments by

$\Omega = \left[x_{\min,} x_{\max}\right] \times [0,T]$ ,where

$x_{\min} = -\ln(4E), x_{\max} = \ln(4E)$, [15].

$$-\frac{1}{2}\sigma^2 \frac{\partial^\alpha u}{\partial t^\alpha} + r\frac{\partial u}{\partial x} + \frac{1}{2}\sigma^2\left(\frac{\partial^2 u}{\partial x^2} - \frac{\partial u}{\partial x}\right) - ru = 0, \quad (2.2)$$

$$\frac{\partial^\alpha u}{\partial t^\alpha} = \frac{2r}{\sigma^2}\frac{\partial u}{\partial x} - \frac{\partial u}{\partial x} + \frac{\partial^2 u}{\partial x^2} - \frac{2r}{\sigma^2}u, \text{ where } k = \frac{2r}{\sigma^2}$$

Therefore $$\frac{\partial^\alpha u}{\partial t^\alpha} = (k-1)\frac{\partial u}{\partial x} + \frac{\partial^2 u}{\partial x^2} - ku, \quad (2.3)$$

$$u(x,0) = \max\left(e^x - E, 0\right), x \in (x_{\min}, x_{\max}),$$

$$u\left(x_{\min,}, t\right) = 0, u\left(x_{\max}, t\right) = e^{x_{\max}} - Ee^{-\int_0^t r(s)ds}, \ t \in [0,T].$$

## 3. Spatial Discretization

The MLPG method construct the local weak form over local subdomain, which is a small region taken for each node in global domain. Multiplying test function $v_i$ into Eq.(2.3) and then integrate over subdomain $(\Omega_s^i)$ which is located inside the global domain ( $\Omega$ ) yields the following expression

$$\int_{\Omega_s^i} \frac{\partial^\alpha u}{\partial t^\alpha} v_i\, d\Omega = \int_{\Omega_s^i} \left(\frac{\partial^2 u}{\partial x^2} + (k-1)\frac{\partial u}{\partial x} - ku\right)v_i\, d\Omega, \quad (3.1)$$

Where $v_i$ is a test function that make significant for each nodes. Rearrange Eq.(3.1), we have

$$\int_{\Omega_s^i} \frac{\partial^\alpha u}{\partial t^\alpha} v_i d\Omega = \int_{\Omega_s^i} u,_{xx} v_i d\Omega + \int_{\Omega_s^i} (k-1)u,_x v_i d\Omega - \int_{\Omega_s^i} kuv_i d\Omega, \tag{3.2}$$

Where $u,_{xx} = \dfrac{\partial^2 u}{\partial x^2}, u,_x = \dfrac{\partial u}{\partial x}$. Substituting trial function $u^h(x,t) = \sum_{j=1}^{N} \phi_j(x)\hat{u}_j(t)$ into $u$ in Eq.(3.2)

$$\int \sum_{j=1}^{N} \phi_j(x) v_i(x) \frac{\partial^\alpha \hat{u}_j(t)}{\partial t^\alpha} d\Omega = \int_{\Omega_s^i} \sum_{j=1}^{N} \phi_{j,xx}(x)\ v_i(x)\hat{u}_j(t) d\Omega$$
$$+(k-1)\int_{\Omega_s^i} \sum_{j=1}^{N} \phi_{j,x}(x) v_i(x)\hat{u}_j(t) d\Omega$$
$$-k\int_{\Omega_s^i} \sum_{j=1}^{N} \phi_j(x) v_i(x)\hat{u}_j(t) d\Omega \tag{3.3}$$

where $N$ is the number of nodes surrounding point $x$ which has the effect on *u(x)* and $\hat{u}_j$ is value of option at time *t*. The shape function, $\phi_j$, is constructed by moving kriging interpolation which has the Kronecker delta property, thereby enhancing the arrangement nodal shape construction accuracy. Rearrange Eq.(3.3) yields the following result

$$\sum_{j=1}^{N} \int_{\Omega_s^i} \phi_j(x) v_i(x) \left( \frac{\partial^\alpha \hat{u}_j(t)}{\partial t^\alpha} \right) d\Omega = \sum_{j=1}^{N} \int_{\Omega_s^i} \phi_{j,xx}(x) v_i(x)\hat{u}_j(t) d\Omega$$
$$+(k-1)\sum_{j=1}^{N} \int_{\Omega_s^i} \phi_{j,x}(x) v_i(x)\hat{u}_j(t) d\Omega$$
$$-k\sum_{j=1}^{N} \int_{\Omega_s^i} \phi_j(x) v_i(x)\hat{u}_j(t) d\Omega \tag{3.4}$$

This research use MLPG2 then the test function is chosen by Kronecker delta function,

$$v_i(x) = \begin{cases} 0, & x \neq x_i \\ 1, & x = x_i \end{cases}, \ i = 1, 2, \ldots, N.$$

The test function will define significance for each node in subdomain. In this case, substituting test function $v_i(x)$ to Eq.(3.4) and then integrate over subdomain $\Omega_s^i$ yields the following result

$$\sum_{j=1}^{N} \phi_j(x_i) \frac{d^\alpha \hat{u}_j(t)}{dt^\alpha} = \sum_{j=1}^{N} [\phi_{j,xx}(x_i) + (k-1)\phi_{j,x}(x_i)$$
$$-k\phi_j(x_i)]\hat{u}_j(t) \tag{3.5}$$

Eq.(3.5) can be written in the matrix form as following

$$A\frac{d^\alpha U}{dt^\alpha} = BU, \tag{3.6}$$

where $A = [A_{ij}]_{N\times N}, A_{ij} = \phi_j(x_i)$,

$B = [B_{ij}]_{N\times N}, B_{ij} = \phi_{j,xx}(x_i) + (k-1)\phi_{j,x}(x_i) - k\phi_j(x_i)$,

$U = [\hat{u}_1\ \hat{u}_2\ \hat{u}_3 \ldots \hat{u}_N]^T$

Since the shape function that is constructed by the moving kriging interpolation satisfy the Kronecker delta property, *A* is the identity matrix. Therefore, Eq.(3.6) can be written as

$$\frac{d^\alpha U}{dt^\alpha} = BU \tag{3.7}$$

## 4. Temporal Discretization

The numerical solution of European option use the implicit finite difference method. By a finite approximation made for the time fractional derivative with notation $\dfrac{\partial^\alpha u(x_i,t_n)}{\partial t^\alpha}$ that approximates the exact solution $u(x_i,t_n)$ at time level n, we restrict attention to the finite space domain $x_{\min} < x < x_{\max}$ with $0 < \alpha < 1$. The time fractional derivative use the implicit finite difference [16], defined by

$$\frac{d^\alpha U}{dt^\alpha} = \sigma_{\alpha,\Delta t} \sum_{j=1}^{n} \left( U^{n-j+1} - U^{n-j} \right) + O(\Delta t) \tag{4.1}$$

where $\sigma_{\alpha,\Delta t} = \dfrac{1}{\Gamma(1-\alpha)} \dfrac{1}{1-\alpha} \dfrac{1}{(\Delta t)^\alpha}$.

Hence, $\dfrac{d^\alpha U}{dt^\alpha} = D_t^{(\alpha)} U_i^n + O(\Delta t)$.

The first-order approximation method for the computation of Caputo's fractional derivative is given by

$$D_t^{(\alpha)} U^n = \sigma_{\alpha,k} \sum_{j=1}^{n} \omega_j^{(\alpha)} \left( U^{n-j+1} - U^{n-j} \right) \tag{4.2}$$

where $\omega_j^{(\alpha)} = j^{1-\alpha} - (j-1)^{1-\alpha}$.

Consider the Eq. (4.2) and substitute time fractional derivative that following

$$\sigma_{\alpha,k}\sum_{j=1}^{n}\omega_j^{(\alpha)}\left(U^{n-j+1}-U^{n-j}\right)+O\left(\Delta t\right)=BU^n,$$

$$\sigma_{\alpha,k}\omega_1^{(\alpha)}\left(U^n-U^{n-1}\right)=-\sigma_{\alpha,k}\sum_{j=2}^{n}\omega_j^{(\alpha)}\left(U^{n-j+1}-U^{n-j}\right)+BU^n,$$

We consider the first case for $n=1$,

Case $n=1$

$$\sigma_{\alpha,k}\omega_1^{(\alpha)}\left(U^1-U^0\right)=BU^1,$$
$$(\sigma_{\alpha,k}\omega_1^{(\alpha)}I-B)U^1=\sigma_{\alpha,k}\omega_1^{(\alpha)}U^0,$$

Case $n\geq 2$

$$\left(\sigma_{\alpha,k}\omega_1^{(\alpha)}I-B\right)U^n=\sigma_{\alpha,k}\omega_1^{(\alpha)}U^{n-1}$$
$$-\sigma_{\alpha,k}\sum_{j=2}^{n}\omega_j^{(\alpha)}\left(U^{n-j+1}-U^{n-j}\right).$$

## 5. Numerical Examples

In this section, we are going to present various numerical results to evaluate proposed meshless approaches. Using the MLPG2 method, the resulting problems for European call options are solved via implicit finite difference method.

The European call option can be modeled by fractional Black-Scholes PDE as following : [17]

$$\frac{\partial^\alpha u}{\partial t^\alpha}=\frac{\partial^2 u}{\partial x^2}+\left(k-1\right)\frac{\partial u}{\partial x}-ku,\ 0<\alpha<1. \tag{5.1}$$

where $k=\dfrac{2r}{\sigma^2}$.

with initial condition given by

$$u\left(x,0\right)=\max\left(e^x-E,0\right),\ x\in\mathbb{R}^+. \tag{5.2}$$

The analytical solution for the European call option is

$$u\left(x,t\right)=\max\left(e^x,0\right)(1-E_\alpha(-kt^\alpha))+\max\left(e^x-1,0\right)E_\alpha(-kt^\alpha) \tag{5.3}$$

where $E_\alpha(-kt^\alpha)=\sum_{n=0}^{\infty}\dfrac{(-kt^\alpha)^n}{\Gamma(n\alpha+1)}$

**Example 1.** We consider the fractional Black-Scholes equation in Eq.(5.1). The numerical simulation was done for European call option with parameters as following:

Case 1. For $\sigma=0.2, r=0.04, \alpha=0.5, T=2, k=2$. In this case , we get the exact solution as following : $u\left(x,t\right)=\max\left(e^x,0\right)(1-e^{-2t})+\max\left(e^x-1,0\right)e^{-2t}$,

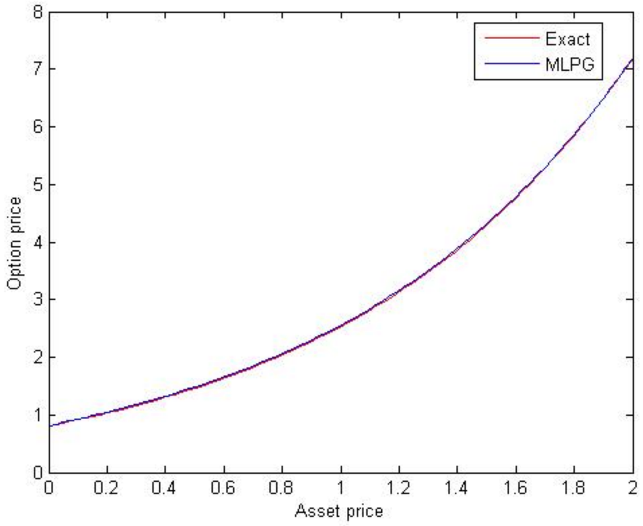


Figure 1. The approximate solution compare with the exact solution for $\sigma=0.2, r=0.04,$ $\alpha=0.5, k=2, t=T.$

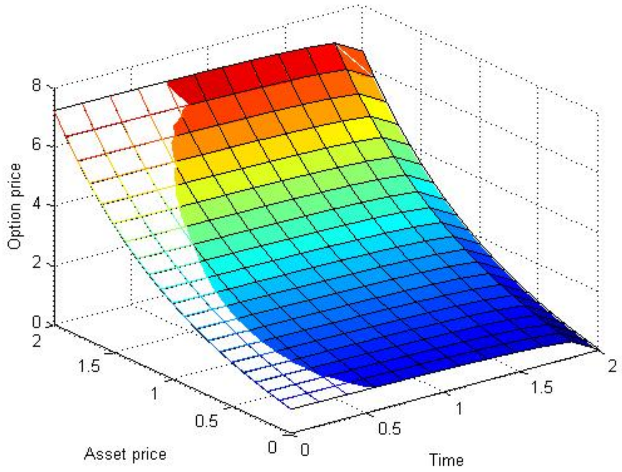


Figure 2. The approximate solution compare with the exact solution for $\sigma=0.2, r=0.04, \alpha=0.5, k=2,$ $0\leq t\leq T.$

Case 2. For $\sigma=0.2, r=0.01, \alpha=0.99, T=4, k=5$ In this case, we get the exact solution as following : $u\left(x,t\right)=\max\left(e^x,0\right)(1-e^{-5t})+\max\left(e^x-1,0\right)e^{-5t}$,

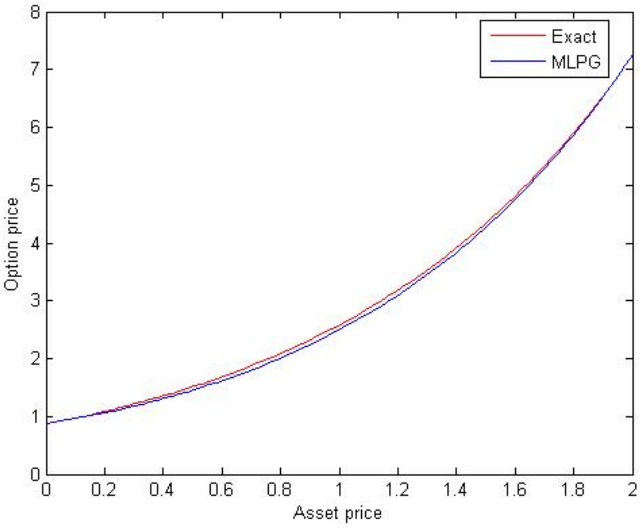


Figure 3. The approximate solution compare with the exact solution for $\sigma = 0.2, r = 0.01, \alpha = 0.99,$

$k = 5, t = T.$

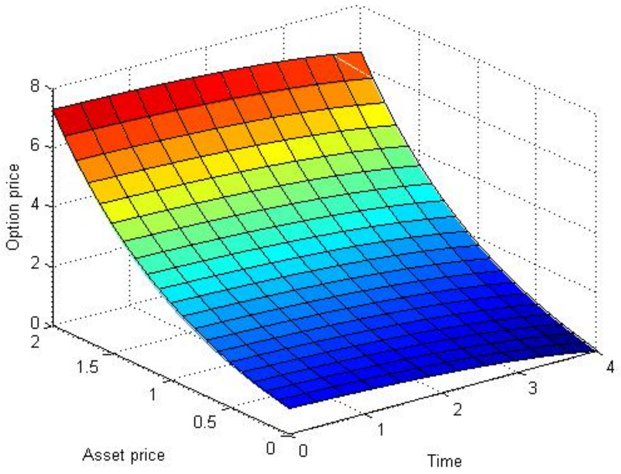


Figure 4. The approximate solution compare with the exact solution for $\sigma = 0.2, r = 0.01, \alpha = 0.99, k = 5,$

$0 \le t \le T.$

Case 3. For $\sigma = 0.1, r = 0.06, \alpha = 0.99, T = 1, k = 12$ . In this case , we get the exact solution as following :

$$u(x,t) = \max\left(e^x, 0\right)(1 - e^{-12t}) + \max\left(e^x - 1, 0\right)e^{-12t},$$

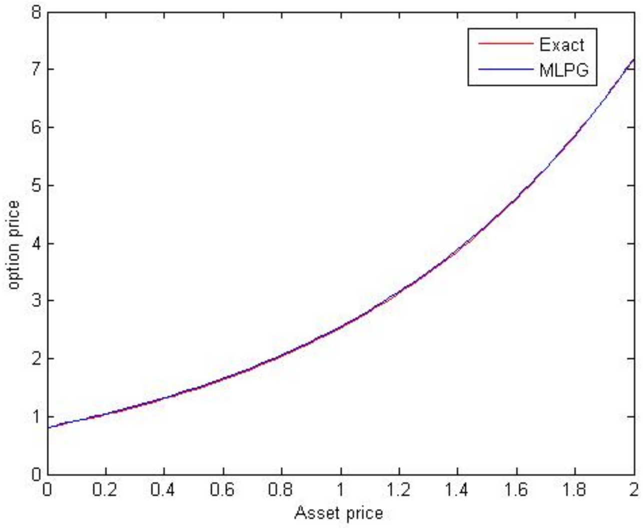


Figure 5. The approximate solution compare with the exact solution for $\sigma = 0.1, r = 0.06, \alpha = 0.99,$

$k = 12, t = T.$

.

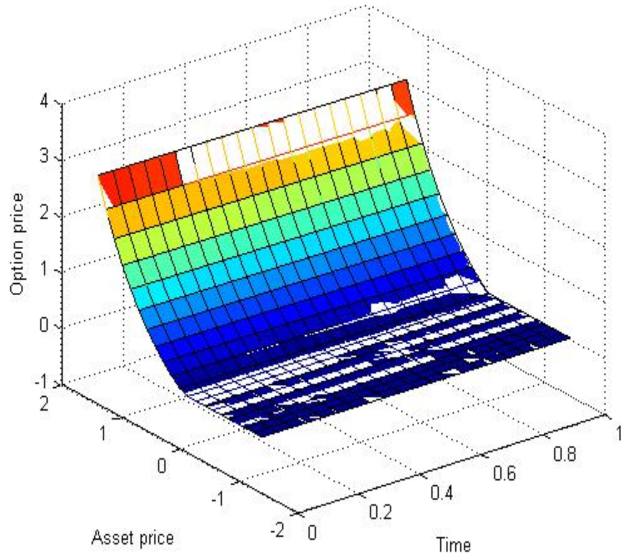


Figure 6. The approximate solution compare with the exact solution for $\sigma = 0.1, r = 0.06, \alpha = 0.99, k = 12$

$0 \le t \le T.$

Case 4. For $\sigma = 0.4(2 + \sin x), r = 0.06,$

$\alpha = 0.99, T = 1, k = \dfrac{2r}{\sigma^2}$ .

In this case , the exact solution is unknown.

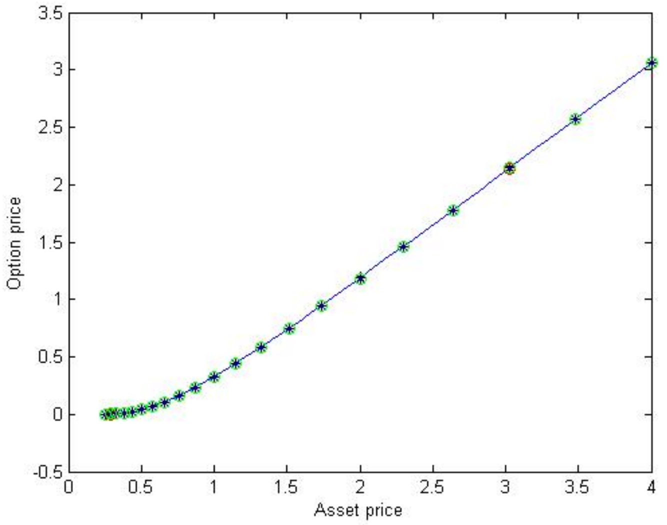


Figure 7. The comparison of the approximate solutions of the fractional and standard Black-Scholes equation

for $r = 0.06, \alpha = 0.99, k = \dfrac{2r}{\sigma^2}, t = T.$

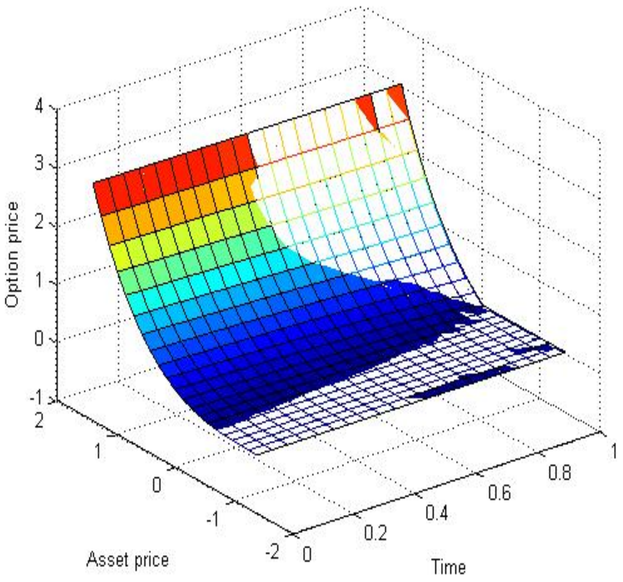


Figure 8. The comparison of the approximate solutions of the fractional and standard Black-Scholes equation $r = 0.06, \alpha = 0.99, k = \dfrac{2r}{\sigma^2}, 0 \le t \le T.$

## 6. Conclusion

In this paper, the fractional Black-Scholes equations are solved by the implicit finite difference method and MLPG2 for discretizing in time variable and option price, respectively. The Caputo partial derivative of fractional order $\alpha$ are used for numerical scheme.

The numerical results are presented in four cases. Case 1 and 2 presents numerical results for varieties of parameters and $\alpha = 0.5, 0.99.$ In figure 1.and 3., we found that the last time have a little difference value of option. In figure 2., the value of option will only differ for initial time and case $\alpha = 0.99,$ the value of option have no difference from exact solution for all time. Case 3 present various parameters and final case show comparison of approximation solutions of the fractional Black-Scholes equation and standard Black-Scholes equation.

**Acknowledgements**

This research is partially supported by King Mongkut's University of Technology Thonburi (KMUTT). The authors would like to thank their advisor for providing advice and taking care of this research and Rajamangala University of Technology Krungthep for providing a scholarship.

## References

[1] Ghandehari MAM, Ranjbar M. European option pricing of fractional Black-Scholes model with new Lagrange multipliers. Computational Methods for Differential Equations. 2014; 2(1): 1-10.

[2] Ishteva MK. Properties and Applications of the Caputo Fractional Operator [Master Thesis]. Sofia: Karlsruhe Univ; 2005.

[3] Black F, Scholes M. The pricing of options and corporate liabilities. Journal of Political Economy. 1973; 81(3): 637-654.

[4] Barles G, Soner HM. Option pricing with transaction costs and a nonlinear Black-Scholes equation. Finance and Stochastics . 1998; 2: 369-397.

[5] Davis MHA, Panas VG, and Zariphopoulou T. European option pricing with transaction costs. SIAM Journal on Control and Optimization. 1993; 470-493.

[6] Merton RC. Option when underlying stock returns are discontinuous. Journal of Financial Economics. 1976; 3: 125-144.

[7] Hull JC, White AD. The pricing of options on aseet with stochastic volatilities. Journal of finance. 1987; 42: 281-300.

[8] Bjork T,Hult H. A note on Wick products and the fractional Black-Scholes model. Finance and stochastics. 2005; 9: 197-209.

[9] Wang XT. Scaling and long-range dependence in option pricing I: Pricing European option with transaction costs under the fractional Black-Scholes model. Physica. 2010; 438-444.

[10] Song L,Wang W. Solution of the Fractional Black-Scholes Option Pricing Model by Finite Difference Method. Abstract and Applied Analysis.2013;

[11] Atluri N, Shen S. The Meshless Local Petrov-Galerkin (MLPG) method: A Simple & Less-costly Alternative to the Finite Element and Boundary Element Methods. *CMES*. 2002; 3(1): 11-51.

[12] Podlubny I. Fractional differential calculus. New York: Academic Press; 1999.

[13] Caputo M. Elasticita e Dissipazione. Bologna: Zani-Chelli; 1969.

[14] Mittag-Leffler Gm. Sur la nouvelle function . CR Acad Sci Paris(Ser. II). 1903; 137: 554-558.

[15] Haung J, Cen Z. Cubic Spline Method for a Generalized Black-Scholes Equation. Mathematical Problems in Engineering. 2014; 7(3): 229–235.

[16] Murio DA. Implicit finite difference approximation for time fractional diffusion equations. Computers and Mathematics with Applications.2008; 56: 1138-1145.

[17] Kumar A, Tripathi LP and Kadalbajoo MK. A numerical study of Asian option with radial basis functions based finite differences method. Engineering Analysis with Boundary Elements. 2014;50: 177-183.

# The Negative Hypergeometric Distribution Revisited with Applications to Gaming

Rohan J. Dalpatadu[1*] and Steven N. Jones[2]
[1]*Department of Mathematical Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154-4020, U.S.A..*
[2]*Department of Mathematical Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154-4020, U.S.A.*

**Abstract**

The negative hypergeometric distribution, which may be viewed as a finite negative binomial distribution, arises when sampling from a population without replacement as opposed to sampling with replacement that gives rise to the negative binomial distribution. This distribution is widely used in the gaming industry. The derivations of the mean and variance of the above distribution use advanced techniques. In this paper, we will derive the expressions for the mean and the variance of the negative hypergeometric distribution using elementary mathematical techniques. Furthermore, we will derive further results that have applications in gaming.

*Keywords*: Negative Hypergeometric Distribution: Gaming: Award: Mean: Variance

*Corresponding Author
E-mail Address: dalpatad@unlv.nevada.edu

## 1. Introduction

The negative hypergeometric distribution is a “forgotten” distribution in the sense that it is not widely used in applied statistics. In this paper, we reintroduce this distribution with fairly elementary derivations for its mean and variance along with applications to gaming.

## 2. The Negative Hypergeometric Distribution

Consider a finite population of size *N* that has *n* objects of a first kind and $N-n$ objects of a second kind. Objects are drawn from this population at random without replacement until *k* objects of the first kind are drawn. Let *X* be the number of the draw in which this happens. Then *X* has a negative hypergeometric distribution with probability function

$$p_X(x) = \Pr\left[(k-1) \text{ of first kind in } (x-1) \text{ draws}\right]$$

$$\times \Pr\left[\text{first kind in } k^{\text{th}} \text{ draw}\right]$$

$$= \frac{\binom{N-n}{x-k}\binom{n}{k-1}}{\binom{N}{x-1}} \frac{(n-k+1)}{(N-x+1)} = \frac{\binom{x-1}{k-1}\binom{N-x}{n-k}}{\binom{N}{n}};$$

$$k \le x \le k+N-n.$$

This distribution is more like a modified truncated negative binomial distribution. The following Lemmas [1] are needed to provide elementary derivations of the expressions for *E*[*X*] and *Var*(*X*).

**Lemma 1:** Let *k*, *n*, and *N* be positive integers such that $k \le n < N.$ Then

$$\sum_{y=0}^{N-n}\binom{y+k-1}{k-1}\binom{N-k-y}{n-k} = \binom{N}{n}.$$

**Proof:**

$$\sum_{x=k}^{k+N-n} p_X(x) = 1 \Rightarrow \sum_{y=0}^{N-n} p_X(y+k) = 1,$$

$$\Rightarrow \sum_{y=0}^{N-n} \frac{\binom{y+k-1}{k-1}\binom{N-y-k}{n-k}}{\binom{N}{n}} = 1.$$

Hence,

$$\sum_{y=0}^{N-n}\binom{y+k-1}{k-1}\binom{N-k-y}{n-k} = \binom{N}{n}.$$

**Lemma 2:** Let *p*, *q*, and *r* be positive integers such that $p > q.$ Then

$$\sum_{y=0}^{p-q}\binom{y+r}{r}\binom{p-y}{q} = \binom{p+r+1}{q+r+1}.$$

**Proof:** Let $N = p+r+1,\ n = q+r+1,\ k = r+1.$ Then *k*, *n*, and *N* are integers such that $1 \le k \le n < N.$ Therefore, $p = N-k, q = n-k, p-q = N-n,\ r = k-1,$ and by Lemma 1,

$$\sum_{y=0}^{p-q}\binom{y+r}{r}\binom{p-y}{q} = \sum_{y=0}^{N-n}\binom{y+k-1}{k-1}\binom{N-k-y}{n-k}$$

$$= \binom{N}{n} = \binom{p+r+1}{q+r+1}.$$

Using Lemma 2, we can now establish the expressions for *E*[*X*] and *Var*(*X*). [1]

**Proposition 1:** The expected value and the variance of *X* are given by:

$$E[X] = \mu_X = \frac{N+1}{n+1}k$$

$$Var(X) = \sigma_X{}^2 = \frac{(N+1)(N-n)}{(n+1)^2(n+2)}k(n-k+1).$$

**Proof:**

$$\sum_{x=k}^{k+N-n} x\binom{x-1}{k-1}\binom{N-x}{n-k}$$

$$= \sum_{y=0}^{N-n} (y+k) \binom{y+k-1}{k-1} \binom{N-k-y}{n-k}$$

$$= \sum_{y=0}^{N-n} (y+k) \frac{(y+k-1)!}{(k-1)!y!} \binom{N-k-y}{n-k}$$

$$= \sum_{y=0}^{N-n} k \binom{y+k}{k} \binom{N-k-y}{n-k}$$

$$= k \binom{(N-k)+k+1}{(n-k)+k+1} = k \binom{N+1}{n+1},$$

using Lemma 2. Hence,

$$E[X] = \sum_{x=k}^{k+N-n} x p_X(x) = \sum_{x=k}^{k+N-n} x \frac{\binom{x-1}{k-1}\binom{N-x}{n-k}}{\binom{N}{n}}$$

$$= \frac{k\binom{N+1}{n+1}}{\binom{N}{n}} = \frac{N+1}{n+1}k.$$

$$\sum_{x=k}^{k+N-n} x(x+1) \binom{x-1}{k-1} \binom{N-x}{n-k}$$

$$= \sum_{y=0}^{N-n} (y+k)(y+k+1) \binom{y+k-1}{k-1} \binom{N-k-y}{n-k}$$

$$= \sum_{y=0}^{N-n} (y+k)(y+k+1) \frac{(y+k-1)!}{(k-1)!y!} \binom{N-k-y}{n-k}$$

$$= \sum_{y=0}^{N-n} k(k+1) \binom{y+k+1}{k+1} \binom{N-k-y}{n-k}$$

$$= k(k+1) \binom{(N-k)+(k+1)+1}{(n-k)+(k+1)+1} = k(k+1) \binom{N+2}{n+2},$$

using Lemma 2. Hence,

$$E[X(X+1)] = \sum_{x=k}^{k+N-n} x(x+1) p_X(x)$$

$$= \sum_{x=k}^{k+N-n} x(x+1) \frac{\binom{x-1}{k-1}\binom{N-x}{n-k}}{\binom{N}{n}}$$

$$= \frac{k(k+1)\binom{N+2}{n+2}}{\binom{N}{n}} = \frac{(N+1)(N+2)}{(n+1)(n+2)} k(k+1).$$

Therefore, the variance is given by

$$Var(X) = E[X(X+1)] - E[X] - \{E[X]\}^2$$

$$= \frac{(N+1)(N+2)}{(n+1)(n+2)} k(k+1) - \frac{N+1}{n+1}k - \left(\frac{N+1}{n+1}k\right)^2$$

$$= \frac{k(N+1)\left[-n^2 + (N+k-1)n - (k-1)N\right]}{(n+1)^2(n+2)}$$

$$= \frac{-k(N+1)(n-N)(n-k+1)}{(n+1)^2(n+2)}$$

$$= \frac{(N+1)(N-n)}{(n+1)^2(n+2)} k(n-k+1).$$

The expected value is linear in *k*, and the variance is quadratic in *k* with a maximum at $k = \frac{n+1}{2}$ if *n* is odd and at both $k = \frac{n}{2}$ and $k = \frac{n}{2}+1$ if *n* is even. The minimum value of the variance is $\frac{n(N+1)(N-n)}{(n+1)^2(n+2)}$ at $k = 1$ and $k = n.$ The above elementary proofs are given in [1]. Other elaborate proofs are given in [2].

**Note:** Suppose $\lim_{N\to\infty} \frac{n}{N} = p \in (0,1).$ Then

$E(X) = \frac{N+1}{n+1}k = \frac{1+1/N}{n/N+1/N}k \to \frac{k}{p}$ as $n \to \infty$ and

$$Var(X)$$

$$= \frac{(1+1/N)(1-n/N)k(n/N - k/N + 1/N)}{(n/N+1/N)^2(n/N+2/N)}$$

$\to \frac{1-p}{p^2 p} kp = \frac{k(1-p)}{p^2}$ as $n \to \infty.$

The above limiting expected value and variance are that of a negative binomial distribution with parameters *p* and *k*.

### 3. Applications to Gaming

In applications to gaming, each draw *i* with $i < x,$ where *x* is the draw in which the *k*th object of the first kind is drawn, is associated with a payout $A_{j(i)}$ if this draw does not reveal an object of the first kind; else it a fixed amount $A_0$. The mapping $i \to j(i)$ is one-to-one. Let us define $Y = X - k,$ where *X* is the negative hypergeometric random variable. Then the expected value and variance of *Y* are given by

$$E[Y] = \mu_Y = E[X-k] = \mu_X - k$$

and

$$Var(Y) = \sigma_Y{}^2 = Var(X) = \sigma_X{}^2.$$

**Proposition 2:** The total winnings on the type of game given above is a random variable

$$W = \sum_{i=1}^{X-k} A_{j(i)} + kA_0$$

whose expected value and variance are given by

$$E[W] = \mu_Y \overline{A} + kA_0 = \mu_X \overline{A} - k\left(\overline{A} - A_0\right)$$

and

$$Var\left(W\right)=\frac{\left(N-n\right)\bar{A}^2+\left(n+1\right)\overline{A^2}}{N+1}\sigma_Y{}^2$$

$$=\frac{\sigma_X{}^2}{\mu_X}\left[k\left(\overline{A^2}-\bar{A}^2\right)+\mu_X\,\bar{A}^2\right],$$

where

$$\bar{A}=\frac{1}{N-n}\sum_{j=1}^{N-n}A_j \text{ and } \overline{A^2}=\frac{1}{N-n}\sum_{j=1}^{N-n}A_j{}^2.$$

**Proof:** Let $V=\sum_{i=1}^{X-k}A_{j(i)}$. We will first show that $E\left[V\right]=\mu_Y\,\bar{A}=\left(\mu_X-k\right)\bar{A}.$ Each outcome $\omega$ of the game is a subset of $\{0,1,2,...,N-n\}$. The number of elements of $\omega$, denoted by $c\left(\omega\right)$ is defined to be the size of $\omega$. Let $\Omega_r=\{\omega\,|\,c\left(\omega\right)=r\}$. Then we see that the set of all possible outcomes is given by $\Omega=\bigcup_{r=0}^{N-n}\Omega_r$.

For each $j=1,2,...,N-n$ define $I_j\left(\omega\right)=\begin{cases}1, & j\in\omega\\ 0, & j\notin\omega\end{cases}$ where $\omega\in\Omega$. Then the total number of outcomes of size *r* containing *j* is given by

$$\sum_{\omega\in\Omega_r}I_j\left(\omega\right)=\binom{N-n-1}{r-1}=\frac{r}{N-n}\binom{N-n}{r}$$

$$=\frac{r}{N-n}c\left(\Omega_r\right);\ r\geq 1.$$

The total number of outcomes of size *r* containing both *j* and *m* is given by

$$\sum_{\omega\in\Omega_r}I_j\left(\omega\right)I_m\left(\omega\right)=\binom{N-n-2}{r-2}$$

$$=\frac{r\left(r-1\right)}{\left(N-n\right)\left(N-n-1\right)}\binom{N-n}{r}$$

$$=\frac{r\left(r-1\right)}{\left(N-n\right)\left(N-n-1\right)}c\left(\Omega_r\right);\ r\geq 2.$$

The probability of an outcome is

$$\Pr\left(\omega\right)=\Pr\left[\omega\,|\,Y=c\left(\omega\right)\right]\Pr\left[Y=c\left(\omega\right)\right]$$

$$=\left\{c\left(Y\right)\right\}^{-1}\Pr\left[Y=c\left(\omega\right)\right]$$

and the winnings on an outcome is

$$w\left(\omega\right)=\sum_{j=1}^{N-n}I_j\left(\omega\right)A_j+kA_0.$$

The sum $v\left(\omega\right)=\sum_{j=1}^{N-n}I_j\left(\omega\right)A_j$ is equal to $V=\sum_{i=1}^{X-k}A_{j(i)}$.

Therefore,

$$E\left[V\,|\,Y=y\right]=\frac{1}{\Pr\left[Y=y\right]}\sum_{\omega\in\Omega_y}v\left(\omega\right)\Pr\left[\omega\right]$$

$$=\frac{1}{\Pr\left[Y=y\right]}\sum_{\omega\in\Omega_y}\left[\frac{\Pr\left[Y=c\left(\omega\right)=y\right]}{c\left(\Omega_y\right)}\sum_{j=1}^{N-n}I_j\left(\omega\right)A_j\right]$$

$$=\left\{c\left(\Omega_y\right)\right\}^{-1}\sum_{j=1}^{N-n}A_j\sum_{\omega\in\Omega_y}I_j\left(\omega\right)$$

$$=\left\{c\left(\Omega_y\right)\right\}^{-1}\sum_{j=1}^{N-n}A_j\frac{y}{N-n}c\left(\Omega_y\right)=y\bar{A}.$$

Hence,

$$E\left[V\right]=E\left[E\left[V\,|\,Y\right]\right]=E\left[\bar{A}Y\right]=\mu_Y\,\bar{A}=\left(\mu_X-k\right)\bar{A}.$$

The first result followsbecause $E\left[W\right]=E\left[V\right]+kA_0$.

In order to prove the second result, we need the identity

$$\sum_{\substack{i,j=1\\ i\neq j}}^{N-n}A_iA_j=\left(N-n\right)^2\bar{A}^2-\left(N-n\right)\overline{A^2},$$

which is easily obtained by expanding $\bar{A}^2$.

$$\sum_{\omega\in\Omega_y}v^2\left(\omega\right)$$

$$=\sum_{\omega\in\Omega_y}\left[\sum_{j=1}^{N-n}I_j\left(\omega\right)A_j{}^2+\sum_{\substack{i,j=1\\ i\neq j}}^{N-n}I_i\left(\omega\right)I_j\left(\omega\right)A_iA_j\right]$$

$$=\sum_{j=1}^{N-n}\left[A_j{}^2\sum_{\omega\in\Omega_y}I_j\left(\omega\right)\right]+\sum_{\substack{i,j=1\\ i\neq j}}^{N-n}\left[A_iA_j\sum_{\omega\in\Omega_y}I_i\left(\omega\right)I_j\left(\omega\right)\right]$$

$$=\sum_{j=1}^{N-n}\frac{A_j{}^2y}{N-n}c\left(\Omega_y\right)+\sum_{\substack{i,j=1\\ i\neq j}}^{N-n}\frac{A_iA_jy\left(y-1\right)}{\left(N-n\right)\left(N-n-1\right)}c\left(\Omega_y\right)$$

$$=c\left(\Omega_y\right)\left[y\overline{A^2}+\frac{y\left(y-1\right)\left[\left(N-n\right)\bar{A}^2-\overline{A^2}\right]}{\left(N-n-1\right)}\right].$$

Therefore,

$$E\left[V^2\,|\,Y=y\right]=\frac{1}{\Pr\left[Y=y\right]}\sum_{\omega\in\Omega_y}v^2\left(\omega\right)\Pr\left[\omega\right]$$

$$=\frac{1}{\Pr\left[Y=y\right]}\sum_{\omega\in\Omega_y}v^2\left(\omega\right)\frac{\Pr\left[Y=y\right]}{c\left(\Omega_y\right)}$$

$$=y\overline{A^2}+\frac{y\left(y-1\right)\left[\left(N-n\right)\bar{A}^2-\overline{A^2}\right]}{\left(N-n-1\right)}.$$

Hence,

$$Var\left(V\,|\,Y\right)$$

$$=Y\overline{A^2}+\frac{Y\left(Y-1\right)\left[\left(N-n\right)\bar{A}^2-\overline{A^2}\right]}{\left(N-n-1\right)}-\left(Y\bar{A}\right)^2$$

$$=\frac{\overline{A^2}-\bar{A}^2}{N-n-1}\left[\left(N-n\right)Y-Y^2\right].$$

It follows that,

$$E\left[Var\left(V\,|\,Y\right)\right]$$

$$=\frac{\overline{A^2}-\bar{A}^2}{N-n-1}\left[\left(N-n\right)\mu_Y-\left(\sigma_Y{}^2+\mu_Y{}^2\right)\right]$$

and using the expressions for $\mu_Y$ and $\sigma_Y{}^2$, this becomes

$$E\left[Var\left(V\mid Y\right)\right]=\frac{n+1}{N+1}\left(\overline{A^2}-\overline{A}^2\right)\sigma_Y{}^2.$$

Hence,

$$\begin{aligned}Var\left(V\right)&=E\left[Var\left(V\mid Y\right)\right]+Var\left(E\left[V\mid Y\right]\right)\\&=\frac{n+1}{N+1}\left(\overline{A^2}-\overline{A}^2\right)\sigma_Y{}^2+\overline{A}^2\sigma_Y{}^2\\&=\frac{\left(N-n\right)\overline{A}^2+\left(n+1\right)\overline{A^2}}{N+1}\sigma_Y{}^2.\end{aligned}$$

Therefore,

$$\begin{aligned}Var\left(V\right)&=\frac{\left(N-n\right)\overline{A}^2+\left(n+1\right)\overline{A^2}}{N+1}\sigma_Y{}^2\\&=\frac{\left(\overline{A^2}-\overline{A}^2\right)\left(n+1\right)}{N+1}\sigma_X{}^2+\overline{A}^2\sigma_X{}^2\\&=\frac{\left(\overline{A^2}-\overline{A}^2\right)}{\mu_X/k}\sigma_X{}^2+\overline{A}^2\sigma_X{}^2\\&=\frac{\sigma_X{}^2}{\mu_X}\left[k\left(\overline{A^2}-\overline{A}^2\right)+\mu_X\overline{A}^2\right]=Var\left(W\right).\end{aligned}$$

## 4. Applications for Bonus Games

We present two different types of applications for bonus games.

**Type 1:** In this type of bonus game, the player is presented with *N* covered locations (squares, bubbles, etc.), with each location containing either an award or a 'devil' (joker). There are *n* locations with the 'devil'. The player uncovers each location until the $k^{\text{th}}$ "devil' is uncovered. The player is then awarded the sum of all the prize money uncovered plus a consolation award for uncovering *k* 'devils'. The latter award may be zero, a fixed award, or an award proportional to the remaining covered locations.

**(a)** The simplest possible game is one in which each of the non-devil locations has a fixed amount of $A$ and each of the 'devil' locations has an amount of $A_0$. The player is awarded all the prize money uncovered, including the awards for the 'devils'. For this game, let *X* be the total number of uncovered locations and let *W* be the total award for the game. Then *W* is given by:

$$W=A_0k+A\left(X-k\right)$$

Since $A_0$ and $A$ are constants,

$$E\left[W\right]=A_0k+A\left(E\left[X\right]-k\right)=A\mu_X-\left(A-A_0\right)k$$

$$Var\left(W\right)=A^2Var\left(X\right).$$

**Example:** In this example, there are 15 locations of which 3 are occupied by 'devils'. The locations may be given by a 3x5 matrix or 15 'eggs' or 'balloons'. Furthermore, let the fixed prize award be 25 and let the consolation award for the 'devil' be 2. Therefore, we have: $N=15,\ n=3,\ A_0=1,$ and $A=10$. The mean and the variance of *X* are given by:

$$\mu_X=\frac{k\left(N+1\right)}{n+1}=\frac{k\left(16\right)}{4}=4k$$

$$\begin{aligned}\sigma_X{}^2&=\frac{k\left(N+1\right)\left(N-n\right)\left(n-k+1\right)}{\left(n+1\right)^2\left(n+2\right)}\\&=\frac{k\left(16\right)\left(12\right)\left(4-k\right)}{\left(4\right)^2\left(5\right)}=\frac{12k\left(4-k\right)}{5}.\end{aligned}$$

The mean and the variance of the bonus game are given by:

$$E\left[W\right]=\left(10\right)\left(4k\right)-\left(10-1\right)k=31k$$

$$Var\left(W\right)=\left(10\right)^2\frac{12k\left(4-k\right)}{5}=240k\left(4-k\right).$$

The following table provides the mean, the variance, and the standard deviation for all possible values of *k*:

| *k* | *E[W]* | *Var(W)* | $\sigma_W$ |
|---|---|---|---|
| 1 | 31 | 720 | 26.8 |
| 2 | 62 | 960 | 31.0 |
| 3 | 93 | 720 | 26.8 |

**(b)**In this game, the (non-devil) prize awards are not fixed. Let these be denoted by: $A_1,A_2,...,A_{N-n}$. The total award for the game is given by:

$$W=A_0k+V\ ,$$

where*V* is the sum of all the prizes that have been uncovered at the non-devil locations. The mean and the variance of *W* are given by:

$$E\left[W\right]=A_0k+E\left[V\right]=\overline{A}\mu_X-\left(\overline{A}-A_0\right)k$$

$$\begin{aligned}Var\left(W\right)=Var\left(V\right)&=\frac{k\sigma_X{}^2\left(\overline{A^2}-\left(\overline{A}\right)^2\right)}{\mu_X}+\left(\overline{A}\right)^2\sigma_X{}^2\\&=\frac{\sigma_X{}^2}{\mu_X}\left[k\left(\overline{A^2}-\left(\overline{A}\right)^2\right)+\mu_X\left(\overline{A}\right)^2\right],\end{aligned}$$

where

$$\overline{A}=\frac{1}{N-n}\sum_{j=1}^{N-n}A_j\ \text{and}\ \overline{A^2}=\frac{1}{N-n}\sum_{j=1}^{N-n}A_j{}^2.$$

**Example:** In this example, there are 20 locations of which 4 are occupied by 'devils'. The locations may be given by a 4x5 matrix or 20 'eggs' or 'balloons'. Furthermore let there be 7 prizes of 10, 2 prizes of 15, 4 prizes of 25, 2 prizes of 50, and 1 prize of 100 along with a consolation award of 2 for each of the 'devils'. For this case, we have: $N$ = 20, $n$ = 4, $A_1,...,A_7=10,$

$A_8=A_9=15,$

$A_{10},...,A_{13}=25,\ A_{14}=A_{15}=50,\ A_{16}=100.$ $\overline{A}$ and $\overline{A^2}$ are given by:

$$\overline{A}=\frac{7\left(10\right)+2\left(15\right)+4\left(25\right)+2\left(50\right)+100}{16}=25$$

$$\overline{A^2} = \frac{7\left(10^2\right)+2\left(15^2\right)+4\left(25^2\right)+2\left(50^2\right)+100^2}{16}$$

$$=1165.625 \approx 1165.6$$

The mean and the variance of *X* are given by:

$$\mu_X = \frac{k(N+1)}{n+1} = \frac{k(21)}{5} = \frac{21k}{5}$$

$$\sigma_X{}^2 = \frac{k(N+1)(N-n)(n-k+1)}{(n+1)^2(n+2)}$$

$$= \frac{k(21)(16)(5-k)}{(5)^2(6)} = \frac{56k(5-k)}{25}.$$

The mean and the variance of the bonus game are given by:

$$E[W] = 25\left(\frac{21k}{5}\right) - (25-2)k = 82k$$

$$Var(W) = \frac{56k(5-k)}{25(21k/5)}\left[k(1165.6-625)+\left(\frac{21k}{5}\right)(625)\right]$$

$$=1688k(5-k).$$

The following table provides the mean, the variance, and the standard deviation for all possible values of *k*:

| *k* | *E[W]* | *Var(W)* | $\sigma_W$ |
|---|---|---|---|
| 1 | 82 | 6752 | 82.2 |
| 2 | 164 | 10128 | 100.6 |
| 3 | 246 | 10128 | 100.6 |
| 4 | 328 | 6752 | 82.2 |

**(c)** In this game, the (non-devil) prize awards are not fixed. Let these be given by: $A_1, A_2, ..., A_{N-n}$. The player is awarded the non-devil prizes that were uncovered along with an award of $A_0$ for each of the remaining ('devil' plus covered) locations. The total award for the game is given by:

$$W = A_0(N-X)+V\ ,$$

whereV is the sum of all the prizes that have been uncovered at the non-devil locations. As in **(b)**, the mean and variance of *V* are given by:

$$E[V] = E[X-k]\overline{A}$$

$$Var(V) = \frac{k\sigma_X{}^2\left(\overline{A^2}-\left(\overline{A}\right)^2\right)}{\mu_X} + \left(\overline{A}\right)^2\sigma_X{}^2$$

where

$$\overline{A} = \frac{1}{N-n}\sum_{j=1}^{N-n} A_j \text{ and } \overline{A^2} = \frac{1}{N-n}\sum_{j=1}^{N-n} A_j{}^2.$$

The mean and the variance of *W* are given by:

$$E[W] = A_0\{N-E[X]\}+E[V]$$

$$= \left(\overline{A}-A_0\right)\mu_X - \left(k\overline{A}-NA_0\right)$$

$$Var(W) = Var(V-A_0X+A_0N) = Var(V-A_0X)$$

$$= Var(V) + A_0{}^2 Var(X) - 2A_0 Cov(V,X)$$

$$= \frac{\sigma_X{}^2}{\mu_X}\left[k\left(\overline{A^2}-\left(\overline{A}\right)^2\right)+\mu_X\left(\left(\overline{A}\right)^2+A_0{}^2\right)\right]$$

$$-2A_0Cov(V,X).$$

The covariance term in the last expression is rather complicated and in most cases could be omitted due to its relative size as compared to the other terms. The actual formula for it will be given in a later study.

**Example:** In this example, there are 20 locations of which 4 are occupied by 'devils'. These locations may be given by a 4x5 matrix or 20 'eggs' or 'balloons'. Furthermore let there be 7 prizes of 10, 2 prizes of 15, 4 prizes of 25, 2 prizes of 50, and 1 prize of 100 along with a consolation award of 2 for each of the 'devils'. For this case: $N = 20,\ n = 4,\ A_0 = 2,\ A_1, ..., A_7 = 10,$

$$A_8 = A_9 = 15,\ A_{10}, ...,\ A_{13} = 25,\ A_{14} = A_{15} = 50,\ A_{16} = 100.$$

$\overline{A}$ and $\overline{A^2}$ are given by:

$$\overline{A} = \frac{7(10)+2(15)+4(25)+2(50)+100}{16} = 25$$

$$\overline{A^2} = \frac{7\left(10^2\right)+2\left(15^2\right)+4\left(25^2\right)+2\left(50^2\right)+100^2}{16}$$

$$=1165.625 \approx 1165.6$$

The mean and the variance of *X* are given by:

$$\mu_X = \frac{k(N+1)}{n+1} = \frac{k(21)}{5} = \frac{21k}{5}$$

$$\sigma_X{}^2 = \frac{k(N+1)(N-n)(n-k+1)}{(n+1)^2(n+2)}$$

$$= \frac{k(21)(16)(5-k)}{(5)^2(6)} = \frac{56k(5-k)}{25}.$$

The mean and the variance of the bonus game are given by:

$$E[W] = (25-2)\left(\frac{21k}{5}\right) - (25k-40)$$

$$= 71.6k+40$$

$$Var(W)$$

$$\approx \frac{56k(5-k)}{25(21k/5)}\left[k(1165.6-625)+\left(\frac{21k}{5}\right)(625+4)\right]$$

$$=1697k(5-k).$$

The following table provides the mean, the variance, and the standard deviation for all possible values of *k*:

| *k* | *E[W]* | *Var(W)* | $\sigma_W$ |
|---|---|---|---|
| 1 | 111.6 | 6788 | 82.4 |
| 2 | 183.2 | 10182 | 100.9 |
| 3 | 254.8 | 10182 | 100.9 |
| 4 | 326.4 | 6788 | 82.4 |

**Type 2.**This type of bonus game is a simple match game: The player first selects an object from apopulation of *N* objects of which there are $n_1$ objects of type 1, $n_2$ objects of type 2,…, $n_r$ objects of type *r*, where $n_1 + n_2 + ... + n_r = N$. The selected object is removed from the population and the player selects objects, one at a time, without replacement, from the remaining set of objects, until a match is found. The player is then awarded a prize that is dependent on the number of the matching draw. An alternative approach would be to have two identical populations and for the player to select an object from the first population and try to match it with an object from the second population.

For the first approach, if object *i* were selected, then there would be $N-1$ objects left in the population out of which $n_i - 1$ objects would be of the selected type. Therefore, if the match, i.e., $k=1$, is made on the $X^{th}$ draw, the mean and variance of *X* would be given by:

$$\mu_X = \frac{(1)\{(N-1)+1\}}{(n_i-1)+1}$$

$$\sigma_X{}^2 = \frac{(1)\{(N-1)+1\}\{(N-1)-(n_i-1)\}\{(n_i-1)+(1)-1\}}{\{(n_i-1)+1\}^2\{(n_i-1)+2\}}.$$

These expressions reduce to:

$$\mu_X = \frac{N}{n_i}, \quad i = 1,...,r$$

$$\sigma_X{}^2 = \frac{N(N-n_i)(n_i-1)}{n_i^2(n_i+1)}, \quad i = 1,...,r \; .$$

For the second approach, if object *i* were selected, then there would be $N$ objects in the second population out of which $n_i$ objects would be of the selected type. Therefore, if the match, i.e., $k=1$, is made on the the $X^{th}$ draw, the mean and variance of *X* would be given by:

$$\mu_X = \frac{N+1}{n_i+1}, \quad i = 1,...,r$$

$$\sigma_X{}^2 = \frac{(N+1)(N-n_i)n_i}{(n_i+1)^2(n_i+2)}, \quad i = 1,...,r \; .$$

The probability functions for both approaches are easily computed using the values of the parameters: $N$, $n$, and $k=1$.

**Example:** In this example, the population is a standard deck of cards and the first approach is used. The player is awarded the prizes that depend on the placement of the match. The distinct objects, in the population are denoted by the ranks of the cards: $2,3,4,5,6,7,8,9,10,J,Q,K,A.$ Hence, $N=52$, $n_i=4$, and we present the following table for a bonus game with an expected value of approximately 25:

| $x$ | $\Pr(X=x)$ | Award | *E*[Award] |
|---|---|---|---|
| 1-10 | 0.488115 | 10 | 4.881150 |
| 11-20 | 0.296039 | 15 | 4.440585 |
| 21-30 | 0.151980 | 30 | 4.559400 |
| 31-40 | 0.055943 | 100 | 5.559430 |
| 41-49 | 0.007923 | 700 | 5.546100 |
| Total | 1.000000 | - | 24.986665 |

The variance of the bonus game could be computed using the ideas in the first part of the presentation, or by simulation.

## 5. Conclusion

The negative binomial is presently used in the applications illustrated in the previous section. However, the expected value and variance of the games are obtained by simulation or by approximating via a geometric or a negative binomial distribution. It is better to use the actual underlying distribution for these bonus games to obtain the exact estimated value of the total slot game.

## 6.Bibliography:

[1] Jones, S. N. A Gaming Application of the Negative Hypergeometric Distribution[Thesis]. Las Vegas: University of Nevada Las Vegas;

[2]Schuster, E. F. and Sype, W. R. (1987). On the Negative Hypergeometric Distribution.International Journal of Mathematical Education in Science and Technology. 1987; 18 (3), 453 – 459.

# Adaptive Interest Rate Models and Their Empirical Performance: A Case Study of Thailand's Policy Interest Rate

Krisada Muangsri[1] and Surapap Rayanakorn[2*]
[1]*Graduate School of Management and Innovation, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand, krisada.m@mail.kmutt.ac.th*
[2]*Graduate School of Management and Innovation, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand, surapap.ray@kmutt.ac.th*

## Abstract

In this study, we present adaptive interest rate models and evaluate their performance with the policy rate data in Thailand during 2001-2014. We apply local parametric estimation to two standard models with a simplified method in determining homogeneous intervals. With locally estimated parameters, the adaptive models show substantial improvement in policy rate forecasting. Their one-step-ahead forecasts result in a 60% reduction in forecasting errors. In addition, we find that the policy rate determination based on the inflation targeting policy is quite effective. The inflation responds to the policy interest rate after one year during most of the 2001-2014 period.

*Keywords*: CIR model, interest rate models, local parametric estimation, policy interest rate

*Corresponding Author
E-mail Address: surapap.ray@kmutt.ac.th

## 1. Introduction

The policy interest rate (or policy rate) is one of the most important variables for the economy and the financial markets. At the macroeconomic level, the central bank uses the policy interest rate to control the money supply. In the financial markets, the policy rate influences various rates, ranging from bank loan interest rates to bond yields. There is a large body of the finance literature that studies the influence of the policy interest rate on the markets. For example, Bernanke and Kuttner in [1] report that a decrease of 0.25% in the US federal funds rate is related to a 1% increase in the equity market. Therefore, better understanding of the policy interest rate and its modeling is beneficial to both policymakers and individuals.

Among the earliest work in interest rate modeling is the Vasicek model. It explains that the interest rate is mean-reverting with a random component that causes noisy movements around its mean [2]. The Cox-Ingersoll-Ross (CIR) model in [3] is another widely used interest rate model. Other popular models include the Merton model and the Brennan-Schwartz model.

Several studies indicate that there is room for improvement for most of the well-known interest rate models due to their restriction on parameter estimates. For example, the constant volatility estimate does not capture the heteroskedasticity of the interest rate change [4-5]. In addition, the constant mean estimate does not work well when the interest rate switches regimes [6].

This paper addresses these challenges by applying the local parametric approach in [7-9] to the CIR and Vasicek models with a simpler method in determining estimation intervals. We then test the adaptive models with the policy rate and relevant data starting from 2001, one year after the beginning of inflation targeting.

The remainder of the paper is organized as follows. Section 2 describes the policy interest rate and other data used in this study. Section 3 outlines the CIR model and the Vasicek model as well as the local parametric estimation used in the adaptive models. The results of the adaptive models in comparison with the benchmark models are then presented in Section 4. Section 5 concludes the important points of this work.

## 2. Data

In estimating model parameters and evaluating the adaptive model performance, we need the following data: policy interest rate, one-month Treasury bill (T-Bill) rate, and inflation.

### *2.1 Data Description*

We obtain the policy interest rate during 2000-2014 from the Bank of Thailand. The one-month T-Bill rate from 2001 to 2014 is gathered from the Thai Bond Market Association. The core consumer price index (core CPI) is collected from the Bureau of Trade and Economic Indices to calculate the core inflation for the same period. In addition, we gather other data that may be relevant, such as the nominal effective exchange rate (NEER), from the Bank of Thailand.

A subset of the data does not have daily observations. Specifically, the policy interest rate changes only on the announcement dates made by the Monetary Policy Committee (MPC). These dates are about 6 weeks apart under normal circumstances. We use the one-month T-

Bill rate as its proxy because the T-Bill rate time series closely tracks the policy rate time series as shown in Figure 1. As for the inflation, the core CPI values are collected at monthly frequency. We use linear interpolation to approximate its daily values so that the inflation on a given day can be calculated.

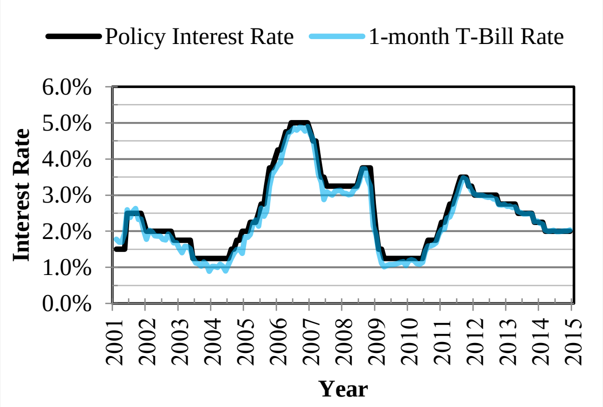


Figure 1: Time Series of Policy Interest Rate and 1-month T-Bill Rate from February 2001 to December 2014

### 2.2 Summary Statistics

The summary statistics of the data are reported in Table 1 for the 2001-2014 period. For each variable, there are 115 observations taken on MPC dates. The policy interest rate and the one-month T-Bill rate have very similar characteristics. For example, they average at 2.3-2.5% per year with the same volatility of 1.0%. They also show the same persistent pattern with almost identical autocorrelations at various lags. They are almost perfectly correlated with a correlation of 0.987. Therefore, the one-month T-Bill rate is a very good proxy for the policy interest rate.

As for the T-Bill rate change, it has a mean of 0.0%, confirming that the interest rate is mean-reverting. It is less autocorrelated than other variables, and it is not correlated with the policy rate.

The inflation averages at 1.3%, which is half of the policy rate mean. It is slightly less volatile than the policy interest rate. The inflation and its lags are quite correlated with the policy rate based on their correlation coefficients in Table 1 and Table 3. The NEER change is more volatile than other time series, but it exhibits a similar autocorrelation pattern to the inflation.

Table 1: Summary Statistics of Policy Interest Rate $r_t^p$, 1-month T-Bill Rate $r_t$, Core Inflation $inf_t$, Nominal Effective Exchange Rate (NEER) Change $r_t^{NEER}$, and Change of 1-month T-Bill Rate $\Delta r_t$ from March 2001 and December 2014

The sample consists of 115 observations (expressed in percentage per year) on MPC dates. Adjacent MPC dates are 45 days apart on average. Symbol * and *** denote statistical significance at 10% level and 1% level, respectively.

| Statistic | $r_t^p$ | $r_t$ | $inf_t$ | $r_t^{NEER}$ | $\Delta r_t$ |
|---|---|---|---|---|---|
| Mean | 2.49% | 2.34% | 1.25% | 0.88% | 0.00% |
| Std. Dev. | 1.03% | 1.01% | 0.88% | 4.74% | 0.23% |
| Skewness | 0.79 | 0.73 | 0.21 | 0.25 | -0.70 |
| Kurtosis | -0.01 | 0.07 | -0.36 | 0.25 | 2.49 |
| Max | 5.00% | 4.88% | 3.65% | 14.90% | 0.53% |
| Min | 1.25% | 0.98% | -1.06% | -9.55% | -0.81% |
| Correlation with $r_t^p$ (p-value) | 1.000 | 0.987<br>***<br>(0.00) | 0.707<br>***<br>(0.00) | 0.332<br>***<br>(0.00) | 0.157<br>*<br>(0.10) |
| Autocorrelation at lag $s$, $\rho_s$: | | | | | |
| $\rho_1$ | 0.969 | 0.972 | 0.932 | 0.876 | 0.524 |
| $\rho_2$ | 0.911 | 0.916 | 0.818 | 0.667 | 0.367 |
| $\rho_3$ | 0.835 | 0.840 | 0.690 | 0.466 | 0.119 |
| $\rho_8$ | 0.419 | 0.410 | 0.023 | -0.356 | 0.004 |
| $\rho_{12}$ | 0.080 | 0.072 | -0.113 | -0.248 | -0.013 |

## 3. Methodology

We describe our methodology in two parts. First, we outline the original interest rate models which serve as a benchmark. Second, we explain the adaptive models which estimate the model parameters locally.

### 3.1 Original Interest Rate Models

The two popular models used in our study are the Vasicek model and the CIR model. The Vasicek model [2] describes interest rate movements with the following stochastic differential equation (SDE):

$$dr_t = \kappa_V(\theta_V - r_t)dt + \sigma_V dZ_t, \tag{1}$$

where: $r_t$ is the interest rate.
$\theta_V$ is the long-term mean of the interest rate.
$\kappa_V$ is the mean reversion rate or the speed of adjustment.
$Z_t$ is a Wiener process.
$\sigma_V$ is the interest rate volatility or standard deviation factor for $Z_t$.

The SDE in (1) says that an incremental interest rate change is driven by its deviation from the steady state value and a risk source.

Unlike the Vasicek model, the CIR model restricts the interest rate to positive values and specifies its volatility to be dependent on its level as follows [3][1]:

$$dr_t = \kappa(\theta - r_t)dt + \sigma\sqrt{r_t}dZ_t, \tag{2}$$

where: $r_t$ is the interest rate.
$\theta$ is the long-term mean of the interest rate.
$\kappa$ is the mean reversion rate or the speed of adjustment.
$Z_t$ is a Wiener process.
$\sigma\sqrt{r_t}$ is the standard deviation factor for $Z_t$.

To calibrate the CIR model with the data, we discretize the SDE in (2) so that the parameters can be

[1] From this point onward, we explain the parameter estimation and the performance comparison primarily for the CIR model. However, a similar discussion applies to the Vasicek model. The Vasicek model is presented at the beginning of this section for completeness.

estimated by regression. Similar to [6], the discrete-time equation is the following:

$$r_{t+1} - r_t = \kappa(\theta - r_t)\Delta t + \sigma\sqrt{r_t}\epsilon_t, \tag{3}$$

where: $\epsilon_t \sim N(0, \Delta t)$.

We reorganize (3) into (4) and estimate the parameters $(\theta, \kappa, \sigma)$ with ordinary least squares (OLS) regression.

$$\frac{r_{t+1} - r_t}{\sqrt{r_t}} = \kappa\theta \frac{\Delta t}{\sqrt{r_t}} - \kappa\sqrt{r_t}\Delta t + \sigma\epsilon_t, \tag{4}$$

The one-month T-Bill rate in the 2001-2014 period is used to estimate the regression coefficients. The interest rate model parameters are then calculated from these coefficients. The estimated parameters for the CIR model are constant as shown in Figure 5 and Figure 6 in Appendix B (dashed blue lines).

### *3.2 Interest Rate Models with Locally Estimated Parameters*

Model parameters change over time, based on the interest rate data and the literature (for example, [5] and [6]). We therefore should be able to refine the interest rate models with locally estimated parameters.

We follow the local parametric approach (LPA)[2] in [7] to divide the sample period into homogenous intervals. We then locally estimate the model parameters via OLS of (4) in these intervals. In this work, the homogenous intervals (HIs) are defined to be time windows in which the interest rate values move within two standard deviations of the window average. This method of HI determination is simpler than that in the literature.[3] Specifically, for each MPC date, the initial homogenous interval (HI) consists of the three observations immediately preceding the MPC date. In each of the next steps, the interest rate observation that immediately precedes the HI is added to the HI unless it deviates from the HI average by more than two standard deviations. When a preceding observation deviates more than this threshold, it is not added to the HI, and the process stops. At this point, the longest HI (LHI) is determined for the MPC date. This process is summarized by the flowchart in Figure 2.

We use the T-Bill rate data during 2001-2014 for the estimation. The estimated parameters for the CIR model with LPA (CIR-LPA) are time-varying as shown in Figure 5 and Figure 6 in Appendix B (solid blue lines with diamond markers).

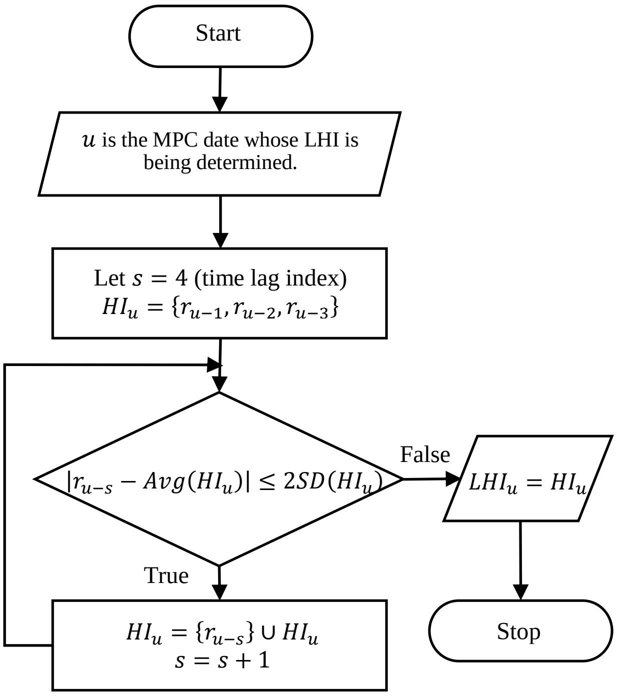

Start
$u$ is the MPC date whose LHI is being determined.
Let $s = 4$ (time lag index)
$HI_u = \{r_{u-1}, r_{u-2}, r_{u-3}\}$
$|r_{u-s} - Avg(HI_u)| \le 2SD(HI_u)$
False
$LHI_u = HI_u$
True
$HI_u = \{r_{u-s}\} \cup HI_u$
$s = s + 1$
Stop

Figure 2: Flowchart Explaining the Longest Homogeneous Interval for a Given MPC Date (adapted from the local parametric approach in [7] and [8])

## 4. Results and Discussion

In this section, we compare the benchmark models and the adaptive models with local parametric estimation in two ways, namely, the model goodness of fit and the model error. Moreover, we discuss the effectiveness of the inflation targeting policy for Thailand based on the results. Finally, we explore some related work.

Figure 3 shows the one-step forecasts of the CIR model (dashed blue line) and the CIR-LPA model (solid blue line with diamond markers) in comparison with the policy interest rate (solid black line). The CIR and CIR-LPA forecasts are the averages of 2,000 Monte Carlo simulations. [4] Although the CIR forecast has an increasing trend, it cannot capture the policy rate movement. With locally estimated parameters, the CIR-LPA tracks the policy rate quite closely with an adjusted-$R^2$ of 87.1%. Note that the CIR-LPA forecast lags the policy rate movement slightly. This small lag likely results from the fact that we use past observations to estimate the model parameters. These parameters are then used to predict one-step-ahead values. The Vasicek model and the Vasicek model with LPA (Vasicek-LPA) show similar results.

[2] Based on this approach, a number of studies obtain improvement on modeling accuracy in different contexts such as US interest rate modeling [8] and S&P500 volatility modeling [9].

[3] For a more rigorous test of homogenous intervals or variants, please refer to [7]-[9].

[4] Instances of the CIR model have different paths due to the Brownian motion term. Hence, the average of multiple Monte Carlo simulations is used.

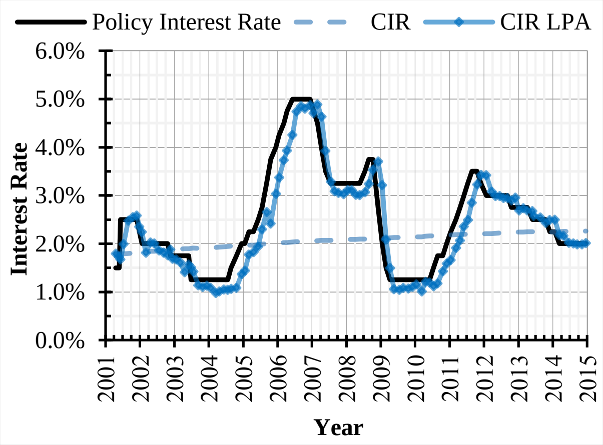


Figure 3: Performance Comparison of Interest Rate Models from April 2001 to December 2014

As for the forecasting error, the mean absolute error (MAE) reduces from 0.008 to 0.003 with the application of the LPA to the CIR model.[5] The Vasicek model also shows a similar decrease of 60% in MAE. The errors as measured by MAE and RMSE are reported in Table 2, along with the $R^2$ and the adjusted-$R^2$ values.

Table 2: Performance Comparison of Interest Rate Models from April 2001 to December 2014

This table reports $R^2$, adjusted $R^2$, and error measures of the interest rate models.

| Model | $R^2$ | Adj-$R^2$ | MAE | RMSE |
|---|---|---|---|---|
| CIR LPA | 0.872 | 0.871 | 0.003 | 0.004 |
| Vasicek LPA | 0.868 | 0.867 | 0.003 | 0.004 |
| CIR | 0.031 | 0.022 | 0.008 | 0.011 |
| Vasicek | 0.028 | 0.020 | 0.008 | 0.011 |

As for the effectiveness of the inflation targeting policy, we examine the inflation time series into the future for a given policy rate date. Figure 4 plots the policy interest rate and the inflation with a 12-month lead. We find that the inflation responds to the policy interest rate approximately 12 months after the announcement date for more than half of the 2001-2014 period. At other times, the inflation changes more quickly or more slowly. To illustrate this pattern, in Figure 4, the level of the lead inflation time series is opposite to that of the policy interest rate time series, i.e., the future inflation level is low when the current policy rate level is high and vice versa. The slope of the lead inflation time series is also opposite to that of the policy rate time series. From 2001:H2 to 2004:H2 when the policy interest rate decreases from 2.5% to 1.25%, the inflation increases from 0.2% to 1.2% one year later.[6] When the policy interest rate increases from 1.25% to 5.0% during the 2004:H2-2006:H2 period, the inflation goes down from 2.0% to 0.8% during the 2006:H2-2007:H2 period or about one to two years later. After 2007, the two time series have an opposite relation similar to the pattern explained earlier. Conclusively, the policy interest rate, determined under the inflation targeting policy, is effective in driving the inflation about one year in the future.

There are a number of studies related to the LPA in various settings. For example, [10] proposes an adaptive dynamic Nelson-Siegel (ADNS) model that locally estimates term structure parameters. Using the US yield curve data between 1998 and 2009, they report a reduction of 20-60% in forecast errors.

In addition to the local parametric estimation, there are also other strands of model development. For example, a weak hidden Markov model (WHMM) is proposed in [11] to model the Canadian 1-month T-Bill rate during a 10-year period. They find that increasing the number of states reduces forecast errors.

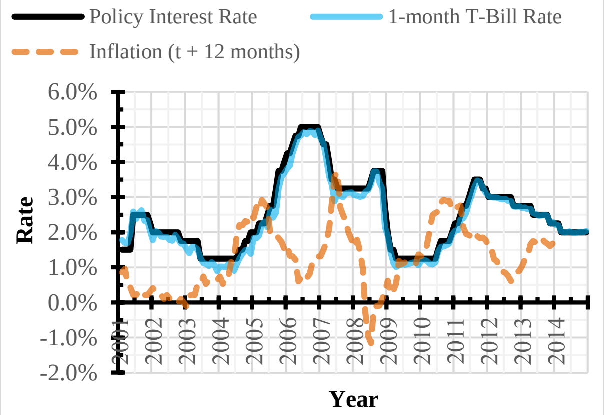


Figure 4: Policy Interest Rate Time Series and Lead Inflation Time Series (with a lead of 12 months)

## 5. Conclusion

In this paper, we apply the local parametric approach to two popular interest rate models, aiming to capture the time-varying nature of the model parameters. Based on the interest rate data one year after the introduction of the inflation targeting policy by the Bank of Thailand, the two benchmark models perform better after the addition of local parametric estimation. They achieve an increase of over 80% in $R^2$ and a 60% reduction in forecast errors. These results add to the literature in two ways. First, our approach in locally estimating model parameters is simpler, but still yield substantial improvement. Second, while most of recent work examines developed markets, we test the adaptive models in an emerging market and find that the inflation targeting policy is effective.

[5] Using the T-Bill rate from 1998 to 2009, [8] reports an improvement of the LPA over the moving window (MW) estimation for the US, especially when the window sizes are larger and the forecasting step is longer.

[6] 2001:H2 and 2004:H2 denote the second half of 2001 and the second half of 2004, respectively.

## References

[1] Bernanke, Ben S. and Kenneth N. Kuttner, 2005, What explains the stock market's reaction to federal reserve policy?, *Journal of Finance* 60, 1221-1257.

[2] Vasicek, Oldrich, 1977, An equilibrium characterization of the term structure, *Journal of Financial Economics* 5, 177-188.

[3] Cox, John C., Jonathan E. Ingersoll, Jr., and Stephen A. Ross, 1985, A theory of term structure of interest rates, *Econometrica* 53, 385-408.

[4] Chan K. C., G. Andrew Karolyi, Francis A. Longstaff, and Anthony B. Sanders, 1992, An empirical comparison of alternative models of the short-term interest rate, *Journal of Finance* 47, 1209-1227.

[5] Furno, Marilena, 1993, Monetary policy and interest rates, *Journal of Economic Dynamics and Control* 17, 571-588.

[6] Sanders, Anthony B. and Haluk Unal, 1988, On the intertemporal behavior of the short-term rate of interest, *Journal of Financial and Quantitative Analysis* 23, 417-423.

[7] Spokoiny, Vladimir, 2009, Multiscale local change point detection with applications to value-at-risk, *The Annals of Statistics* 37, 1405-1436.

[8] Guo, Meng and Wolfgang K. Härdle, 2010, Adaptive interest rate modelling, *SFB 649 Discussion Paper*.

[9] Chen, Ying, Wolfgang K. Härdle, and Uta Pigorsch, 2010, Localized realized volatility modeling, *Journal of the American Statistical Association* 105, 1376-1393.

[10] Chen, Ying and Linlin Niu, 2014, Adaptive dynamic Nelson-Siegel term structure model with applications, *Journal of Econometrics* 180, 98-115.

[11] Xi, Xiaojing and Rogemar Mamon, 2014, Capturing the regime-switching and memory properties of interest rates, *Computational Economics* 44, 307-337.

## Appendix A: Pairwise Correlations between Policy Interest Rate and Other Rate Lags

Table 3: Pairwise Correlations between Policy Interest Rate and Rate Lags (Lagged Policy Interest Rates, Lagged Core Inflation, and Lagged NEER Change)

Symbol *** denotes statistical significance at 1% level. P-values are in parentheses.

| | $r_t^p$ | $r_{t-1}^p$ | $r_{t-2}^p$ | $inf_{t-1}$ | $inf_{t-2}$ | $r_{t-1}^{NEER}$ |
|---|---|---|---|---|---|---|
| $r_t^p$ | 1.00 | | | | | |
| $r_{t-1}^p$ | 0.97***<br>(0.000) | 1.00 | | | | |
| $r_{t-2}^p$ | 0.92***<br>(0.000) | 0.97***<br>(0.000) | 1.00 | | | |
| $inf_{t-1}$ | 0.73***<br>(0.000) | 0.71***<br>(0.000) | 0.67***<br>(0.000) | 1.00 | | |
| $inf_{t-2}$ | 0.71***<br>(0.000) | 0.73***<br>(0.000) | 0.71***<br>(0.000) | 0.93***<br>(0.000) | 1.00 | |
| $r_{t-1}^{NEER}$ | 0.30***<br>(0.001) | 0.34***<br>(0.000) | 0.36***<br>(0.000) | -0.06<br>(0.501) | 0.00<br>(0.977) | 1.00 |

## Appendix B: Estimated Parameters of Interest Rate Models

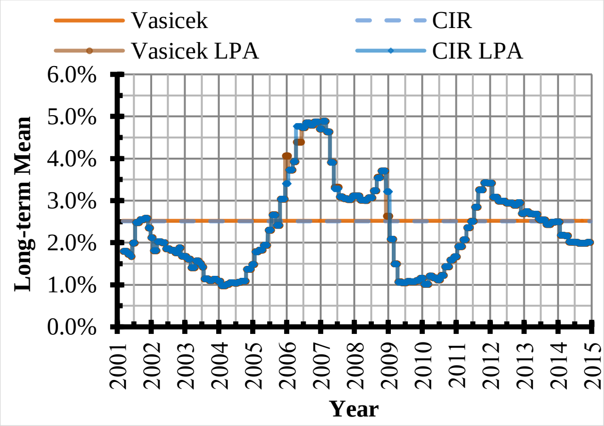


Figure 5: Long-term Mean $\theta$

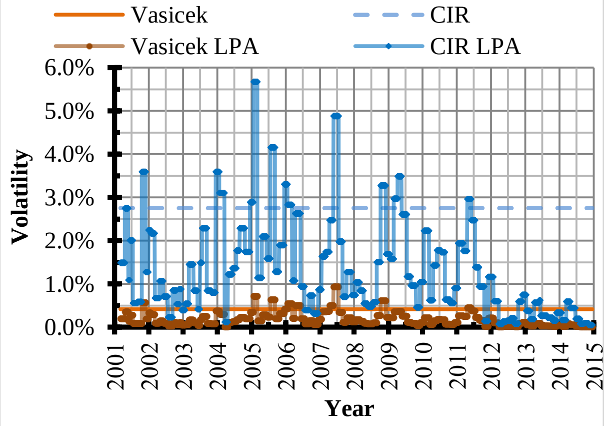


Figure 6: Volatility $\sigma$

# Development of Attention, Cognitive and Affective Skills Through Mindfulness Training Among Young Children in a Classroom

Yau Yan Wong*

*International Program, Kasetsart University Laboratory School, Chatuchak, Bangkok 10900, Thailand, yauyanw34@gmail.com*

## Abstract

The purpose of this study is to investigate if daily mindfulness practice in the classroom can improve the cognitive skills, emotional awareness and concentration. A group of 30 fourth graders from 9 to 10 years old in an elementary school in Thailand participated in a 15-session-mindfulness-training-course from November 2014 to February 2015. The course aimed at cultivating students' awareness in their body sensations, emotions, thoughts and judgments in their daily life so that they would attend more to the present moment and make better decisions. These 15 sessions included topics like mindful eating, mindful breathing, active listening, conflict resolution, etc. In addition to the training course, participants practiced mindful breathing for 3 minutes per day during those four months. Participants took a pretest and a posttest in the study. Two self-report scales were used, including the Cognitive and Affective Mindfulness Scale (CAMS) and Mindful Attention Awareness Scale adapted for Children (MAAS-C). The quantitative results show significant differences in both scales. This means the participants felt a general improvement in their own awareness, sustained attention, and acceptance of their thoughts and feelings. Most participants were more able to enjoy and focus on the present tasks in class after the 3-minute-mindful-breathing exercise during those four months of training. The implication of the study is that educators may use mindfulness training as a tool to enhance students' attention in learning, ability to handle stress, interpersonal communication skills and school satisfaction.

*Keywords*: education; mindfulness training; classroom management; self-awareness; attention; interpersonal communication skills

*Corresponding Author
E-mail Address: yauyanw34@gmail.com

## 1. Introduction

### 1.1 Purpose of Study

Educators have been studying the factors contributing to the resilience, happiness and success of students for years. The main role of educators or parents is to guide children towards happiness and success. However, many schools nowadays are still putting emphasis on academic performance rather than the cultivation of emotional and social skills in their curricula. The lack of emotionally-focused school programs may lie in the fact that researchers have developed no comprehensive conception of usefulness of emotion for purposes of academic learning and instruction (Goleman, 1995; Price, 1998; Salovy & Sluyter, 1997; Sonnier, 1989; cited from Amy, M.B. et al., 2003). Many young people are expected to pick up these skills on their own in an often stressful and competitive school environment. In 1967, the average age of onset major depression was 31 to 33 years of age. Over a 30 year period, this slipped down to 13 to 15 years of age according to the American Journal of Psychiatry, 2007. According to Chris Cullen, co-founder of Mindfulness in Schools Project (MiSP), "The average levels of anxiety for teenagers today are equivalent to those inpatient psychiatry patients in the 1950s. We really do face a mental health epidemic." The World Health Organization (WHO) states that mental health problems will be one of the biggest burden on a society out of all other health conditions. As educators, we cannot afford to let happiness and success of the students happen by chance. There is clearly a need for educators to support young people early on by providing them with the tools to navigate life skillfully.

Mindfulness practice has been recognized as one of the ways to help people reduce depression, anxiety and stress in recent years. It is the practice of present moment awareness of the body and the mind. Although it originated from Buddhism, many psychologists have learned the techniques from mindfulness practice and developed therapies to improve their patients' mental health. Besides its application in counselling, many mindfulness programs start to emerge in schools, such as Mind Up, Mind Matters, Mindfulness in School Projects, etc. There is a huge increase in research publications on mindfulness from 1980 to 2013. These research findings have expanded educators' contemporary belief that intelligences are fixed qualities since people were born. We can actually enhance the capacity of our brains, our ability to focus, and capacity for emotional wellbeing through mindfulness practice. Some brain imaging studies on adults have shown that mindfulness practice profoundly altered the structure and functions of the brain resulting in improved intelligence and mental health. It produced greater blood flow and thickening of the cerebral cortex in areas associated with attention and emotional integration (Davidson & Luz, 2008).

The purpose of this paper is to explore the possibility of developing attention, cognitive and affective skills through mindfulness practice among young children.

This paper was organized as follows: theory background and hypotheses in section 1.2 and 1.3, followed by the research methodology in section 2, results and discussion in sections 3 respectively. The last section is the conclusions and suggestions for further investigation.

### *1.2 Define Mindfulness Practice*

Mindfulness Practice originated from the Buddhist traditions. However, much of mindfulness practice today takes place within a secular context, such as the Mindfulness Based Stress Reduction (MBSR) program developed by Dr. Kabat-Zinn. Rather than seeing it as a religious practice, many people see it as a life skill, a lifestyle or a philosophy of living. Many scientists and educators saw the opportunities given by mindfulness practice and they are looking for ways to apply it practically in their professions.

Mindfulness is paying attention to the present moment through the objective observation of the mind and the body. There are many ways to train our attention. For example, we can use our breathing as the anchor of our mind. Whenever the mind wanders which always happens, we can notice the change of the mind and bring our attention gently back to our breathing. We can use other things as the anchor of the mind besides breathing. We may use movements of our body when we are performing some daily activities, such as walking, eating, sitting or even lying down. The main objective of mindfulness practice is to enhance one's awareness of how the thoughts and emotions affect one's choices and actions. Someone with a high level of mindfulness usually are less likely to react impulsively towards one's thoughts and emotions. Such a person is more likely to slow down and see situations clearly from a distance. This ability allow a person to experience less stress, more peace and happiness.

### *1.3 The Impact for Mindfulness on Children*

According to a paper entitled "Evidence for the Impact of Mindfulness on Children and Young People" by Katherine Weare (2012), there are a lot of studies showing promising benefits associated with mindfulness practice among young children. To name just a few, Napoli, Krech & Holley (2005) reported that children from 5 to 8 showed significant decrease in both test anxiety an ADHD behaviors, and also an increase in the ability to pay attention after 12 sessions of mindfulness practice.

Another study conducted by Semple et al (2010) reported significant improvements on measures of attention and reduction in anxiety and behavior problems among 9 to 13 year old children, which took a 12-week group program based on MBCT, compared to the control group.

Saltzman and Goldin (2008) conducted an 8-week MBSR intervention with 31 children, aged 9 to 11, with their parents. Based on the self-reports of children and parents' feedback as well as objective measures, significant improvement in attention, emotional reactivity and some areas of metacognition was found.

As a result, mindfulness training has been offered as a series of PSHE (Social and Emotional Learning lessons) in schools or integrated informally with learning in a curriculum. Research studies have shown that children did better academically when the schools attend to the emotional and social development (Gewartz, 2003).

From the above findings, the following hypotheses were proposed.

**Hypothesis:**

1. There is a positive relationship between mindfulness training and students' attention, cognitive and affective capacities.
2. There is a positive relationship between mindfulness level and students' academic performance.

## 2. Research Methodology

### 2.1 Samples and Data collection

This action research is a pretest posttest longitudinal study conducted in an elementary English program in Bangkok, Thailand. The socio-economic-status of the students ranged from middle to upper class. All of the lessons are conducted in English, except subjects such as Social Culture and Thai Language. Students have a high level of English proficiency. The mindfulness training course was conducted in English by a foreign teacher.

The target group is grade 4 elementary students ranging from 9 to 10 years old, which starts to become aware of social dynamics and experience more stress from academics. There were in total 31 participants (18 girls and 13 boys) in the mindfulness training, but one of the boys' scores were omitted because he was absent from school for many days and missed almost half of the mindfulness training sessions.

### *2.2 Mindfulness training*

From November 2014 to February 2015, students participated in a mindfulness training course. There were in total 15 sessions (not including the pretest and the posttest) and each session lasted about 15 to 20 minutes. The first eight sessions focused on the cultivation of one's awareness and attention towards the body movements and the nature of the mind. The first part included activities, such as, mindful breathing, bell meditation, eating contemplation, understanding the monkey mind, mindfulness jar, and walking meditation. The last seven sessions focused on the cultivation of interpersonal skills, such as, compassion towards oneself and the others, active listening, and conflict resolution skills. The second part included activities, such as, beginning anew, active listening, watering a flower, and conflict role play. The lessons of the training course were adapted from Thich Nhat Hanh's booked entitled Planting Seeds: Practicing Mindfulness with Children, 2011. (Table 1)

In addition to the training course, students also practiced 3-minute-breathing-exercise before every math or science class on a daily basis for consolidating their practice. The breathing exercise was also used occasionally in-between classes or during the transition between activities in order to improve students' concentration and reduce stress.

*Table 1 Descriptions of Mindfulness Training Sessions*

| **Lesson** | **Title** | **Tasks** |
|---|---|---|
| **1** | Belly Breathe | Observe how the belly moves while breathing |
| **2** | Ring a bell | Observe the sense of sound |
| **3-4** | Catch me if you can | Be aware of the movement of hands and teacher's instructions |
| **5-6** | Food contemplation | Observe the sense of taste and be aware of food consumption |
| **7** | Walking game | Cultivate concentration by observing the movement of legs |
| **8** | The Monkey Mind | Observe how often the mind gets distracted by using breathing as an anchor |
| **9** | The Mindfulness Jar | Learn how difficult emotions affect the mind |
| **10** | Beginning Anew | Let go of the past, apologize for one's mistakes and start anew |
| **11** | Watering Flowers | Identify the positive qualities in the others and express appreciation |
| **12-13** | Active Listening | Define deep listening, take turns to listen to a partner without interruption |
| **14-15** | Conflict Role Play | See a conflict from different perspectives and develop constructive strategies to resolve it |

*2.3 Measurements*

Thirty participants took a pretest before the intervention and a posttest after the intervention. Two scales from the Ohio State University Mental Health Outcome Measures were used. The first one was a self-report scale called the Cognitive Affective Mindfulness Scale-Revised (CAMS-R) (Feldman et al., 2003) with 10 items on a four-point Likert-type scale. It measures the participants' trait mindfulness qualities, like the ability to regulate attention, orientation towards present experience, awareness of the experience, and an open attitude towards the experience. This scale is suitable for participants with no prior experience in mindfulness training and not specific to any particular way of mindfulness training.

Like CAMS-R, the second scale is also a self-report measurement called Mindful Awareness Attention Scale adapted for children (MAAS-C) (Molly et al., 2013). All of the 15 items of the MAAS loaded on only one factor (Acting with awareness/automatic pilot/concentration/non-distraction). It is a child-friendly 6-point Likert-type scale without specialized jargons for experienced meditators. Unlike CAMS-R, this scale focuses mainly on attention regulation and self-awareness, but does not cover attitudinal aspect.

## 3. Research Results and Discussion

### 3.1 Data Descriptions

There were 18 female students (60%) and 12 male students (40%) participating in this study. However, when analyzing the data of the Cognitive Affective Mindfulness Scale-Revised (CAMS-R), two outliers are omitted. As a result, the data of only 10 male students (35.7%) and 18 female students (64.3%) are included.

Female students perceived themselves having a higher level of cognitive and affective capabilities from M=27.78, SD=3.42 in pretest to M=30.44, SD=4.11 out of 40. Like the female students, the mean score of CAMS-R of the male students increases from M=27.60, SD=6.40 in pretest to M=28.50, SD=6.29 in posttest. (Table 2) This means the male students also perceive themselves achieving a higher level of cognitive and affective abilities in terms of their ability to regulate their attention, orientation towards the presence, and awareness and openness towards current experiences.

Most of the male students have a higher posttest score, except two students (outliers) with exceptionally low posttest scores. These two students had low academic performance and some social problems before the course, but they scored quite high in the pretest.

After the training, they reported noticing a lot of distractions in their mind and how they reacted impulsively to some stressful situations. Hence, their low posttest scores may actually reflect that they achieved higher level of self-awareness after the mindfulness training. However, an objective emotional-laden scale may be needed to assess their genuine cognitive and affective abilities.

A t-test is performed to compare the means of CAMS-R between male and female students. No significant difference can be found between the two genders in both pretest scores (t=-0.096, p=0.924) and posttest scores (t=-0.879, p=0.395).

*Table 2 Mean and Standard Deviations of Cognitive Affective Mindfulness Scale-Revised between male and female students*

| | **GENDER** | **STATISTIC** | |
|---|---|---|---|
| **CAMS-R PRETEST** | Male | N | 10 |
| | | Mean | 27.60 |
| | | Std. Deviation | 6.398 |
| | Female | N | 18 |
| | | Mean | 27.78 |
| | | Std. Deviation | 3.422 |
| **CAMS-R POSTTEST** | Male | N | 10 |
| | | Mean | 28.50 |
| | | Std. Deviation | 6.294 |
| | Female | N | 18 |
| | | Mean | 30.44 |
| | | Std. Deviation | 4.1055 |

Three outliers are omitted from the analysis of the Mindful Attention Awareness Scale adapted from Children (MAAS-C scores, resulting in 9 male (33.3%) and 18 female students (66.7%). The mean score of MAAS-C of the male students increases from (M=70, SD=8.426) to (M=71.56, SD=11.812). The mean score of MAAS-C of the female students increases from (M=64.28, SD=10.71) in the pretest to (M=71.44, SD=8.18) in the posttest out of 90. Both male and female students perceive themselves having a higher level of attention and self-awareness after the mindfulness training.

A t-test with independent means is performed to compare the mean scores of male and female students. No significant gender difference can be found for this scale in the pretest (t=1.515, p=0.145) and the posttest scores (t=0.025, p=0.98).

*Table 3 Mean and Standard Deviations of Mindful Attention Awareness Scale adapted for Children between male and female students*

| | **GENDER** | **STATISTIC** | |
|---|---|---|---|
| **MAAS C PRETEST** | Male | N | 9 |
| | | Mean | 70.00 |
| | | Std. Deviation | 8.426 |

| | | | |
|---|---|---|---|
| | Female | N | 18 |
| | | Mean | 64.28 |
| | | Std. Deviation | 10.709 |
| **MAAS C POSTTEST** | Male | N | 9 |
| | | Mean | 71.56 |
| | | Std. Deviation | 11.812 |
| | Female | N | 18 |
| | | Mean | 71.44 |
| | | Std. Deviation | 8.176 |

3.2 Test of Hypothesis

3.2.1 Pretest and Posttest Comparison

All statistical tests have been performed within a 90% confidence interval. According to the results, the posttest scores were significantly more than the pretest scores for both scales. The minimum score of CAMS-R pretest was 17 out of 40 and the highest score was 36 out of 40. The minimum score of CAMS-R pretest was 17 out of 40 and the highest score of CAMS-R posttest was 38 out of 40.

A t-test with dependent means was performed to compare the pretest scores (M=27.71, SD=4.585) and posttest scores (M=29.75, SD=4.971) of CAMS-R. The average difference is significant, which means students' assessment of their own level of self-awareness, emotional wellbeing, willingness to attend to the present, and ability to regulate attention have been enhanced (t=-2.00, p=0.56). (Table 4)

The minimum score of MAAS-C pretest was 40 out of 90 and the maximum score was 85 out of 90. The minimum score of MAAS-C posttest was 50 out of 90 and the maximum score was 86 out of 90.

A t-test with dependent means was performed to compare the pretest scores (M=66.19, SD=10.22) and posttest scores (M=71.48, SD=9.31) of MAAS-C. The average difference is significant, which implies that students perceive themselves having achieved a higher ability to regulate attention and higher awareness of their own thoughts and emotions through the mindfulness training (t=-2.88, p=0.008). (Table 4)

*Table 4 T-test results of CAMS-R Pretest-Posttest Scores and MAAS-C Pretest-Posttest Scores*

| | N | M | SD | t | df | Sig |
|---|---|---|---|---|---|---|
| **CAMS-R Pretest Posttest** | 28 | -2.04 | 5.39 | -2 | 27 | .056 |
| **MAAS-C Pretest Posttest** | 27 | -5.30 | 9.57 | -2.88 | 26 | .008 |

Two students with attention deficit showed significant higher scores in both scales. Although the sample size was too small to draw a solid conclusion, this result may suggest possible use of mindfulness practice in helping students with attention deficit to enhance their cognitive ability. However, no significant improvement in their affective abilities can be found through observations by teachers.

3.2.2 Attention, Cognitive, Affective Skills and GPA

Pearson correlation coefficient is calculated between students' first semester GPA and CAMS-R pre-and post-test scores. There is a moderately significant correlation between GPA and pre-test scores ($R^2$=0.515, p<0.005). However, no significant correlation can be found between the CAMS-R posttest scores and GPA ($R^2$=0.349, p>0.05). This may imply a positive relationship between students' academic performance and cognitive and affective abilities, but further investigation is needed to explore this relationship.

Pearson correlation coefficient is calculated between students' first semester GPA and MAAS-C pre- and post-test scores. A moderately significant correlation is found between MAAS-C posttest scores and GPA ($R^2$=0.563, p<0.005). However, no significant relationship can be found between MAAS-C pretest scores and GPA ($R^2$=0.373, p>0.05). This may suggest a positive relationship between student's academic performance and attention and self-awareness. But more research is needed to explore this relationship.

### 3.3 Qualitative Explanation

#### 3.3.1 Lesson Reflections

Throughout the training, many participants reported that they enjoyed their experiences and were able to apply some techniques, for example loving-kindness meditation and belly breathe, in their daily life. Some students liked to do the 3-minute-breathing-exercise before the math tests without any prompting from the teacher. They believed this activity helped them reduce anxiety and enhance their test performance. The 3-minute-breathing-exercise was conducted in the beginning of every math class. Some students reported that they felt more focused and could finish their mad minute multiplication quiz faster after the 3-minute-breathing-exercise.

During lessons 6 and 7, after the food contemplation exercise, a lot of students reported that they had never appreciated a piece of fruit mindfully before. They noticed that they used to gobble food too quickly without appreciating what they ate. One of the participants said he noticed that he breathed in whenever he took a bite of his food at dinner. This shows that the participant became more aware of the autopilot working of the body and the mind. He became more aware of the food that he was consuming and how his consumption affected his body.

During lesson 9, The Monkey Mind, students were asked to sit quietly and observe how the mind get distracted within 3 minutes. One student said he felt time had slowed down while he was observing his breathing and he noticed more thoughts in his mind. This shows that he was more aware of the changes happening in his mind.

During lesson 10, students were asked to add some colored sand to a glass of clear water. Then they observed how the sand swirled around in the water and gradually settled down. The teacher asked them to imagine the sand as difficult emotions and the water as their mind. Students talked about their observations with their partners after that. One pair of students mentioned that the difficult emotions were like the colored sand that always existed in the mind, but would show up whenever the mind was disturbed. Another pair said that when the mind was calm, the difficult emotions would settle down. One pair of students also added that difficult emotions were like the colored sand in the water that clouded their vision. At the end of the session, many students suggested using belly breathe, taking a mindful walk or loving-kindness meditation to calm themselves whenever facing difficult emotions.

About 3 students reported that the loving-kindness meditation in lesson 11, Beginning Anew, helped them understand their friends more and forgive their mistakes more easily. During lessons 13 and 14, students had to role play a conflict happened before and discussed how they could resolve the conflict. There were in total 6 groups. All of the groups agreed that being kind, patient, honest and forgiving was the key to resolve a conflict. One group added a short presentation about different techniques, which they learned from the Mindfulness Training course, can be used to resolve a conflict. Below is the list of steps that they suggested:

"Number 1: Cool off! Count down backwards, Close your eyes and take a deep breath. Number 2: Figure out what's bothering you. Number 3: Deal with the issue. Talk in private. Keep your voice calm. Share how you feel. Number 4: Listen without judgment. Number 5: Work it out. Talk about ways to settle the conflict that will meet both of your needs. Be willing to change and keep an open mind. Be willing to say you're worry, forgive, and move on."

In summary, mindfulness training may be a useful tool for children to reduce stress, alleviate anxiety, manage and communicate difficult emotions effectively, and bring their focus back to the current task. It can be used by educators for cultivating a positive learning atmosphere and reducing the bullying problem in schools.

#### 3.3.2 Parents' Feedback

A voluntary survey, adapted from CAMS-R was conducted electronically among parents and only 6 parents responded. There were 10 questions on a 4-point-Likert-scale. All of these 6 parents noticed some or significant improvement in their children's ability to pay attention to the present moment and remember thing in greater details. 5 parents noticed some significant improvement in their children's tolerance and acceptance to emotional stress or changes, but 1 parent noticed no change at all. All of the 6 parents noticed some or significant improvement in their children's ability to stay focused while doing homework. 5 parents noticed a significant or great increase level of school satisfaction, but 1 parent did not notice any change. 3 parents noticed some improvement in their children's ability to think before reacting to a situation, 2 parents noticed a significant improvement, and 1 parent noticed a great improvement. All of the 6 parents noticed significant or great improvement in their children's ability to relax and focus under stress.

Although the response rate of this survey was low, the feedback from these parents gave some insight about how students used the skills they gained from the mindfulness training in their daily life and the positive impact of the training.

## 4. Conclusion

In this study, thirty children from 9 to 10 years of age participated in a 4-month-mindfulness training course together with daily 3-minute-breathing exercise. The main goal of this training is to enhance students' attention regulation, emotional wellbeing, intentional awareness and acceptance of present experiences. The results showed a significant difference between the pretest and posttest scores of both measures used in the study, which are Cognitive Affective Mindfulness Scale and Mindful Attention Awareness Scale for Children. Participants showed a higher level of attention, happiness, self-awareness, and resilience towards emotional stress. The results from the study imply that mindfulness can be trained and enhanced among young children. Students can use mindfulness practice as a tool to manage their attention, emotional stress, social problems, and anxiety from study more effectively. Educators may use mindfulness practice to improve the learning atmosphere in the classroom and cultivate a healthy learning community.

There were some methodological difficulties in this study, including small sample size, lack of control group comparison, and a great reliance on self-report. So the conclusions can only be tentative. There were other logistical difficulties, such as interruption of other school activities which caused cancellation of some lessons and occasional absence of some students. However, as there are not many mindfulness research focusing on children, it can be used as a pilot study for future investigation about the possibility and effect of mindfulness training conducted among children.

The data analysis of this study found a somewhat positive relationship between children's academic performance and mindfulness training. However, more investigation is needed to look into whether and how mindfulness training enhances students' academic performance. Future research can also look into whether mindfulness training can be used to help children with special educational needs (SEN), such as Autism, ADHA, ADD or anxiety disorder, enhance their cognitive and affective abilities.

## 5. Acknowledgement

This study was based upon work supported by International Program, Kasetsart University Laboratory School, Thailand and the technical assistance from Assistant Prof. Chawalit Jeenanunta of SIIT, Thammasart University, Thailand.

## 6. References

[1] Brett, A., Smith, M., Price, E., & Huitt, W. (2003*)*. Overview of the affective domain. *Educational Psychology Interactive*. Valdosta, GA: Valdosta State University.

Available from:
http://www.edpsycinteractive.org/brilstar/chapters/affectdev.pdf

[2] Davidson, R. & Lutz, A. (2008). Buddha's brain: neuroplasticity and meditation. IEEE Signal Process Mag.25(1):176-174.
Available from:
http//www.ncbi.nlm.nih.gov/pmc/articles/PMC2944261/accessed 30th January 2012.

[3] Feldman, G. C., Hayes, A. M., Kumnar, S.M., & Greenson, J. M. (2003). Clarifying the construct of mindfulness: Relations with emotional avoidance, over-engagement, and change with mindfulness training. Paper presented at the Association for the Advancement of Behavior Therapy, Boston, MA.

[4] Gewertz, C. (2003). Hand in hand. *Education Week, 23*(1), 38-42.

[5] Goleman, D. (1995). *Emotional Intelligence: Why It Can Matter More Than IQ*, Bantam Books.

[6] Katherine, W. (2012). Evidence for the Impact of Mindfulness on Children and Young People, University of Exeter, UK.

[7] Molly, S. L., Kimberly, A. S. R., Anne, M. G. & Bruno, D. Z. (2013). A Validation Study of the Mindful Attention Awareness Scale Adapted for Children. Springer Science and Business Media, NY, USA

[8] Napoli, M., Krech, P.R., & Holley, L.C. (2005). Mindfulness Training for Elementary School Students. Journal of Applied School Psychology, 21(1), 99-125.

[9] Price, E. (1998). Instructional systems design and the affective domain. *Educational Technology, 38*(6), 17-28.

[10] Saltzman, A., & Goldin, P. (2008). Mindfulness based stress reduction for school-aged children. In S.C. Hayers & L. A. Greco (Eds.), Acceptance and mindfulness intervention for children adolescents and families (pp. 139-161). Oakland, CA: Context Press/ New Harbinger.

[11] Salovy, P., & Sluyter, D. (Eds.). (1997). *Emotional development and emotional intelligence - educational implications*. New York: Basic Books.

[12] Semple, R. Lee, J., Dinelia, R., & Miller, L. (2010). A randomized trial of mindfulness-based cognitive therapy for children promoting mindful attention to enhance social-emotional resilliency in children. Journal of Child and Family Studies, 19(2), 218-229.

[13] Sonnier, I.L. (1989). *Affective education: Methods and techniques*. Englewood Cliffs, NJ: Educational Technology Publications.

[14] Thich Nhat Hanh & the Plum Village Community. (2011). Planting Seeds-Practicing Mindfulness with Children. Parallax Press. Berkeley, CA, U.S.A.

# The Role of Manager's Knowledge and Experience on Hotel Innovation and Performance

Narongsak Pongsathornwiwat*, Chawalit Jeenananta, Thanakorn Leelangkura, Sarun Leelaphiwat,
Patcharasit Phunpetch and Teerapat Vongpakorn
*School of Management Technology, Sirindhorn International Institute of Technology,*
*Thammasat University, Pathum-Thani, 11120, Thailand,*
*e-mail address: npongsathornwiwat@yahoo.com*

Abstract

Innovation is one of the main sources for sustaining long-term competition in all businesses as intensively investigated in manufacturing and production based industries. Firms that have more innovations can gain better performance. However, there is a lack of research empirically investigating what are the key indicators (factors) affecting the successful innovation especially in hospitality sectors in the extent of the literature. This study therefore aims to explore and empirical analysis such innovating factors and also investigate the relationships between innovation and firm performance in Thai hotel industry. Data were collected form a stratified sample of 150 hotel firms in Bangkok, Thailand using cross-sectional survey. The statistical methods include confirmatory factor analysis and two-stage least square regression. The findings indicate that the managerial role of top management as manager's knowledge backgrounds has positively impact not only innovation but so firm performance. The results provide the significantly understanding on innovation performance in hotel industry.

*Keywords*: Innovation, Firm performance, Internal factor, Hotel industry, Survey

*Corresponding Author

E-mail Address: npongsathornwiwat@yahoo.com

## 1. Introduction

The importance of service industry is becoming one of the main sources of economic growth as shown in the GDP [13]. Among the service based businesses, the tourism sector is the highest industries that play significantly impact to the economy around the world, especially in Thailand [13]. In tourism sector, there are many firms/organizations that can be various roles in the tourism value chain [15], however, most are in the hotel industry [7, 13]. Since the numbers of travelers have been dramatically in past five years [13], hotel firms then are seeking for the way to increase and improve their capacities in order to satisfy the needs of travelers favorably and also to gain the competitive advantage over the competitors. Innovation is one of the main management practices aiming to improve the performances such as productivity and operational excellences to services successfully [10, 15]. The example of the innovation practices in hotel industry can be described as the additional (valued-added) services such as clean and healthy foods, free internet access and the fasten reception process at the front desk, which are positively to influence the higher performance [7].

As mentioned earlier, the benefits of innovation is very sweet to all organizations, especially in hotel firms. Nevertheless, most are failing in introducing the successful innovation. The question is what are the critical factors that initiate the innovation in hotel industry is still scant since the extent of innovation literature has been more investigated over the past decades. Also, the previous researches have intensively examined the factors such as human capital, knowledge integration and relationships between tourism organizations and also hotel's characteristics [7, 13-14]. The managerial role of top management, however, is less investigated. To fulfill the literature, the main motivation of this research is to explore the factor that positively impact innovation using hotel industry as our stratified sample.

This paper is structured in the following section. Section 2 is described the theory background and building a set of hypotheses. Next, an empirical study is analyzed. The discussion of the results is then presented in the section 4. The last section is provided the concluding remarks and possible question for future research.

## 2. Theory backgrounds and hypotheses

In this section, we shall briefly the foundation of our study and also the set of proposed hypotheses.

### *2.1 Innovation in hotel industry*

Innovation generally refers a change or creation more effective processes, products and ideas [7]. Innovation is linked to performance and growth through improvements in efficiency, productivity, quality, competitive position and also market share. It typically added values by changing old organizational forms and practices. Since the services is intangible by itself, it is therefore difficult to measure innovation directly like the numbers of technological products in manufacturing and production based industries such as automotive and electronics.

In this study, we adapt and bring this idea to hotel businesses, it could be meant to implement new ideas, creating dynamic services or improving not only your existing services but also managerial practices [3-4, 6-7]; for example the additional services form the regular by adopting the information technology and to provide a specific food for the particular nationality. Based on the previous studies, chawalit et al. [8] have explored the innovation factors by comparing between service sector and manufacturing industry using a cross-sectional data. They indicate that the innovation process for production linkages flows for customer to supplier. However, in service industry the internal factors are affects to the innovation. Similarly, Narissara and Chawalit [13] have also founded that internal factors such as human assets, technological developments are positively influence to the successful innovation in hospital industry.

### *2.2 Role of manager's knowledge and experience*

Based on the resource based view theory [6, 13], the competitive advantage of firm is very relying on the internal resource they have. It can be categorized the resources into two perspectives including tangible assets and intangibles assets. Most studies are more investigates the tangible assets such human capital and physical assets [6, 13-14]; however, the innovation in service based industry the knowledge as intangible assets is significantly very important to sustain the long-term competition which are lacks of empirical analysis.

As mentioned earlier, the manager's knowledge as construct has been discussing in the current research in various perspectives [14]. The sources of innovation in hotel industry are strongly based on the internal sources in the organizations as refers in [6. 13-14]. Nieves and Segarra-Cipres [6] have stated that the human capital, the capacity of knowledge integration and manager's relationship are the positively significant factors influencing the innovation. Similarly, the study of Wagner [13] pointed out that the knowledge sharing has positive linkage to service innovation. However, both related studies did not consider the role of manager's knowledge that can bring the knowledge not only form their experiences (internal sources) but also from outside of firms (knowledge spillover). Referring to the background, in this study we formulate the hypotheses as below;

Hypothesis: There is a positive relationship between the knowledge, manager's experience and the type of management style on innovation

Hypothesis 1: The manager's knowledge has positive impact on innovation
Hypothesis 2: The manager's experiences of manager has positive impact on innovation
Hypothesis 3: The type of management has positive impact on innovation

### *2.3 Innovation and firm performance*

Firms that have more innovativeness will gain better performance. This concept is empirically supported in many studies especially in manufacturing and production based industries [2-3]. Qao and Chang [2] is explored and empirically investigated the positive linkage innovation on firm performance in USA's car manufacturing. However, few are explored and tested such constructs in hotel industry. For example, Ramayah et al,. [11] have shown that the network collaboration between tourism organizations has positive effect on firm performance in Malaysia. But they did not explored how tourism supply chain members react to each other such as how knowledge transfer or spillover form firm to firm and also whether how such knowledge has positive impact on not only innovation and performance or not. It is emphasized here that the innovation performance and firm performance should be distinct in order to see the direct effect between both constructs [10]. Based on the literature backgrounds, we shall propose the hypothesis as below;

Hypothesis 4: Innovation has positive impact on performance

## 3. Research methodology

### *3.1 Samples and data collection*

This research creates from case study and literature review to understand about which factor that influence to innovation in hotel in Thailand. Then the questionnaire was designed to refine and collect the data. The data was collected using cross-sectional multiple methods including online survey, mail survey and electronic mail based on stratified sampling technique. Moreover, researcher also use the purposive sampling to be a target people to answer questions such as hotel manager who has experiences and knowledge about management. The survey was sent to 1,000 hotel firms in January 2015 to April 2015. There are 150 hotels had responsed survey which is the rate of 15% that divide into 103 hotels from interview, 7 hotels from online survey, 35 hotels from e-mail and 5 hotels from mails.

Table 1 shows the characteristics of our sample data set. The major attribute is the independent operating hotel at 65.3%. Most are invested by local investor

Table 1: Statistical descriptions of hotel structure

| Construct and items | Frequency | Percent |
|---|---|---|
| Hotel type | | |
| • Operating under the company network. | 46 | 30.7 |
| • Independent opening | 98 | 65.3 |
| • Independent operating under franchise agreement | 6 | 4 |
| Hotel Structure | | |
| • 100% of Thai Shareholders. | 129 | 86 |
| • Proportion of shares divided between Thai and foreigners | 21 | 14 |
| Employee | | |
| • Less than 50 | 57 | 38 |
| • 51 - 100 | 49 | 32.7 |
| • 201 - 300 | 6 | 4 |
| • 301 - 400 | 11 | 7.3 |
| • 401 - 500 | 9 | 6 |
| • More than 500 | 18 | 12 |

about 86%, and the numbers of employees are less than 50 and 50-100 which are 38% and 32.7%, respectively.

### *3.2 Measures*

A questionnaire using a five-point Likert scales was used to gather data for all of each construct. All instruments were adapted from previous literatures and were modified to measure the performance as shown in Table 2.

Table 2 shows the descriptive statistic of each construct and measurement items. Firm performances including the financial and service capabilities are proposed to measure the research model.

### *3.3 Goodness of measures*

In this study, hotel types and the numbers of employees are defined as control variables that can confound in the regressions analysis since there is no consensus in the literature about the impacts of these variables on the relationships e.g. innovation and performance [4, 6]. To testing the consistency, both validity and reliability analysis are used for testing goodness of our proposed [9]. We used principle component analysis (PCA) and VARIMAX rotation to extract the factors. All sampling adequacy values are range over 0.60. It means that the factors are appropriately validity [3, 19]. Also, the Cronbach's alpha coefficient was assessed the inter-item consistency of our measurement items as shown in Table 2. As seen from Table 2, all alpha values of all constructs are above 0.70 [3, 19]. As such we can conclude that the measurement scales are reliable.

## 4. Empirical result and Discussion

In order to test the three hypotheses generated, we applied econometric approaches a two-stage least square regression (2SLS) as a technique to examine the proposed hypotheses. Next, we discussed in details intensively.

### *4.1 Effects of manager's knowledge on innovation using OLS*

The analysis has shown that the manager's education and types of management style are positively significant impact on innovation statistically at 0.05 confident levels as shown in Table 3

Table 2: Statistical descriptions of hotel firms

| Construct and items | Frequency | Mean | S.D. | KMO | Factor Loading | α |
|---|---|---|---|---|---|---|
| Manager knowledge | | | | 0.610 | | 0.702 |
| *Level of education* | | 2.50 | 0.496 | | 0.848 | |
| • Below undergraduate | 6 | | | | | |
| • Undergraduate | 65 | | | | | |
| • Higher education | 79 | | | | | |
| *Which country graduated* | | 1.34 | 0.490 | | 0.843 | |
| • Thailand | 98 | | | | | |
| • Oversea | 52 | | | | | |
| Knowledge experiences | | | | | | |
| *Working experiences* | | 3.23 | 1.163 | | 0.750 | |
| • Lower than 1 year | 3 | | | | | |
| • 1-5 years | 45 | | | | | |
| • 6-10 years | 48 | | | | | |
| • 11-15 years | 23 | | | | | |
| • More than 16 years | 31 | | | | | |
| *Graduate from faculty of tourism and hotel management* | | 1.58 | 0.496 | | 0.744 | |
| • Yes | 62 | | | | | |
| • No | 88 | | | | | |
| Types of management | | 1.072 | 0.260 | | 0.937 | |
| • Manage by owner | 137 | | | | | |
| • Hiring outsource | 13 | | | | | |
| Construct and items | Frequency | Mean | S.D. | KMO | Factor Loading | α |

| | | | | | |
|---|---|---|---|---|---|
| Innovation | | | 0.840 | | 0.829 |
| • New services to customer | 3.45 | 0.809 | | 0.846 | |
| • Competitor adopts your provide services | 3.19 | 0.865 | | 0.618 | |
| • Always additional services to customer | 3.49 | 0.822 | | 0.727 | |
| • Develop new internal managerial processes | 3.34 | 0.847 | | 0.793 | |
| • Adopt new managerial process from competitor | 3.20 | 0.870 | | 0.759 | |
| • Develop the relationship with other tourism supply chain members. | 3.27 | 0.850 | | 0.662 | |
| Firm performance | | | 0.784 | | 0.776 |
| • Higher profit | 3.51 | 0.730 | | 0.743 | |
| • Lower cost | 3.30 | 0.842 | | 0.635 | |
| • Fasten services processes | 3.60 | 0.849 | | 0.742 | |
| • Flexible and specific services delivery | 3.52 | 0.887 | | 0.783 | |

Notes: 1 = least, 2 = little, 3 = moderate, 4= very, 5very much.

Table 3: Managers' knowledge and innovation

| Independent Variable | Dependent Variable | |
|---|---|---|
| | (1) | (2) |
| Manager's knowledge | 0.449* | 0.296* |
| Manager's experiences | -0.097 | 0.147 |
| Types of management | 0.171* | 0.133 |
| R-squared | 0.225 | 0.264 |
| F-statistics | 14.134* | 10.754* |
| Hotel type | No | Yes |
| Employees numbers | No | Yes |
| Constant | Yes | Yes |
| Observation | 150 | 150 |

Notes: Robust standard errors in parentheses. *p<0.05.

In addition, the control variable was included in the equation in order to robust analysis as shown in column (2) in Table 3. The empirical evidence provided that only the *manager's knowledge* is significant support robustly. But, in case of manager's experiences and types of management style, the further evaluations should be investigated in order to examine whether such factors have positive influence to the innovation or not. The result is similar and consistent with the related studies [6-7, 12-13]. It can be concluded that the level of education and country where graduation has positive impact to innovation. The knowledge is flowed within organization because the manager bring such the benefits of knowledge not only form their education but also from the outbound sources that knowledge is spillover.

### *4.2 Effects of innovation and firm performance using OLS*

The OLS estimations were used to examine the impact of innovation on firm performance. The OLS estimation result shows in Table 4. The finding indicates that the estimated coefficient of innovation is positively significantly on firm performance at the 5% level statistically. Our result is consistent with the related literature [3-4]. It can be pointed out that when firms have more innovativeness, they can gain better performance such as higher profit and can reduce the operation cost. The result provides the reflect contribution to the literature not only hotel industry but also service industry.

Table 4: Innovation and firm performance using OLS

| Independent Variable | Dependent Variable (Firm performance) |
|---|---|
| Innovation | 0.384* |
| R-squared | 0.141 |
| F-statistics | 23.330* |
| Hotel type | Yes |
| Employees numbers | Yes |
| Constant | Yes |
| Observation | 150 |

Notes: Robust standard errors in parentheses. *p<0.05.

### *4.3 Effects of innovation and firm performance using 2SLS*

As suggested in management literature, the measurement error is a main problem when utilizing the regression analysis [1, 9]. In order to robust the empirical analysis, in this session we reformulate an assumption that more intense better firm performance will cause firm to pursue the innovation would not be determined jointly with the independent variable. Again, we also assume that firm achieving better performance might be more likely to be influenced by the role of manager performance. Therefore, if the reverse causality problem exits, OLS will yield inconsistent and biased estimates. We will solve this problem by performing two-stage least squares (2SLS). It should be emphasized here that in order to produce 2SLS, the number of instruments is greater than or equal to the number of independent variables [1]. The procedure will discuss in the following paragraph.

Table 5 shows the coefficient results estimated by 2SLS. Robustly, there is positively significant between innovation and firm performance. However, the value of R-Squared at 0.032 is lower than the previous analysis in Table 4. This may caused by the effect of *manager's knowledge and experiences* as the independent variable. The further analysis should be carried out. To robustness, the robust regression F of endogeneity rejects the null hypotheses that the variable for innovation is exogenous in all of estimations for the regression of firm performance. The score chi-square test statistics are exceeding the critical $\chi^2$ value in 8425.42 as the reference value [20] at the 5% level, indicating that the instruments are not weak and valid [20]. Based on the empirical evidences, we can conclude that the manager' knowledge and experience has positive influence on firm performance by mediating of innovation. The overall analyses are summarized in Table 6.

Table 5: Innovation and firm performance using 2SLS

| Independent Variable | Dependent Variable (Firm performance) |
|---|---|
| Innovation | 0.377* |
| R-squared | 0.032 |
| Hotel type | Yes |
| Employees numbers | Yes |
| Constant | Yes |
| *Endogeneity test* | |
| Robust regression F | 5.430* |
| *Overidentifying restriction test* | |
| Score $\chi^2$ | 8425.42* |
| Observation | 150 |

Notes: Robust standard errors in parentheses. *p<0.05

Table 6: Robust empirical evidence conclusion

| Relationship | Conclusion |
|---|---|
| Manager's knowledge → Innovation | Support |
| Innovation → Firm performance (OLS) | Support |
| Manager → Innovation → Firm performance (2SLS) | Support |

Notes: Robust standard errors in parentheses. *p<0.05.

## 5. Concluding remarks and limitations

This study aims at examining whether the role of managers' knowledge has a positive impact on innovation and also on firm performance in hotel industry. The findings show the robust positively significant relationships between managers' knowledge and innovation, and also firm performance by mediating of innovation. The results of statistical analysis are empirically analyzed using OLS method and 2SLS method, consecutively. For practical implications, it should be emphasized that if hotels want to gain more innovativeness, they need to invest in knowledge assets such as to support their manager to have more training to gain more oversea knowledge. The example of managerial implication is useful not only for Thai hotel industry, but also Thai tourism industry.

While our study contributes considerably to the innovation management literatures, especially for Southeast Asia, there are some limitations.

Firstly, the important limitation is the small sample size. This may lead the weak results of our analysis. Hence, we will use a bigger sample sizes to robust our research model.

The second limitation is the 2SLS estimator is separately estimates the measurement model and structural model. Thus, it is strongly necessary to further estimate the effect of these variables simultaneously using the advance techniques such as Structural Equation Modeling (SEM) and maximum likelihood (ML) estimator to robust the study [1, 10].

Finally, we only explore key impacts: the role of managers' knowledge and innovation in hotel industry. However, there are other tourism business such tour operator and also other types of innovation; for example internal and organizational innovation that must need further investigation to provide more understanding. Additionally, in this study do not consider the effect of these indicators on financial and quantitative performances such as profit and gross margin which are shown extensively in literature. These mentioned limitations are challenging opportunities for future investigation.

Acknowledgments

This research was conducted under a grant in the SIIT-JAIST-NECTEC Dual Doctoral Degree Program and also this paper was partially supported by Center of Excellence in Intelligent Informatics, Speech and Language Technology and Service Innovation (CILS), Thammasat University.

References

[1] Bollen KA, Biesanz JC. A note on two-stage least square estomator for higher-order factor analyses, Socio. Meth. Res. 2002, 30 (4); 568-579. 2002.

[2] Cao M., Zhang Q. Supply chain collaboration: Impacts on collaboative advantage and firm performance, Journal of Operation Management. 2011 29: 163-180.

[3] U. Grissemann, A. Plank and A. Brunner-Sperdin, Enhancing business performance of hotels: the role of innovation and customer orientation, Inter. J. Hosp.Manage., vol. 33, pp. 347-356. 2013.

[4] J.-S. Chen, H.-T., Tsou, and R.K.H. Ching, "Co-production and its effects on service innovation," Ind. Market. Manage., vol. 40, pp. 1331-1346. 2011.

[5] L. Crevani, K. Palm and A. Schilling, "Innovation management in service firms: a research agenda," Serv. Bus., vol. 7, pp. 177-193. 2011.

[6] J. Nieves and M. Segarra-Cipres, "Management innovation in the hotel

industry," Tourism. Manage., vol. 46, pp. 51-58. 2015.

[7] F. Orfila-Sintes and J. Mattsson, "Innovation behavior in hotel industry," OMEGA. Int. J. Manage. S., vol. 37, pp. 280-394. 2009.

[8] E. Phong-Arjarn and J. Chawalit, "Explorinng supply chain collaboration in Thai Major Industries," Naresuan. University. J., vol. 19, no. 3, pp. 2011.

[9] P.M. Podsakoff, S.B. MacKenzie and N.P. Podsakoff, "Common method biases in behavioral research: a critical review of the literature and recommended remedies," J. Appl. Physiol., vol.88, no. 5, pp. 879-903. 2003.

[10]Pongsathornwiwat N, Jeenananta C, Theeramunkong T, Hyunh VN. Supply chain collaboration and service innovation in Thai hotel industry, Proceeding of the 5th International Industrial Engineering and Operations Management Conference; 2015 March 3-5; Dubai, UAE. 2015.

[11] T. Ramayah, J.W.C. Lee and J.B.C. In, "Network collaboration and performance in the tourism sector," Serv. Bus., vol. 5, pp. 411-428. 2011.

[12] T.-W. Tang, M.C.-H. Wang and Y.-Y. Tang, " Developing service innovation capability in the hotel industry," Serv. Bus., in press.

[13] Visavanont N, Jeenananta C. Critical factors influencing innovation in Thai hospitals. Nida Business Journal. 2014. 54(2); 229-267.

[14] S.M. Wagner, "Partners for Business-to-Business Service Innovation," IEEE Trans. Eng. Manage., vol. 60, no. 1, pp.113-123., Feb. 2013.

[15] X. Zhang, H. Song and G.Q. Huang, "Tourism supply chain management: A new research agenda," Toursim. Manage., vol. 30, no. 3, pp. 345-358. 2009.

# Crew Rostering in Thai Airways Case by Using Greedy Algorithm

Thanaphat Limgitnuwat and Chawalit Jeenanunta*
*Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, 12121, Thailand*

## Abstract

Airline Crew Rostering Problem (ACRP) or crew scheduling is the complex and hard to solve problem because it consists of many constrains such as; block time, rest time period, aviation regulation, perdiem or salary, and workload. Normally, the methods for solving crew rostering problem that applied in various airlines are: assigning high priority employees to high priority pairing, assigning by consider to crew member requests and assigning without consider to crew member requests. Thus, the aim of this paper is to balance the workload and perdiem for each crew member. This paper presents the Greedy Algorithm method for constructing the crew timetable and minimize standard deviation of workload and perdiem. The proposed Greedy Algorithm developed to solve this problem compose of two phases: construction phase and improvement phase. The construction phase present, the simple crew member schedule which will revise to be more balanced and effective in improvement phase.

*Keywords*: Crew Rostering, Greedy Algorithm, Heuristic, Scheduling

*Corresponding Author
E-mail Address: chawalit@siit.tu.ac.th

## 1. Introduction

Today airline transport is a very popular mode for travel. Thus, the number of flight and number of crew are so high that it makes scheduling a very difficult problems. The good efficiency of crew scheduling can reduce a lot of operation cost, which is the second largest expense for airline [1].

Crew scheduling can divided into 2 main phases; crew pairing, and crew rostering. Crew pairing is the process that manages flight legs within the same fleet that start and ends at the same crew base. The meaning of crew base is the home or station that crew member resides. Another phase of crew scheduling is crew rostering. The purpose of crew rostering is to assign individual crew member to crew pairing, usually on a monthly basis. As the above information, there are three major methods to schedule or roster the crews such as; assign high priority employees to high priority pairing, developing monthly rosters for individual crew member based on their requests, and developing monthly rosters for each day of the month without considering the crew request. Normally, cockpit aircrew member and cabin aircrew member are assigned roster differently. The reason is cockpit aircrew members may require licenses to fly with specific type of aircraft. In general, crew rostering will develop as weekly basis instead of monthly rosters because it is easier and less complex to solve.

The objective of this experiment is to balance the workload and perdiem simultaneously. The reasons is to reduce any inequality amongst the crew members. The method to balance the workload and perdiem is to minimize Standard Deviation (SD). The SD is a measurement of variation, if the value of SD close to zero that means good balance solution. While this paper try to balance workload and perdiem many paper propose several method to minimize a whole operation cost of crew section.

This paper uses Greedy algorithm for solving crew rostering problem. The algorithm will choose optimal path way to solve the problem. The advantages of greedy algorithm are simple to construct, and fast computational time. The disadvantage is that, it is not guaranteed of global optimal solution.

## 2. Literature Review

Crew rostering is the complex and hard-to-solve problem that compose of many constraints. So, there have been many researchers who proposed several methods in order to get the optimal solution. Metaheuristics algorithm is key method that used to solve complex and combinatorial optimization problem that cannot be solved by traditional heuristics. The example of metaheuristics are: genetic algorithm, particle swam optimization, ant colony, tabu search and simulated annealing is one of many applied solving problem technique. The examples of metaheuristic approach for solving crew rostering problem are;

Michel et al. develops graph coloring model for feasibility problem with preferential bidding. They also propose tabu search algorithm for determining if the problem contains at least one feasible solution [2]. Panta and Dusan attempts to solve multi objective aircrew rostering problem by using simulated annealing, genetic algorithms, and tabu search techniques [3]. Broos and Mario present the scatter search algorithm and use meta-heuristic with combination of crew scheduling and improvement principles [4]. Zhang et al. present genetic algorithm and simulated annealing algorithms to solve multi objective crew rostering problem [5]. Nadia and Jacques try to solve crew scheduling problem by applied genetic algorithm. They decomposed the problem into two sub-problems; crew pairing and crew rostering. But they fail to solve crew pairing with large number of possible pair [6]. Walid et al. aim to minimize operational cost based on crew satisfaction. Genetic algorithm applied to solve medium size crew rostering problem [7]. Azadeh et al. present particle swarm optimization (PSO) method which achieved more successful result than other algorithms [8]. Guang-Feng formulated airline crew scheduling problem as traveling salesman problem and then introduce ant colony optimization algorithm to solve it. In addition, the performance was evaluated by performing computational tests regarding real cases as the test problems. The objective is to minimize total crew costs. Traveling Salesman Problems (TSP) and Ant Colony Optimization (ACO) applied to this paper. The result have shown that ACO-based algorithm can be a potential technique for airline crew scheduling [9].

Heuristic approach is also the method that several paper proposed such as Wagner and Nicolau propose integrated solution of crew assignment problem with heuristic procedures. Their objectives are to balance the flight credit and attend to crew demand for day off or duty period on specific date. As a result, this experiment provided very useful information especially for problems with small and medium size instances [10]. Yindong et al. present hybrid genetic algorithm with not-fixed chromosome to solve problem. The techniques performed than fuzzy genetic algorithm and achieved the result and computation time similar to that of linear programming [11]. Herbert et al. experiment on European airlines with an aim to minimize total cost and develop an optimal time table for crew in computationally short time by applying branch-and-bound technique to enhanced model with downgrading method [12].

Constraint Programming (CP) is a powerful tool to find feasible solutions of crew rostering. The technique is very useful for a lot of constraints problem. Even though, this technique is not good enough for finding optimal solution. Column Generation technique was proposed to improve cost saving [13, 14]. Jenny et al. also use column generation technique to solve crew scheduling problem. The result shown as an integrated solution with aircraft maintenance routing problem which results in short computational time and tight schedule bounds. This mean a solution was success for reducing operational cost [15].

Artificial intelligence which include fuzzy set theory, and expert systems also applied in many paper for solving crew rostering problem such as: Dusan and Panta present fuzzy control method by applied day-by-day heuristic to solve crew scheduling problem [16]. Ernesto and Joao use AI-based tool application, ESCALAS to solve crew scheduling problem. It can perform crew assignment to each one of the schedule produced and modify both schedule and assignment produced [17].

Some of researchers applied mathematical programming which normally formulated as linear programs or general mathematical. Budi et al. proposed differential evolution (DE) which is difference from common DE by random swap as mutation operator. This algorithm produces near optimal solution accurately for optimization problem. The result of small and medium dataset are satisfied, and reasonable result for large dataset [18]. Shangyao and Yu-Ping propose pure network model to solve crew rostering problem of Taiwan's China Airlines [19].

Greedy algorithm is the method to solve the problem by choosing the most obvious and immediate choice. Usually greedy algorithm is more suitable for simple problem. Greedy algorithm not support for reconsiders data therefore it has probability to found only local optimal solution but not global optimal solution. Even tough, there have been several paper that applied Greedy algorithm for solve scheduling problem such as: Imma propose an iterated Greedy algorithm for solving the blocking flowshop scheduling problem aim to minimize makespan. The experiment has shown that the algorithm can make efficient result [20]. Nearchou et al. propose two greedy heuristic algorithms for solving manpower shift planning. One for single and another for multiple workday shifts. Their result demonstrates very satisfactory performance in terms of both solution time and quality [21].

## 3. Problem Description

Thai Airways is the biggest airline in Thailand with services in domestic, regional, and international route. The international route can be divided into four regions such as; Asia, Australia and New Zealand, America and European. The distance of each region are different, which consequently have its own amount of workload and perdiem that this paper aims to balance. There are two main constrains that involves as follows;

1. Flight time or block time is the period that start from the plane moving to runway for take-off until it arrived at destination airport and stop every engines.

2. Flight Duty Period is the period of crews operate their duty count from one hour before take-off until thirty minutes after landing.

There also have the flight time limitation and rest time period as shown in Table 1. The rest time period constrain determine by department of civil aviation of Thailand. Flight time limitations are such as;

1. Every 7 days each crew member must accumulate less than 34 hours of flight time.

2. Every 28 days each crew member must accumulate less than 110 hours of flight time.

3. Every 365 days each crew member must accumulate less than 1,000 hours of flight time.

Table 1: Rest period constraint

| Flight duty period<br>(hour) | Rest period<br>(hour) |
|---|---|
| 0 - 8 | ≥ 8 |
| 8 - 10 | ≥ 10 |
| 10 - 12 | ≥ 12 |
| 12 - 14 | ≥ 14 |
| 14 - 16 | ≥ 16 |
| 16 - 20 | ≥ 24 |

Thai Airways aims to create equality of every crew member. Thus, the objective to balance workload and perdiem were set. That means minimizing the Standard Deviation (SD) was the target, the measurement of variation, was set as a target. Due to a lot of constraints; rest time, flight duty period, and fixed operate date make this problem difficult and complex to solve.

## 4. Proposed Algorithms

This paper applied greedy algorithm to solve crew rostering with an aim to balance the workload and perdiem simultaneously. The objective of this experiment is to minimize Standard Deviation (SD).

Greedy algorithm for solving this problem can be divided into two mains phases; construction phase and improvement phase. Construction phase use to create simple crew schedule table. Improvement phase will sort, change, or swap the crew table that came from construction by using greedy algorithm for reduce SD.

Construction phase composed of a set of crew C = {$C_1$, $C_2$,…, $C_i$}, a set of day D = {$D_1$, $D_2$, …, $D_{14}$}, and a set of pairing P = {$P_1$, $P_2$,…, $P_i$} which also have variables of start working date StD = {$StD_1$, $StD_2$,..., $StD_{14}$}. The state ST = {$ST_1$, $ST_2$,…, $ST_i$} was declare for crew schedule comparison. In this experiment, Thai Airways data incorporate rest time and number of working day into operation day already.

The method to reduce workload and perdiem is to create the normalized value by using equation (1). This value can be calculated by dividing workload with maximum workload and dividing perdeim with maximum perdiem and add them together. Then, we combine both values of workload and perdiem to be total normalized workload and perdiem ($T_{wp}$).

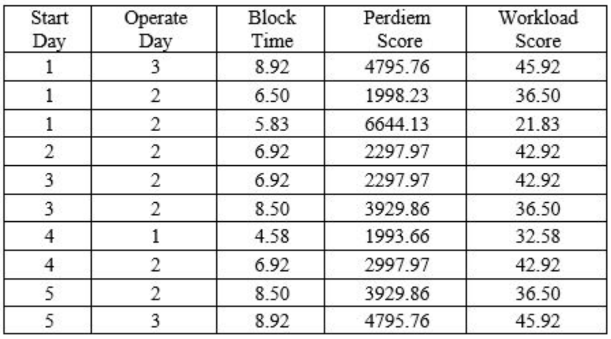
$$Twp_c = \frac{W_c}{Mw} + \frac{P_c}{Mp} \qquad (1)$$

Where *T* is the total normalized value of workload *w* and perdiem *p* at index number *c*. The *M* stand for maximum value.

This experiment also created workload and perdiem bound limitation. The process to limits workload bound occurred after we sorted the workload at first iteration, the result was limited by the value of minimum and maximum total workload for all of the next iterations. The iterations after first iteration will be sorted by perdiem.

The perdiem bound limitation is similar with workload bound limitation. The process to limits perdiem bound occurred after we sorted the perdiem at first iteration, the result was limited by the value of minimum and maximum total perdiem for all of the next iterations. The iterations after first iteration will be sorted by workload.

### *4.1 Construction Phase*



| Start Day | Operate Day | Block Time | Perdiem Score | Workload Score |
|---|---|---|---|---|
| 1 | 3 | 8.92 | 4795.76 | 45.92 |
| 1 | 2 | 6.50 | 1998.23 | 36.50 |
| 1 | 2 | 5.83 | 6644.13 | 21.83 |
| 2 | 2 | 6.92 | 2297.97 | 42.92 |
| 3 | 2 | 6.92 | 2297.97 | 42.92 |
| 3 | 2 | 8.50 | 3929.86 | 36.50 |
| 4 | 1 | 4.58 | 1993.66 | 32.58 |
| 4 | 2 | 6.92 | 2997.97 | 42.92 |
| 5 | 2 | 8.50 | 3929.86 | 36.50 |
| 5 | 3 | 8.92 | 4795.76 | 45.92 |

Figure 1: Simple test data

This phase use to create simple crew time table or crew scheduling with limitation of constraint. This phase still ignore SD. It will be revise again in next phase. The example test data shown in Figure 1. The steps of construction phase are presented as the following;

**Step 1:** Normalize workload and peridem.

**Step 2:** Check available day at crew $C_1$ to $C_i$ for execute pairing $P_{1.}$

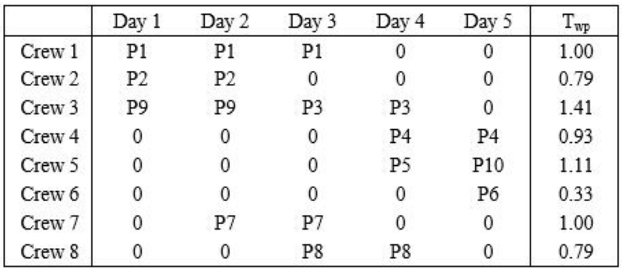
**Step 3:** If crew $C_1$ have available day for work then, assign $P_1$ to him from $D_1$ to $D_n$, where n is maximum operation day of $P_1$. The value of block time and $T_{wp}$ of $P_1$ also add to $C_1$ for calculate total block time and total $T_{wp}$ of each crew.

If not, the program will delete assigned pairing, block time, and $T_{wp}$ from $C_1$. Then, program will focus to next crew and execute step 2 again until $C_{i.}$

**Step 4:** Calculate SD of $T_{wp}$ to create initial SD for comparison in improvement phase.

**Step 5:** After $P_1$ was execute, it will process $P_2$ until $P_i$. Then the construction of crew scheduling or crew timetable is completed.

Figure 2, shows simple crew schedule result of construction phase which have SD about 14.0693. P represents pairing that each crew execute. For example, crew $C_1$ operates pairing $P_1$ which have operation task for three days, and execute pairing $P_5$ which have operation task only one day.



| | Day 1 | Day 2 | Day 3 | Day 4 | Day 5 | $T_{wp}$ |
|---|---|---|---|---|---|---|
| Crew 1 | P1 | P1 | P1 | 0 | 0 | 1.00 |
| Crew 2 | P2 | P2 | 0 | 0 | 0 | 0.79 |
| Crew 3 | P9 | P9 | P3 | P3 | 0 | 1.41 |
| Crew 4 | 0 | 0 | 0 | P4 | P4 | 0.93 |
| Crew 5 | 0 | 0 | 0 | P5 | P10 | 1.11 |
| Crew 6 | 0 | 0 | 0 | 0 | P6 | 0.33 |
| Crew 7 | 0 | P7 | P7 | 0 | 0 | 1.00 |
| Crew 8 | 0 | 0 | P8 | P8 | 0 | 0.79 |

Figure 2: Simple Construction table

*4.2 Improvement Phase*

The improvement phase is method to reduce SD of $T_{wp}$ by sort, change, and swap pairing in crew scheduling. This phase can divide into four main techniques such as;

4.2.1 Change pairing directly

4.2.2 Change pairing descending

4.2.3 Change pairing ascending

4.2.4 High workload and perdiem distribution

4.2.1 Change pairing directly

This method will change pairing of every crew in crew scheduling. The idea is to find and select minimum SD for every one step changing, which expected result will be minimum SD. This method will process from $P_1$ to $P_i$. Start from initial pairing to final pairing. The processes of this method are;

**Step 1:** Use crew schedule from construction phase as starter or initial solution. Find operation day that each pairing really executed.

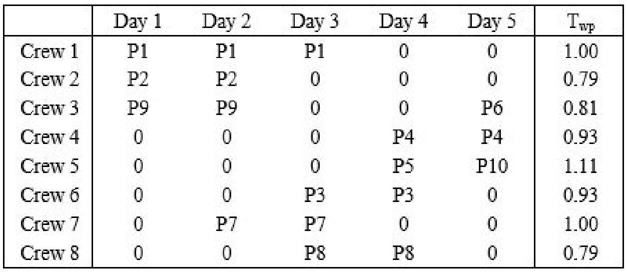
**Step 2:** Check available day of $C_{n+1}$, where n is id of crew that execute $P_1$ which have size equal or greater than operation day of $P_1$ and not have over total block time. This step also checks the bound limitation of workload and perdiem.

**Step 3:** If it available, copy initial table to $ST_x$, where ST is state, and x is number of pairing assignment successes. Remove pairing, block time, and $T_{wp}$ of $C_n$ and add to $C_{n+1}$. Calculate SD of each $ST_x$ table. Continue operate this step until $C_{n-1}$.

If not, the program will focus on $C_{n+2}$ and execute step 2 again until $C_{n-1}$.

**Step 4:** Compare SD of $T_{wp}$ from $ST_1$ until $ST_i$, then selected minimum $ST_x$ and set it to be initial solution then, execute step 2 again until $P_i$.

Figure 3, shows crew schedule after we applied the change pairing directly technique, which reduce SD to 1.0711.



| | Day 1 | Day 2 | Day 3 | Day 4 | Day 5 | $T_{wp}$ |
|---|---|---|---|---|---|---|
| Crew 1 | P1 | P1 | P1 | 0 | 0 | 1.00 |
| Crew 2 | P2 | P2 | 0 | 0 | 0 | 0.79 |
| Crew 3 | P9 | P9 | 0 | 0 | P6 | 0.81 |
| Crew 4 | 0 | 0 | 0 | P4 | P4 | 0.93 |
| Crew 5 | 0 | 0 | 0 | P5 | P10 | 1.11 |
| Crew 6 | 0 | 0 | P3 | P3 | 0 | 0.93 |
| Crew 7 | 0 | P7 | P7 | 0 | 0 | 1.00 |
| Crew 8 | 0 | 0 | P8 | P8 | 0 | 0.79 |

Figure 3: Change pairing table

4.2.2 Change pairing descending

This method is similarly to previous method. The idea is sort SD of $T_{wp}$ from largest to smallest first. Then, change pairing which have largest SD of $T_{wp}$ until smallest one. The process of this method are;

**Step 1:** Extend construction phase crew schedule to be initial solution. Find operation day that each pairing really executed.

**Step 2:** Applied bubble sort by sorting SD of workload and perdiem, depends on the bound limitation, from largest to smallest.

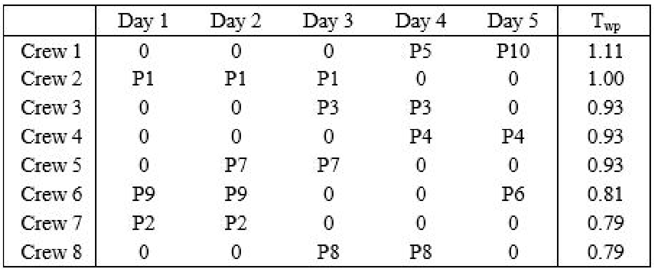
**Step 3:** Use sorted crew schedule. Then, check available day of $C_{n+1}$, where n is id of crew that execute $P_1$ which have size equal or greater than operation day of $P_1$ and not have over total block time. This step also checks the bound limitation of workload and perdiem.

**Step 4:** If available, copy initial table to $ST_x$, where x is number of pairing assignment successes. Remove pairing, block time, and $T_{wp}$ of $C_n$ and add to $C_{n+1}$. Calculate SD of each $ST_x$ table. Continue operate this step until $C_{n-1}$.

If not, the program will focus on $C_{n+2}$ and execute step 3 again until $C_{n-1}$.

**Step 5:** Compare SD of $T_{wp}$ from $ST_1$ until $ST_i$, then selected minimum $ST_x$ and set it to be initial solution then, execute step 3 again until $P_i$.

Figure 4, shows result from change pairing descending technique. The solution have similar result with pervious technique with equal SD but with different crew sequence.



| | Day 1 | Day 2 | Day 3 | Day 4 | Day 5 | $T_{wp}$ |
|---|---|---|---|---|---|---|
| Crew 1 | 0 | 0 | 0 | P5 | P10 | 1.11 |
| Crew 2 | P1 | P1 | P1 | 0 | 0 | 1.00 |
| Crew 3 | 0 | 0 | P3 | P3 | 0 | 0.93 |
| Crew 4 | 0 | 0 | 0 | P4 | P4 | 0.93 |
| Crew 5 | 0 | P7 | P7 | 0 | 0 | 0.93 |
| Crew 6 | P9 | P9 | 0 | 0 | P6 | 0.81 |
| Crew 7 | P2 | P2 | 0 | 0 | 0 | 0.79 |
| Crew 8 | 0 | 0 | P8 | P8 | 0 | 0.79 |

Figure 4: Change pairing descending table

4.2.3 Change pairing ascending

This method adjusted from previous method. We use ascending SD method, sort SD of $T_{wp}$ from smallest to largest. This also aim to observe and analyse different of both unsorted and descending sort effect. Thus, the processes are similarly to previous method such as;

**Step 1:** Extend construction phase crew schedule to be initial solution. Find operation day that each pairing really executed.

**Step 2:** Applied bubble sort by sorting SD of workload and perdiem, depends on the bound limitation, from largest to smallest.

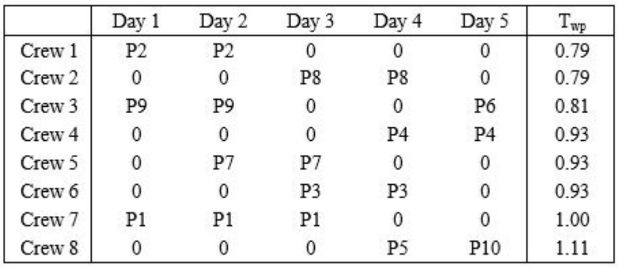
**Step 3:** Use sorted crew schedule. Then, check available day of $C_{n+1}$, where n is id of crew that execute $P_1$ which have size equal or greater than operation day of $P_1$ and not have over total block time. This step also checks the bound limitation of workload and perdiem.

**Step 4:** If available, copy initial table to $ST_x$, where x is number of pairing assignment successes. Remove pairing, block time, and $T_{wp}$ of $C_n$ and add to $C_{n+1}$. Calculate SD of each $ST_x$ table. Continue operate this step until $C_{n-1}$.

If not, the program will focus on $C_{n+2}$ and execute step 3 again until $C_{n-1}$.

**Step 5:** Compare SD of $T_{wp}$ from $ST_1$ until $ST_i$, then selected minimum $ST_x$ and set it to be initial solution then, execute step 3 again until $P_i$.

Figure 5, shows change pairing ascending technique. This technique have equal SD with both previous methods but different in crew sequence.



| | Day 1 | Day 2 | Day 3 | Day 4 | Day 5 | $T_{wp}$ |
|---|---|---|---|---|---|---|
| Crew 1 | P2 | P2 | 0 | 0 | 0 | 0.79 |
| Crew 2 | 0 | 0 | P8 | P8 | 0 | 0.79 |
| Crew 3 | P9 | P9 | 0 | 0 | P6 | 0.81 |
| Crew 4 | 0 | 0 | 0 | P4 | P4 | 0.93 |
| Crew 5 | 0 | P7 | P7 | 0 | 0 | 0.93 |
| Crew 6 | 0 | 0 | P3 | P3 | 0 | 0.93 |
| Crew 7 | P1 | P1 | P1 | 0 | 0 | 1.00 |
| Crew 8 | 0 | 0 | 0 | P5 | P10 | 1.11 |

Figure 5: Change pairing ascending table

4.2.4 High workload and perdiem distribution

In this section, the crew schedule are improved by disperse pairing that contributes to high number of $T_{wp}$. The idea is to disperse highest pairing of each high $T_{wp}$ crew to the lowest $T_{wp}$ crew. The process of this method are;

**Step 1:** Sort set of pairing P = {$P_1$, $P_2$,…, $P_i$} from construction phase descending by workload and perdiem, depends on the bound limitation.

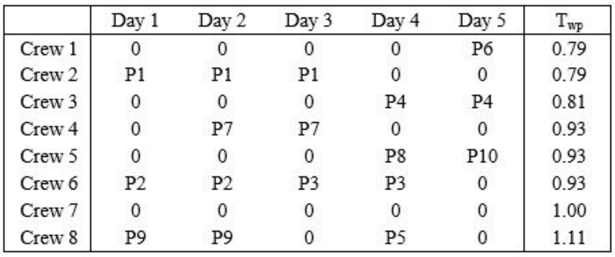
**Step 2:** Find highest pairing P in all set of crew C = {$C_1$, $C_2$,…, $C_i$}.

**Step 3:** Divided set of crew C from $C_1$ to $C_{n/2-1}$ and $C_{n/2}$ to $C_i$.

**Step 4:** Assign highest selected pairing $P_1$ to $P_i$ from smallest $T_{wp}$ crew $C_{n/2}$ to $C_i$. This pairing must not have value higher than block time period and result inside the bound limitation. Then, calculate SD of table.

**Step 5:** Compare every SD of table to find minimum $T_{wp}$ SD and set to be initial crew schedule. Then, execute step 4 again until crew $C_{n/2-1}$.

Figure 6, shows high workload and perdiem distribution technique. It yield higher SD than other technique about 7.6481.



| | Day 1 | Day 2 | Day 3 | Day 4 | Day 5 | $T_{wp}$ |
|---|---|---|---|---|---|---|
| Crew 1 | 0 | 0 | 0 | 0 | P6 | 0.79 |
| Crew 2 | P1 | P1 | P1 | 0 | 0 | 0.79 |
| Crew 3 | 0 | 0 | 0 | P4 | P4 | 0.81 |
| Crew 4 | 0 | P7 | P7 | 0 | 0 | 0.93 |
| Crew 5 | 0 | 0 | 0 | P8 | P10 | 0.93 |
| Crew 6 | P2 | P2 | P3 | P3 | 0 | 0.93 |
| Crew 7 | 0 | 0 | 0 | 0 | 0 | 1.00 |
| Crew 8 | P9 | P9 | 0 | P5 | 0 | 1.11 |

Figure 6: High Workload and Perdiem Distribution table

## 5. Testing Instances

Table 2 shows test instances that based on data from Thai Airways. These data can be divided into 3 categories; small, medium, and large according to number of pairing. These categories can also be divideded into sub-categories such as; short, long, and mix based on range of pairing. Short represents pairing that service around Thailand region, for example: Singapore, Manila, and Seoul. Mix represents pairing that service in regional Asia, Australia and New Zealand, and European region. Long represents pairing that service long range of distance such as; America, Australia and New Zealand, and European. The number of pairing and Inflight Manager (IM) or crew are also different.

Table 2: Test Instances.

| Instance | No. of Pair | No. of Inflight Manager (IM) | Duration of Pairs | Duration of Schedule |
|---|---|---|---|---|
| TA84S | 84 | 30 | Short haul | 14 Days |
| TA84M | 84 | 30 | Various haul | 14 Days |
| TA84L | 84 | 45 | Long haul | 14 Days |
| TA150S | 150 | 60 | Short haul | 14 Days |
| TA140M | 140 | 65 | Various haul | 14 Days |
| TA146L | 146 | 65 | Long haul | 14 Days |
| TA330S | 330 | 150 | Short haul | 14 Days |
| TA334M | 334 | 150 | Various haul | 14 Days |
| TA238L | 238 | 120 | Long haul | 14 Days |

From Table 2, number of pair represents task flight, number of Inflight Manager (IM) represent number of available crew, duration of pairs represent range of each flight, and duration of schedule represent number of experiment day. In addition, each pair have specific start day of flight and rest period. Thus, the different of pair, crew, start work date, perdiem, workload, and duration of flight in each instance make this problem difficult and complex to solve.

## 6. Research Result and Discussion

This experiment performs on Intel® Core™ i5-3210M CPU @ 2.50GHz 2.30GHz installed memory (RAM) 4.00 GB machine running under Windows environment. The solution coded in Microsoft Visual C++ 2010 Express.

The objective is to balance workload and perdiem by minimize Standard Deviation (SD). This test working under flight time limitation constraint for 14 days and rest time of 34 hours for 7 days. The result shows SD and percentage changed of $T_{wp}$ SD. The percentage changed comes from percent of improvement phase SD reduced from construction phase SD. The result of construction phase with workload bound limitation shown in the form of table (Table 3).

The result of construction phase yielded in high SD however, the aims is only constructs crew schedule or crew timetable.

For the Change pairing directly technique, this is good solution for TA146L, and TA330S which are 0.3353, and 0.4033 respectively. While, the average percent SD changed is 36.8269%. This is quite simple technique that change every pairing to every crew to get minimum SD.

For the Change pairing descending technique, this result has acceptable SD on TA84S, TA84M, TA146L, and TA334M which are 0.5865, 0.5358, 0.3353 and 0.4204 respectively. While, the average percent SD changed is 36.7604%. This method mostly yielded a good solution in various haul flight such as 84, and 334 pairing

For the Change pairing ascending technique, this yielded acceptable SD on TA84M, TA150S, TA140M, and TA146L instance which are: 0.5358, 0.3345, 0.4387, and 0.3353. While, the average percent SD changed is 37.5994%. Most of the good results came from medium number of pairing instances.

The high workload and perdiem distribution technique, yielded only two good solutions such as; TA84L, and TA238LL which are 0.4619, and 0.4347 respectively. The interesting of this result is it deal a minimum SD in mostly long haul flight instances. On the other hand, it produces many worst results such as; TA84S, TA84M, TA150S, TA140M, TA146L, TA330S, and TA334M. . While, the average percent SD changed is only 25.4294%.

From table 4, it shows the SD and SD percentage changed result of crew rostering that limited by perdiem bound. This experiment produced similar result with the previous workload bound limitation. The different occurred with the Change pairing descending technique and the Change pairing ascending technique. The Change pairing descending technique produced only two satisfied results such as: TA84S, and TA146L. The Change pairing ascending technique produced better result than previous method which is TA334M.

The result from improvement phase with workload and perdiem shows that the change pairing ascending technique can produced satisfied percentage SD changed which are 37.5994% for workload bound limitation and 37.0972% for perdiem bound limitation. The high percentage SD changed shows an effectiveness of workload and perdiem balancing.

Table 3: Crew Rostering by minimize SD of $T_{wp}$ with workload bound

| Instance | SD of Construction phase | SD and percent changed of each Improvement Phase with workload bound | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | SD Tech 1 | SD Changed Tech 1 (%) | SD Tech 2 | SD Changed Tech 2 (%) | SD Tech 3 | SD Changed Tech 3 (%) | SD Tech 4 | SD Changed Tech 4 (%) |
| TA84S | 1.3080 | 0.6555 | 49.8820 | 0.5865 | 55.1572 | 0.6476 | 50.4857 | 0.9960 | 23.8504 |
| TA84M | 1.6257 | 0.5431 | 66.5925 | 0.5358 | 67.0398 | 0.5358 | 67.0423 | 0.6460 | 60.2672 |
| TA84L | 0.6710 | 0.4883 | 27.2245 | 0.5153 | 23.1982 | 0.4883 | 27.2245 | 0.4619 | 31.1582 |
| TA150S | 0.7183 | 0.3770 | 47.5144 | 0.3806 | 47.0041 | 0.3345 | 53.4225 | 0.5166 | 28.0709 |
| TA140M | 0.7169 | 0.4397 | 38.6579 | 0.4404 | 38.5709 | 0.4387 | 38.7980 | 0.4678 | 34.7511 |
| TA146L | 0.4713 | 0.3353 | 28.8630 | 0.3353 | 28.8630 | 0.3353 | 28.8630 | 0.4424 | 6.1450 |
| TA330S | 0.5998 | 0.4033 | 32.7740 | 0.4174 | 30.4074 | 0.4040 | 32.6471 | 0.5202 | 13.2781 |
| TA334M | 0.6059 | 0.4244 | 29.9556 | 0.4204 | 30.6242 | 0.4245 | 29.9335 | 0.5116 | 15.5627 |
| TA238L | 0.5162 | 0.4647 | 9.9784 | 0.4647 | 9.9784 | 0.4647 | 9.9784 | 0.4347 | 15.7811 |
| AVG | 0.8037 | 0.4590 | 36.8269 | 0.4552 | 36.7604 | 0.4526 | 37.5994 | 0.5552 | 25.4294 |

Table 4: Crew Rostering by minimize SD of $T_{wp}$ with perdiem bound

| Instance | SD of Construction phase | SD and percent changed of each Improvement Phase with perdiem bound | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | SD Tech 1 | SD Changed Tech 1 (%) | SD Tech 2 | SD Changed Tech 2 (%) | SD Tech 3 | SD Changed Tech 3 (%) | SD Tech 4 | SD Changed Tech 4 (%) |
| TA84S | 1.3080 | 0.6329 | 51.6149 | 0.6027 | 53.9212 | 0.6461 | 50.6058 | 0.9899 | 24.3201 |
| TA84M | 1.6257 | 0.5682 | 65.0484 | 0.5682 | 65.0484 | 0.5347 | 67.1104 | 0.6297 | 61.2657 |
| TA84L | 0.6710 | 0.4883 | 27.2245 | 0.5108 | 23.8637 | 0.4883 | 27.2245 | 0.3928 | 41.4489 |
| TA150S | 0.7183 | 0.3903 | 45.6548 | 0.3751 | 47.7754 | 0.3640 | 49.3147 | 0.5044 | 29.7806 |
| TA140M | 0.7169 | 0.4397 | 38.6579 | 0.4418 | 38.3663 | 0.4379 | 38.9143 | 0.4664 | 34.9469 |
| TA146L | 0.4713 | 0.3353 | 28.8630 | 0.3353 | 28.8630 | 0.3353 | 28.8630 | 0.4328 | 8.1690 |
| TA330S | 0.5998 | 0.4044 | 32.5826 | 0.4176 | 30.3864 | 0.4091 | 31.7912 | 0.4918 | 18.0156 |
| TA334M | 0.6059 | 0.4266 | 29.5976 | 0.4250 | 29.8507 | 0.4237 | 30.0722 | 0.5074 | 16.2597 |
| TA238L | 0.5162 | 0.4647 | 9.9784 | 0.4647 | 9.9784 | 0.4647 | 9.9784 | 0.4195 | 18.7296 |
| AVG | 0.8037 | 0.4612 | 36.5802 | 0.4601 | 36.4504 | 0.4560 | 37.0972 | 0.5372 | 28.1040 |

## 7. Conclusion

This paper use greedy algorithm to solve crew rostering problem for Thai Airways. The objective of this experiment is to minimize Standard Deviation (SD) of $T_{wp}$ in order to balance the workload and perdiem. The paper proposed two phases of crew rostering; Construction, and Improvement phase. Construction phase is the simple phase that used to assign pairing to each crew. Improvement phase used to adjust pairing for workload and perdiem balancing. The experiment tested on two different bound limitation: workload, and perdiem. The result shown in table of SD of $T_{wp}$ and percentage SD changed from SD of construction phase. Due to nature of greedy algorithm, it can find only optimal solution at some cases, and not always guarantee global solution. Consequently the result yielded high SD. Even though, the result are high SD but the technique required less compilation time than other method. Thus, for further work, we will continue to refine the improvement method with other techniques, in order to achieved optimal solution.

## Acknowledgements

This paper was partially supported by Logistics and Supply Chain Systems Engineering (LogEn) Research Unit, Sirindhorn International Institute of Technology, Thammasat University.

## References

[1] Bazarga M. Airline Operations and Scheduling. Burlington; Ashgate Publishing Ltd; 2008.

[2] Michel G, Alain H, Jerome O. A graph coloring model for a feasibility problem in monthly crew scheduling with preferential bidding. Computers & Operations Research. 2007; 34: 2384-2395.

[3] Dusan T, Panta L. Metaheuristics approach to the aircrew rostering problem. Annals of Operations Research. 2007; 155: 311-338.

[4] Broos M, Mario V. A hybrid scatter search heuristic for personalized crew rostering in the airline industry. European Journal of Operational Research. 2010; 206: 155-167.

[5] Zhang Y, Rao Y, Zhou M. GASA Hybrid Algorithm Applied in Airline Crew Rostering System. Tsinghua Science and Technology. 2007; 225-259.

[6] Nadia S, Jacques T. Genetic algorithm based approach for the integrated airline crew-pairing and rostering problem. European Journal of Operational Research. 2009; 199: 674-683.

[7] Walid E, Carlos A, Felix M. An Intelligent Approach for Solving the Airlines Crew Rostering Problem. IEEE. 2001; 73-79.

[8] Azadeh A, Hosseinabadi M, Eviazy H, Nazari S, Asadipour G. A hybrid meta-heuristics algorithm for optimization of crew scheduling. Applied Soft Computing. 2013; 158-164.

[9] Guang-Feng D, Woo-Tsong L. Ant colony optimization-based algorithm for airline crew scheduling problem. Expert Systems with Applications. 2011; 38: 5787-5793.

[10] Wagner G, Nicolau G. Heuristics to solve the integrated airline crew assignment problem. The Journal of Transport Literature. 2015; 9.

[11] Yindong S, Kunkun P, Kai C, Jingpeng L. Evolutionary crew scheduling with adaptive chromosomes. Transportation Research Part B. 2013; 56: 174-185.

[12] Herbert D, Johannes K, Christine S. An enhanced rostering model for airline crews. Computers & Operations Research. 2001; 28: 671-688.

[13] Anne M, Francois S. An integrated aircraft routing, crew scheduling and flight retiming model. Computer & Operations Research. 2007; 34: 2251-2265.

[14] Claude P, Nidhi S. Airline crew scheduling from planning to operations. European Journal of Operational Research. 2007; 183: 1013-1027.

[15] Jenny D, Jose I, Federico T. Aircraft maintenance, routing, and crew scheduling planning for airlines with a single fleet and a single maintenance and crew base. Computers & Industrial Engineering. 2014; 75: 68-78.

[16] Dusan T, Panta L. A fuzzy set theory approach to the aircrew rostering problem. Fuzzy sets and systems. 1996; 95: 261-271.

[17] Ernesto M, Joao M. An AI-Based Approach to Crew Scheduling. IEEE. 1993; 71-77.

[18] Budi S, Andiek S, Arief R. Airline Crew Scheduling using the Differential Evolution Method. Applied Mathematics. 2010; 316-325.

[19] Shangyao Y, Yu-Ping T. A network model for airline cabin crew scheduling. European Journal of Operational Research. 2002; 140: 531-540.

[20] Imma R, Ramon C, Xavier T. An iterated greedy algorithm for the flowshop scheduling problem with blocking. Omega. 2011; 39:293-301.

[21] Nearchou, A.C., Giannikos, I.C., Lagodimos A.G. Efficient greedy algorithms for economic manpower shift planning. Taylor & Francis; 2013.

# Modified Differential Evolution for Permutation Flow Shop Scheduling

Tawun Remsungnen[1]*Pornpan Hemtong[2] and Pannarat Guayjarernpanishk[3]
[1]*Faculty of Applied Science and Engineering, NongKhai Campus, KhonKaen University,*
*NongKhai, 43000, Thailand, rtawun@kku.ac.th*
[2]*Department of Mathematics, Faculty of Science KhonKaen University,*
*KhonKaen, 40002, Thailand, Pronpanhem13@gmail.com*
[3]*Faculty of Applied Science and Engineering, NongKhai Campus, KhonKaen University,*
*NongKhai, 43000, Thailand, Pannarat7881123@gmail.com*

## Abstract

The differential evolution with position order (PODE) algorithm is modified for the permutation flow-shop scheduling. The changes are made to the mutation and crossover operators while the selection operator is applied immediately for each individual population of generation. This new proposed method is denoted by MPODE algorithm. The standard Taillard's benchmark which is composed by several sizes of instances i.e. from 20 jobs with 5 machines to 500 jobs with 20 machines, is used in the numerical experiment for the comparison between PODE and MPODE methods. The results show clearly that the proposed MPODE method is better than the original PODE method.

*Keywords*: differential evolution, evolution operators, flow shop scheduling, numerical experiment

*Corresponding Author
E-mail Address: rtawun@kku.ac.th

## 1. Introduction

The permutation flowshop scheduling problemhas been worldwide studied with varieties of objective functions andthe optimization techniques, due to its roles in many industrial applications. The regular problem composed by two main sets,i.e. a set of $m$ machines and a set of $n$ jobs thatmust be processed by the same ordering of machines. At a time,each job can be processed on only one machine and each machine can process only one job. The operations are not allowed preemptable. The problem is to find the best sequences of $n$ jobs to be processed on the same order of $m$ machines in order to get the optimal objectives. One of the desired objective functions is the makespanor completion time ($C_{max}$). The problem with minimizing of$C_{max}$ on two machines can be solved exactly by using the Johnson's rule, while the problem with more than two machines has been proved to be strongly NP-complete [1-3].

Themetaheuristics method which is an iterative generation process by applying the combinations of smart concepts and strategies in order to search the optimal or near optimal solutions in the feasible space.There are several metaheuristics, e.g. genetic algorithm (GA), simulated annealing, (SA), tabu search (TS) algorithm and differential evolution (DE) algorithm [4-7]. The DE algorithm is one of the evolutionary computational methodsthat had outstanding performance in the first contest of IEEE evolutionary computational. In principle, this method will generate the next solutions from the combination of previous solutions using the evolution-like operators, which are crossover, mutation and selection operators. However, the DE is first developed for the real space problem. The applying of DE for the permutation space problem needs the modification of these operators.

In this study, the DE method with an additional operator that is called position ordering (PO) which is applied for traveling salesman problem is modified and applied for the flowshop scheduling problem. The makespan is used as the objective functions.

## 2. Methods

### *2.1The PODE method*

The DE method with the position ordering(PO) operator[8], denoted here as PODE (Fig. 1),is described as following steps:

i) Set the parameters NP, NG, Fac and Crfor the population size, the number of generation, the mutation factor and the crossover rate, respectively, for the problem of $n$ jobs.

ii) Randomly generate the individuals of start generation $g=0$which are vectors of solution, $X_{j,i,g} = (x_{1,i,g}, x_{2,i,g},\ldots, x_{n,i,g})$, where $j$ and $i$ indicate the $j^{\text{th}}$ job and an $i^{\text{th}}$ individual, respectively.

iii) Mutation operator: generate the mutation vectors$V_{i,g}$ from the $X_{i,g}$ when $i = 1,2,\ldots,n$ using the equation (1).

$$V_{i,g} = X_{r1,g} + \text{Fac}(X_{r2,g} - X_{r3,g}) \qquad (1)$$

The index $r1$, $r2$ and $r3$ is uniform random from $\{1,2,\ldots,n\}\backslash\{i\}$, respectively.

iv) Crossover operator: generate the crossover vectors $U_{j,i,g} = (u_{1,i,g},u_{2,i,g},\ldots,u_{n,i,g})$ from the vectors of $V_{i,g}$ and $X_{i,g}$ as equation (2).

$$u_{j,i,g}=\begin{cases} v_{j,i,g}\text{ , if rand }(0,1)\leq \text{Cr or } j=j_{rand} \\ x_{j,i,g}\text{, otherwise}\end{cases} \qquad (2)$$

The rand(0,1)is a uniform random variable on [0,1] and $j_{rand}$is a random integer from 1 to *n* and renew again for every *i*. The obtained vectors $U_{j,i,g}$ now are the sequences of mixed integer and real numbers and are not in the space of solutions. The position ordering operator is added in this step for converting the mixed sequences into the permutation sequences. Every members of $U_{j,i,g}$ are first increasing sorted, thenreplace each member by its position index in the sorted sequence.

v) Selection operator: select the better solution between $X_{j,i,g}=(x_{1,i,g},x_{2,i,g},\ldots,x_{n,i,g})$ and $U_{j,i,g}=(u_{1,i,g},u_{2,i,g},\ldots,u_{n,i,g})$ to be the new $i^{\text{th}}$ individual in the next g+1 generation as equation (3).

$$X_{i,g+1}=\begin{cases} U_{i,g}\text{,if } f(U_{i,g})\leq f(X_{i,g}) \\ X_{i,g}\text{, otherwise}\end{cases} \qquad (3)$$

vi) Repeat: set g = g+1 then go back to step iii) until *g=NG*. The best solutions are the best that found in the NG generation.

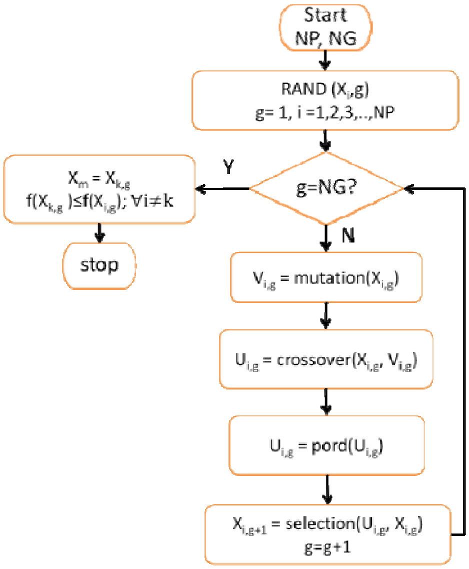


Figure 1: Flowchart of the PODE method.

### 2.2The proposed MPODE method

The equation (1) in the PODE methodgeneratesvectors $V_{i,g}$ which are the random sequences of mixedreal and integernumbers from the previous three random solutions of $X_{j,i,g}$ which later are the source of vectors $U_{i,g}$. After the ordering of vectors $U_{i,g}$ by PO operator, the obtained ordered $U_{i,g}$ may not close to any solution of previous $X_{j,i,g}$even all those three of its parent. For the permutation space problems, this make the DE method has high efficiency in the global searching but may not for the local searching. In this study, the modification of PODE method is proposed and is denoted by MPODE (Fig. 2) in order to increase the efficiency of the local searching. First, the equation (1) is modified to be as an equation (4),

$$V_{i,g}=X_{s1,g}+\text{Fac}(X_{s2,g}-X_{s3,g}) \qquad (4)$$

where, $X_{s1,g}$, $X_{s2,g}$ and $X_{s3,g}$ are order after sorting of the three random $X_{r1,g}$, $X_{r2,g}$ and $X_{r3,g}$ form the best to worst. This makes the mutation operator generates the $V_{i,g}$ more closer to the best of its three parents, $X_{s1,g}$. The PO operator is now applied on $V_{i,g}$ instead of $U_{i,g}$ as in step (iii), of PODE. Then the crossover operator in step (iv) is replaced by a new crossover criteria. First, the number of positions, *nX* in $U_{i,g}$ that are set to be as same as $X_{i,g}$, obtained by, $nX = n - max(\lfloor Cr\times n\rfloor, 1)$. Then the *nX* positions *j* of $U_{i,g}$ are random and set $u_{j,i,g}=x_{j,i,g}$. The rest positions of $U_{i,g}$ are filled with the members of $V_{i,g}$ that are not yet in $U_{i,g}$. Then,the obtained $U_{i,g}$has already been permutation sequences which most of its members are in the same position as $X_{i,g}$ and the rest members are in the same order as $V_{i,g}$.

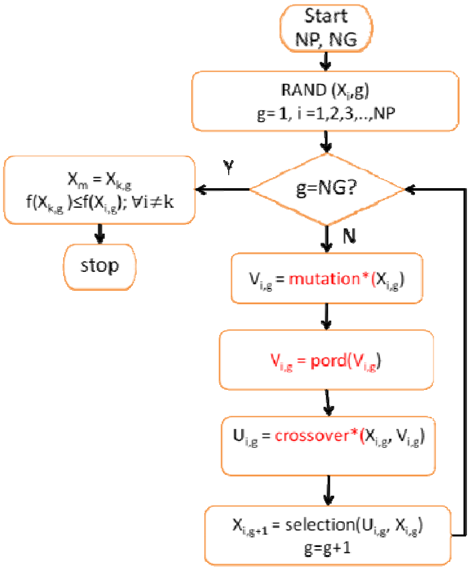


Figure 2: Flowchart of the MPODE method.

### 2.3The objective function, $C_{max}$

The makespan or the complete time, $C_{max}$, is the amount of time that needed to finish the last job on the last machine. All machines had assumed to start at the same time that when the first job is start processed on the first machines. If a time $t_{ij}$ is given for the amount of time that a machine *i* needs for process a job *j*, the $T_{ij}$ is the time that job *j*finished on machine *i*(Fig. 3).

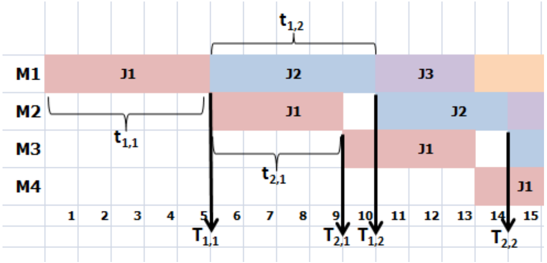


Figure 3: Grant's chart shows the meaning of $t_{ij}$ and $T_{ij}$ .

We will have,
$T_{1,1} = t_{1,1}$,
$T_{1,j} = T_{1,j-1}+t_{1,j}$,
$T_{i,1} = T_{i-1,1}+t_{i,1}$,
$T_{i,j} = max(T_{i-1,j} , T_{i,j-1},)+ t_{i,j}$.

Then the $C_{max}$ is set to$T_{m,n}$.

### *2.4The numerical experiments*

The PODE and MPODE methods are coded with the freeware SCILAB [9].The standard benchmark problemsof Taillard[10] which are composed by several sizes of instances i.e. from 20 jobs with 5 machines to 500 jobs with 20 machines are used in the numerical experiment. Since the main aim of this study is to compare the efficiency between PODE and MPODE methods, the reported best solutions of the instances are not considered here.

Two experiments are setup and are called short time test (STT) and long time test (LTT). In the STT, one of 20 jobs with 5 machines and one of 50 jobs with 10 machines instances of the benchmark are selected. The population sizes parameters, NP are set to the number of jobs ($n$) in each instance and the number of generation, NG are set to 50, 100 and 200. There are 10 iterations applied for each numbers of NG. While in the LTT, all instances of the benchmark are used. The parameters NG and NP are set to 2$n$ and 200$n$, respectively, except for the 500 jobs with 20 machines, the NG is set to 50000. There is only one iteration is applied for each instances. The average percentage of derivation (APD) is proposed to measure the different of the obtained solutions of PODE and MPODE from their best one as equation (5),

$$\mathrm{APD} = \frac{1}{\mathrm{R}}\sum_{\mathrm{r=1}}^{\mathrm{R}}(100 \times \frac{\mathrm{S_{r,M}-S_b}}{\mathrm{S_b}}) \qquad (5)$$

where R=10, is the number of iterations in the STT, and is the number of instances in each problem sizes of the benchmark in LTT. The $S_{r,M}$ refer to the best/average solutions that obtained in each run ($r$) of PODE or MPODE methods, and the $S_b$ is the best obtained solution of all $R$ run.

## 3. Results

All APDs for STT and LTT experiments are summarized in Table 1 and Table 2, respectively. In the STT, the overall best solution of the 20×5 instance is obtained by PODE method but just only one of ten times.

Table 1: The APDs in the STT experiment.

| Problems | | PODE | | MPODE | |
|---|---|---|---|---|---|
| $n$×$m$ | NG | Best | Average | Best | Average |
| 20×5 | 50 | 1.35 | 1.17 | 1.02 | 0.56 |
| | 100 | 2.65 | 1.27 | 1.90 | 0.39 |
| | 200 | 1.88 | 1.10 | 0.78 | 0.35 |
| 50×10 | 50 | 1.69 | 1.13 | 0.64 | 0.17 |
| | 100 | 2.47 | 1.71 | 1.32 | 0.30 |
| | 200 | 2.18 | 2.04 | 0.76 | 0.37 |

In overall of STT experiment, the APDs shown that the MPODE method is better than the PODE method for both Best and Average solutions, i.e. the MPODE method is better than the PODE method in both points of local and global searching.

Table 2: The APDs in the LTT experiment.

| Problems | PODE | | MPODE | |
|---|---|---|---|---|
| $n$×$m$ | Best | Average | Best | Average |
| 20×5 | 0.65 | 0.44 | 0.02 | 0.08 |
| 20×10 | 0.22 | 0.43 | 0.53 | 0.10 |
| 20×20 | 0.45 | 0.29 | 0.40 | 1.24 |
| 50×5 | 0.69 | 0.92 | 0.05 | 0.00 |
| 50×10 | 2.13 | 0.95 | 0.06 | 0.00 |
| 50×20 | 0.93 | 1.02 | 0.08 | 0.00 |
| 100×5 | 0.50 | 0.80 | 0.06 | 0.00 |
| 100×10 | 1.42 | 1.48 | 0.00 | 0.00 |
| 100×20 | 0.63 | 0.71 | 0.08 | 0.00 |
| 200×10 | 1.46 | 1.50 | 0.01 | 0.00 |
| 200×20 | 6.90 | 6.96 | 0.00 | 0.00 |
| 500×20 | 0.14 | 0.30 | 0.07 | 0.00 |
| Average | 1.34 | 1.32 | 0.11 | 0.12 |

In the LTT experiment, just only one problem size of 20×20 instances, that the Average ADP of PODE method is better than those of MPODE method. By the way, in overall problem sizes, the MPODE method is again better than the PODE method, especially for the big size problems.

## 3. Conclusion

Since most of the meta-heuristic methods are based on the random searching of solutions in the feasible space, the short time running, i.e. small number of random solutions may not give enough information of the efficiency. However, in this study, both short time and long time experiments show very clearly that the proposed MPODE method is much better than the PODE method. By the way, the finding of best solutions for each instances in order to comparing with other heuristic methods are needed.

**Acknowledgements**

This research is supported by the Higher Education Research Promotion and National Research University

Project of Thailand, Office of the Higher Education Commission, through the Cluster of Research to Enhance the Quality of Basic Education.

**References**

[1] Jonhson SM. Optimal two and three-stage production schedule with setup times included. Naval Research Logistics Quarterly. 1954;1(1):61–68.

[2] Palmer DS. Sequencing jobs through a multistage process in the minimum total time: a quick method of obtaining a near optimum. Operations Research.1965;16:101–107.

[3] Garey MR, Jonhson DD, Sethi R. The complexity of flowshop and jobshop scheduling. Mathematics of Operations Research, 1976;1:117–129.

[4] Osman IH, Potts CN.Simulated Annealing for Permutation Flow-shop Scheduling. OMEGA, 1989; 17; 551–557.

[5] Srikanth K, et al. Improved genetic algorithm for the permutation andflowshop scheduling problem. Computers and Operations Research 2004; 31: 593–606.

[6] Nowicki E, Smutnicki C.A fast tabu search algorithm for the permutation flow-shop problem.European Journal of Operational Research.1996; 91: 160–175.

[7] Storn R, Price K. Differential evolution Ca simple and efficient heuristic for global optimization over continuous spaces. Journal of global optimization, 1997;11:341–359.

[8] Xiang W, Guoyi X. Hybrid Differential Evolution Algorithm for Traveling Salesman Problem. Procedia Engineering, 2011;15:2716 – 2720.

[9] ENTERPRISES, Scilab, et al. Scilab: Free and Open Source software for numerical computation. Scilab Enterprises, Orsay, France, 2012, 3.

[10] Taillard E. Benchmarks for basic scheduling problems. European Journal of Operational Research.1993;64:278–285.

# Using R statistical software for visualizing the epidemiological data of Dengue disease in Chiang Mai, Thailand

Veerasak Punyapornwithaya[1*], Sukolrat Boonyayatra[1] and Peninnah Oberdorfer[2]

*[1]Faculty of Veterinary Medicine, Chiang Mai University, Chiang Mai, Thailand,*
*veerasak.p@cmu.ac.th*
*[1]Faculty of Veterinary Medicine, Chiang Mai University, Chiang Mai, Thailand,*
*bsukolrat@yahoo.com*
*[2] Department of Pediatrics, Faculty of Medicine, Chiang Mai University, Chiang Mai, Thailand,*
*aoberdor@med.cmu.ac.th*

**Abstract**

Visualization is a powerful mechanism for extracting information from data. Effective visualization provides users in analyzing and reasoning about data and evidence. The aim of this study was to use information graphics to visualize the epidemiological data for Dengue disease in Chiang Mai, Thailand. Data consisted of number of Dengue cases in human from 25 districts in Chiang Mai serially collected daily for 6 years. R statistical software (version 3.1.2) with *ggplot2*, *openair, lattice* and *ggmap* package were used for creating informative graphs including Cleveland dot plots, calendar heatmap, trellis graphs and geographical map. The Cleveland dot plots produced by *ggplot2* package efficiently presented the number of cases by district provided the overall epidemiological information of the disease that could be used by public health personals. The calendar heatmap from *openair* package and calendarHeat function presented an understandably information suitable for general people to recognize the “high risk” period of disease outbreak whereas the spatial information produced by *ggmap* package suggested the “hot spot” area for the disease. A trellis system of *lattice* created multiple graphs for the number of cases per day for each 25 districts each year at the same time powerfully. All graphs and maps created by R and selected packages in this study provided a dynamic information for a different level of users to understand the epidemiology of Dengue disease. In conclusion, the using R and packages created understandable statistical graphics in an efficient, time-saving, elegant and systematic manner.

*Keywords*: data visualization, R statistical software, Dengue, epidemiology

*Corresponding Author
E-mail Address: veerasak.p@cmu.ac.th

## 1. Introduction

Basically, the epidemiological data is typically summarized using numerous quantitative presentations such as frequencies and percentage in many cells in table. However, there is a limitation of using table format to communicate with audiences. For instance, trend of disease, relationship between disease and risk factor, and visualization of hierarchical data are difficult to perceive from table data. Alternatively, graphs can depict complex relationships not easily presented by table format. In many instances, graphs can reveal result more precisely than conventional descriptive summarize statistics [1].

R statistical software [2], an open source computer program, has a great capability to produce a high quality display of graph. With some extraordinary R packages such as *ggplot2* [3] and *lattice* [4], to produce simple, complex or even very complex graph with high quality is very efficient.

The aim of this study was to use R and some package to visualize the epidemiological data for Dengue disease in Chiang Mai, Thailand using R statistical software.

## 2. Research Methodology

### *2.1 Dengue data set*

Dengue dataset from Chiang Mai Provincial public Health Office were used. Data consists of hospital patient ID, patient address, sick date, diagnostic date, hospital name where patients were diagnosed, patient gender and others information collected for 6 years from 2008-2013 from 25 districts in Chiang Mai. Originally, data were kept as Microsoft Excel file. The dataset has 25,910 rows and 18 columns.

### *2.2 Data management*

Data were imported to R using *read.table* function. The type of variables such as date, string and integer were specified by R before doing the further steps.

We use *dplyr* package [5] to summarize 1) the number of Dengue cases in 2013 by districts, 2) the number of Dengue cases for each month in 2013 by districts and 3) the total number of Dengue cases each day from 2008-2013 aggregated from all districts.

### *2.2 Data visualization*

A dot plot was created by *ggplot2* package to present the frequency of Dengue cases in 2013 from 25 districts. The *lattice* package was use for making a multi-panel graphs showing a line graph presenting the number of monthly Dengue cases in 2013 for each district. The *CalendarHeat* function [6] was used to produce calendar heatmap showing the number of Dengue cases daily from 2008 to 2013. Also, the calendar heatmap for the number of Dengue cases in 2013 was produced by *openair* package [7]. The *ggmap* package [8] was used to create a map showing the geographical objects and the number of Dengue cases in 2013 for Muang, Mae Rim, San Sai and Hang Dong district. The maps for such districts were downloaded from google cooperation via internet.

## 3. Research Results and Discussion

Normally, the epidemiological data were presented as table and some graphs such as bar graph, pie graph and line graph. However, such graphs had its advantage and disadvantage. In this study, we created various graphs and map as alternative tools to visualize Dengue disease data using R and some packages.

For epidemiology data, bar graph is generally used for presenting prevalence data as the x-axis was the label of categorical data and y-axis was the number of cases. However, the Cleveland dot plot should be used alternatively. Cleveland dot plot is a two-dimension graphical display of objects. Basically, the horizontal axis of dot plot is a scale covering the range of quantitative values to be shown whereas the vertical axis shows the descriptive labels associated with each of quantitative values [9]. For Dengue data, bar graph can be used to display the data. However, there is a limitation for bar graph to be displayed on a commonly used document size like A4 paper. With a list of the 25 district name labelled on x-axis, a bar graph could not well fit the paper. In contrast, the dot plot shown in Figure 1 fitted well in the paper and also it provides the same information as bar graph. In addition, the dot plot (figure 1) is very easy to interpret. The label on y-axis showed the name of district and each point show the frequency of cases. The farther to the right, the higher number of cases; the farther to the left, the lower number of cases.

Because Dengue cases was recorded daily for each district therefore the trend of disease could be displayed for each year which is useful for comparing the incidence of disease by place and time. There are two possibilities to show the number of cases monthly by districts. One can create a single graph showing all trends of Dengue cases from 25 districts or create 25 graphs to present the trend for each district. The common spreadsheet software offers the capability for making a single line graph as the x-axis is the sequential of calendar month, y-axis is a range value of cases and each line shows the trend for each district. However, such graph is hard to read because there are 25 lines will be shown on a single graph. Alternatively, data can be presented for each 25 districts by producing each graph for each district. Unfortunately, it is a time consuming for making a graph for each district because many software typically produces a single graph instead of simultaneously plotting multiple graphs. Nevertheless, with trellis system from *lattice*, multiple graphs can be automatically employed via multi-panel conditioning.

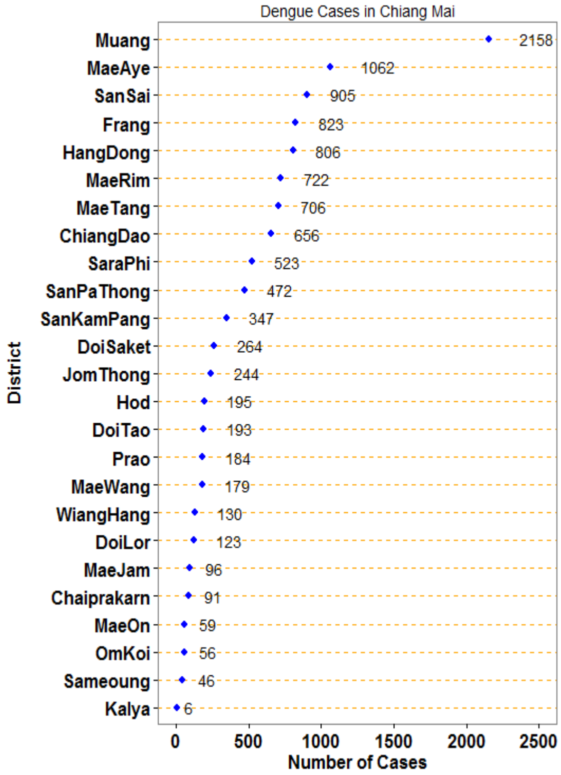


Figure 1: Cleveland dot plot showed the number of Dengue cases in Chiang Mai by districts

Figure 2 illustrated the number of cases by time, conditional on district. This figure offered a powerful visualization for public health professionals to compare the disease by time and place. Although a large number of graphs are necessary to be presented, it is not difficult to handle because the number of graph by row, column and page can be specified to fit the presentation medium. Moreover, Trellis Graphics has an excellent mechanism for the details of rendering graphs such as line type, line color, plotting symbols, aspect ratio and scales. It was note that, for demonstration in this study, we adjusted the y-axis scale not exceed 200 therefore data will not show on the graph if y-value is greater 200.

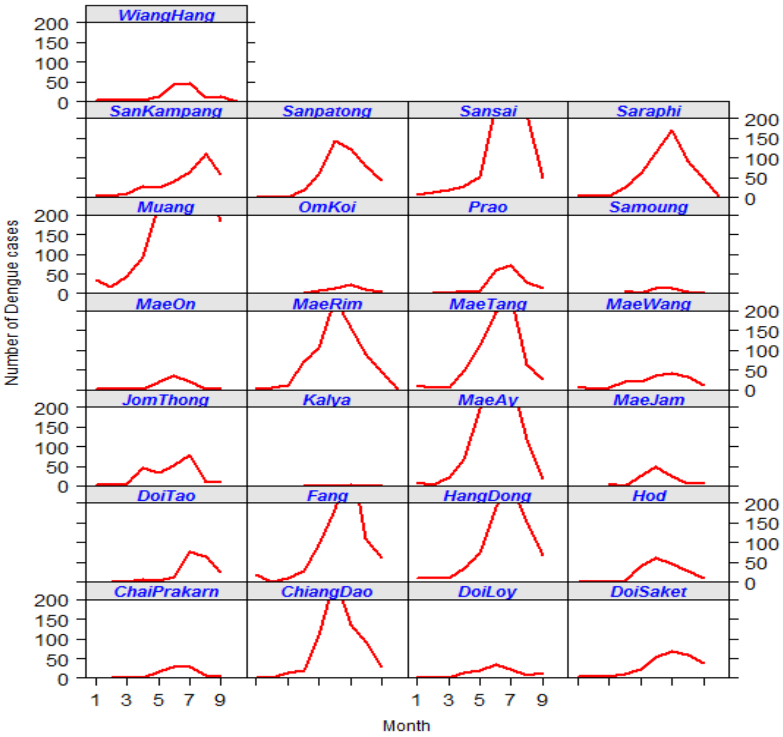


Figure 2: Multiple graphs simultaneously produced by trellis system

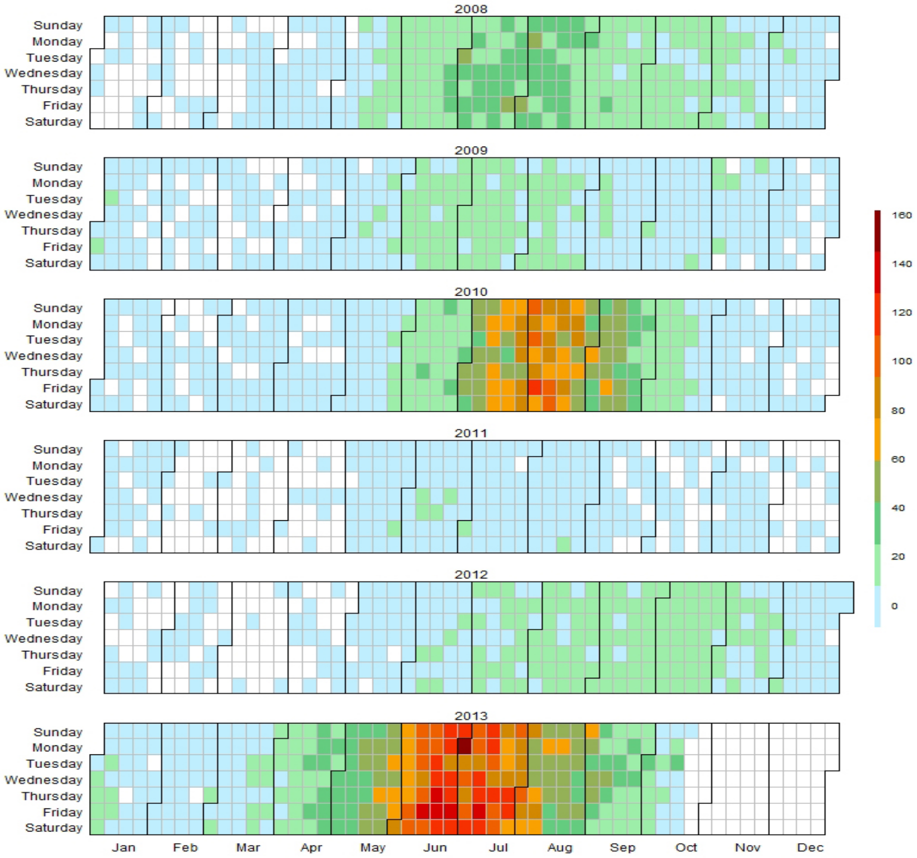


Figure 3: Calendar heat map for Dengue cases for 6 year period

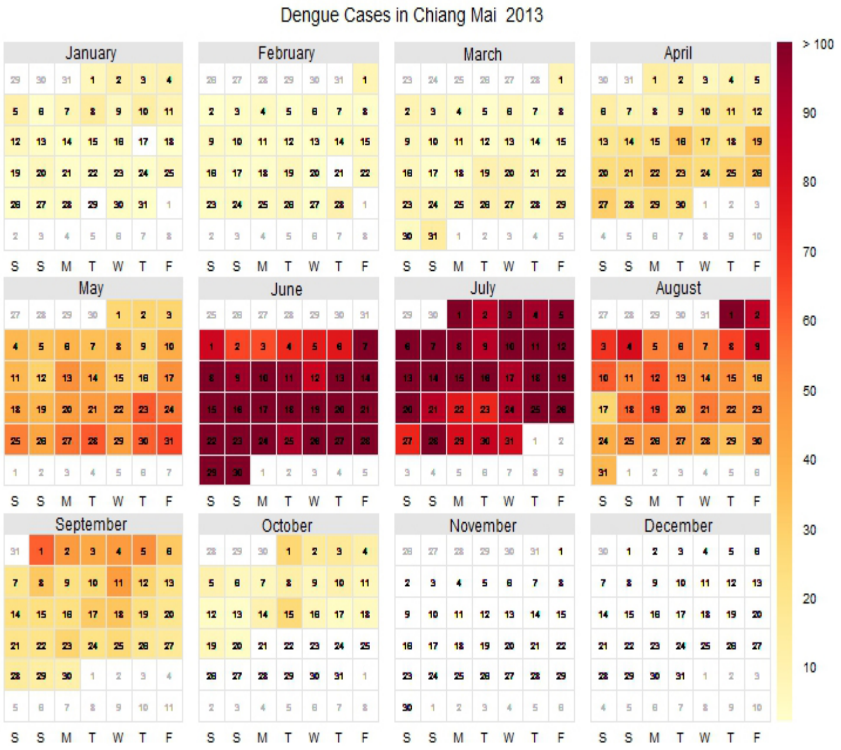


Figure 4: Calendar heat map for Dengue cases in 2013

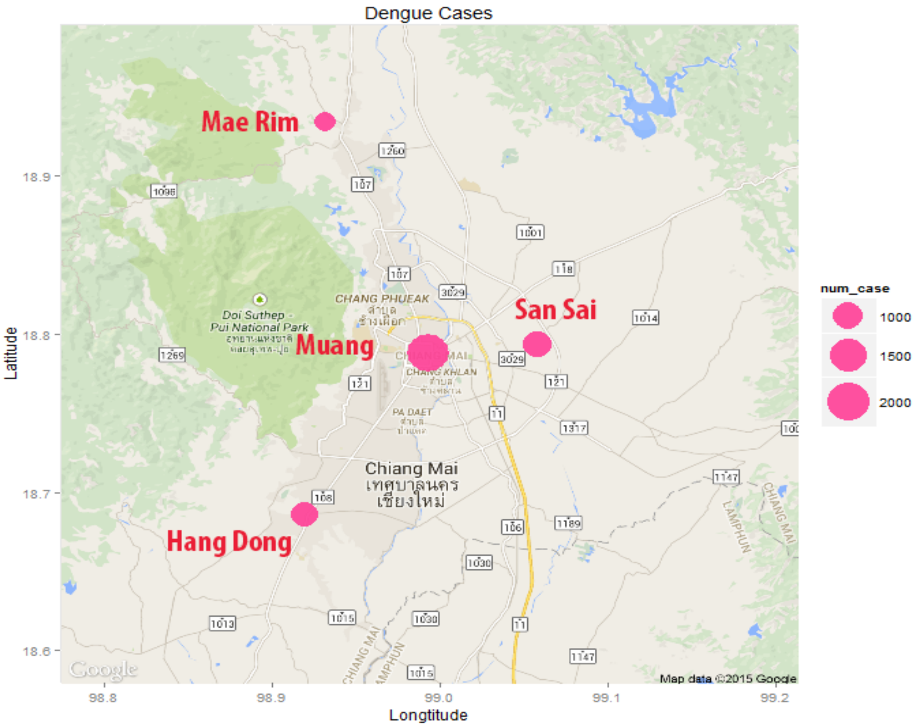


Figure 5: The map with number of Dengue cases in 4 districts

Calendar heatmaps are an interesting alternative view of time series data. The quantitative value is displayed as color mapped cells in a calendar [10]. The magnitude of value can be presented by using different shades of grays or different color schemes. We used R with *calendarHeat* function to create calendar heatmap for Dengue cases from 2008 -2013 and with *openair* to produce Dengue cases in 2013. The calendar heatmaps of Dengue data provides an efficient communication to people because people generally familiar with calendar and color. The calendar heat map for Dengue data is very understandable. People can recognize that the darker color represents the high number of cases and the lighter color reflects to the low number of cases. In other words, people will know about when is the “high risk” period and “low risk” period. This map might be used for disease surveillance activities such as the communication among medical doctors, health care professionals and people who are at risk for Dengue disease.

Generally, epidemiological data is a collection of event, time and place. For instance, investigators would like to explain the outbreak of disease to answer the question about place that disease occurs, time related with disease outbreak and the number of infected patient. To describe disease by places, one of the most powerful presentations is disease mapping. It provides beneficial information about places especially for the places that the audiences never have an experience. R and various package related with spatial analysis able to show epidemiological data on map. We produce a very simple graph to show number of Dengue cases in 2013 for some district on a map (Figure 5) using *ggmap* package. The map can be freely downloaded from various providers such as *Google*, *OpenStreetMap* and *stamen*. Once the specific map is already downloaded, we used functions from *ggmap* to connect the geo-coordinate from text file with the map and add a new layer of circle with a different size to proportionally present the number of cases for each geo-location. The map provided the important information about the location and the number of patients. For example, people can simply perceive that the “hot spot” area for Dengue disease is Muang district by looking at the size of circle. The map also give an overall information about place. Comparing to create graphs shown in the study, to make a graph is easier because a few commands are required and the programming skill is not necessary. The user can use command scripts as a template to produce map for other disease and other areas.

The disadvantage of using R is that R software is difficult to learn compared to other statistical package because R is command based software. An R user has to learn and practice for building the programing knowledge and skills.

## 4. Conclusion

Epidemiological data for Dengue disease were summarized and visualized as graphs and a map suitable for disease communication in term of cases, time and place using R and some packages. Conclusively, R statistical software has a powerful capability to create understandable statistical graphics in an efficient, time-saving, elegant and systematic manner.

## 5. Acknowledgements

We would like to thank Chiang Mai Provincial public Health Office, Ministry of Public Health for allowing us to use the data. This study was funded by Faculty of Medicine, Chiang Mai University.

## 6. References

[1] Tufte ER. Visual Display of Quantitative Information. Cheshire, CT: Graphic Press; 1983

[2] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [Internet]. 2015[cited 2015 Apr 14]. http://www.R-project.org/.

[3] Wickham H. ggplot2: Elegant graphics for data analysis. Springer, New York, 2009.

[4] Sarkar D. Lattice: multivariate data visualization with R. Springer, New York, 2008

[5] Wickham H, Francois R. (2015). dplyr: A Grammar of Data Manipulation. R package version 0.4.1 [Internet]. 2015[cited 2015 Apr 14]. http://CRAN.R-project.org/package=dplyr

[6] Bleicher, P. CalendarHeat: An R function to display time-series data as a calendar heatmap [Internet]. 2015 [cited 2015 Apr 17]. Available from:https://github.com/jbryer/makeR/blob/master/R/calendarHeat.R

[7] Carslaw DC, Ropkins K. openair: an R package for air quality data analysis. Environmental modelling & software. 2012; 27-28:52-61

[8] Kahle D, Wickham H. ggmap: Spatal visualization with ggplot2. The R Journal. 2013; 5:144-162

[9] Cleveland W. Graphical Methods for Data Presentation: Full Scale Breaks, Dot Charts, and Multibased Logging. The American Statistician. 1984; 38(4) 270-280

[10]Ohri A. R for business analystics. Springer, New York, 2012

# An alternative attribute selection to construct decision trees

W. Wongrin[1], P. Thongchan[1], W. Bodhisuwan[1]* and L. Saothayanun[2]

[1]*Department of Statistics, Kasetsart University, Chatuchak, Bangkok, 10903, Thailand,*

[2]*Department of Statistics, University of the Thai Chamber of Commerce, Dindang, Bangkok, 10400, Thailand,*

*e-mail: weerinradaj@gmail.com, panu.tho@hotmail.com,fsciwnb@ku.ac.th, luckhana sao@utcc.ac.th*

## Abstract

High dimensionality of the attribute space in big data is one of the most important concerns in decision tree problems due to number of attributes and the accuracy considerations. To classify data, the suitable attributes and the number of branches in tree is important. In early, a traditional criterion which is an information gain, is used to select attributes and create decision trees. The maximal information coefficient has been a novelty measuring associations for $21^{st}$ century. This study is focused on the use of the maximal information coefficient to split data to select suitable attributes so as to construct decision trees. Four data sets are considered to build decision trees with information gain and the maximal information coefficient. In summary, we found that the maximal information coefficient can be used as an alternative criterion for building a decision tree.

*Keywords:* Correlation, Decision trees, Information theory, Maximal information coefficient

*Corresponding author.
E-mail address: fsciwnb@ku.ac.th

## 1 Introduction

One of the most popular classification technique is the decision trees technique. Decision trees are systematic to build classification model with learning algorithm to identify the best relationship model between the attributes (variables) set and the label from input data. The structure of decision tree is consisting of node and direct edge. The tree has three types of nodes: root node, internal node and leaf (terminal) nodes. The important process of decision tree is find out attribute for the best separate split data. In general, algorithm provides a method for specifying the criterion for each attributes as well as objective measure for selecting the best split information. These measures are defined in terms of the class information of before and after splitting. The attribute selection criterion has three types in general.

$$\text{Gain} = \text{Entropy[before split]-Entropy[after split]}, \tag{1}$$

$$\text{Gini index} = 1 - \sum_{t=0}^{n}[p(x_i)]^2, \tag{2}$$

$$\text{Classification error} = 1 - \max[p(x_i)], \tag{3}$$

where $p(x_i)$ is the probability of $X$.

As in decision tree learning, iterative Dichotomiser 3 (ID3) is an algorithm invented by Quinlan in 1986 [1], used to generate a decision tree from a data set. It begins with the original set $S$ as the root node. On each iteration of the algorithm, it iterates through every unused attribute of the set $S$ based on the entropy and information gain to select the attribute.

When using ID3 algorithm, a data set may contain noises that lead to construction of an inaccurate decision tree; this is a known weakness associated with decision trees [2]. To address this problem, we propose to use a new attribute selection criterion that is called Maximum information coefficient (MIC).

The MIC is a nonparametric method introduced by Reshef *et al.* [3]. The method has strength for measuring linear, non-linear and undiscovered relationship between variables. The MIC also detects the correlation between variables when the data have noise [4]. The concept of MIC is partitioned to grid and measure max information between variables. In this paper, we show how to apply the MIC in the particular case of ID3 algorithm.

## 2 Information theory

This section provides information theory and new measuring association of dependence variables.

Information theory is a study of an amount of random variables or collection of random variables. In statistical sense, we get information from observing the occurrence of an event link to probability theory. In 1948, Shannon introduced the concept of information being an uncertainty [5]. Shannon's information measures refer to entropy and mutually information that are information measure.

### 2.1 Entropy and mutually information

Entropy is an uncertainty measure of a random variable. Let $X$ is a random variable in a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ as

$$H(X) = -\sum_{i=1}^{n_x} p(x_i) \log_2 p(x_i), \tag{4}$$

and joint entropy is

$$H(X,Y) = -\sum_{i=1}^{n_x} p(x_i, y_i) \log_2 p(x_i, y_i). \tag{5}$$

Mutually information is the measure of information between one random variable to another defined by

$$I(X,Y) = H(X) + H(Y) - H(X,Y). \tag{6}$$

### 2.2 Maximal information coefficient

Consider two random variables, $X$ and $Y$. Suppose that we want to know how much variable $X$ tells us about $Y$ or how much $X$ relates to $Y$. To answer this question, we could compute the correlation coefficients, but there are limitations. For instance, the random variables must be defined in real-valued and should be a linear function.

In 2011, Reshef *et al.* [3] introduced a new measuring association of two variables: maximal information coefficient (MIC). It is based on the idea that if a relationship exists between two variables, then a grid can be drawn on the scatterplot of the two variables that partitions the data to cover that relationship. The statistics MIC is obtained by normalising mutually information (MI) into range of [0,1]. Next, we will briefly introduce some concepts of the MIC.

To compute the MIC, let us define a finite set $D$ of pair $(X, Y)$. Reshef *et al.* [3] considered one of the dimensions as $x$-values and the other as $y$-values. They called $x$-by-$y$ grid as the partition of the pair $(X, Y)$ in $x$ bins for the variable $X$ and $y$ bins for the variable $Y$.

Let $D_{|G}$ be the distribution of $D$ which is divided by one of $x$-by-$y$ grids as $G$, and $I(G)$ $(I(X,Y))$ is the mutually information. The highest mutual information score for each grid is normalized by $\log_2 \min\{x, y\}$ and stored in the characteristic matrix [6--8]. The MIC is defined by

$$MIC \;=\; \max\left\{\frac{I^*(X,Y)}{\log_2 \min\{x,y\}}\right\}, \tag{7}$$

where $I^*(X,Y) = \max\{I(X,Y)\}$, and $M(X,Y) = \frac{I^*(x,y)}{\log_2 \min\{x,y\}}$ is called the characteristic matrix.

The MIC can be used to discover relationships between two random variables in a simple way measure, whereas the correlation coefficients cannot measure. For this reason, the MIC has been called ``A correlation of the $21^{st}$ century" [9, 10].

It is used to be a novel measuring association in many fields such as gene expression, global indicators from the world health organization and bacterial species in the gut microbiota of Mice [3]. It is also applied to many algorithms for attribute selection such as attributes clustering [6] and model selection [7, 11].

#### 2.2.1 Some properties of maximal information coefficient

Some properties of the MIC was proven by Reshef *et al.* [3] including:

- MIC assigns a perfect score of 1 to all never-constant noiseless functional relationships

- MIC assigns scores that tend to 1 for a larger class of noiseless relationships (including superpositions of noiseless functional relationships)

- MIC assigns a score of 0 to statistically independent variables

For a pair of random variables $X$ and $Y$, (i) if $Y$ is a function of $X$ that is not constant on any open interval, then data drawn from $(X, Y)$ will receive an MIC tending to 1 with probability one as sample size grows; (ii) if the support of $(X, Y)$ is described by a finite union of differentiable curves of the form $c(t) = (x(t), y(t))$ for $t$ in [0,1], then data drawn from $(X, Y)$ will receive an MIC tending to 1 with probability one as sample size grows, provided that $dx/dt$ and $dy/dt$ are each zero on finitely many points; (iii) the MIC of data drawn from $(X, Y)$ converges to zero in probability as sample size grows if and only if $X$ and $Y$ are statistically independent.

## 3 Decision tree

A decision tree is used as classifiers on the raw data sets to predict the classes of the data. It can be constructed by many algorithms. An important algorithm for machine learning is ID3. The ID3 was developed with the purpose of producing knowledge from a limited training set; it builds a decision tree through inductive steps.

Throughout the ID3 algorithm, the decision tree is constructed with each non-terminal node representing the selected attribute on which the data was split, and terminal nodes representing the class label of the final subset of this branch.

In ID3, information gain can be calculated for each remaining attribute. The attribute with the largest information gain is used to split the set S on this iteration.

### 3.1 Selection attribute with MIC

For this study, we apply the MIC being a novel measuring association between two variables to construct decision trees. The MIC is applied to the decision tree with ID3 algorithm. The data set is split by the highest MIC score, and the node is created using package `data.tree` [12] in R Programming Language [13]. The MIC value is calculated by using package `minerva` [14] in R Language.

**Attribute selection algorithm**

```
Input: a data set
Output: Selection attributes
1. for all pairs of attributes
       Calculate the MIC values of every attribute
       in data set
   end for
2. Sort the values of MIC in descending order
3. Split the set $\mathcal{D}$ into subsets using the attribute
   for which the MIC value is maximum
4. Make a decision tree node containing
   that attribute
5. Recur on subsets using remaining attributes.
```

## 4 Evaluating model

The model efficiency for classification data can be considered by the accuracy index. It is considered like an error from hypothesis testing. For this study, the confusion matrix is used to present the error from prediction. The table of confusion matrix is shown in Table 1.

**Table 1:** The confusion matrix

| | | Actual | |
|---|---|---|---|
| | | True | False |
| Prediction | True | TP | FP |
| | False | FN | TN |

Where

- TP is true positive being the number of positive correct prediction
- FP is false positive being the number of positive incorrect prediction
- FN is false negative being the number of negative incorrect prediction
- TN is false negative being the number of negative correct prediction

The accuracy index can be obtained by

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \qquad (8)$$

## 5 Real data

In this study, the data sets considered to classify with decision trees in study have been used for many machine learning studies.

### 5.1 Mushroom data set

When we want to decide the mushroom, the target classification is ``which is toxic mushroom?" which can be a *poisonous* or *edible*. The attributes that can be classify the mushroom are 22 attributes such as shape, size, points, and colour. The data set can obtain from UCI machine learning repository which is a center for machine learning and intelligent systems.

### 5.2 Iris data set

Iris data is a data set from the University of California Irvine (UCI). It consists of 150 iris flowers and 50 each from one of three iris species: sentosa, versivolour and virginica each flower have four characteristics: sepal length, sepal width, petal length and petal width. The data set can be obtained from package `data.tree`.

### 5.3 Voting data set

This data set came from 1984 United Stated Congressional Voting Records. It includes votes for each of the U.S. House of Representatives Congressmen on the 16 attribute votes identified by the Congressional Quarterly Almanac (CQA). The data has been classified as Republican or Democrat. The data set was presented in UCI machine learning repository which is a center for machine learning and intelligent systems.

### 5.4 Playing baseball data set

When we want to decide whether the weather is amenable to playing baseball. Over the course of 2 weeks, data was collected to build a decision tree. The target classification is ``should we play baseball?" which can be *yes* or *no*. The weather attributes are outlook, temperature, humidity, and wind speed. Baseball data can be received from package `data.tree`.

## 6 Results

This section presents the results of classification for four real data sets. The classifications are built from IG and MIC criterions. Its results are shown in Tables 2- 4.

The examples of confusion matrices will be shown in Table 2 and 3.

**Table 2:** The confusion matrix from IG for iris data set

| Iris | Raw data | | |
|---|---|---|---|
| Predict | setosa | versicolor | virginica |
| setosa | 50 | 50 | 50 |
| versicolor | 0 | 0 | 0 |
| virginica | 0 | 0 | 0 |

**Table 3:** The confusion matrix from MIC for iris data set

| Iris | Raw data | | |
|---|---|---|---|
| Predict | setosa | versicolor | virginica |
| setosa | 50 | 50 | 50 |
| versicolor | 0 | 0 | 0 |
| virginica | 0 | 0 | 0 |

**Table 4:** The accuracy classification with four data sets

| Data sets | Accuracy | |
|---|---|---|
| | IG (branches) | MIC (branches) |
| Mushroom | 70% (32) | 70% (32) |
| Iris | 33% (133) | 33% (100) |
| Voting | 97% (31) | 97% (31) |
| Base ball | 100% (12) | 100% (12) |

The results show that the accuracy of decision tree predictions from IG and MIC criterions have the same accuracy index with high accuracy. For iris data set, the accuracy index is not high because iris data is a data set that is difficult to classify by any methods. From MIC criterions, the decision tree has 100 branches, while IG has 133 branches. In decision trees, we do not want too many the number of branches because overfitting problem will be concerned with the number of branches.

## 7 Conclusions and discussions

For selection of attributes or random variables, it is based on correlation of attributes which is a traditional topic. However, a novelty measuring association between dependent attributes gives the concept of new correlation meanings. The aim of this study is in order to select suitable attributes to construct a decision tree. The idea is followed by the iterative Dichotomiser 3 (ID3) of which information gain is a powerful measuring association to split data.

For different data sets, the effectiveness of the maximal information coefficient is investigated and compared against information gain. The criterions are an accuracy index and the number of branches.

Maximal information coefficient shows the same accuracy of prediction, but the number of branches based on the maximal information coefficient may be less than those of information gain. Therefore, the maximal information coefficient seems to be much better alternatives for selection a wide variety of relationships that attributes might exhibit, and can be as useful as original tools.

## Acknowledgement

The authors would like to thank the anonymous reviewers for the careful reading and valuable comments. This work was partially supported by Department of Statistics, Faculty of Science, Kasetsart University and Graduate School Kasetsart University.

## References

[1] Quinlan JR. Induction of Decision Trees. Mach Learn. 1986 Mar;1(1):81--106.

[2] Eusebi C, Gliga C, John D, Maisonave A. Data Mining on a Mushroom Database. 2. Proceedings of Student-Faculty Research Day, CSIS, Pace University; 2008. p. B2.1--B2.9.

[3] Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, et al. Detecting Novel Associations in Large Data Sets. 2011;334(6062):1518--1524.

[4] Reshef DN, Reshef YA, Mitzenmacher M, Sabeti PC. Cleaning up the record on the maximal information coefficient and equitability. 2014;111(33):E3362--E3363.

[5] Shannon CE. A Mathematical Theory of Communication. Bell System Technical Journal; 27(3).

[6] Zhao X, Deng W, Shi Y. Feature Selection with Attributes Clustering by Maximal Information Coefficient. Procedia Computer Science. 2013;17(0): 70--79. First International Conference on Information Technology and Quantitative Management.

[7] Qiuheng T, Hangjin J, Yiming D. Model selection method based on maximal information coefficient of residuals. Acta Mathematica Scientia. 2014;34(2):579 -- 592.

[8] Kinney JB, Atwal GS. Equitability, mutual information, and the maximal information coefficient. 2014;111(9):3354--3359.

[9] Speed T. A Correlation for the 21st Century. 2011;334(6062):1502--1503.

[10] Murphy KP. Machine Learning: A Probabilistic Perspective. The MIT Press; 2012.

[11] Hu G, Wang J, Feng W. In: Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing 2012. vol. 443 of Studies in Computational Intelligence; 2013. .

[12] Glur C. data.tree: Hierarchical Data Structures; 2015. R package version 0.1.6.

[13] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2015.

[14] Filosi M, Visintainer R, Albanese D. minerva: minerva: Maximal Information-Based Nonparametric Exploration R package for Variable Analysis; 2014. R package version 1.4.1.

# Statistical analysis for the effectiveness of tuberculosis tests

Shishen Xie[1*] and Yingda L. Xie[2]

[1]*Department of Mathematics and Statistics, University of Houston-Downtown, Houston, Texas 77002, USA, xies@uhd.edu*
[2]*Institute of Allergy and Infectious Diseases, National Institute of Health, Bethesda, Maryland 20892, USA, yingda.xie@nih.gov*

**Abstract**

Tuberculosis (TB) is a critical human health concern. Approximately one third of the world's population has latent tuberculosis, which becomes transmissible active tuberculosis in approximately 10% of those infected over a lifetime. Effective testing is important to discover and treat individuals who carry the latent infection before it turns into the active disease. Currently, the two primary methods available to diagnose latent TB infection are the tuberculin skin test (TST) and the newer Interferon-Gamma Release Assays (IGRA). Both of these tests are based on assessing immune response to TB, but do not always agree. In this paper, clinical data from healthcare workers in a US hospital system is used to analyze the agreement and correlation of the IGRA test in comparison with those of the tuberculin skin test (TST) when used in serial testing. Data analysis is performed to assess the concordance between IGRA and TST results at baseline, as well as among conversions and reversions in IGRA and TST.

*Keywords*: latent tuberculosis infection; interferon gamma release assay; data analysis; correlation

*Corresponding Author
E-mail Address: xies@uhd.edu

## 1. Introduction

Tuberculosis is a critical human health concern. It is estimated by the World Health Organization (WHO) that approximately 9 million people developed tuberculosis in 2013 alone, and 1.34 million people died as a result of this disease [1]. However, in most people, the mycobacteria that causes tuberculosis (*M.tb*) is contained by the host immune response and remains latent. During the lifetime of about 10% of these latently infected individuals, *M.tb* is able to overcome the immune response and lead to highly transmissible and often debilitating active disease. According to estimations, approximately one third of the world's population has latent tuberculosis. One of the biggest hurdles is finding and treating individuals who carry the latent bacteria before it is able to turn into an active disease. The two widely used tests to diagnose latent tuberculosis are the tuberculin skin test (TST) and interferon-gamma release assays (IGRA). In contrast to the much older TST, where *M.tb* proteins are injected directly into the person to stimulate a measurable immune response, IGRA involves stimulating and measuring an immune response from a blood sample using more specific *M.tb* antigen. Countries including the US have implemented both tests in many settings to diagnose latent TB infection and have started to face programmatic challenges with reproducibility and discordancy of these tests [2]. These challenges can lead to clinical dilemmas of whether to provide the several month treatment courses for TB. In this study, employee health data among health-care workers at a large US urban medical center will be statistically analyzed to further characterize these operational limitations in a non-TB endemic population where both tests are used. The data set consists of each employee's identification number, and the corresponding test results of TST performed before 2007, IGRA performed between 2007 and March 2009, and IGRA performed after March 2009. Specifically, the concordance between IGRA and TST will be investigated, as well as their variations on serial testing.

## 2. Data from TST and IGRA

First, we calculated the number of employees who took each test as well as the number of employees who retested. In order to make our calculations, we needed to determine the number of duplicate positives, negatives and indeterminate. We then consolidated the data using SAS based on the specified range. Once the data set range was specified, we were able to write code to combine the duplicates.

There were 2330 health care workers who took the TST as shown in Figure 1. Among them, 201(8.63%) workers tested positive and 2129 (91.37%) tested negative.

There are 3000 health care workers who took the Interferon-Gamma Release Assays between 2007 and March 2009 (labeled as IGRA_1). It is shown in Figure 2 that among the 3000 workers, 93 (3.10%) tested positive, 2828 (94.27%) tested negative, and 79 (2.63%) workers' tests were indeterminate.

2839 workers took IGRA after March 2009 (labeled as IGRA_2). Figure 3 shows that among them, 175 (6.16%) workers tested positive, 2610 (91.93%) tested negative, and 54 (1.90%) workers' tests were indeterminate.

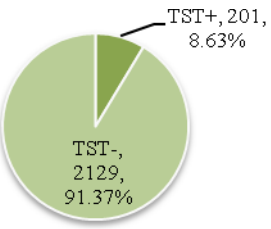


Figure 1. Data of Tuberculin Skin Test (TST)

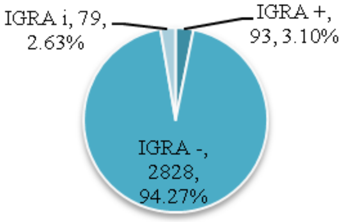


Figure 2. Data of Interferon-Gamma Release Assays 2007 – 3/2009 (IGRA_1)

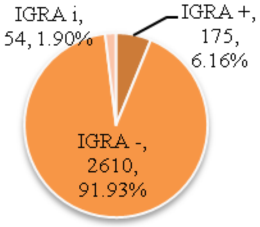


Figure 3. Data of Interferon-Gamma Release Assays (IGRA_2) post 3/2009

## 3. Data Analysis for Reproducibility

40 health care workers had the serial TST test performed. Table 1 shows that among them, 11 (27.50%) changed from positive to negative (reversion) and 29 (72.50%) changed from N (negative) to P (positive) (conversion). This means that all (100%) of the 40 workers who were retested had different test results and none of them kept the same results. The reproducibility of TST is very poor.

Table 1: Tuberculin Skin Test (TST) Retest

| Frequency<br>Percent<br>Row Pct<br>Col Pct | Table of First_TST by Second_TST | | | |
|---|---|---|---|---|
| | | Second_TST | | |
| | First_TST | N | P | Total |
| | N | 0<br>0.00<br>0.00<br>0.00 | 29<br>72.50<br>100.00<br>100.00 | 29<br>72.50 |
| | P | 11<br>27.50<br>100.00<br>100.00 | 0<br>0.00<br>0.00<br>0.00 | 11<br>27.50 |
| | Total | 11<br>27.50 | 29<br>72.50 | 40<br>100.00 |

The reproducibility of IGRA is also poor. There were 93 health workers who had the serial IGRA (2007-3/2009) testing. Table 2 shows that among them, all 16 who tested positive the first time were negative on the retest (reversion). Meanwhile, among the 38 workers who first tested negative, 22 became positive (conversion) and 16 became indeterminate (I) on the retest. This means that all workers who were retested had different test results and none of them kept the same results. The reproducibility of IGRA is very poor. Furthermore, there were 39 indeterminate cases (41.94% of all retested cases) in the first test, of whom 1 became positive and the rest became negative on retest.

Table 2: IGRA_1 (2007-3/2009) Retest

| Frequency<br>Percent<br>Row Pct<br>Col Pct | Table of First_IGRA_1 by Second_IGRA_1 | | | | |
|---|---|---|---|---|---|
| | First_<br>IGRA_<br>1 | Second_IGRA_1 | | | |
| | | I | N | P | Total |
| | I | 0<br>0.00<br>0.00<br>0.00 | 38<br>40.86<br>97.44<br>70.37 | 1<br>1.08<br>2.56<br>4.35 | 39<br>41.94 |
| | N | 16<br>17.20<br>42.11<br>100.00 | 0<br>0.00<br>0.00<br>0.00 | 22<br>23.66<br>57.89<br>95.65 | 38<br>40.86 |
| | P | 0<br>0.00<br>0.00<br>0.00 | 16<br>17.20<br>100.00<br>29.63 | 0<br>0.00<br>0.00<br>0.00 | 16<br>17.20 |
| | Total | 16<br>17.20 | 54<br>58.06 | 23<br>24.73 | 93<br>100.00 |

A Chi-square test of association is used to determine if there is a correlation (association) between the results (P or N) of the first and second IGRA tests. In other words, is it possible to predict the results (P or N) of the second IGRA test from those of the first IGRA test? Null hypothesis $H_0$ in this Chi-square test is that there is an association between the results of the first and second IGRA tests, while the alternative hypothesis ($H_a$) is that there is no association between the results of the first and second IGRA tests.

Table 3: Statistics for Table of First_IGRA_1 by Second_IGRA_1

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 4 | 89.0847 | <.0001 |
| Likelihood Ratio Chi-Square | 4 | 118.2686 | <.0001 |
| Mantel-Haenszel Chi-Square | 1 | 0.0259 | 0.8721 |
| Phi Coefficient | | 0.9787 | |

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Contingency Coefficient | | 0.6995 | |
| Cramer's V | | 0.6921 | |
| WARNING: 22% of the cells have expected counts less than 5. Chi-Square may not be a valid test. | | | |

The small *P* value in Table 3 indicates that the null hypothesis should be rejected and there is no association between the test results (P or N) of the first and second IGRA (2007 to 3/2009) test. Of note, due to the low percentage of the cells that have expected counts less than 5 the Chi-square test may not be valid.

There are 85 health workers who took the serial IGRA (post 3/2009) test (Table 4). Among them, all 39 workers with an initial positive test had a negative result on the retest (reversion). Meanwhile, 19 employees with initial negative test became positive on the retest (conversion) while the remaining 4 were indeterminate on retest. This means that all 85 health care workers who had done the retest had different test results and none of them kept the same results. The reproducibility of IGRA (post 3/2009) is very poor. Furthermore, there were 27 indeterminate cases (31.76% of all retested cases) in the first test, and 1 of them became positive and the rest (26) became negative on the retest.

Table 4: Number of Employees who were retested IGRA_2 (post 3/2009)

| Frequency<br>Percent<br>Row Pct<br>Col Pct | Table of First_IGRA_2 by Second_IGRA_2 | | | | |
|---|---|---|---|---|---|
| | First_IGRA_2 | Second_IGRA_2 | | | |
| | | I | N | P | Total |
| | I | 0<br>0.00<br>0.00<br>0.00 | 26<br>30.59<br>96.30<br>40.00 | 1<br>1.18<br>3.70<br>6.25 | 27<br>31.76 |
| | N | 4<br>4.71<br>21.05<br>100.00 | 0<br>0.00<br>0.00<br>0.00 | 15<br>17.56<br>78.95<br>93.75 | 19<br>22.35 |
| | P | 0<br>0.00<br>0.00<br>0.00 | 39<br>45.88<br>100.00<br>60.00 | 0<br>0.00<br>0.00<br>0.00 | 39<br>45.88 |
| | Total | 4<br>4.71 | 65<br>76.47 | 16<br>18.82 | 85<br>100.00 |

Table 5: Statistics for Table of First_IGRA_2 by Second_IGRA_2

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 4 | 79.7434 | <.0001 |
| Likelihood Ratio Chi-Square | 4 | 84.6562 | <.0001 |
| Mantel-Haenszel Chi-Square | 1 | 0.5179 | 0.4717 |
| Phi Coefficient | | 0.9686 | |
| Contingency Coefficient | | 0.6957 | |
| Cramer's V | | 0.6849 | |
| WARNING: 22% of the cells have expected counts less than 5. Chi-Square may not be a valid test. | | | |

The *P* value in the chi-square test in Table 5 is also small. This indicates that the null hypothesis should be rejected and there is no association between the test results (P or N) of the first and second IGRA (post 3/2009) test. Again the warning is noticed that chi-square may not be a valid test.

## 4. Analysis for Concordance between TST and IGRA

For the analysis of the concordance between TST and IGRA, the Kappa correlation coefficient is used. Cohen's Kappa statistic, *κ*, can be used to assess the agreement between alternative methods (TST and IGRA in this case) of categorical assessment (P or N) when a technique is under study [18, 19]. In order to find *κ* the formula: $\kappa = \frac{\Pr(a)-\Pr(e)}{1-\Pr(e)}$ is used with Pr(*a*) being the observed percentage of agreement between the methods and Pr(*e*) the expected percentage of agreement. The observed percentage of agreement Pr(*a*) implies the proportion of ratings where the assessments agree, and the expected percentage Pr(*e*) is the proportion of agreements that are expected to occur by chance as a random result. The Kappa correlation coefficient *κ* can be interpreted in the following scale:

| $\kappa$ Range | Interpretation |
|---|---|
| if $\kappa \le 0$ | No agreement |
| if $\kappa \le 0.20$ | Poor agreement |
| if $0.20 < \kappa \le 0.40$ | Fair agreement |
| if $0.40 < \kappa \le 0.60$ | Moderate agreement |
| if $0.60 < \kappa \le 0.80$ | Good agreement |

There were a total of 2100 health workers who had both TST and IGRA_1 (2007-3/2009) tests. Among them, there were 46 employees whose IGRA test results were indeterminate. With these 46 employees excluded, a total of 2054 employees were included in the analysis of the concordance between TST and IGRA_1 (Table 6).

Table 6 shows that there were 153 (7.45%) employees who tested positive by TST but negative by IGRA_1 (2007-3/2009), and 8 (0.39%) who tested negative in TST but positive in IGRA_1 (2007-03/2009). The discordance rate is 7.84% (161 out of 2054 cases), which is in the acceptable range.

Table 6: Concordance between TST and IGRA_1 (2007-3/2009)

| Frequency<br>Percent<br>Row Pct<br>Col Pct | Table of TST by IGRA_1 | | | |
|---|---|---|---|---|
| | | IGRA_1 | | |
| | TST | N | P | Total |
| | N | 1871<br>91.09<br>99.57<br>92.44 | 8<br>0.39<br>0.43<br>26.67 | 1879<br>91.48 |
| | P | 153<br>7.45<br>87.43<br>7.56 | 22<br>1.07<br>12.57<br>73.33 | 175<br>8.52 |
| | Total | 2024<br>98.54 | 30<br>1.46 | 2054<br>100.00 |

The Kappa correlation coefficient $\kappa$ is also applied to the above data to determine the agreement between the results of TST and IGRA_1 (2007-3/2009). Note that there were 22 cases that were positive, determined by both TST and IGRA (2007-3/2009), and 1871 cases that were negative determined by both tests. Thus, the observed agreement is $\Pr(a) = \frac{22+1871}{2054} = 0.921616.$

To calculate Pr($e$) (the expected percentage of agreement) it is noted that TST found 8.52% positive and 91.48% negative, and IGRA (2007-3/2009) found 1.46% positive and 98.54% negative.

Therefore, the probability that both tests found positive randomly is 8.52% · 1.46% = 0.1244% and the probability that both found negative is 91.48% ·98.54% = 90.1444%. As a result the overall probability of random agreement is Pr($e$) = 0.1244% + 90.1444% = 90.2688%.

Finally, the Cohen's Kappa correlation coefficient can be determined:

$$\kappa = \frac{92.1616\% - 90.2688\%}{1 - 90.2688\%} = 0.1945$$

This result is confirmed by SAS output in Table 7. Also in Table 7, with asymptotic standard error (ASE) = 0.0369, the 95% confidence interval for $\kappa$ is (0.1222, 0.2669).

Table 7: Statistics for Table of TST by IGRA_1

| Simple Kappa Coefficient | |
|---|---|
| Kappa | 0.1945 |
| ASE | 0.0369 |
| 95% Lower Conf Limit | 0.1222 |
| 95% Upper Conf Limit | 0.2669 |

The Kappa Correlation coefficient $\kappa = 0.1945$ is near the borderline of poor and fair agreement between TST and IGRA (2007-3/2009).

Similarly the Kappa Correlation coefficient can be applied to determine the agreement between TST and IGRA (post 2009) (with the 39 IGRA indeterminate excluded in Table 8).

Table 8: Concordance between TST and IGRA_2 (post 3/2009)

| Frequency<br>Percent<br>Row Pct<br>Col Pct | Table of TST by IGRA_2 | | | |
|---|---|---|---|---|
| | | IGRA_2 | | |
| | TST | N | P | Total |
| | N | 1612<br>89.41<br>96.82<br>94.05 | 53<br>2.94<br>3.18<br>59.55 | 1665<br>92.35 |
| | P | 102<br>5.66<br>73.91<br>5.95 | 36<br>2.00<br>26.09<br>40.45 | 138<br>7.65 |
| | Total | 1714<br>95.06 | 89<br>4.94 | 1803<br>100.00 |

In Table 8, there were 102 (5.66%) health care workers who were TST positive but IGRA negative (post 3/2009), and 53 (2.94%) who were TST negative but IGRA positive (post 03/2009). Therefore, the discordance rate is 8.60% (155 out of 1803 cases).

Table 9: Statistics for Table of TST by IGRA_2

| Simple Kappa Coefficient | |
|---|---|
| Kappa | 0.2736 |
| ASE | 0.0414 |
| 95% Lower Conf Limit | 0.1925 |
| 95% Upper Conf Limit | 0.3547 |

In Table 9 the Kappa correlation coefficient is determined to be $\kappa$ = 0.2736, which indicates that there is a fair agreement between TST and IGRA (post 3/2009). Compared Table 9 with Table 7 it is concluded that IGRA performance after 2009 has better agreement with TST than IGRA performed between 2007-3/2009.

## 5. Result and Discussion

1. Agreement between TST and IGRA:

Concordance between the TST and IGRA have widely varied in published studies, with variability attributed to factors such as test interpretation, recent TST, demographics, coinfections, immunosuppression, and background prevalence of tuberculosis [3-7]. As per the results above we can conclude that the Cohen's Kappa measurement for TST and IGRA 2007-3/2009) is 0.1945 and considered poor agreement but near the fair agreement. Cohen's Kappa measurement for TST and IGRA (post 3/2009) is 0.2736 and considered fair agreement. One possible reason for improved agreement after March 2009 may be related to operational characteristics of introducing the IGRA test into the employee-screening program. As IGRA had been newly implemented into their laboratory system around 2007, more technical acclimation was initially required, partially evident in the higher number of indeterminate results. More speculative factors may be related to changes in the employee population that was tested over time. The majority of the disagreements in all years were TST+/IGRA-, which are commonly linked to older age and populations enriched with foreign-borne individuals who had previous BCG vaccination [6.8].

2. Reversions and Conversions:

Both TST and IGRA responses commonly fluctuate on serial testing, to a magnitude where they may change from negative to positive (conversion) or from positive to negative (reversion) [9, 10].

Among our set of employees who received multiple TSTs, there were 11 cases of reversion and 29 cases of conversion (Table 1). Conversion typically signifies the development of an immune response to mycobacterial antigens following new infection with M. tb, other mycobacteria, or BCG vaccination and serial testing is typically used to investigate of actively infected TB cases. However, in the absence of a known exposure between the two tests, it is also often difficult to distinguish true conversion from a boosting phenomenon where the first TST primes the immune system to react to the second TST and create a positive result [11]. Therefore, a proportion of the 29 “conversions” may have not had exposure at all but manifested with a boosting phenomenon. Time between the two tests and degree of increase in response are used as general guidelines to distinguish conversion from the boosting phenomenon, but this is subject to inter-individual variability. Reversion is more common in older adults and in those manifesting the booster phenomenon [ 12 ]. It may also occur following treatment for TB infection in some patients, especially when their baseline TST was borderline positive [13] .

Among our set of employees who received multiple IGRA tests, there were 16 cases of reversion and 22 cases of conversion prior to March 2009 (Table 2) and 39 cases of IGRA reversion and 15 cases of IGRA conversion after March 2009 (Table 4). Conversions, or increases in IGRA responses, may signify a new infection, but could also occur due to laboratory variations or a boosting phenomenon following TST placement [14, 15]. Reversions, or decreases in the IGRA response, can be even more difficult to interpret. In a minority of cases, IGRA reversions have been attributed to clearance of infection, whether spontaneous or after treatment, but are more commonly due to variance in testing. Notably, conversions and reversions are commonly seen when baseline TST/IGRA results were discordant, implying a weaker baseline result, or in IGRA responses near cut points dividing positive and negative results [10, 16, 17]. Therefore, it is often unclear whether these fluctuations or changes on serial testing are attributed to internal test qualities and interpretation or to true responses in the patient.

The statistical analysis highlights operational challenges with current tests used for diagnosing latent TB. Specifically, disagreement between the TST and IGRA tests, as well as poor reproducibility of these tests especially evident by high reversion rates, confirms the need to interpret these tests results with caution. Ultimately, cut-off points to distinguish intrinsic variability in the test and new TB infection need to be optimized based on specific host and population characteristics. In the meantime, patient-specific risk factors for developing severe infection as well as new TB exposure should also be taken into consideration when making treatment decisions for latent TB infection.

## References

[1] Organization, W.H., Global Tuberculosis Report 2014, Executive Summary xi – xiii. 2014.

[2] Menzies, D., M. Pai, and G. Comstock, Meta-analysis: new tests for the diagnosis of latent tuberculosis infection: areas of uncertainty and recommendations for research. Ann Intern Med, 2007. 146(5): p. 340-54.

[3] Brock, I., et al., Comparison of tuberculin skin test and new specific blood test in tuberculosis contacts. Am J Respir Crit Care Med, 2004. 170(1): p. 65-9.

[4] Jepsen, M., et al., Comparison of QuantiFERON and tuberculin skin test in possible suspected tuberculosis infection. Ugeskr Laeger, 2009. 171(37): p. 2625-31.

[5] Loywyck, V., et al., A comparison of two methods for prediction of response and rates of inbreeding in selected populations with the results obtained in two selection experiments. Genet Sel Evol, 2005. 37(3): p. 273-89.

[6] Nienhaus, A., et al., Evaluation of the interferon-gamma release assay in healthcare workers. Int Arch Occup Environ Health, 2008. 81(3): p. 295-300.

[7] Tripodi, D., et al., Evaluation of the tuberculin skin test and the interferon-gamma release assay for TB screening in French healthcare workers. J Occup Med Toxicol, 2009. 4: p. 30.

[8] Nienhaus, A., A. Schablon, and R. Diel, Interferon-gamma release assay for the diagnosis of latent TB infection--analysis of discordant results, when compared to the tuberculin skin test. PLoS One, 2008. 3(7): p. e2665.

[9] Pai, M., et al., Serial testing of health care workers for tuberculosis using interferon-gamma assay. Am J Respir Crit Care Med, 2006. 174(3): p. 349-55.

[10] Perry, S., et al., Reproducibility of QuantiFERON-TB gold in-tube assay. Clin Vaccine Immunol, 2008. 15(3): p. 425-32.

[11] Menzies, D., Interpretation of repeated tuberculin tests. Boosting, conversion, and reversion. Am J Respir Crit Care Med, 1999. 159(1): p. 15-21.

[12] Van den Brande, P. and M. Demedts, Four-stage tuberculin testing in elderly subjects induces age-dependent progressive boosting. Chest, 1992. 101(2): p. 447-50.

[13] Johnson, D.F., et al., Tuberculin skin test reversion following isoniazid preventive therapy reflects

diversity of immune response to primary Mycobacterium tuberculosis infection. PLoS One, 2014. 9(5): p. e96613.

[14] Leyten, E.M., et al., Effect of tuberculin skin testing on a Mycobacterium tuberculosis-specific interferon-gamma assay. Eur Respir J, 2007. 29(6): p. 1212-6.

[15] Naseer, A., S. Naqvi, and B. Kampmann, Evidence for boosting Mycobacterium tuberculosis-specific IFN-gamma responses at 6 weeks following tuberculin skin testing. Eur Respir J, 2007. 29(6): p. 1282-3.

[16] Lucet, J.C., et al., Interferon-gamma Release Assay vs. Tuberculin Skin Test for Tuberculosis Screening in Exposed Healthcare Workers: A Longitudinal Multicenter Comparative Study. Infect Control Hosp Epidemiol, 2015. 36(5): p. 569-574.

[17] Veerapathran, A., et al., T-cell assays for tuberculosis infection: deriving cut-offs for conversions using reproducibility data. PLoS One, 2008. 3(3): p. e1850.

[18] Bakeman, R.; Gottman, J.M., Observing interaction: An introduction to sequential analysis (2nd ed.). Cambridge, UK: Cambridge University Press, 1997.

[19] Fleiss, J.L., Statistical methods for rates and proportions (2nd ed.). New York: John Wiley, 1985.

[20] Ronald P. Cody and Jefferey K. Smith, Applied Statiistics and the SAS Programming Language, 5th ed. Pearson Prentice Hall; 2006.

# Principal Components Analysis of Socio- Economic Factors of Renewable Energy Users of Countries such as Nepal

Jyoti U. Devkota[1*], Kamala Upreti[2], and Deepa Gurung[3]

[1]*Department of Natural Science, Mathematics Group, Kathmandu University, Dhulikhel, Kavre, Nepal,*
*drjdevkota@ku.edu.np*
[2]*Department of Natural Science, Mathematics Group, Kathmandu University, Dhulikhel, Kavre, Nepal,*
*kamalaupreti123@gmail.com*
[3]*Department of Environmental Science and Engineering, Kathmandu University, Dhulikhel, Kavre, Nepal,*
*deepagurung2000@hotmail.com*

## Abstract

Detailed analysis of the socio economic status of renewable energy consumers namely biogas and micro hydro is done for Nepal. The questionnaire of the two surveys was same except for questions pertaining to biogas and MHP. Direct questions relating to income fails to reflect the socioeconomic status. Thus proxy asset indicators are identified. This paper is based on the survey of 400 households using biogas as a source of renewable energy, which was conducted in three different rural settings of Nepal during September to November 2010. Out of 467 variables studying various socio-economic and performance parameters in the consumer profile database 47 proxy asset variables are chosen. The dimension is reduced to ten orthogonal variables explaining 60 percent of the variability using Principal Components Analysis (PCA). A survey of 51 households of Micro Hydro consumers was done in January 2014. The number of variables studying different socio-economic and performance parameters was 386, out of which 49 were identified as proxy asset indicators. Principal components analysis reduced the dimension to eight orthogonal variables explaining 61 percent of the total variation. Multinomial regression is applied to minutely analyze the dynamics of change in response of time saved with respect to independent variables income status, distance travelled and occupation. This paper aims to stimulate interest in the interdisciplinary applications of various statistical methodologies to a problem from renewable energy. The results will help understand the socio-economic status of the consumers which will help make an optimum market strategy that will popularize renewable energy.

*Keywords*: Principal Components Analysis, Multinomial Regression, Biogas, Micro Hydro, Consumer Profile Database, Survey, Questionnaire, Socioeconomic, Asset Variables, Age Distribution

*Corresponding Author
E-mail Address: drjdevkota@ku.edu.np

## 1. Introduction

Nepal has an agriculture based economy. Biogas plants are ideal source of renewable energy as they run on cow dung and other farm animals waste. Cattles are integral part of such economy. The energy shortage is marked by up to fourteen hours of load shedding in Nepal. Biogas and Micro hydro are plausible answers to this shortage. Nepal is rich in water resources and Micro Hydro can be installed in any nook and corner of a mountainous terrain. Micro hydro, the technology for small sized hydropower projects of up to 100 kW, can not only bring light into people's lives in the off grid locality but also give energy and water security to population.

Renewable energy sources are mainly used by farmers of Nepal’s agriculture based economy. In such economies most of the socio economic transactions are carried outside the market. Hence many benefits cannot be evaluated monetarily. These socioeconomic transactions also cannot be easily quantified and are therefore subjective in nature. Direct questions on income and expenditure are offensive and are of sensitive nature. Quantitative assessment of socio economic indices requires special attention. Hebert et al. [7] reduced the dimension of a number of predictor variables with minimal loss of information. The seven variables related to socioeconomic conditions included in computing the socioeconomic status principal components (SESPCs) were as follows: GNP, life expectancy at birth, infant mortality rate, proportion of the population without safe water or excreta disposal facilities, and the number of physicians or hospital beds per 10 000 population. A survey using a structured questionnaire to collect data on socio-economic characteristics and malaria beliefs and practices among more than 400 net-owning and non-net-owning households was conducted by Howard [8]. A composite socioeconomic index was created is, and survey households were divided into socio-economic quartiles. Vyas and Kumranayake [10] reviewed various issues related to choice of variables and data preparation to Brazil and Ethiopia Demographic Health Survey data. Filmer and Pritchett [6] studied wealth by constructing a linear index from asset ownership indicators. Different methods to be adopted to improve the running of PCA on discrete data were suggested by Kolenikov and

Angeles [10]. In modern days, data based research requires sound statistical analysis and its interdisciplinary application. The mortality and fertility data of developing countries especially from Nepal are analyzed by using deterministic and mixed effect models by Devkota [4] [5]. Several problems associated with the continual collection of vital statistics especially in countries with limited and defective data like Nepal are discussed by Devkota [3]. Similarly the importance of digitization of survey data is elucidated in details by Devkota [2].

In this paper the analysis of assets done as a part of socioeconomic analysis, deals essentially with intangible factors and their impact with reference to the use of renewable energy. We attempt to do the difficult task of objective quantification of socioeconomic benefits of two sources of renewable energy where the definition and evaluation of such impacts is not possible in numerical terms. The remainder of this paper is arranged as follows. In section 2 called methods and materials, the steps followed for design as well as implementation of the survey and the resulting dataset that motivates this study are described along with a brief literature review. Section 3 gives the results and discussions of this study and the concluding section 4 gives the overall discussion of the conclusions.

## 2. Research methodology

### *2.1 Mathematical Background*

Principal component analysis is a statistical approach that can be used to analyse the interrelationship among a large number of variables such that the information contained in a number of original variables is condensed into a smaller set of variates (factors) with minimum loss of information. This data summarization helps identify the underlying dimension or factor, estimates of factors and contribution of each variable to the factors (termed loadings). Unrotated factor matrix comprising of factor loadings is used when the main objective of research is in best linear combination of variables where the a particular combination of original variables account for more of variance in the data as a whole than any other linear combination.

Suppose we have a set of N variables, a*$_{1j}$ to a*$_{Nj}$, representing the ownership of N assets by each household j. Further, let us standardize each variable by its mean and standard deviation: for example, $a_{ij} = \frac{a^*_{1j} - a^*_1}{s^*_1}$, where $a^*_1$ is the mean of $a^*_{1j}$ across households and $s^*_1$ is its standard deviation. These selected variables are expressed as linear combination of a set of underlying components for each household j:

$$a^*_{1j} = v_{11} * A_{1j} + v_{12} * A_{2j} + \cdots + v_{1N} * A_{Nj}$$

.

$$a^*_{Nj} = v_{N1} * A_{1j} + v_{N2} * A_{2j} + \cdots + v_{NN} * A_{Nj} \text{ ....... (1)}$$

where, j=1,….J. A's are the components and v's are the coefficient on each component for each variable. The "scoring factors" from the model are recovered by inverting the system implied by eq. (1), and yield a set of estimates for each of the N principal components:

$$A_{1j} = f_{11} * a_{1j} + f_{12} * a_{2j} + \cdots + f_{1N} * a_{Nj}$$

.

.

$$A_{Nj} = f_{N1} * a_{1j} + f_{N2} * a_{2j} + \cdots + f_{NN} * a_{Nj} \text{..........(2)}$$

where j=1,…J. The asset index expressed in terms of the original (unnormalized) variables, is therefore an index for each household based on the expression

$$A_{1j} = f_{11}\left(\frac{a^*_{1j} - a^*_1}{s^*_1}\right) + \ldots + f_{1N}\left(\frac{a^*_{Nj} - a^*_N}{s^*_N}\right) \text{........................(3)}$$

where j=1,….J. With respect to the study of Micro hydro consumers of the study there are 49 asset variables and 51 households. So N = 49 and J = 51. Whereas with respect to Biogas consumers N = 47 and J = 400.

Multinomial logistic regression is used to predict categorical placement in or the probability of category membership on a dependent variable based on multiple independent variables. It is a simple extension of binary logistic regression that allows for more than two categories of the dependent or outcome variable. Let J denote the number of categories of Y. Here Y is a multinomial response variable. Let {$\pi_1$, $\pi_2$…….$\pi_j$} denote the response probabilities, satisfying the condition that their sum is equal to 1. Logit models for multinomial response pair each category with a baseline category. When the last category (J) is the baseline, the baseline-category logits are

$$\log\left(\frac{\Pi_i}{\Pi_j}\right), j = 1, \ldots J - 1$$

Given that the response falls in the category j or J, this is the log odds that the response is j. The baseline category logit model with predictor x is

$$\log\left(\frac{\Pi_i}{\Pi_j}\right) = \alpha_j + \beta_j x, j = 1, \ldots J - 1$$

The model has J-1 equations with separate parameters for each. The effects vary according to the category paired with the baseline. When J = 2, this model simplifies to single linear equation for $\log\left(\frac{\pi_1}{\pi_2}\right) = logit(\Pi_1)$, resulting in ordinary logistic regression for binary responses.

There is an odds ratio associated with each predictor. It is denoted by Exp(B). It is more than 1 in cases where predictors increase the logit, Exp(B) is equal to 1 in cases where predictor don't have any influence on the logit and Exp(B) is less than 1 in cases where predictors decrease the logit.

### *2.2 Sample survey and consumer profile database*

A survey of 400 households was conducted in three different regions of Nepal. A comprehensive database of 467 variables was constructed on the basis of questions related to their socio economic standards, life style change, health impacts and design expectations. This survey was conducted during September to November 2010. The questionnaire was designed with an objective of keeping biogas use in the core and getting all the possible information about a typical middleclass Nepalese family inhabiting in rural areas, its economic and social background and change after biogas was used in their household. In 59 questions information was collected on amount of landholdings, livestock and fuel wood expenses before and after the installation of plants. These households had 2272 individuals of different age groups. So with the structure of the questions information can be obtained about the households that haven't installed biogas as a source of renewable energy. Among biogas consumers of Nepal inhabiting in the rural area who are primarily using cow dung for their plant, income and expenditure transactions are carried out outside the market.

Similarly a sample survey of 51 households was conducted from a population of 179 of Micro hydro project (MHP) consumers of Karam Danda MHP. Karam Danda Micro hydro is located in ward no. 2 of Vimkhori village development committee (VDC) of Kavre district. It is located about 50 Kilometers from Kathmandu. Despite its close geographical proximity to the capital it is a relatively backward with respect to electricity connection to the national grid. The inhabitants of this VDC have an access to electricity since 2010 through the construction of 17 KW Karam Danda MHP. They are primarily farmers by profession. These households can light 100 watt of bulb from 6 PM to 6 AM from this electricity. A comprehensive database of 386 variables was constructed on the basis of questions related to their socio economic standards, life style change, health impacts and design expectations. This survey was conducted during January – February 2014. The questionnaire was designed with an objective of keeping micro hydro use in the core and getting all the possible information about a typical middleclass Nepalese family inhabiting in rural areas, its economic and social background and change after micro hydro was used in their household. In order to compare biogas users with MHP users the questionnaire designed was very similar to the questionnaire for biogas consumers. Thus in 60 questions information was collected on topics such as the age distribution of 51 households, amount of landholdings, livestock, the fuel wood expenses before and after the installation of plants etc. These households comprises of 278 individuals of different age groups. So here also with the structure of the questions information can be obtained about the households that haven't installed micro hydro as a source of renewable energy. Most of the questions were yes/no in nature.

This article gives some results in the task of objective quantification of socio economic structure of biogas and micro hydro consumers. Out of 467 variables 47 are indicator asset variables that can be classified into four groups. In the first group there were 8 questions related to asset ownership which comprised of land, house, means of transportation and electronic equipments such as computer, television, radio, mobiles, telephones and refrigerator. There were 10 indicator variables in this category. Characteristics of house dwellings and toilet were the second group and comprised of 6 questions and 10 indicator variables. Here questions on materials used in the construction of house were asked. Open latrine, latrine far away from the house or close to the house etc were queries related to the type of latrine. Water source and needs was the third group with 6 questions and 20 indicator variables. Detailed question on the use of water such as for bathing or cooking etc were asked. Different sources of water such as private well, open well, community water supply etc were asked. Fourth and the last group was the amount of land owned and comprised of 5 indicator variables Similarly out of 386 variables of consumer profile database based on a sample survey data of 51 households of Micro hydro consumers, 49 proxy asset variables are identified. They are closely interlinked to the socio economic structure of the micro hydro consumers and are more accurate and reliable indicators than the exact monetary accounts. These assets variables are classified into four categories namely Material Assets, Type of House, Amount of Land Possession and Source of Water. These variables resulted from a questionnaire comprising of 60 questions with answers classified into several categories.

## 3. Research results and discussions

Principal components analysis was conducted. Here 61 percent of underlying correlation between 49 assets variables of MHP consumers was explained by 8 orthogonal principal components. First two principal components explain 24 percent of the total variation. The first component has high absolute value of factor scores with respect to indicators of wealth such as ownership of refrigerator, own sources of water for cooking, bathing, washing and drinking. Thus the first component is an indicator of wealth of the family. Access to own water source for drinking, cooking, bathing is the main indicator of the economic well-being of the families. Then the socioeconomic categorization is done classifying the households into three socio-economic groups, where rich comprised of top 20% of the socio-economic group of the society, the middle income group comprised of middle 40% and the lowest 40% are economically most deprived. This was on the basis of asset index which is linear combination of factor loadings of first principal component and normalized asset ownership variables. Hence it is a dependent variable constructed on the basis of factor scores of first principal component. The robustness of this classification is tested by the data on those assets which are conventionally owned by people who are more economically well off. Thus the dimensionality of the data comprising of 49 interrelated asset variables is reduced to 8 orthogonal variables. It is also seen that ownership of own water source is the main factor differentiating economically well off households with poor households. Income status is the variable resulting from this asset index. Similarly for biogas consumers, PCA of 47 asset variables from 467 variables of entire socio-economic survey extracted 10 components. This explained 60% of the total variance. First two principal components explained 19% of the total variation. Then the socio-economic categorization is done classifying the households into three socio-economic groups, where rich comprised of top 20% of the socio-economic group of the society, the middle income group comprised of middle 40% and the lowest 40% are economically most deprived. This was on the basis of asset index which is linear combination of factor loadings and normalized asset ownership variables [1]. The distribution of households of MHP consumers and biogas consumers according to the assets index is shown in Figure 1 and Figure 2. These asset indices based on the factor loadings for first principal component has classified these families into three income groups.

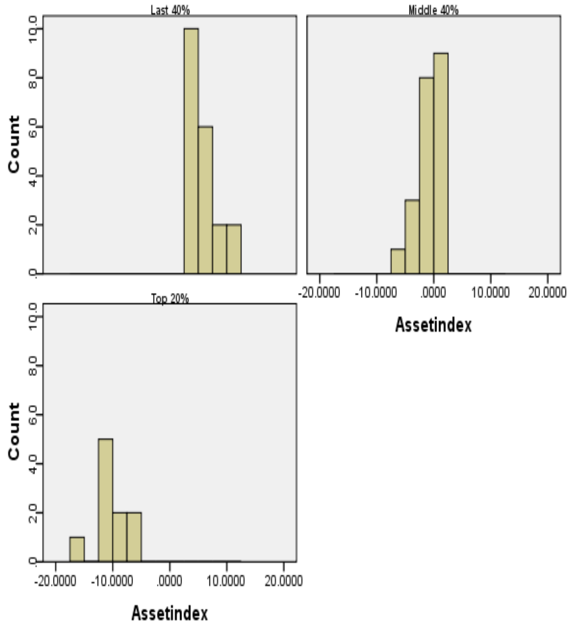


Figure 1: Income classification of Micro Hydro Consumers

Further multinomial logistic regression is fitted to these variables for MHP and biogas consumers separately. Time saved is regressed on Distance travelled, Income status and Occupation. Time saved, Distance travelled and Income Status are categorical ordinal data whereas Occupation is a nominal data. Here the dependent variable namely Time Saved quantifies the amount of time saved per household. It is an ordinal data where for MHP consumers five options were provided, which are the following no time, 15 – 30 min, 30 – 45 min, 45 – 60 min, and 60 – 120 min saved with ranks 0, 1, 2, 3, 4 respectively.

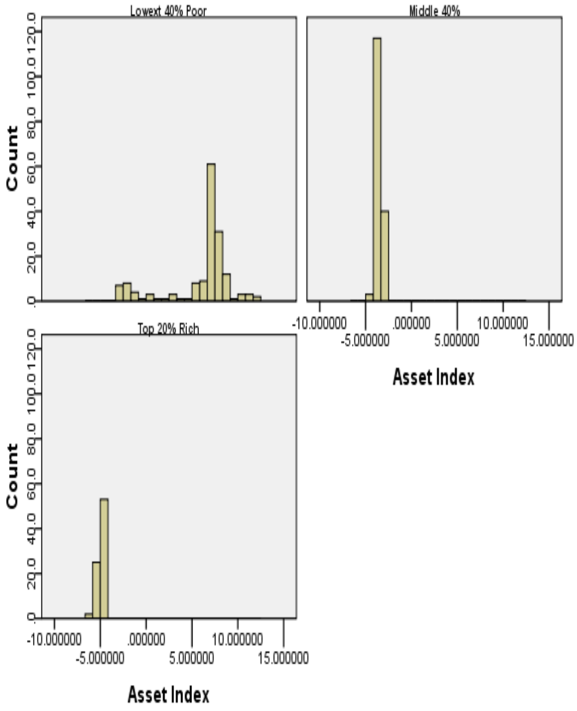


Figure 2: Income classification of Biogas Consumers

Similarly Distance travelled quantifies the distance covered for the collection of firewood. It is an independent variable. Five options were provided as answers namely No distance travelled, less than 100m, 100 – 200m, 200- 500m and more than 500 m that were ranked as 0, 1, 2, 3 and 4 respectively. Income Status is another independent variable categorised into 1, 2 and 3 representing bottom 40%, middle 40% and top 20% of the socioeconomic classification. The variable Occupation gives first option chosen by the household as response and it has three categories namely Agriculture, Business and Livestock that are categorised as 1, 2 and 3 respectively. For biogas consumers the options provided as answers to the question on Time Saved per day is no time, less than 60 min, 60 – 180 min, 180 – 300 min, more than 300 min with ranks 0, 1, 2, 3 and 4. As the biogas generated by the plant was mainly used in cooking the amount of time saved by a biogas consumer is substantial in comparison to an MHP consumer; they use the electricity generated for lighting their houses in the evening. The options provided as answers for Distance Travelled, Income Status and Occupation are same for both type of consumers. Questionnaire for both these surveys had been designed in such a way that the questions asked along with the options provided were same except for the questions pertaining to the performance of MHP and biogas. The results of model fitting are summarized in Table1. The regression is highly significant with a p value of 0.013 for MHP consumers. The value of Chi square test statistics is 52.501. The value of -2logliklihood is 47.78. The efficiency of the model is given by Nagel Kerke pseudo $R^2$ value of 0.709. The benefits of installation of MHP in a community of 179 households can be quantified with the help of odds ratio. Installation of micro hydro had a significant effect of the families claiming that farming is their major occupation; families engaged in farming have 73.6% higher odds of claiming that micro hydro saved their time by 45 min – 1 hour. Similarly those in the poor income group are 23.6% more likely to claim that micro hydro saved their time by 1 – 2 hours. These values of odds ratio help quantify the extent of benefit of this source of renewable energy to the farmers and livestock keepers. Similarly households covering 200 – 500m daily for the collection of firewood are 3472 times more likely to claim that MHP saved their time by 45 min – 1 hour. The results of regression are shown in Table1. This shows a very good fit with 80.4% of accuracy. Similarly Time saved response after a switch over biogas is regressed upon Distance travelled, Income Status and Occupation. The regression is highly significant with a p value of 0. The value of Chi Square is 114.954 at 36 degrees of freedom. The value of -2logliklihood is 123.604. The efficiency of the model is given by Nagel Kerke pseudo $R^2$ value of 0.274. Ordinary least squares (OLS) are based on minimising the difference between observed and predicted value. But this is not true for logistic regression. The efficiency of OLS models is measured with the help of coefficient of determination $R^2$, but in the case of logistic models Nagel Kerke $R^2$ is more suitable as it compares the value of $R^2$ with its upper bound ($R^2$ max). This is obtained by dividing $R^2$ by $R^2$ Max. Comparison of both $R^2$ shows that logistic regression is more successful in the case of MHP data than biogas data. This fact is validated by higher value of Nagel Kerke $R^2$ (0.709) for MHP consumers than for Nagel Kerke $R^2$ (0.274) for biogas consumers. The benefits of installation of Biogas in a community of 400 households can be quantified with the help of odds ratio. Households covering 100 – 200m for the collection of fire wood before the construction of biogas plants were 487800 times more likely to claim that it saved their time by 1- 3 hours. As seen from Table 2 and Table 3, the prediction of amount of time saved using logistic regression model shows 80.4% and 56.2% accuracy in case of MHP and biogas respectively.

Table1: Time saved is regressed on Distance travelled, Income Status and Occupation

| Renewable Energy Source | -2logliklihood | Chi Square | Pseudo $R^2$ |
|---|---|---|---|
| Biogas | 123.604 | 52.501 | 0.274 |
| MHP | 47.78 | 114.954 | 0.709 |
| Dependent Variable | Independent Variable | | |
| MHP | | | |
| Time Saved | Occupation | Distance Travelled | Income Status |
| 45 min – 60 min | Farming<br>Odds ratio<br>1.736<br>Wald's Statistics<br>0.006 | | |
| 1 - 2 hours | | | Poor<br>Odds Ratio<br>1.215<br>Wald's Statistics<br>0.001 |
| 45 min – 1hour | | 200-500m<br>Odds Ratio<br>3472<br>Wald's Statistics<br>0.045 | |
| Biogas | | | |
| 3-5 hours | 100 – 200 m<br>Odds Ratio<br>487800<br>Wald's Statistics<br>0.000 | | |
| 3-5 hours | | | Poor<br>Odds Ratio<br>0.00000634<br>Wald's Statistics<br>0.001 |
| 30 min | | Farmer<br>Odds Ratio<br>4.446E-16<br>Wald's Statistics<br>0.000 | |

Table 2: Classification table MHP

| | Predicted | | | | | |
|---|---|---|---|---|---|---|
| Observed | No time saved or NA | 15-30min | 30-45min | 45min - 1hour | 1hour-2 hour | Percent Correct |
| No time saved or NA | 2 | 0 | 0 | 0 | 0 | 100.0% |
| 15-30min | 0 | 12 | 1 | 2 | 1 | 75.0% |
| 30-45min | 0 | 3 | 2 | 0 | 0 | 40.0% |
| 45min - 1hour | 0 | 0 | 0 | 2 | 0 | 100.0% |
| 1hour-2 hour | 0 | 2 | 1 | 0 | 23 | 88.5% |
| Overall Percentage | 3.9% | 33.3% | 7.8% | 7.8% | 47.1% | 80.4% |

## 4. Conclusion

Table 3: Classification table Biogas

| | Predicted | | | | | |
|---|---|---|---|---|---|---|
| Observed | 0 | 30 mins | 1-3 hr | 3-5 hr | >5 hr | Percent Correct |
| 0 | 0 | 4 | 6 | 0 | 3 | .0% |
| 30 mins | 0 | 63 | 58 | 0 | 11 | 47.7% |
| 1-3 hr | 0 | 43 | 149 | 0 | 3 | 76.4% |
| 3-5 hr | 0 | 8 | 14 | 1 | 0 | 4.3% |
| >5 hr | 0 | 9 | 16 | 0 | 12 | 32.4% |
| Overall Percentage | .0% | 31.8% | 60.8% | .2% | 7.2% | 56.2% |

This paper has attempted the difficult task of objective quantification of socio-economic parameters of biogas and micro hydro consumers. Difficulty in quantification of such parameters which are real and intangible overshadows its great importance in such research. These studies are very important; better knowledge of such studies will help develop optimum market strategies according to the income level. Such studies involving data generation, verification, rectification and prediction are very crucial for countries overshadowed by inaccurate official records. Principal components analysis is used for the analysis of socio-economic status; these are intangible and hence cannot be quantified. Asset index is also calculated on the basis of the first principal component. This component is an indicator of wealth with high absolute value of factor scores on households with own water source. Thus having own water source is the most important indicator of affluence in such groups. This is plausible as they are all farmers living in close proximity to each other. Their access to the water source such as private tap, public tap, common well, river/ spring for meeting their cooking, bathing and irrigation needs plays a key role. On the basis of the values of the asset index the consumers were classified into rich that were the top 20 % which was followed by middle 40% and bottom 40%. It can also be classified as rich, middle income and poor respectively is an indicator of wealth having high loadings on expensive assets. Multinomial regression has quantified the dynamics of response of time saved due to the use of renewable energy with respect to income status, distance travelled for the collection of firewood and occupation. The results show that although renewable energy has been beneficial to all type of consumers; but consumers from lower income group were more vocal about greater benefits of MHP. Similarly those covering longer distance for the collection of firewood claimed that the use of renewable energy saved their time considerably. Thus the extent of benefit could be quantified with the help of odds ratio. The results of such detailed studies can be generalized to many countries in South Asia and Africa including Bangladesh, Pakistan.

### Acknowledgements

This research is funded Renewable Nepal Project Grant No: RENP-10-06-PID-379. We acknowledge Renewable Nepal for supporting our research endeavors and providing research funding for the project activities.

### References

[1] Devkota, J. U., Singh S. , Prajapati C. , Hada B., 2014. Principal components analysis of the socioeconomic conditions of biogas users - with

example from Nepal. International Journal of Renewable Energy Research.

[2] Devkota J. U., Hada B., Prajapati C., Singh S., 2012a. The importance of research data digitization and its statistical analysis-with examples of biogas consumers of Nepal. International Journal for Environmental Science and Development, 3 (2), 103-108.

[3] Devkota J. U., 2012b., Mortality and Fertility Models for Countries with Limited Data - Results Based on Demographic Data of Nepal, India and Germany. Lambert Academic Publishing, Saarbruecken, Germany.

[4] Devkota J. U., Singh R. S., 2011. Mathematical modeling of mortality for countries with limited and defective data. Journal of Applied Statistical Sciences, 19 (1), 24-36.

[5] Devkota J. U., Singh R. S., 2010. Deterministic and probabilistic models with applications to modeling fertility data, Journal of Applied Statistical Sciences, 18 (2), 161-176.

[6] Filmer, D., and Pritchett, L. 2001. Estimating wealth effect without expenditure data - or tears: An application to educational enrollments in states of India. Demography, 38, 115-132.

[7] Hebert, J. R., Hurley, T. G. et. al., 1998. Nutritional and Socioeconomic Factors in Relation to Prostate Cancer Mortality: a Cross-National Study. Journal of National Cancer Institute, 90(21), 1637-1647.

[8] Howard, N., Chandramohan, D., et al., 2003. Socio-economic factors associated with the purchasing of insecticide-treated nets in Afganisthan and their implications for social marketing.Tropical Medicine and International Health, 8(12), 1043-1050.

[9] Kolenikov, S., and Angeles, G., 2009. Socioeconomic status measurement with discrete proxy variables: is principal components analysis a reliable answer? Review of Income and Wealth, 55(1), 128-165.

[10]Vyas, S., Kumaranayake, L., 2006. Constructing Socio-Economic status indices: how to use principal component analysis. Health Policy and Planning, 21, 459-468.

# Brand Identity and Brand Loyalty in Thai Social Marketing

KwanchanokChumnumporn, ChawalitJeenanunta*, NarongakPongathornwiwat,
AtcharaParsont, KanisornLunsai, and RamidaPiyapaneekul
*School of Management Technology, Sirindhorn International Institute of Technology,*
*Thammasat University, Pathum-Thani, Thailand*

## Abstract

A social identity is an important perspective of customer–brand relationship. Customers are comfortable to buy the products through social network. However, they could rapidly change their mind to other competitors as well. For this reason, most of sellers have to find the suitable techniques in order to gain the competitive advantage to maintain customers' loyalty. A diversity of online products are become a major competition particular with product design, price and promotion in E-commerce market. These factors are combined as Brand identity. The objective of the study is therefore to examine and provide the empirical analysis whether Brand identity factors has positive impacted on brand loyalty particularly in the social commerce. The studied population focuses on the socio-demographic of the online consumer. The statistical analysis was analyzed using factor analysis and multiple linear regression. The results have shown that Brand identity; including product, price, and promotion, have positive influence significantly on brand loyalty. The main contribution of this study is that this study is the first to empirically test the effect of such factors on brand loyalty in social commerce in Thailand.

*Keywords*: Brand identity, Social marketing, Customer behavior, Brand loyalty, E-Commerce

*Corresponding Author
E-mail Address: chawalit@siit.tu.ac.th

## 1. Introduction

As the numbers of internet users are still increasing in the last ten years, this gains the opportunity of merchandisers to increase the sale channels such as social-media. There are 1.5 billion internet users, and still ongoing by five-percent year-on-year [6]. The social channels are turned advantage for small businesses by communicating directly to customers and could take market share from the large business [12]. Therefore, it is necessary that company has to become brand loyalty to gaincompetitive advantage. As the number of the purchasing through the social media is popular, the products are become diversity on online market especially the skin care products. The issue of brand loyalty for E-commerce has been intensively studying in various literatures such in the South Korea [9]. Online customer base their base their repurchase on previous experience through full service include product quality and online transaction service. There is intense competition on Product, Price and Promotion. As the previous study, Price and Promotionare effect to young shoppers' behavior [5], and Product assortmentis also important to customer [20], these factors are integrate as Brand identity. Brand identity is the starting point of building brand reputation and integrated marketing communicationsfor the purpose of building brand loyalty [14]. As guided intensively in literature; however, there is lack of empirical study testing such Brand identity on brand loyalty. The purpose of this research is therefore to explore and examine the factors that affect online loyalty in context of Thailand. The research questions are (1) what are the main factors influencing the successful online loyalty and (2) what are the impacts of such factors on the online loyalty

This paper was organized as follows; theory background and hypotheses is in the section 2, followed by the research methodology and data analysis is in section 3 and 4 respectively. The last section is the conclusions and suggestions for further investigation.

## 2. Theory Background and Hypothesis

### *2.1Brand loyalty in E-business*

Nowadays brand are not define as in the past. In classical definition (1960), the American Marketing Association (AMA) defined the brand as "a name, term, sign, symbol, or design, or a combination of them which is intended to identify the goods or services of one seller or a group of sellers and to differentiate them from those of competitors.", [7, p.9]. But for today's economy, brand becomes meaning in many intangible elements, not only tangible representations of a product, service or company as the AMA had mentioned. A brand is not only meaning to product, name, or logo but it combines communication, action, and organization. All process of the brand affect to behavior and action of customer. From the research on repurchase behavior, when customers become loyalty to the brand, customers havecommitment and attachment towards their loyalty brand, resistance to switch and willingness to pay more [19].The importance of loyalty has been a critical issue in the study of online market[16, 22]. Brand loyalty helps to increasing in market share, higher profits and better goodwill among consumers which is direct impact on the revenue and profitability of a company.

In this study, Brand loyalty in E-business is defined as "customer's favorable attitude and commitment towards the online retailer that results in repeat purchase behavior"[18]. Furthermore in this research is consist of three main point study of brand loyalty which are repurchase, word-of-mouth, and support the new product and brand activities.

### *2.2 Brand identity*

Different individuals have different levels of motivation and need for distinctiveness in their identities [21]. A brand with more distinctive identity is advantageous in obtaining consumers' attitudinal and action supports [8]. From pass researched brand identity has a direct effect on brand relationship and builds brand reputation [14, 17]. Some scholars define brand identity as the distinctive and relatively enduring characteristics of a focal brand (or company) [3, 4]. In this study, brand identity defined as the distinctive elements and communication style of online brand that present through the marketing mix (4P's).

Borden (1964) identified 12 marketing elements as the marketing strategy ormarketing mix which can help business to operate in a more profitably way [2]. Afterward, McCarthy and Brogowicz (1981)studied in basic marketing then determined marketing mix into 4 essential elements which are Product, Price, Promotion, and Place [13].Marketing mix is mentioned as a valuable tool for develop planning for marketing activities [1]. These four elements involve different subcategories that depend on the target customer of the brand which can make brand identity. For example, some online brand can impressed customers by product design, packaging of parcel delivering, or low price product with high quality. This study would questionnaire focus on Product, Price, and Promotion. For Place we interest in online transaction through social media channel (include: Facebook, Line, IG (Instagram), Google+, Twitter, and YouTube).

**Hypothesis:** There is a positive relationship between Product, Price, and Promotion on Brand loyalty

## 3. Research Methodology

### 3.1 Samples and Data collection

The survey method in this study is the purposive sampling (non-probability sampling method) by two ways approaches including online and face-to-face survey. The target group is to have higher percentage of female than male and higher percentage of age range 18-40 years old than less than 18 and more than 40 years old. The targetrespondent is those who had an online purchase experience, then, the online survey is posted on Facebook in the cosmetic fan page and follower of skin care/cosmetic selling page. While face-to-face survey, data is collected at public facilities in Bangkok and out skirting city of Bangkok including five university locations, a public library, and ten local shopping malls.There are 200 responseswhere there are 50 responses from online and 150 responses from face-to-face survey. This study only focuses onrespondents who had online transaction and thereare 146 valid responses for data analysis.

### *3.2 Measures Goodness of measure*

All of constructs and measurement items were adopted form the intensively review of the literature. The survey questions apply from the study ofHe et al (2012), BakewellandMitchell (2006), and the review of Thai online shopping [8, 5]. There were three components to measureBrand identity as follows, 7 items for Product, 2 items for Price, 6 items for Promotion, and 4 items for Brand loyalty.Each construct was measured by a five-point Likert type scale rang in from 1 (Strongly Disagree) to 5 (Strongly Agree). In this study, the data analysed in second steps. First step, we checked reliability of all measurement items, then used factor analysis to ensure the measurement items in each component (E-tail quality and Brand loyalty) can group by using Kaiser-Meger-Olkin (KMO) technique, the component are inter correlation if KMO should be greater than 0.60, then used reliability analysis to check internal consistency of components by Cronbach's alpha coefficient, which is a good method for Likert scale data [11]. Cronbach's alpha recommended threshold over 0.70 [15]. Second step, we conducted a multiple regression analysis to test hypothesis, one independent to one dependent each test. The relationship of two components represent by F-test and t-test. The beta (β) number show level of relationship of each item in the component and the adjust $R^2$ number show how items can explained the different of items and that component [11].

## 4. Research Results and Discussion

### *4.1 Data description*

From 146 samples, the majority of respondents were female with 79.45% (male 20.55%). The respondents consisted large number of studying age people (39.04% of 18-23 years old), and working age people (27.40% of 24-29 years old, and 21.92% of 30-40 years old), which trend to have well-education (69.18% had a bachelor's degree). The time of shopping online was generally 1-5 times per month (86.99%) and more than half of respondent had experience purchase skincare/cosmetic online (63.70%).

The online shopping channel that most popular is Facebook, IG (Instagram), Line, Other website/social network, Google+, YouTube, Twitter, and lastly WhatsApp.

Data description show in Table 2, Brand identity composes of three groups, the first group is Product has high mean from 3.75 to 4.36. Second group, Price has mean from 2.89 to 3.19. And the last group, Promotion has mean from 3.28 to 4.02.

Table 1: Demographic of the respondents

| | | Frequency | % |
|---|---|---|---|
| Gender | Male | 30 | 20.55 |
| | Female | 116 | 79.45 |
| Age | < 18 | 2 | 1.37 |
| | 18-23 | 57 | 39.04 |
| | 24-29 | 40 | 27.40 |
| | 30-40 | 32 | 21.92 |
| | 40+ | 15 | 10.27 |
| Education | High school graduate | 9 | 6.16 |
| | Associate degree, occupational | 9 | 6.16 |
| | Bachelor's degree | 101 | 69.18 |
| | Higher than Bachelor's degree | 27 | 18.49 |
| Personal income per month | 0-10,000 bath | 47 | 32.19 |
| | 10,001-20,000 bath | 39 | 26.71 |
| | 20,001-30,000 bath | 34 | 23.29 |
| | > 30,000 bath | 26 | 17.81 |
| Time of | 1–5 times | 127 | 86.99 |
| shopping online | 6–10 times | 15 | 10.27 |
| | 11–20 times | 1 | 0.68 |
| | 21–30 times | 1 | 0.68 |
| | >30 times | 2 | 1.37 |
| Shopping channel | Facebook | 88 | 60.27 |
| | IG (Instagram) | 66 | 45.21 |
| | Line | 58 | 39.73 |
| | Other | 22 | 15.07 |
| | Google+ | 11 | 7.53 |
| | YouTube | 7 | 4.79 |
| | Twitter | 3 | 2.05 |
| | WhatsApp | 2 | 1.37 |
| Shopping cosmetic online experience | Yes | 93 | 63.70 |
| | No | 53 | 36.30 |

The mean of Brand loyalty are from 3.36 to 3.64. The activities to get reward from brand include like brand's photo/page on website/social network, visit brand's website, watching a video product trial, referring friends, share/mention brand's website/social network, and checking in to a location via social network (ex.Facebook, IG (Instagram)).

### *4.2 Reliability test and factor analysis*

The reliability of our data is teat usingCronbach's alpha (α). The three factors of Brand identityhaveCronbach's alpha of 0.871, and Brand loyalty has Cronbach's alpha of 0.848. As Cronbach's alpha recommended threshold to be over 0.70 (Mouakket and Al-hawari, 2012). Since, allof the factors have the Cronbach'salpha more than 0.70, then the measurement are reliable.

Factor analysis was then considered. The KMO value is recommended to be more than 0.60[11]. All of threefactors (Product, Price, and Promotion) of Brand identity have KMO of 0.837and Brand loyalty has KMO of 0.758. Therefore,all items can be grouped as shown in Table 2.

### *4.3 Test of hypothesis*

Testing hypothesizes by multiple regression analysisto examine the relationship between the independent variable and the dependent variable. Simple linear regression equation show as follow:

$$Brand\widehat{loy}alty = a + b_1(Product) + b_2(Price) + b_3(Promotion)$$

At the level of significance of 0.10, the value of F-test should over2.71and the value of t-test should over 1.28for acceptable hypothesis [11]. The result show in table 3,Hypothesis has F-test value of 13.794; Productinteraction on Brand loyalty was significant (b=0.379,t=4.621, sig.=0.000) which means Product has positive effect on Brand loyalty; Price interaction on Brand loyalty was not significant (b=0.026, t=0.320, sig.=0.749)which meansthere are not has enough evident to support the relationship between Price and Brand loyalty at significant level of 0.10;Promotion interaction on Brand loyalty was significant (b=0.160, t=1.803, sig.=0.073) which mean Promotion has positive effect on Brand loyalty.These results show that Product and Promotion are supported. However, there isnot enough evidence to support Hypothesis.Product and Promotion are effect customer decision to Brand loyalty as the literature review, the study of He et al (2012) [8]. The result of Promotion also supports the study of Krishnamurthy (2009) that Promotion help to create a brand on online community [10].Krishnamurthy (2009) studied how the members of Firefox community used Promotion behaviors to create a brand [10].For Price, the research of Aghaei et al (2014) found that Price has positive effect to customers' decision [1], and also support the study of Yeu et al (2012) that customer draw their own mental picture of what a product is worth. Seller should be aware of product quality and its price when brand expects to create strong relationship with customers [23, p.1057].With the increased connectivity afforded by the internet, the growth of online communities help sellers and customers interact to develop more engaging product and build Brand loyalty.

Table 2: Reliability test, and factor analysis

| | Mean | Std. Deviation | Factor Loading | KMO | α |
|---|---|---|---|---|---|
| Brand identity | | | | 0.837 | 0.871 |
| Product: | | | | | |
| Product is worth for money | 4.36 | 0.837 | 0.663 | | |
| Food and Drug Administration (FDA) approved | 4.29 | 0.848 | 0.748 | | |
| Quality of product | 4.12 | 0.854 | 0.646 | | |
| A parcel is appreciate with product | 4.09 | 0.975 | 0.801 | | |
| Clear seller address on parcel | 3.99 | 0.939 | 0.829 | | |
| Beautiful handwriting on parcel | 3.75 | 1.056 | 0.734 | | |
| Design/packaging | 3.74 | 0.918 | 0.682 | | |
| Price: | | | | | |
| Lower priced products. | 3.19 | 0.897 | 0.811 | | |
| Products have reasonable price | 2.89 | 1.127 | 0.809 | | |
| Promotion: | | | | | |
| Free products sample | 4.29 | 0.848 | 0.777 | | |
| Discount | 4.02 | 1.060 | 0.654 | | |
| Membership discount | 3.84 | 1.031 | 0.731 | | |
| Promotional gift | 3.68 | 1.043 | 0.804 | | |
| Cash back | 3.53 | 1.164 | 0.804 | | |
| Reward point | 3.28 | 1.113 | 0.643 | | |
| Brand loyalty: | | | | 0.758 | 0.848 |
| Customers choose their brand loyalty before others | 3.64 | 0.974 | 0.864 | | |
| Customers are willing to try brand loyalty's new product | 3.58 | 0.988 | 0.797 | | |
| Customers support the offerings from their brand loyalty | 3.47 | 0.977 | 0.889 | | |
| Activities to get reward from the brand | 3.51 | 0.98 | 0.765 | | |

Table 3: Result for linear regression analysis on hypothesized

| Independent variables | F-test | t-test | Beta | Adjust $R^2$ | Sig. | Std. Error of the Estimate | Conclusion |
|---|---|---|---|---|---|---|---|
| Brand identity | 13.794 | | | 0.209 | | 0.889 | |
| Product | | 4.621 | 0.379 | | 0.000 | | Support |
| Price | | 0.320 | 0.026 | | 0.749 | | Not Support |
| Promotion | | 1.803 | 0.160 | | 0.073 | | Support |

Note Dependent variable: Brand loyalty

Father more Hypothesis has the adjust $R^2$ value of 0.209; this means Brand identity (Product, Price, and Promotion) can explain Brand loyalty 20.9 percent. And The standard error of the Estimate is 0.889.

## 5. Discussion and Conclusion

In this empirical study, we are interested in people who are loyalty to skincare/cosmetic brand and also have an experience on online transaction. Our goal in this study focus on the influences of Brand identity: Product, Price, and Promotion, on Brand loyalty in Social Commerce in Thailand. There are in total 200 responses where 50 responses are from online and 150 responses are from face-to-face survey. However,there are 146 valid responsesthathave experience in Social Commerce that we are using for further analysis.

The first important awareness is Product. Customers are willing to pay for product that worth for money. FDA approved and quality of product is also important factors for the customers. When the product has delivered to customer, the parcel should suitable with the product and its price, with clear address of seller and customers on parcel which can help product delivery arrive on time.

However, there arenotenough evidence to show that Price would influence the Brand loyalty. We may explain thatwhen customers become loyalty to the Brand, they do not carein price of product. Moreover,customers do notbuy the product just because that product has reasonable price.

Our study shows that loyalty customers are attractive to promotion. The attractive promotions are free products sample, discount, membership discount, promotional gift, cash back, and lastly reward point.

In addition, customers prefer feel connect with the seller through social network channel (i.e. Facebook, Line, Twitter etc.). As long as customers are loyal, they will not switch to other brand and willing to support the new product and the offerings from their loyaltybrand. By applyingthese finding factors, seller couldimprove theirBrand identity that will maintain their customer loyalty.

This research has several limitations that suggest promising avenues for future research.The research examines only Brand identity as the independent variable that effect Brand loyalty. Only Brand identity may notfully explain Brand loyalty on the social commerce. Future research should examine other effects, such as E-tail quality, E-satisfaction, and E-trust on Brand loyalty.

## 6. Acknowledgements

This research was financially supported by underSirindhorn International Institute of Technology (SIIT), Thammasat University.

## References

[1] Aghaei M, Vahedi E, Kahreh MS, Pirooz M. An examination of the relationship between Services Marketing Mix and Brand Equity Dimensions. Procedia - Social and Behavioral Sciences. 2014; 109: 865 – 869.

[2] Borden NH. The concept of the marketing mix. Journal of Advertising Research 1964; 4: 2–7.

[3] BalmerJMT. Corporate identity, corporate branding and corporate marketing—seeing through the fog. Eur J Mark. 2001; 35(3/4): 248–91.

[4] Bhattacharya CB, Sen S. Consumer–company identification: a framework for understanding consumers' relationships with companies. J Mark 2003; 67(2): 76–88.

[5] Bakewell C, Mitchell VW. Male versus female consumer decision making styles. Journal of Business Research. 2006; 59: 1297–1300.

[6] Bennett, C. Social Media's Role Will Soon Shift From Driving Awareness to Creating Revenue[Internet]. 2014 [updated 2014 Nov 19; cited 2014 Nov 27]. Available from: http://www.entrepreneur.com/article/239764

[7] Heding T, Knudtzen CF, &Bjerre M. Brand Management Research, theory and practice. Routledge, New York, NY. 2009.

[8] He H, Li Y, Harris L. Social identity perspective on brand loyalty. Journal of Business Research. 2012; 65: 648-657.

[9] Koo DM. The fundamental reasons of e-consumers' loyalty to an online store. Electronic Commerce Research and Applications. 2006; 5: 117-130.

[10] Krishnamurthy S. CASE: Mozilla vs. Godzilla—The launch of the Mozilla Firefox Browser. Journal of Interactive Marketing. 2009; 23(3): 259–271.

[11] KanlayaVanichbuncha. Statistic for research. Bangkok: Chulalongkorn University Bookshop. 2012.

[12] Kemp S. Social, Digital & Mobile Worldwide in 2014[online]. 2014[update 2014 Jan 23; cited 2014 Oct 3]. Available from: http://wearesocial.net/blog/2014/01/social-digital-mobile-worldwide-2014/

[13] McCarthy EJ, Brogowicz AA. Irwin series in marketing, Basic marketing: a managerial approach. R.D. Irwin. 1981).

[14] Madhavaram S, Badrinarayanan V, McDonald RE. Integrated marketing communication (IMC) and brand identity as critical components of brand equity strategy: a conceptual framework and research propositions. J Advert. 2005; 34(4): 69–80.

[15] Mouakket S, Al-hawari MA. Examining the antecedents of e-loyalty intention in an online reservation environment. Journal of High Technology Management Research. 2012; 23: 46-57.

[16] Park CH, Kim YG. Identifying key factors affecting consumer purchase behavior in an online

shopping context. International Journal of Retail and Distribution Management. 2003; 30 (1): 16–29.

[17] Schmitt BH, Pan Y. Managing corporate and brand identities in the Asia-Pacific region. Calif Manage Rev. 1994; 36(4): 32–48.

[18] Srinivasan SS, Anderson R, Ponnavolu K. Customer loyalty in e- commerce: an exploration of its antecedents and consequences. Journal of Retailing. 2002; 78 (1): 41–50.

[19] Shankar, VS, Smith AK, Rangaswamy A. Customer satisfaction and loyalty in online and offline environments. International Journal of Research in Marketing. 2003; 20 (2): 153–175.

[20] Souitaris V, Balabanis G. Tailoring Online Retail Strategies to Increase Customer Satisfaction andLoyalty. 2007; 40(2): 244-261

[21] Tian KT, Bearden WO, Hunter GL. Consumers' need for uniqueness: scale development and validation. J Consum Res. 2001; 28(1): 50–66.

[22] Yang, Z., Peterson, R.T., 2004. Customer perceived value, satisfaction, and loyalty: the role of switching costs. Psychology & Marketing. 2004; 21 (10): 799–822.

[23] Yeu CS, Leong KC, Tong LC, Hang S, Tang Y, Bashawir A, Subhan M. A comparative study on international marketing mix in China and India: The case of McDanal's. Procedia - Social and Behavioral Sciences. 2012; 65: 1054-1059

# Analyzing the dual long memory in Thailand stock market

Panupong Oonsupwilai[1], Pairote Satiracoo[2]* and Renato Alencar Adelino da Costa[3]
[1]*Department of Mathematics, Faculty of Science, Mahidol University, Thailand*
*Center of Excellence in Mathematics, Faculty of Science, Mahidol University, Thailand*
*eakaslowhand@hotmail.com*
[2]*Department of Mathematics, Faculty of Science, Mahidol University, Bangkok, Thailand*
*Center of Excellence in Mathematics, Faculty of Science, Mahidol University, Thailand*
*pairote.sat@mahidol.ac.th*
[3]*Department of Mathematics, Faculty of Science, Mahidol University, Bangkok, Thailand*
*Center of Excellence in Mathematics, Faculty of Science, Mahidol University, Thailand*
*renato.cos@mahidol.ac.th*

## Abstract

Over the past decades, a variety of models including ones taking into account autoregressive fractionally integrated effects and different distribution alternatives have been applied to forecast changes in stock market prices. Those recent studies, however, have neglected the presence of long memory in both returns and volatility, indicating that the present data depend on distant information which can be predictable. In this present study, the long memory property in returns and volatility of Thailand SET50 index, based on daily closing prices from June 17, 1996 to December 27, 2014 has been investigated. In addition, conditional variance models have been used to account for serial correlations in squared residuals. The preliminary results indicate that there is no evidence in long memory property in returns based on the ARFIMA model. The in-sample results also show that the volatility would be adequately modeled by using the ARMA-FIEGARCH, which proves the existence of long memory and asymmetry in volatility.

*Keywords*: Long memory; ARFIMA; ARFIMA-FIGARCH; ARFIMA-FIEGARCH; Thailand stock market; Efficient-market hypothesis

*Corresponding Author
E-mail Address: pairote.sat@mahidol.ac.th

## 1. Introduction

Many investors currently focus on investment in Thailand capital market or stock market due to financial market development, competitiveness and economic growth. Individual investors, proprietary investors, foreign investors and institutional investors invest in various securities. Especially retail investors expect greater profits from many markets such as securities, mutual funds and derivative instruments because their benefits are considerably better than deposit interest rates.

Thailand SET50 index has become underlying assets for SET50 index futures and SET50 index options contracts which were recently opened for trading on April 28, 2006 and October 29, 2007, respectively. Investors use options and futures contracts to earn profits and hedge their investments against loss. However, the common characteristic that all stock markets have in common is uncertainty, which is undesirable for the investors. Therefore, prediction or forecasting of SET50 Index is necessary for investors to reduce uncertainty and investment risks and assist in decision making.

The presence of long memory in financial time series has become a popular research topic in financial area [2, 6, 9, 15-17, 22]. Long memory in returns implies that the market value does not quickly respond to new information, but reacts to the historical information. If asset returns exhibit long memory, then the historical information can be used to predict future asset returns and ones can use this advantage to speculate the profits. On the other hand, long memory in volatility implies that risk is an important factor of the behavior of the asset prices.

To account for the long memory property in financial time series, Granger and Joyeux [11] and Hosking [14] proposed the class of fractionally integrated ARMA or ARFIMA model as a statistical approach to test the long memory.

Although both ARMA and ARFIMA models assume a constant variance, the autoregressive conditional heteroscedastic (ARCH) model proposed by Engle [10] can be used to model changes in volatility. The generalization of ARCH model, namely the general autoregressive conditional heteroskedasticity (GARCH) model was introduced by Bollerslev [5]. The error variance in the GARCH model is assumed to follow an autoregressive moving average model.

The existence of asymmetric volatility in market prices can be observed in many markets [1]. Asymmetric volatility phenomenon is a market dynamic in which stock returns volatility is larger when the market price is falling than when it is rising [4]. However, the standard GARCH model could not be used to describe asymmetric volatility. To account for asymmetric volatility, Nelson [19] proposed the exponential GARCH (EGARCH) model.

Ding et al. [9] discovered that autocorrelation coefficients of the squared returns of S&P 500 index decayed hyperbolically to zero, which showed strong evidence of long memory dynamics in volatility. In the case of long memory or persistence in volatility, Baillie et al. [3] introduced the fractionally integrated generalized autoregressive conditional heteroskedasticity (FIGARCH) to model the long memory in volatility of returns. Moreover, Bollerslev and Mikkelsen [6] developed the fractionally integrated exponential general autoregressive conditional heteroskedasticity (FIEGARCH) model by taking into account main features from both the FIGARCH and EGARCH models. The model provides greater flexibility for modeling asymmetry and long memory in volatility.

Kang and Yoon [15] examined the long memory of Japan, South Korea, Hong Kong and Singapore stock markets. The results showed evidence of the long memory property in returns by using the ARFIMA model. However, the standardized residuals and squared residuals still have serially correlation. Hence, the investigation of long memory in returns is not clear enough to present long memory property in stock market. Kasman and Torun [16-17] investigated dual long memory property in both returns and volatility of Turkish stock market. The results indicated that the long memory dynamics in both returns and volatility can be adequately estimated by the jointed ARFIMA-FIGARCH model.

A number of studies focus on forecasting changes in in Thailand stock market prices by means of conditional mean and conditional variance models. Supornjag J. [21] studied the fluctuation of rate of returns of stock index futures in the derivative markets using ARIMA-EGARCH model. The study investigated the stock index futures in 4 countries including Thailand, the United States of America, Japan and Hong Kong based on the time-series data of closing prices reported from April 28, 2007 to May 31, 2008. The results revealed that the ARIMA-EGARCH was the most appropriate model for forecasting the rate of return of stock index futures in Thailand stock index futures

Yawirach J. [23] investigated conditional volatilities of index returns in Thailand, Singapore, Malaysia, Indonesia and the Philippines stock markets by using the GARCH and FIGARCH models where the error terms were assumed to follow normal, student's-*t* and normal inverse Gaussian distributions. The results revealed that the squared standardized residuals were not serially correlation in both GARCH and FIGARCH, indicating the estimated GARCH and FIGARCH specifications were correctly specified. Furthermore, the FIGARCH model with non-normal error distributions performed better than GARCH and FIGARCH with the normal error distribution.

Although both ARIMA-EGARCH and FIGARCH are appropriate models for forecasting the rate of returns of Thailand stock market, the results of previous studies indicate the existence of long memory in both stock returns and volatility in some international markets [15-17]. Therefore, the aim of this study is to provide additional information on the long memory property in both returns and volatility of the Thailand SET50 index by using the ARFIMA-FIEGARCH model. The model provides a greater flexibility to analyze the long memory in returns and volatility and takes into account asymmetry in volatility. In addition, the study also considers the distributional properties of stock returns using the normal, skewed student's-*t* and general error distribution (GED).

The article is organized as follows: Section 2 presents the model framework. The data and empirical results are discussed in Section 3. Section 4 provides a conclusion.

## 2. Model framework

In this section, the models used to investigate long memory in index returns and volatility are provided.

### *2.1 Autoregressive fractionally integrated moving average model (ARFIMA)*

The ARFIMA model was introduced by Granger and Joyeux [11] and Hosking [14]. The model assumes the fractionally integrated process in the conditional mean. The ARFIMA $(p, \zeta, q)$ model can be expressed as a generalization of the ARIMA model as follow:

$$\psi_p(B)(1-B)^{\zeta}(r_t-\mu)=\theta_q(B)\varepsilon_t\,, \tag{1}$$

where $\psi_p(B)=1-\psi_1 B-\psi_2 B^2-...$ and $\theta_q(B)=1+\theta_1 B+\theta_2 B^2+...$ are the AR and MA polynomials with standing in outside of unit roots, respectively. $\mu$ is an unconditional mean and $\varepsilon_t$ is independently and identically distributed random variable with zero mean and variance $\sigma_t^2$. The fractional differencing term $(1-B)^{\zeta}$ is the binomial expansion

$$(1-B)^{\zeta}=\sum_{k=0}^{\infty}\binom{\zeta}{k}(-B)^k$$

$$=1-\zeta B-\frac{\zeta}{2!}(1-\zeta)B^2-... \tag{2}$$

The ARFIMA process exhibits the property of long memory when $\zeta\in(0,0.5)$. If $\zeta=0$, the ARFIMA process reduces to a short memory process, namely ARMA process.

### *2.2 Lo's modified R/S statistic*

Lo's modified R/S statistic was introduced by Lo [18]. It can be used to discriminate between long-range and short-range dependence. The Lo's modified R/S statistic $Q_n$ is defined as

$$Q_n = \frac{1}{\hat{\sigma}_n^2(q)} \left[ \max_{1 \le k \le n} \sum_{j=1}^{k} (r_j - \overline{r}_n) - \min_{1 \le k \le n} \sum_{j=1}^{k} (r_j - \overline{r}_n) \right] (3)$$

where

$$\hat{\sigma}_n^2(q) = \frac{1}{n}\sum_{j=1}^{n}(r_j - \overline{r}_n)^2 + \frac{2}{n}\sum_{j=1}^{q} w_j(q)\left[\sum_{i=j+1}^{n}(r_i - \overline{r}_n)(r_{i-j} - \overline{r}_n)\right]$$

and

$$w_j(q) = 1 - \frac{j}{q+1} \quad , \quad q < n$$

The null and alternative hypotheses of Lo's modified statistic test are

$H_0$ : no long memory and

$H_1$: there exists long term dependence.

The confidence intervals at various significance levels are given in Table1.

Table 1: Confidence intervals for Lo's modified R/S test

| Significance level | Confidence interval |
|---|---|
| 1% | (0.721,2.098) |
| 5% | (0.809,1.862) |
| 10% | (0.861,1.747) |

### *2.3 General autoregressive conditional heteroskedasticity model (GARCH)*

Bollerslev [5] introduced the general autoregressive conditional heteroskedasticity model. The GARCH*(p,q)* model can be written as follows:

$$r_t = \mu + \varepsilon_t$$
$$\varepsilon_t = z_t \sigma_t$$
$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$$
$$= \omega + \alpha(B)\varepsilon_t^2 + \beta(B)\sigma_t^2 \qquad (4)$$

where $r_t$ is the daily return, $\omega > 0$, $\alpha_i, \beta_j \ge 0$. $\varepsilon_t$ is independently and identically distributed random variable with zero mean and variance $\sigma_t^2$. $\alpha(B)$ and $\beta(B)$ are the lag polynomials of order q and p, respectively.

### *2.4 Exponential general autoregressive conditional heteroskedasticity model (EGARCH)*

The exponential general autoregressive conditional heteroskedasticity model proposed by Nelson [19] and Bollerslev and Mikkelsen [6] identifies the symmetric effect in conditional variance. The EGARCH*(p,q)* model can be written as follows:

$$r_t = \mu + \varepsilon_t$$
$$\varepsilon_t = z_t \sigma_t$$
$$g(z_t) = \gamma_1 z_t + \gamma_2(|z_t| - E|z_t|)$$
$$\ln \sigma_t^2 = \omega + (1-\beta(B))^{-1}(1+\alpha(B))g(z_{t-1}) \qquad (5)$$

for $\gamma_1 < 0$, the positive shocks generate less volatility than negative shocks. This case supports the asymmetric volatility phenomenon that explained by Black [4].

### *2.5 Fractionally integrated general autoregressive conditional heteroskedasticity model (FIGARCH)*

An extension of the ARFIMA representation in $\varepsilon_t{}^2$ leading to the FIGARCH model proposed by Baillie et al. [3]. The FIGARCH*(p,d,q)* can be expressed as follows:

$$r_t = \mu + \varepsilon_t$$
$$\varepsilon_t = z_t \sigma_t$$
$$(1-\beta(B))\sigma_t^2 = \omega + (1-\beta(B) - \phi(B)(1-B)^d)\varepsilon_t^2 \qquad (6)$$

where $(1-\beta(B))$ and $\phi(B)$ are the polynomials with standing in outside the unit root. $0 < d < 1$ is a fractional integrated parameter which characterizes the long memory property in volatility. If $d = 0$, Eq.(6) reduces to the GARCH model.

### *2.6 Fractionally integrated exponential general autoregressive conditional heteroskedasticity model (FIEGARCH)*

The fractionally integrated EGARCH (FIEGARCH) model proposed by Bollerslev and Mikkelsen [6] extends the asymmetric EGARCH model of Nelson [19] to long memory. The FIEGARCH*(p,d,q)* model can be described as

$$r_t = \mu + \varepsilon_t$$
$$\varepsilon_t = z_t \sigma_t$$
$$g(z_t) = \gamma_1 z_t + \gamma_2(|z_t| - E|z_t|)$$
$$\ln \sigma_t^2 = \omega + \phi(B)^{-1}(1-B)^{-d}(1+\alpha(B))g(z_{t-1}) \qquad (7)$$

If $-0.5 < d < 0.5$, the FIEGARCH model is covariance-stationary and invertible. In addition, long memory property will dissipate for all $d < 1$.

## 3. Data and empirical results

### *3.1 Data*

The data set used in this study includes 4544 daily closing observations for SET50 price index covering the period from June 17, 1996 to December 27, 2014 . The data set was obtained from the Thailand Stock Exchange as shown in Figure 1. The SET50 return on day *t* ( $r_t$ ) is defined by

$$r_t = 100 \cdot \ln(P_t / P_{t-1})$$

where $P_t$ is the closing price at day *t*.

Figure 2 plots SET 50 index return series. The descriptive statistics for SET50 daily returns are reported in Table 1. Both skewness and excess kurtosis statistics indicate that the return series tend to have a higher peak and fatter-tail than a normal distribution (see Figure 3). In addition, the Jarque-Bera test for normality significantly rejects the null hypothesis of a normal distribution.

The hypothesis of a white-noise process for the sample return series is examined based on the Ljung-Box Q-statistics of returns and squared returns. The results indicate that the null hypothesis of non-serial correlation of return and squared returns has been rejected. The return residuals and the squared return residuals fail to be independent and identically distributed process because the returns and squared residuals are highly correlated up to 5th and 10th lags.

Table 2: Descriptive statistics of SET50 daily returns series

| | SET50 |
|---|---|
| Sample size | 4543 |
| Mean | 0.000754041 |
| Standard deviation | 1.9068 |
| Maximum | 12.58858697 |
| Minimum | -17.23093467 |
| Skewness | 0.20937*** |
| Kurtosis | 9.73479*** |
| Jarque-Bera | 8619.0*** |
| Q(5) | 41.2781*** |
| Q(10) | 69.7516*** |
| $Q_s(5)$ | 853.879*** |
| $Q_s(10)$ | 1295.21*** |

Note : Jarque-Bera test for normality has a chi-squared distribution with two degrees of freedom. Q(q) and $Q_s(q)$ are the Ljung-Box Q-statistics of order q computed on returns and squared returns respectively.(***) denoted the significance levels at 1%

### *3.2 Empirical results*

The results of Augmented Dickey-Fuller (ADF) test are reported in Table 3. The test statistics for lag length 0 are calculated, based on the lowest value of Schwartz criterion. Since all test statistics are less than MacKinnon's critical values at 1%, the null hypothesis of non-stationarity of the SET50 return series should be rejected. Hence, the SET50 return series is stationary and can be used to analyze the long memory property in returns and volatility.

Table 3: ADF test result

| | Term included | Test statistic | McKinnon's critical value |
|---|---|---|---|
| SET50 returns | none | -61.9074 | -2.57 |
| | intercept | -61.9006 | -3.43 |
| | Intercept and time trend | -61.9503 | -3.96 |

Note: max length = 31

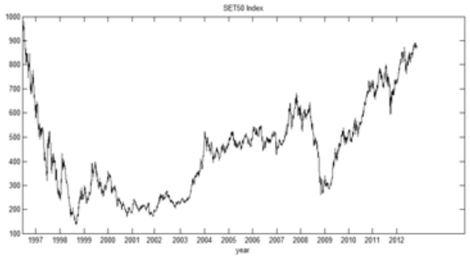


Figure 1:SET50 daily closing price from 17 June 1996 to 27 December 2014

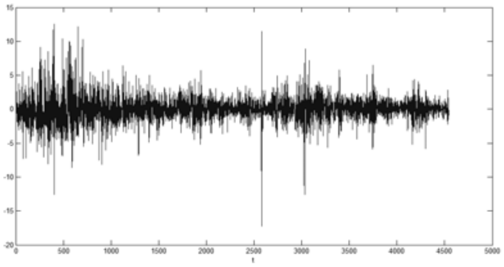


Figure 2:SET 50 index returns from 17 June 1996 to 27 December 2014

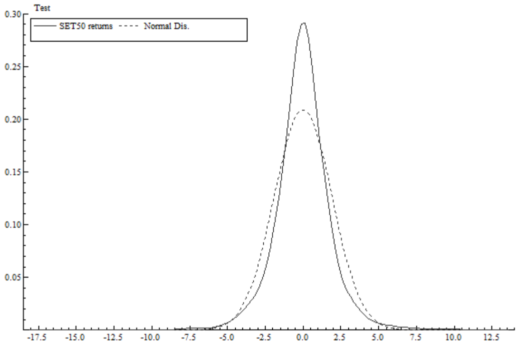


Figure 3:SET50 return distribution

### *3.3 Long memory in return*

The estimation results of the ARFIMA model under the assumption of normal distribution are reported in Table 4a and Table 4b. The lag order of $p$ = 0, 1, 2 and $q$ = 0, 1, 2 are used for the ARFIMA model as suggested by Chueng [8]. $ARFIMA(0,\zeta,2)$ model has been chosen for the SET50 index returns based on the Akaike Information Criterion (AIC). However, the long memory parameter $\zeta$ is not significantly different from zero, indicating that there is no evidence of long memory but short memory in index returns. The high values of the Jarque-Bera statistic indicate that the residuals tend to have a fat-tailed distribution. Ljung-Box Q-statistics of standardized residuals and squared residuals cannot reject the null hypothesis of

Table 4a: Estimation results for the ARFIMA models

| | ARFIMA(0, $\zeta$ ,0) | ARFIMA(0, $\zeta$ ,1) | ARFIMA(0, $\zeta$ ,2) | ARFIMA(1, $\zeta$ ,0) | ARFIMA(1, $\zeta$ ,1) |
|---|---|---|---|---|---|
| $\mu$ | -0.003727 | -0.001411 | 0.000431 | -0.001204 | -0.000051 |
| | (0.048433) | (0.040129) | (0.033605) | (0.038927) | (0.035290) |
| $\zeta$ | 0.063568*** | 0.035173 | 0.006066 | 0.030027 | 0.012658 |
| | (0.023786) | (0.032452) | (0.037929) | (0.036921) | (0.041139) |
| $\omega$ | 3.611547*** | 3.608289*** | 3.604886*** | 3.607852*** | 3.606682*** |
| | (0.15781) | (0.15683) | (0.15627) | (0.15669) | (0.15647) |
| $\psi_1$ | - | - | - | 0.052841 | 0.295491 |
| | | | | (0.048357) | (0.16749) |
| $\psi_2$ | - | - | - | - | - |
| $\theta_1$ | - | 0.045600 | 0.076307 | - | -0.224146 |
| | | (0.039986) | (0.049111) | | (0.16270) |
| $\theta_2$ | - | - | 0.040139 | - | - |
| | | | (0.029952) | | |
| Log-likelihood | -9363.156 | -9361.106 | -9358.96 | -9360.83 | -9360.09 |
| AIC | 4.123335 | 4.122873 | 4.122370 | 4.122752 | 4.122868 |
| Jarque-Bera | 8473.6*** | 8237.5*** | 8123.1*** | 8205.7*** | 8158.2*** |
| | (0.00000) | (0.00000) | (0.00000) | (0.00000) | (0.00000) |
| Q( 8) | 25.7076*** | 19.1725*** | 11.2907* | 18.1610** | 14.6169* |
| | [0.0011784] | [0.0076637] | [0.0797979] | [0.0112642] | [0.0234552] |
| $Q^2$( 8) | 1052.34*** | 1084.23*** | 1110.30*** | 1089.46*** | 1099.25*** |
| | [0.0000000] | [0.0000000] | [0.0000000] | [0.0000000] | [0.0000000] |
| ARCH(10) | 80.247*** | 83.496*** | 85.455*** | 83.989*** | 84.744*** |
| | [0.0000] | [0.0000] | [0.0000] | [0.0000] | [0.0000] |

Table 4b: Estimation results for the ARFIMA models (continue)

| | ARFIMA(0, $\zeta$ ,0) | ARFIMA(2, $\zeta$ ,0) | ARFIMA(2, $\zeta$ ,1) | ARFIMA(2, $\zeta$ ,2) |
|---|---|---|---|---|
| $\mu$ | 0.000093 | 0.000444 | 0.000194 | -0.000446 |
| | (0.034388) | (0.033779) | (0.034275) | (0.036455) |
| $\zeta$ | 0.009815 | 0.005722 | 0.008501 | 0.017980 |
| | (0.036725) | (0.039765) | (0.038401) | (0.047687) |
| $\omega$ | 3.604660*** | 3.605361*** | 3.605006*** | 3.604270*** |
| | (0.15623) | (0.15638) | (0.15628) | (0.15625) |
| $\psi_1$ | -0.143950 | 0.075930 | -0.132555 | -0.149598 |
| | (0.26696) | (0.050647) | (0.22895) | (0.87745) |
| $\psi_2$ | - | 0.031944 | 0.047353 | -0.146268 |
| | | (0.027312) | (0.034186) | (0.36007) |
| $\theta_1$ | 0.216237 | - | 0.205817 | 0.213444 |
| | (0.26459) | | (0.22909) | (0.83904) |
| $\theta_2$ | 0.048572 | - | - | 0.188654 |
| | (0.033851) | | | (0.37928) |
| Log-likelihood | -9358.82 | -9359.26 | -9359.04 | -9358.57 |
| AIC | 4.122747 | 4.122502 | 4.122843 | 4.123079 |
| Jarque-Bera | 8115.5*** | 8147.3*** | 8125.8*** | 8127.1*** |
| | (0.00000) | (0.00000) | (0.00000) | (0.00000) |
| Q( 8) | 11.5328* | 11.9897* | 11.8222* | 11.9094** |
| | [0.0417821] | [0.0621984] | [0.0373073] | [0.0180375] |
| $Q_s$( 8) | 1111.03*** | 1105.37*** | 1108.66*** | 1109.70*** |
| | [0.0000000] | [0.0000000] | [0.0000000] | [0.0000000] |
| ARCH(10) | 85.495*** | 85.040*** | 85.287*** | 85.471*** |
| | [0.0000] | [0.0000] | [0.0000] | [0.0000] |

Note: The numbers in the parentheses are quasi-maximum likelihood asymptotic standard error which corresponding parameter estimates. Q(10) are Ljung-Box Q-statistics of orders 10 computed on standardized residuals. (***), (**) and (*) denoted the significance levels at the 1%, 5% and 10% respectively. P-value statistics are reported in square brackets.

Table 5: Lo’s modified R/S test

| | Test statistic | Significant level |
|---|---|---|
| SET50 return | 1.46301 | - |

no serially correlation, indicating that the residuals are not independent. The ARCH statistics are highly significant to reject the null hypothesis of no-ARCH effects, indicating that there exist the ARCH effects in standardized residuals.

From the result of Lo’s modified R/S test in Table 5, the value of test statistic lies within the confidence interval at significant level of 1%, indicating the null hypothesis of no long memory in returns cannot be rejected. Hence, the Lo’s modified R/S statistic test supports the results of ARFIMA models in Table 4a and Table 4b that there is no evidence of long memory in SET50 index returns. The ARMA(0,2) model will be used in parameter estimation for the jointed ARMA-GARCH, ARMA-EGARCH, ARMA-FIGARCH and ARMA-FIEGARCH models.

### *3.4 Estimation results of jointed ARMA-GARCH class model*

In this section, we estimate each of the jointed *ARMA-GARCH* class models. Due to the high values of Jarque-Bera test of residuals, the residuals tend to be a fat-tailed distribution. Parameter estimation are performed for ARMA-GARCH, ARMA-EGARCH, ARMA-FIGARCH and ARMA-FIEGARCH under three distributions including the normal distribution, general error distribution (GED) and the skewed student’s-*t* distribution based on the quasi maximum likelihood estimation (QMLE) method. Estimation results for ARMA-GARCH, ARMA-EGARCH, ARMA-FIGARCH and ARMA-FIEGARCH are reported in Tables 6, 7, 8 and 9, respectively.

According to the results from each model, ARCH statistics strongly reject the null hypothesis of no ARCH effect. Consequently, all test models are well-specified.

The comparison of the performance of the ARMA-GARCH and ARMA-FIGARCH models is investigated. As shown in Table 6, the sum of the parameters $\alpha_1 + \beta_1$ in the ARMA-GARCH model under the three distributions are close to 1. Consequently, there exists high volatility persistence in returns. Furthermore, the estimates of the long memory parameter $d$ in ARMA-FIGARCH in Table 8 are significantly different from zero. It can be concluded that the ARMA-GARCH model should be rejected.

Comparing results of the ARMA-EGARCH and ARMA-FIEGARCH models in Table 7 and Table 8, the estimates of asymmetry parameter $\gamma_1$ in both of the ARMA-EGARCH and ARMA-FIEGARCH models are negative and significant at 1% level. This implies the existence of leverage effect in returns. The estimates of long memory parameter $d$ in ARMA-FIEGARCH are significantly different from zero, indicating that the ARMA-EGARCH should be rejected. Furthermore, ARMA-FIEGARCH performs slightly better than ARMA-EGARCH based on Akaike Information Criterion (AIC).

In addition, comparing results of ARMA-FIGARCH and ARMA-FIEGARCH, the estimates of long memory parameter $d$ in both of the ARMA-FIGARCH and ARMA-FIEGARCH models from Table 8 and Table 9 are greater than zero and significantly different from zero, indicating that long memory property is prevalent in volatility for all distributions. The asymmetric parameter $\gamma_1$ in ARMA-FIEGARCH are negative and significant at the 1% level, indicating that existence of leverage effect in volatility. Moreover, ARMA-FIEGARCH models perform better than ARMA-FIGARCH models. It can also be concluded that the ARMA-FIEGARCH model under fat-tailed distribution performs better than normal distribution based on AIC.

## 4. Conclusion

In this study, we have analyzed the dual long memory property of Thailand SET50 index. The results of the ARFIMA model and Lo’s modified R/S test show that there is no evidence of long memory property in SET50 index returns. It can be concluded that the market value at present time does not respond to historical information, but adjust quickly to new information, which follow the weak form efficient-market hypothesis (EMH). Due to serial correlation in squared residuals and ARCH effect, the conditional variance model is a suitable model for time varying volatility. Investigation of long memory property in volatility has also examined. The jointed ARMA-GARCH classes including ARMA-GARCH, ARMA-EGARCH, ARMA-FIGARCH and ARMA-FIEGARCH. The results also reveal that long memory and asymmetry in volatility can be adequately estimated by the joint ARMA-FIEGARCH model under fat-tailed distribution.

## 5. Acknowledgement

The authors would like to thank the Centre of Excellence in Mathematics, Mahidol University.

Table 6: Estimation result for the ARMA(0,2)-GARCH models

| | Normal | GED | Skewed |
|---|---|---|---|
| $\mu$ | 0.083287*** | 0.056056*** | 0.078409*** |
| | (0.023191) | (0.023178) | (0.021245) |
| $\theta_1$ | 0.078457*** | 0.049929*** | 0.061075*** |
| | (0.018239) | (0.018037) | (0.015269) |
| $\theta_2$ | 0.048370*** | 0.039360** | 0.044431*** |
| | (0.017380) | (0.018877) | (0.016002) |
| $\omega$ | 0.083827 | 0.046962*** | 0.036121*** |
| | (0.056038) | (0.017110) | (0.010006) |
| $\alpha_1$ | 0.117110*** | 0.118137*** | 0.118619*** |
| | (0.015555) | (0.014263) | (0.014269) |
| $\beta_1$ | 0.862934*** | 0.875335*** | 0.878420*** |
| | (0.023508) | (0.014714) | (0.013826) |
| $\nu$ | - | - | 7.147041*** |
| | | | (0.95160) |
| $\ln(k)$ | - | - | 0.056857*** |
| | | | (0.020329) |
| Log-likelihood | -8659.37 | -8495.06 | -8449.61 |
| AIC | 3.814824 | 3.742928 | 3.723361 |
| Q(10) | 18.7049 | 25.8305 | 20.9244 |
| | [0.0165201]** | [0.0011228]*** | [0.0073509]*** |
| ARCH(10) | 0.15673 | 0.12847 | 0.12362 |
| | [0.9987] | [0.9995] | [0.9995] |

Note: The numbers in the parentheses are quasi-maximum likelihood asymptotic standard error which corresponding parameter estimates. Q(10) are Ljung-Box Q-statistics of orders 10 computed on standardized residuals. (***),(**) and (*) denoted the significance levels at the 1%, 5% and 10% respectively. P-value statistics are reported in square brackets.

Table 7: Estimation result for the ARMA(0,2)-EGARCH models

| | Normal | GED | Skewed |
|---|---|---|---|
| $\mu$ | 0.063088*** | 0.036008*** | 0.053074*** |
| | (0.013646) | (0.010744) | (0.0094180) |
| $\theta_1$ | 0.069870*** | 0.046663*** | 0.056020*** |
| | (0.011187) | (0.015622) | (0.0085320) |
| $\theta_2$ | 0.041042** | 0.037796*** | 0.044095*** |
| | (0.016870) | (0.0061950) | (0.0092569) |
| $\omega$ | 1.291557*** | 0.893500*** | 1.805493*** |
| | (0.14115) | (0.13361) | (0.41410) |
| $\alpha_1$ | -0.288117 | -0.166294 | -0.109299 |
| | (0.19039) | (0.12973) | (0.12711) |
| $\beta_1$ | 0.972359*** | 0.978500*** | 0.980121*** |
| | (0.011550) | (0.0059562) | (0.0051023) |
| $\gamma_1$ | -0.064356* | -0.050684*** | -0.047924*** |
| | (0.034296) | (0.014259) | (0.010858) |
| $\gamma_2$ | 0.277607*** | 0.241234*** | 0.229235*** |
| | (0.065229) | (0.030873) | (0.024826) |
| $\nu$ | - | 1.345082*** | 7.266455*** |
| | | (0.10069) | (0.95910) |
| $\ln(k)$ | - | - | 0.057783*** |
| | | | (0.019907) |
| Log-likelihood | -8636.359 | -8479.768 | -8435.623 |
| AIC | 3.805573 | 3.737076 | 3.718082 |
| Q(10) | 23.5915 | 30.3802 | 25.0386 |
| | [0.0026823]** | [0.0001810]*** | [0.0015313]*** |
| ARCH(10) | 0.056643 | 0.066206 | 0.071486 |
| | [1.0000] | [1.0000] | [1.0000] |

Note: The numbers in the parentheses are quasi-maximum likelihood asymptotic standard error which corresponding parameter estimates. Q(10) are Ljung-Box Q-statistics of orders 10 computed on standardized residuals. (***),(**) and (*) denoted the significance levels at the 1%, 5% and 10% respectively. P-value statistics are reported in square brackets.

Table 8: Estimation result for the ARMA(0,2)-FIGARCH models

| | Normal | GED | Skewed |
|---|---|---|---|
| $\mu$ | 0.073230*** | 0.050991** | 0.073442*** |
| | (0.023153) | (0.023512) | (0.021228) |
| $\theta_1$ | 0.077307*** | 0.049520*** | 0.060320*** |
| | (0.019121) | (0.017234) | (0.015485) |
| $\theta_2$ | 0.049443*** | 0.041159** | 0.047083*** |
| | (0.016614) | (0.018169) | (0.016265) |
| $\omega$ | 0.391221 | 0.172926 | 0.096999** |
| | (0.30661) | (0.11960) | (0.049284) |
| $d$ | 0.303343*** | 0.366827*** | 0.397777*** |
| | (0.060084) | (0.059958) | (0.063503) |
| $\phi_1$ | -0.175741 | -0.016650 | 0.041492 |
| | (0.15769) | (0.23366) | (0.15861) |
| $\beta_1$ | -0.025218 | 0.226391 | 0.327961 |
| | (0.21738) | (0.28269) | (0.20700) |
| $\nu$ | - | 1.346083*** | 7.857362*** |
| | | (0.10853) | (1.1235) |
| $\ln(k)$ | - | - | 0.052843*** |
| | | | (0.020090) |
| Log-likelihood | -8647.884 | -8485.578 | -8437.447 |
| AIC | 3.810207 | 3.739194 | 3.718445 |
| Q(10) | 19.1455 | 26.4389 | 21.3406 |
| | [0.0141006]** | [0.0008832]*** | [0.0062957]*** |
| ARCH(10) | 0.073544 | 0.092538 | 0.10018 |
| | [1.0000] | [0.9999] | [0.9998] |

Note: The numbers in the parentheses are quasi-maximum likelihood asymptotic standard error which corresponding parameter estimates. Q(10) are Ljung-Box Q-statistics of orders 10 computed on standardized residuals. (***),(**) and (*) denoted the significance levels at the 1%, 5% and 10% respectively. P-value statistics are reported in square brackets.

Table 9: Estimation result for the ARMA(0,2)-FIEGARCH models

| | Normal | GED | Skewed |
|---|---|---|---|
| $\mu$ | 0.061300** | 0.033917 | 0.044931** |
| | (0.025260) | (0.023526) | (0.022208) |
| $\theta_1$ | 0.068774*** | 0.046488*** | 0.054056*** |
| | (0.018369) | (0.013643) | (0.015371) |
| $\theta_2$ | 0.039055** | 0.037613** | 0.044357*** |
| | (0.017021) | (0.015986) | (0.016032) |
| $\omega$ | 1.431293*** | 0.767677*** | 1.574466* |
| | (0.22055) | (0.28795) | (0.82682) |
| $d$ | 0.440988*** | 0.519917** | 0.497402** |
| | (0.099513) | (0.22807) | (0.19850) |
| $\alpha_1$ | -0.521558 | -0.175575 | -0.131452 |
| | (0.32626) | (1.9352) | (1.2833) |
| $\phi_1$ | 0.784028*** | 0.625771 | 0.674657 |
| | (0.17405) | (1.2153) | (0.72130) |
| $\gamma_1$ | -0.086112** | -0.064814*** | -0.057415*** |
| | (0.042496) | (0.020588) | (0.012457) |
| $\gamma_2$ | 0.279648*** | 0.234750*** | 0.217630*** |
| | (0.071976) | (0.047690) | (0.041667) |
| $\nu$ | - | 1.364238*** | 7.768329*** |
| | | (0.10934) | (1.2173) |
| $\ln(k)$ | - | - | 0.031894 |
| | | | (0.034186) |
| Log-likelihood | -8611.501 | -8464.727 | -8420.800 |
| AIC | 3.795070 | 3.730895 | 3.711997 |
| Q(10) | 23.0236 | 29.8889 | 25.3218 |
| | [0.0033341]*** | [0.0002211]*** | [0.0013708]*** |
| ARCH(10) | 0.058528 | 0.062664 | 0.072670 |
| | [1.0000] | [1.0000] | [1.0000] |

Note: The numbers in the parentheses are quasi-maximum likelihood asymptotic standard error which corresponding parameter estimates. Q(10) are Ljung-Box Q-statistics of orders 10 computed on standardized residuals. (***),(**) and (*) denoted the significance levels at the 1%, 5% and 10% respectively. P-value statistics are reported in square brackets.

## References

[1] Alberg DH, Shalit R, Yosef R. Estimating stock market volatility using asymmetric GARCH models. Applied Financial Economics. 2008; 18: 1201-1208.

[2] Baillie RT. Long memory processes and fractional integration in economics. Journal of Econometrics. 1996; 73: 5-59.

[3] Baillie RT, Bollerslev T, Mikkelsen HO. Fractionally integrated generalized autoregressive conditional heteroskedasticity. Journal of Econometrics. 1996; 74: 3 – 30.

[4] Black F. Studies of stock price volatility changes, Proceedings of the 1976 Meetings of the American Statistical Association. Business and Economic Statistics. 1976. p. 177-181.

[5] Bollerslev T. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics. 1986; 31: 307-327.

[6] Bollerslev T, Mikkelsen HO. Modeling and pricing long memory in stock market volatility. Journal of Econometrics. 1996; 73: 151-184.

[7] Bowden N, Payne JE. Short term forecasting of electricity price for MISO hubs: evidence from ARIMA-EGARCH models. Energy Economics. 2008; 30: 3186-3197.

[8] Cheung YW. Test for fractional integration: a Monte Carlo investigation. Journal of Time Series Analysis. 1993; 14: 331-345.

[9] Ding Z, Granger CWJ, Engle RF. A long memory property of stock market returns and a new model. Journal of Empirical Finance. 1993; 1: 83-106.

[10]Engle RF. Autoregressive conditional heteroscedasticity with estimates of the variance of the United Kingdom inflation. Econometrica. 1982; 50(4): 987-1007.

[11] Granger CWJ, Joyeux R. An introduction to long-memory time series models and fractional differencing. Journal of Time Series Analysis. 1980; 1: 15-29

[12] Hassan M, Lixian Su. International evidence on crude oil price dynamics: applications of ARIMA-GARCH models. Energy Economics. 2010; 32: 1001-1008.

[13] Heping Liu, Jing Shi. Apply ARIMA-GARCH approaches to forecasting short-term electricity price. Energy Economics. 2013; 37: 152-166.

[14] Hosking JRM. Fractional differencing. Biometrika. 1981; 68(1): 165-176.

[15] Kang SH, Yoon SM. Asymmetric long memory feature in the volatility of Asian stock markets. Asia-Pacific Journal of Financial Studies. 2006; 35(5): 175-198.

[16] Kasman A, Torun E. Long memory in the Turkish stock market return and volatility. Central Bank Review. 2007; 2: 13-27.

[17] Kasman A, Kasman S, Torun E. Dual long memory property in returns and volatility: evidence from the CEE countries' stock markets. Emerging Markets Review. 2009; 10: 122-139.

[18] Lo AW. Long-term memory in stock market price. Econometrica. 1991; 59(5): 1279-1313.

[19] Nelson DB. Conditional heteroskedasticity in asset returns: a new approach. Econometrica. 1991; 59(2): 347 – 370.

[20] Shamiri A, Hassan A. Modeling and forecasting volatility of the Malaysian and the Singaporean stock indices using asymmetric GARCH models and non-normal densities. Malaysian Journal of Mathematical Science. 2007; 1: 83-102.

[21] Supornjag J. Volatility analysis of rate of returns on futures market price index in Thailand, the United States of America, Japan and Hong Kong [Dissertation]. Chiang Mai: Chiang Mai Univ; 2009.

[22] Ural M, Kucukozmen CC. Analyzing the dual long memory in stock market returns. Ege Academic Review. 2011; 11: 19-28.

[23] Yawirach J. Conditional volatility and price index returns distribution of Thailand, Singapore, Malaysia, Indonesia and the Philippines using GARCH and FIGARCH Models with normal inverse Gaussian distribution [Dissertation]. Chiang Mai: Chiang Mai Univ; 2009.

[24] Xiaodong Lv, Xian Shan. Modeling natural gas market volatility using GARCH with different distributions. Physica A. 2009; 392: 5685 – 5699.

# Modelling impatient length of stay through mixture distributions. The case of HIV/AIDS length of stay in Portuguese hospitals

Martins, Maria Rosário O. [1*], Dias, Sara S.[2,3]

[1]*GHTM- Instituto de Higiene e Medicina Tropical da Universidade Nona de Lisboa, Portugal, mrfom@ihmt.unl.pt*

[2]*CEDOC – Faculdade de Ciências Médicas da Universidade Nova de Lisboa, Portugal, sara.dias@fcm.unl.pt*

[3]*UIS – ESSLeiria, IPLeiria, Leiria, Portugal*

## Abstract

Inpatient length of stay (LOS) is an important measure of hospital activity, but its empirical distribution is often positively skewed, representing a challenge for statistical analysis. Taking this feature into account, we seek to identify factors that are associated with HIV/AIDS length of stay in Portuguese hospitals through a hierarchical finite mixture model. We used data on adult HIV/AIDS diagnosis-related group data (DRG) in two periods: before and after the economic crisis. The model accounts for the demographic and clinical characteristics of the patients, as well the inherent correlation of patients clustered within hospitals (random effects). We also included a variable to take into account the possible effect of the economic crisis in impatient length of stay. It was found that a Gaussian mixture model with two-components had the best fit, resulting in two subgroups of LOS: a short-stay subgroup and a long-stay subgroup. Associated risk factors for both groups were identified as well as some statistical differences in the hospitals. Within the actual context of severe resource constraints in Portuguese public health sector, our findings provide important information for policy makers in terms of discharge planning and the efficient management of LOS. The presence of "atypical" hospitals also suggests that hospitals should not be viewed or treated as homogenous bodies.

*Keywords*: mixture models, random effects, hospital length of stay, modelling, HIV/AIDS

*Corresponding Author

E-mail Address: mrfom@ihmt.unl.pt

## 1. Introduction

Hospitalizations are an important component of the costs of providing health care in governments; specifically the study of length of stay (LOS) is essential for the management, budgeting and funding hospitals. In Portugal, is possible to obtain LOS data from the diagnosis related groups (DRG) classification since their implementation in 1989. The DRG determines the payment allocated to the hospital, and is based on the characteristics of patients consuming similar quantities of resources, as a result of a process of a similar care. The main assumption in funding is that patients with very long inpatient LOS have different resource consumption patterns from those assumed to have an usual LOS. For these reasons, it is crucial to understand and model the distribution of LOS.

Skewness and heterogeneity of LOS represents a challenge for statistical analysis (1). In recent literature, several approaches have been adopted to analyse LOS, such as survival models (2-4), frequentist and Bayesian frameworks (5), and latent class models (6, 7). However, none of these approaches recognized that hospitalizations from the same hospital are more likely to be related. Neglecting the dependence of clustered (multilevel) data may result in spurious associations and misleading inferences. Some authors have attempted to accommodate this risk (8, 9), although not addressing the issues of skewness or the heterogeneity of LOS. Recognizing that there may be subgroups of patients regarding LOS and the multilevel structure of the DRG data, we propose a hierarchical modelling approach to overcome the challenge derived from these two features. The model includes variables at both levels (patients and hospitals) which allows us to estimate differences in outcome that are not fully explained by observed patient or other specific and known conditions.

## 2. Hierarchical finite mixture model

LOS data are skewed and contain atypical observations. For this reason a mixture distribution with two components will be performed to model LOS data.

This analysis resulted from a previous work (10, 11) considering a maximum of 3-components distribution where a model with two components had the best fit. The suitable number of components was chosen via Akaike (AIC) and Bayesian (BIC) information criteria calculated for distributions with different numbers of components.

After the analysis of the LOS distribution, and taking into account two types of heterogeneity (within patients and among hospitals), a hierarchical finite mixture model will be fitted.

Consider $y_{ij}$ ($i = 1, ...,m$; $j = 1, ...,n_i$) the logarithm of LOS for the $j^{th}$ patient in the $i^{th}$ hospital, where $m$ is the number of hospitals, $n_i$ is the number of patients within hospital $i$ resulting in $N = \sum_{i=1}^{m} n_i$ total patients. A finite mixture model (12, 13) for the probability density function of $y_{ij}$ takes the form:

A finite mixture model (12, 13) for the probability density function of $y_{ij}$ takes the form:

$$f\left(y_{ij} \mid x_{ij};\Theta\right) = \sum_{k=1}^{g} \pi_k f_k\left(y_{ij} \mid x_{ij};\beta_k,\phi_k\right) \qquad (1)$$

where $\pi_k$ denotes the proportion of patients belonging to the $k^{th}$ component and $\sum_{k=1}^{K} \pi_k = 1$, $K$ is the number of components in the mixture; $f_k$ describes the normal distribution of the $k^{th}$ component with mean expressed by,

$$\mu_{k,ij} = x_{k,ij}\beta_k + \phi_{ki} \qquad (2)$$

where $x_{k,ij}$ is the covariate matrix for the $k^{th}$ component, $\beta_k$ is the vector of linear effect regression parameters that can be different for the $K$ components; $\phi_{ki}$ is the random effect parameters that capture the hospital heterogeneity in each component through the specification of a Gaussian density function with zero mean and variance-covariance equal to $\sigma_{\phi_k}^2 \mathrm{I}_m$ ($\mathrm{I}_m$ denotes an $m$x$m$ identity matrix).

The vector $\Theta$, which contains all the unknown parameters in model (1), is estimated by maximum likelihood approach through the application of EM algorithm (14). Once $\Theta$ has been estimated, estimates of the posterior probabilities $\left(\hat{\tau}_k\right)$ of population membership can be formed by:

$$\hat{\tau}_k\left(y_{ij};\hat{\Theta}\right) = Pr\left(y_{ij} \in g \mid x_{ij};\hat{\Theta}\right) = \frac{\hat{\pi}_k f_k\left(y_{ij} \middle| x_{ij;}\hat{\beta}_k,\hat{\phi}_{ki}\right)}{\sum_{k=1}^{g} \hat{\pi}_k f_k\left(y_{ij} \middle| x_{ij;}\hat{\beta}_k,\hat{\phi}_{ki}\right)}$$

$$(k = 1,\dots,g) \qquad (3)$$

In order to form no overlapping components (clusters), each $y_{ij}$ will be assigned to the population that has the highest estimate posterior probability. All of the statistical analyses were performed using the statistical software R (15) and its package Flexmix (16, 17).

## 3. Application to HIV/AIDS LOS

This study is based on HIV/AIDS DRG 714 (infection with human immunodeficiency virus, with significant diagnosis related), occurred between 2009 and 2013, which represent the DRG with more hospitalizations amongst the 17 DRG created for HIV/AIDS patients.

### *3.1 Data*

The DRG database were provided by the Central Health System Administration (ACSS). All the hospitalizations meeting the following criteria were analysed: patients aged 18 years or older, hospitalizations from hospitals with more than 10 discharge episodes and transfers to another hospital were eliminated (to avoid including the inpatient episode twice, as the cause of the transfer was often lack of procedure. facilities). The above selection criteria resulted in 5,548 hospitalizations in 40 hospitals.

The outcome variable was the logarithm of the number of days between the hospital admission and discharge dates. Patient's demographic characteristics (age, gender and death), health relevant factors (urgent admission, number of secondary diagnoses, number of procedures, AIDS as principal diagnosis, and presence of pneumonia) and crisis (a dummy variable that has the value 1 for the years 2012 and 2013, reflecting the period of economic crisis in Portugal) were considered in the analysis.

### *3.2 Results*

Estimation results for the two component model are shown in Table 1. The first component can be understood as the latent short-stay group and the second component the long-stay group. As can be seen in Table 1, some differences exist between covariate effects for short-stay and long-stay groups. The estimated proportion for the short-stay is greater (78%) than the corresponding value for long-stay (22%). Although gender is not statistically significant at 5% level, it is retained in the model in order to control for possible confounding. At the 5% level, Age is only statistically significant for the second component; death at the end of hospitalization, number of procedures, pneumonia and AIDS as principal diagnosis are statistically significant in both components. Urgent admission is only significant at 10% level in both components. Number of secondary diagnoses and crisis are relevant variables, at the 5% level, but only in the first component. All the coefficients have the expected signs.

We also plot hospital random effects in order to explore differences across hospitals not measured directly by the model. These include unmeasurable hospitals characteristics as medical expertise, health care, health management among others (Figure 1). There are thirteen hospitals more efficient and six hospitals less efficient than the mean.

### *3.3 Discussion*

In austerity context, the growth of public expenditure on health increases the need for cost reduction by gains in efficiency and the elimination of waste. For this reason the Portuguese Court of Auditors published an audit report about the financing system of the public hospitals belonging to the Portuguese National Health Service. They concluded that the methodology used to analyse DRG LOS is insufficiently grounded and is not the most appropriate (18). One of the recommendations relative to the duration of the hospitalizations was that the finite mixture models should be an alternative against the use of one component distribution.

This paper presents a finite mixture model to analyse HIV/AIDS LOS which emphasizes the heterogeneity of the DRGs by specifying two latent subgroups. Moreover a dummy variable to explore the possible effect of economic crisis in LOS is also included.

The information of the relevant covariates is important for analysing the LOS similarities and dissimilarities between the short-stay and long-stay latent subgroups. Different sets of significant factors for each latent subgroup provides useful guidance to hospital managers with respect to economic decisions such as budgeting expenditures in targeted groups (19).

The discharges variables included in the model behaved as expected. Patient's gender is not statistically significant and age is only significant in the second component in explaining HIV/AIDS LOS, revealing that older patients tend to be hospitalized for longer periods (20), holding other variables unchanged. Adjusting for other factors, deaths are associated with the early days of the hospitalization in both subgroups, meaning that there is high mortality among those patients who arrive at hospital in more severe and advanced states of AIDS-related illness. In Portugal, urgent admissions do not necessarily reflect emergency situations, as it is common that patients seek attendance at a hospital emergency room due to difficulties in making an appointment elsewhere. Holding other variables unchanged, in our study this type of admission contributed to a prolongation of the hospitalization in short-stay latent subgroup and to a reduction in the long-stay latent subgroup, reflecting the hospitalization behaviour actually occurring in Portugal, although this variable is only significant at 10% level.

The number of secondary diagnoses and number of procedures have a positive coefficient, indicating as presumed, a long-duration hospitalization, holding other variables unchanged. A great number of diagnoses or procedures usually indicates a more severe condition of the patient and consequently, a delayed discharge (7, 21). Adjusting for other variables, Pneumonia as a co-morbidity reduces the duration of the hospitalization in both latent subgroups; when AIDS is identified as the principal diagnosis, the hospitalizations tend to be prolonged in both subgroups, as these patients are in more acute situations and more likely to need a longer in-hospital treatment. Everything else being constant and for short-stay latent subgroup (representing 78% of the hospitalizations), we found that in economic crises period, HIV/AIDS LOS has been reduced, suggesting efficiency gains. This increased efficiency has been achieved namely by the restructuring of the primary care network (health family health units); the restructuring of the secondary care network (hospitals); and the development of the integrated continued care network. Further operational highlights include the reinforcement of the strategic planning processes of hospitals and the use of shared services (National Health Plan 2011-16).

However, the hierarchical finite mixture modelling framework presented in our study suggests differences among hospitals, with some of them being more performant than other; this result needs more in depth research. In DRG data, patients are nested within hospitals on the basis of their own choices (place of residence, trust in a particular doctor, the hospital's reputation, etc.), thus violating the independence assumptions of classical regression analysis. As a result, hierarchical modelling is strongly advocated as a more appropriate statistical method for dealing with multilevel structured data, such as patients clustered within hospitals (22-25). Moreover, the hospital random effects, which acknowledge unmeasured factors that are nonetheless important, should be interpreted as differences in hospital quality/performance.

Table 1: Estimates of two-component hierarchical model for HIV/AIDS DRG 714

| Parameter | $1^{st}$ component | | | $2^{nd}$ component | | |
|---|---|---|---|---|---|---|
| | Coefficient | SE | p-value | Coefficient | SE | p-value |
| Male | -0.035 | 0.026 | 0.179 | 0.054 | 0.068 | 0.428 |
| Age/10 | 0.003 | 0.009 | 0.784 | 0.054 | 0.024 | 0.022 |
| Death | -0.252 | 0.043 | <0.001 | -0.878 | 0.103 | <0.001 |
| Urgent | 0.066 | 0.035 | 0.006 | -0.159 | 0.083 | 0.055 |
| Nº diagnoses | 0.049 | 0.005 | <0.001 | 0.012 | 0.008 | 0.152 |
| Nº procedures | 0.106 | 0.004 | <0.001 | 0.085 | 0.009 | <0.001 |
| Pneumonia | -0.073 | 0.027 | 0.007 | -0.312 | 0.072 | <0.001 |
| AIDS | 0.131 | 0.029 | <0.001 | 0.288 | 0.062 | <0.001 |
| Crisis | -0.108 | 0.026 | <0.001 | -0.005 | 0.060 | 0.935 |
| Random effect variance | | 0.048 | | | 0.094 | |
| Mixture proportion | | 0.780 | | | 0.220 | |
| AIC | | 13,520.28 | | | | |
| BIC | | 13,685.81 | | | | |

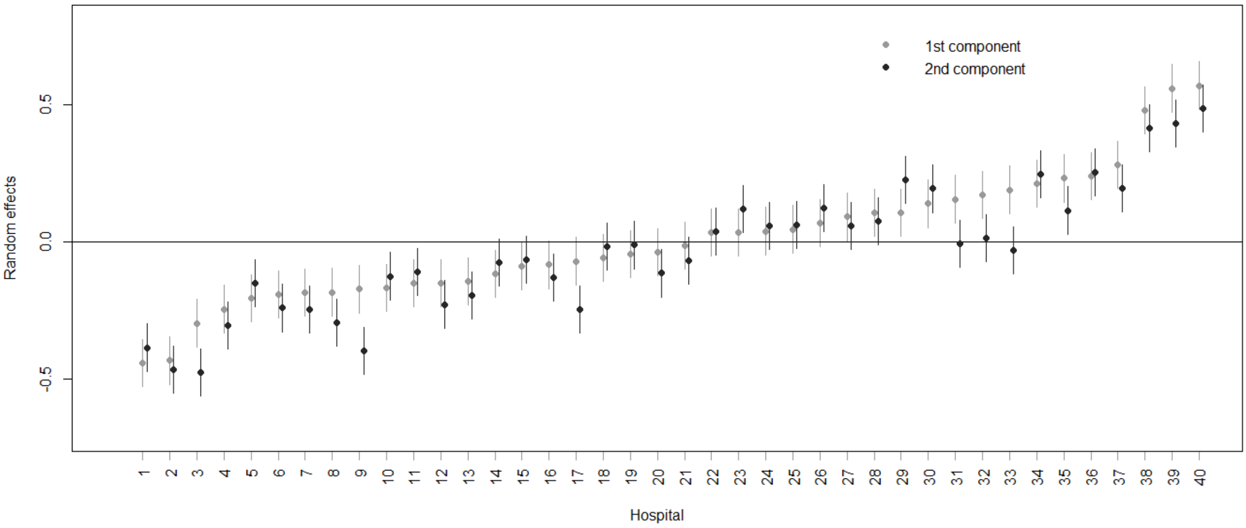


Figure 1: Hospital random effects

## 4. Conclusions

Length of stay in hospitals is often used as an indicator of efficiency. Reducing the time spent in hospital reduces the cost per patient and allows more patients to be treated in a given period. Moreover, a shorter stay in hospital allows treatment to be shifted from expensive inpatient care to less expensive outpatient care. The present study analyses HIV/AIDS LOS in Portugal, a diseases which reveals elevated figures for costs and bed-days when compared to worldwide data.

This analysis is imperative in comparing different hospitals, patients, and places in terms of costs, hospital activity, and efficiency. The proposed methodology is particularly useful within this context, suggesting the presence of a finite mixture of two subpopulations in the 714 HIV/AIDs related DRG. This stands in contrast to linear regression (one component model), which imposes a common average effect of LOS predictors for short-stay and long-stay latent subgroups, thereby leading to misdirected interpretations. Moreover, accounting for clustered observations our framework provides correct inferences about the regression coefficients (β). It also gives estimates of β which are efficient under some regularity conditions.

Our findings are of great interest namely for discharge planning and efficient management of LOS. For healthcare policy purposes, the identification of “atypical” hospitals should caution policymakers not to regard all hospitals equally. Appropriate policies can be developed to manage the hospital care and its resources, as well as promote the early prediction of HIV/AIDS patients requiring a longer period of hospitalization, and the higher costs thus incurred. Costs and funding formulae may also be adjusted according to the relevant factors influencing HIV/AIDS. The results obtained are also very useful for evaluating the impact of policies related to economic crisis.

As a whole, our findings confirm the multifactor nature of influence of health on LOS, and provide advice to the policymaker who is looking for easy ways to reduce NHS hospital expenses.

The drawback to this study resides in the use of the LOS logarithm transformation, which in future research will be analysed in the original scale, considering a finite mixture model with distributions belonging to the exponential family such as gamma or Poisson including random effects.

## References

[1]. Lee A.H., Gracey M., Wang K., Kelvin K.W. A robustified modeling approach to analyze pediatric length of stay. Annals of Epidemiology. 2005; 15(9): 637-77.

[2]. Dias S.S., Andreozzi V., Martins M.O., Torgal J. Predictors of mortality in HIV-associated hospitalizations in Portugal: a hierarchical survival model. BMC Health Services Research. 2009;9.

[3]. Pérez-Hoyos S., Ballester F., Tenías J.M., Marelles A., Rivera M.L. Length of stay in a hospital emergency room due to asthma and chronic obstructive pulmonary disease: Implications for air pollution studies. European Journal of Epidemiology. 2000; 16: 455-63.

[4]. Nordström P., Gustafson Y., Michaëlsson K., Nordström A. Length of hospital stay after hip fracture and short term risk of death after discharge: a total cohort study in Sweden. BMJ 2015-02-20 13:24:46.

[5]. Saez-Castillo A.J., Olmo-Jimenez M.J., Sanchez J.M.P., Hernandez M.A.N., Arcos-Navarro A., Diaz-Oller J. Bayesian Analysis of Nosocomial Infection Risk and Length of Stay in a Department of General and Digestive Surgery. Value in Health. 2010; 13(4): 431-9.

[6].Singh C.H., Ladusingh L. Inpatient length of stay: a finite mixture modeling analysis. European Journal of Health Economics. 2010; 11(2): 119-26.

[7].Xiao J., Lee A.H., Vemurri S.R. Mixture distribution analysis of length of stay for efficient funding. Socio-Economic Planning Sciences. 1999; 33: 39-59.

[8].Leyland A.H., Boddy F.A. Measuring performance in hospital care - Length of stay in gynaecology. European Journal of Public Health. 1997; 7(2): 136-43.

[9].Leung K.M, Elashoff R.M., Rees K.S., Hasan M.M., Legorreta AP. Hospital- and patient-related characteristics determining maternity length of stay: A hierarchical linear model approach. American Journal of Public Health. 1998; 88(3): 377-81.

[10]. Dias S.S., Andreozzi V., Martins M.R.O. Analysis of HIV/AIDS DRG in Portugal: a Hierarchical Finite Mixture Model. European Journal of Health Economics; 2012.

[11]. Dias S.S., Andreozzi V., Martins M. Hierarchical Normal Mixture Model to Analyse HIV/AIDS LOS. In: Pacheco A, Santos R., Oliveira M.dR., Paulino C.D., editors. New Advances in Statistical Modeling and Applications: Springer International Publishing; 2014: 219-27.

[12]. Everitt B.S, Hand D.J. Finite Mixture Distribution. London: Chapman and Hall; 1981.

[13]. McLachlan G, Peel D. Finite Mixture Models. Statistics WSiPa, editor: John Wiley & Sons; 2000.

[14]. Dempster A.P., Laird N.M., Rubin DB. Maximum likelihood from incomplete data via EM algorithm. Journal of the Royal Statistical Society Series B-Methodological. 1977; 39(1): 1-38.

[15]. Team R.D.C. R: A language and Environment for Statistical Computing. 2.8.0 ed. Vienna, Austria: The R foundation for statistical computer; 2008.

[16]. Leisch F. FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. Journal of Statistical Software. 2004; 11(8): 1-18.

[17]. Grun B., Leisch F. Fitting finite mixtures of generalized linear regressions in R. Computational Statistics & Data Analysis. 2007; 51(11): 5247-52.

[18]. Tribunal, de, Contas, Portugal. Auditoria ao sistema de pagamentos e de formação dos preços pagos às unidades hospitalares do Serviço Nacional de Saúde Lisboa; 2011.

[19]. Ng S.K., Yau K.K.W., Lee A.H. Modelling inpatient length of stay by a hierarchical mixture regression via the EM algorithm. Mathematical and Computer Modelling. 2003; 37(3-4): 365-75.

[20]. Crystal S, Lo Sasso A.T., Sambamoorthi U. Incidence and duration of hospitalizations among persons with AIDS: An event history approach. Health Services Research. 1999; 33(6): 1611-38.

[21]. Wang K., Yau K.K.W., Lee A.H. A hierarchical Poisson mixture regression model to analyse maternity length of stay. Statistics in Medicine. 2002; 21: 3639-54.

[22]. Greenland S. Principles of multilevel modelling. International Journal of Epidemiology. 2000; 29: 158-67.

[23]. Diez-Roux A.V. Multilevel analysis in public health research. Annual Review of Public Health. 2000; 21: 171-92.

[24]. Leyland A., Goldstein H. Multilevel modelling of health statistics. New York: Wiley; 2001.

[25]. Bingenheimer J.B., Raudenbush S.W. Statistical and substantive inferences in Public Health: Issues in the Application of Multilevel Models. Annual Review of Public Health. 2004; : 53-77.

# Evaluating Beaches for Tourism Business Opportunity

Peerapat Chankerd*, Sorawit Sangpotirat, Dheeruth Leelawat, Panpawee Chavandij, and Morrakot Raweewan
*School of Management Technology, Sirindhorn International Institute of Technology,*
*Thammasat University, PathumThani, 12000, THAILAND*
*Email Address: peerapat7654@gmail.com*

## Abstract

This study evaluates tourism business opportunity for beaches where behavior and satisfaction of tourists concurrently with environment are first assessed. The locations chosen are Pattaya, Bang Saen, Sor Beach and Toey Ngam Beach located 200-300 kilometer away from Bangkok. While Pattaya and Bang Saen are popular already, Sor Beach and Toey Ngam Beach are not fully developed for tourism business yet. The target groups of Thai and foreign tourists are randomly sampled; 385 of them answer a questionnaire. The questionnaires are collected from four beaches for both tourism part and environment part. However, the beach with 1) good environment but low number of tourists and 2) poor environment but high number of tourists will be focused in order to point out the problems and recommendation for improvement of tourism management and environmental management. Morgan and Statistic Package for the Social Sciences (SPSS) program were used. Morgan is a tool with 40 variables to measure and classify the quality of the beach environment. Evaluated by 1) General information of the tourists, 2) Tourist behavior, 3) Satisfaction of the tourists, and 4) Environment of the beaches, SPSS identifies significant difference between developed and undeveloped beaches.

*Keywords:* Tourism management, Environmental management

*Corresponding Author
E-mail Address: peerapat7654@gmail.com

## 1. Introduction

Tourism in Thailand plays the important role in nowadays making a lot of money from travelers coming from all around the world. The economic impact of business supports people in the area and creates employment, for example, hotel service, guest house, manufacturing industry, and handicraft and souvenir. Here, a study is conducted to compare between beaches that are popular and beaches that are not popular. To support the tourism in the undeveloped beaches, results of comparison are used; so we can see which direction or criteria that we should focus more in order to develop these beaches. Two popular beaches – Pattaya beach and Bangsaen beach – and two undeveloped beaches – Sor beach and Toey-Ngam beach – are selected for the study. The reason Pattaya beach and Bangsaen beach were selected is because they are very famous among Thai and foreign tourists. We can say that Pattaya beach and Bangsaen beach are the center of tourism in Chonburi province. Additionally, Sor beach and Toey-Ngam beach locating not too far from Pattaya and Bangsan have very beautiful coast and environment but the number of tourists is still low.

## 2. Objective of study

2.1 To study and compare the trait of the tourists in Chonburi province including behavior and satisfaction for supporting and promoting the structure of tourism.

2.2 To study and measure the beach environmental quality whether they can be developed by using Morgan tool including physical, social, pollution, and biological aspects.

2.3 To compare and identify data on weak point and strong point in order to find the possibility to develop the beach and to become well-known.

## 3 Literature Review

Katanyu Hiransomboon (2012) studied marketing mix effecting decisions of backpacker tourist on accommodation service buying. This is to study the buying behavior and the level of the marketing mix on the accommodation service buying decision of backpacker tourists. The questionnaire had three parts; general information about the respondents, behavior of backpackers on the accommodation service buying, and the levels of marketing mix consisting of characteristics; accommodation, rental expense, channels of contact, marketing promotion, service process, service personnel and physical evidence in accommodation service buying decision. The results of the important levels of marketing mix ranking respectively are in the following: service personnel factor in having a good manner with the friendly personality, physical evidence factor in having a clean place, accommodation product and service factor in being able to go to the tourist sites conveniently, channel of contact factor in being able to contact through agency, process factor in having a convenient reservation process, marketing promotion in having been advised in travel guidebook column, and price factor in being suitable to their budgets. Additionally, accommodation for short-term and long-term stay have different location according to the different target market. Backpacker tourists prefer the low cost accommodation and accept the reasonable price.

Park et al. (2014) studied the successful managerial behavior on farm-based tourism in South Korea. This research took place in a rural area, indicating how overall tourism grow and how to improve the management performance. This research monitor on six parts: product/service development, business planning and evaluation, promotions, human resource management, networking, and cost reduction. The result were indicates by gross sales and multiple regression.

Sealuem and Chankaew (1999) studied the criteria for recreational beach for measuring of beach quality. The assessment criterion that was designed by the researchers and also considered and improved through specialist's opinion. It was based on 50 specific indicators covering aspects of physical, social, pollution, and biological. All of these were analyzed, marked and totaled according to the Simple Weighting Score Equation of Morgan, 1999, and classified beach quality into 4 Classes, A, B, C and D as a result. This criterion can be applying to other recreational beaches in Thailand. Furthermore, both the local government and concerned organizations would virtually set up the development policy and environmental management for sustainable tourism.

## 4 Methodology

*4.3 Tourism and business*

*4.1.1 Tools for data collection*

Questionnaires are used in this study for tourists in four beaches: Pattaya, Bangsaen, Sor beach, and Toey-Ngam. The questionnaire consists of three parts as following:

Part 1: General information about the respondents

Part 2: Behavior of tourists during their trip

Part 3: The satisfaction of tourists consists of products, prices and costs of tourism, marketing promotion, service providing to tourists, physical evidence, and product and quality.

*4.1.2 Population and sample group*

The population used in this study are both Thais and foreigners who visited coastal area in Chonburi Province, which are counted for 2,675,563 people (during January to December 2013 by Tourism Authority of Thailand, Ministry of sport and tourism). A sample group contained 385 people by randomly selecting without probability involved (non – random Sample). Because the actual number of the population (number of tourists visit) is not known in this case (Boomchom Srisa-ard, 1995, p.185), so the following formula is used to calculate a number of sample size:

$$n = \frac{p\,(1-p)z^2}{e^2}$$

where $n = sample\ size$

p = *expected value is 0.5 or 50 %*

$z = at\ 95\%\ so,\ z = 1.96$

$e = $ *Acceptable tolerances set to 0.05 or 5% to represent the population of the formula*

Therefore, $n = \frac{0.5\,(1-0.5)1.96^2}{0.05^2} = 385$

*4.1.3 Analysis of data*

Descriptive statistics is conducted by the frequency and percentage that composed of the objectives for visiting, the sources that informed them about the locations, kind of traveling, transportation, time period for the trip, costs of the trip, sex, age, address, education, status, occupation, and average income by using SPSS program. Regression analysis is used to analyze the possibility to improve the undeveloped beaches.

*4.2 Environment*

*4.2.1 Tools for data collection*

Environmental quality is measured by the simply equation of Morgan which classified into four groups which each contains different number of indicators and weighted. The checklist was used to collect data from four beaches during winter, summer, and rainy seasons. A checklist was used to collect the data for four separated parts:

1) Physical criteria contains 9 indicators, total mark of 9 - 44 points.
2) Social criteria contains 22 indicators, total mark of 22-110 points.
3) Pollution criteria contains 13 indicators, total mark of 13-42 points.
4) Biological criteria contains 6 indicators, total mark of 6-20 points.

*4.2.2 Analysis of data*

After the checklist was done by observing and collecting data from related organizations. Total score from each criteria were substituted into the equation which weight differently by rank 4, 3, 2, and 1 to summarize the score in physical, social, pollution, and biological.

$$BQ = W_p \sum R_{i-n} + W_s \sum R_{i-n} + W_{pol} \sum R_{i-n} + W_b \sum R_{i-n}$$

where

BQ: Beach Quality

$W_p, W_s, W_{pol}, W_b$: Weighting score of physical and biological factor (4, 3, 2, and 1, respectively)

$R_{i-n}$: Rating scale of indicator 1, 2, 3,…, n

All rating score of environmental factors were sum, so possibility mark was 610 maximum and 134 minimum. This score represents the environmental beach quality, that were classified into four Class, A, B, C and D ((maximum - minimum) / 4). (See Table 1.)

Table 1: Classification of beach quality

| Class | Score | Beach quality standard |
|---|---|---|
| A | 491.0-610.0 | Excellent environmental beach quality |
| B | 372.0-490.0 | Very good environmental beach quality |
| C | 253.0-371.9 | Fair environmental beach quality |
| D | 134.0-252.9 | Poor environmental beach quality |

## 5 Results and Discussion

### *5.1 Tourism and business*

### *5.1.1 General information about the tourists*

Most tourists in the sample group in Pattaya are female (64%), age 21-30 years old (52%), employee (31%), monthly income less than 15000 baht (39%), come from central region (69%).

The majority of tourists in Bang-Saen are female (53%), age 21-30 (51%), employee (31%), monthly income 15001-20000 baht (47%), come from central region (65%).

Most tourists in Sor are female (53.85%), age 31-40 (40.66%), public officer (24.18%), monthly income 15001-20000 baht (32.97%), come from central region (53%).

Most tourists in Toey Ngam are female (67.02%), age 31-40 (55.56%), public officer (41.67%), monthly income 15001-20000 baht (35.11%), come from central region (56.38%). (See Table 2.)

Table 2: Descriptive statistics of the respondents

| **Personal** | **Pattaya** | | **Bang-Saen** | | **Sor** | | **Toey Ngam** | |
|---|---|---|---|---|---|---|---|---|
| **Information** | **No.** | **%** | **No.** | **%** | **No.** | **%** | **No.** | **%** |
| **Gender** | | | | | | | | |
| Male | 36 | 36 | 47 | 47 | 42 | 46.15 | 31 | 32.98 |
| Female | 64 | 64 | 53 | 53 | 49 | 53.85 | 63 | 67.02 |
| **Age** | | | | | | | | |
| <20 | 13 | 13 | 22 | 22 | 14 | 15.38 | 9 | 9.57 |
| 21-30 | 52 | 52 | 51 | 51 | 26 | 28.57 | 13 | 13.38 |
| 31-40 | 20 | 20 | 14 | 14 | 37 | 40.66 | 50 | 55.56 |
| 41-50 | 10 | 10 | 10 | 10 | 10 | 10.99 | 16 | 17.02 |
| 51-60 | 5 | 5 | 3 | 3 | 4 | 4.4 | 6 | 4.02 |
| **Occupation** | | | | | | | | |
| Student | 28 | 28 | 15 | 15 | 16 | 17.58 | 11 | 11.46 |
| Employee | 31 | 31 | 31 | 31 | 19 | 20.88 | 26 | 27.08 |
| Public officer | 11 | 11 | 21 | 21 | 22 | 24.18 | 40 | 41.67 |
| Business person | 5 | 5 | 10 | 10 | 15 | 16.48 | 7 | 7.29 |
| Freelance | 17 | 17 | 15 | 15 | 16 | 17.58 | 4 | 4.17 |
| Other | 8 | 8 | 8 | 8 | 3 | 3.3 | 6 | 8.33 |
| **Income (per month)** | | | | | | | | |
| <15000 | 39 | 39 | 23 | 23 | 21 | 23.08 | 20 | 21.28 |
| 15001-20000 | 23 | 23 | 47 | 47 | 30 | 32.97 | 33 | 35.11 |
| 21001-30000 | 22 | 22 | 19 | 19 | 29 | 31.87 | 25 | 26.6 |
| 30001-40000 | 10 | 10 | 6 | 6 | 5 | 5.49 | 7 | 7.45 |
| 40001-50000 | 4 | 4 | 3 | 3 | 4 | 4.39 | 4 | 4.26 |
| >50000 | 2 | 2 | 2 | 2 | 2 | 2.2 | 5 | 5.3 |
| **Domicile** | | | | | | | | |
| Central region | 69 | 69 | 65 | 65 | 77 | 84.62 | 53 | 56.38 |
| Northeast region | 15 | 15 | 19 | 19 | 10 | 10.94 | 28 | 29.79 |
| Northern region | 12 | 12 | 13 | 13 | 2 | 2.22 | 9 | 9.57 |
| Southern region | 4 | 4 | 3 | 3 | 2 | 2.22 | 4 | 4.26 |

### *5.1.2 Behavior of the tourists*

Most tourists in Pattaya stay for 2 days (35%). Main purpose are holiday/vacation/leisure (51%), accommodation outstanding (26%), and want to come back (97%). There are variety of hotel in Pattaya for tourists to choose, starting from 1 to 5 stars. People from central region prefer spending their weekend in Pattaya.

Most tourists in Bang Saen stay for 2 days (41%), main purpose are holiday/vacation/leisure (37%), water activities outstanding (34%), and want to come back (91%). Variety of water activities could be found there include; Banana boat, Jet-ski, sailing and etc. Additional, Thailand Water Sport Festival is held at Bang Saen annually.

Most tourists in Sor stay for 1 days (52.75%), main purpose are holiday/vacation/leisure (62.64%), water activities outstanding (47.25%), and want to come back (97.80%). Few hotels are found at Sor. Thus, tourists are mostly came for one day trip. Water activities are outstanding in kayaking, scuba diving, and planting corals.

Most tourists in Toey Ngam stay for 1- 2 days (38.30%) at the same rate, main purpose are holiday/ vacation/ leisure (61.70%), water activities outstanding (60.64%), want to come back (96.81%). Toey Ngam located near so and has the same outstanding water activities.

All in all, most of tourists prefer staying 2 days for holiday/ vacation/ leisure. The most outstanding of four beaches are water activities. And most of tourists want to come back again. (See Table 3.)

Table 3: Descriptive statistics of visit

| Behavior | Pattaya | | Bang-Saen | | Sor | | Toey-Ngam | |
|---|---|---|---|---|---|---|---|---|
| | No. | % | No. | % | No. | % | No. | % |
| **Day(s) of visit** | | | | | | | | |
| 1 day | 28 | 28 | 18 | 18 | 48 | 52.75 | 36 | 38.30 |
| 2 days | 35 | 35 | 41 | 41 | 22 | 21.98 | 36 | 38.30 |
| 3 days | 23 | 23 | 37 | 37 | 16 | 17.58 | 16 | 17.02 |
| More than 3 days | 14 | 14 | 4 | 4 | 5 | 7.69 | 6 | 6.38 |
| **Purpose** | | | | | | | | |
| Holiday/vacation/leisure | 51 | 51 | 37 | 37 | 57 | 62.64 | 58 | 61.70 |
| Beach/water activities | 26 | 26 | 34 | 34 | 31 | 34.07 | 17 | 18.08 |
| Meeting/seminar | 8 | 8 | 8 | 8 | 3 | 3.29 | 11 | 11.7 |
| Education | 3 | 3 | 2 | 2 | 0 | 0 | 4 | 4.26 |
| Other | 12 | 12 | 19 | 19 | 0 | 0 | 4 | 4.26 |
| **Most outstanding** | | | | | | | | |
| Accommodation | 26 | 26 | 10 | 10 | 2 | 2.2 | 3 | 3.19 |
| Food/Restaurant | 19 | 19 | 23 | 23 | 2 | 2.2 | 3 | 3.19 |
| Gift/Souvenir shop | 2 | 2 | 9 | 9 | 11 | 12.09 | 2 | 2.13 |
| Water activities; diving, fishing | 3 | 3 | 34 | 34 | 43 | 47.25 | 57 | 60.64 |
| Peacefulness of the beach | 15 | 15 | 2 | 2 | 19 | 20.88 | 9 | 9.57 |
| Community participated | 0 | 0 | 7 | 7 | 1 | 1.1 | 1 | 1.06 |
| Eco-tourism | 2 | 2 | 4 | 4 | 13 | 14.28 | 19 | 20.22 |
| Other(pub/night market) | 33 | 33 | 11 | 11 | 0 | 0 | 0 | 0 |
| **Come back?** | | | | | | | | |
| Yes | 97 | 97 | 91 | 91 | 89 | 97.80 | 91 | 96.81 |
| No | 3 | 3 | 9 | 9 | 2 | 2.2 | 3 | 3.19 |

### *5.1.3 Satisfaction on marketing mix*

This part studies on marketing mix to see the satisfaction of tourists consists; products of tourism, pricing fairness for tourists, market promotions, tourism servicing process, physical evidence, and tourism effectiveness and quality.

For Pattaya, the highest average value is on tourism effectiveness and quality (3.95) because the place is famous and having variety activities. The lowest average value is on physical evidence (2.01) because the beach is unclean and noisy.

For Bang Saen, the highest average value is on physical evidence (3.26) because the water is clean, the beach is peaceful, and transportation is convenient. The lowest average value is on pricing fairness for tourists (2.37) because sometime the price rates are distorted.

For Sor and Toey Ngam, the highest average value is on pricing fairness for tourists (3.51) because they are under supervision of the Thai Royal Navy. The lowest average value is market promotion (2.26) because the places are not well-known and the advertisings are not good. (See Table 4.)

In products of tourism factor, freshness of the ingredients in food is most important in Pattaya and Bang Saen because there are jetties and many famous seafood restaurants. Safety standard of the service providers on marine tourism activities is most important in Sor because it implies to service-mind performance. Cleanness and condition of equipment on marine tourism activities of Toey Ngam is the most important because marine activities are the selling point.

In pricing fairness for tourist factor, fairness of the price for beach/water activities is the most important in Pattaya and Bang Saen because famous activities such as banana boat, jet-ski and etc. are attracted tourists to the location. Fairness of the gift/ souvenir price is the most important of Sor and fairness of the accommodation price is the most important of Pattaya because standard price encouraging returns.

In market promotion factor, hotel's attractiveness of sale promotions is the most important in Pattaya and Bang Saen because of high competitive in the area. Advertisements on newspapers, radio, TV, and online communication throughout websites, or email are important for Sor and Toey Ngam because this is a way to make them become well-known to increase the

Table 4: Summary of marketing mix

| Marketing Mix | Pattaya | | | Bang-Saen | | | Sor | | | Toey-Ngam | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Avg. | S.D. | Rank | Avg. | S.D. | Rank | Avg. | S.D. | Rank | Avg. | S.D. | Rank |
| 1.Products of tourism | 3.02 | 0.86 | 3 | 2.61 | 0.64 | 5 | 3.13 | 1.02 | 4 | 3.13 | 1.02 | 4 |
| 2.Pricing fairness for tourists | 2.44 | 0.80 | 5 | 2.37 | 0.66 | 6 | 3.51 | 0.93 | 1 | 3.51 | 0.93 | 1 |
| 3.Market promotions | 3.66 | 1.03 | 2 | 2.80 | 0.87 | 3 | 2.26 | 0.81 | 6 | 2.26 | 0.81 | 6 |
| 4.Tourism servicing process | 2.69 | 0.92 | 4 | 2.64 | 0.73 | 4 | 3.20 | 0.92 | 3 | 3.20 | 0.92 | 3 |
| 5.Physical evidence | 2.01 | 1.00 | 6 | 3.26 | 0.98 | 1 | 3.28 | 0.93 | 2 | 3.28 | 0.93 | 2 |
| 6.Tourism effectiveness and quality | 3.95 | 0.89 | 1 | 2.94 | 1.00 | 2 | 2.81 | 1.03 | 5 | 2.81 | 1.03 | 5 |

Table 5: Mean and standard deviation of marketing mix in detail

| | **Pattaya** | | **Bang Saen** | | **Sor** | | **Toey Ngam** | |
|---|---|---|---|---|---|---|---|---|
| **Products of tourism** | **Avg.** | **S.D.** | **Avg.** | **S.D.** | **Avg.** | **S.D.** | **Avg.** | **S.D.** |
| Cleanness and security of accommodation | 3.01 | 0.98 | 2.66 | 0.74 | 2.88 | 1.01 | 3.90 | 0.70 |
| Condition of facilities in the room of accommodation | 2.68 | 0.79 | 2.79 | 0.71 | 3.32 | 1.13 | 2.93 | 0.71 |
| Cleanness and tidiness of restaurant | 4.02 | 0.82 | 2.40 | 0.78 | 2.74 | 1.36 | 3.53 | 0.98 |
| Freshness of the ingredients in food | 4.28 | 0.74 | 2.99 | 1.11 | 2.99 | 1.10 | 3.45 | 0.91 |
| Safety standard of the service providers on Marine tourism activities | 1.87 | 0.96 | 2.43 | 0.62 | 3.57 | 0.73 | 3.86 | 0.74 |
| Cleanness and condition of equipment on Marine tourism activities | 2.24 | 1.03 | 2.39 | 0.68 | 3.25 | 0.78 | 3.95 | 0.83 |
| **Pricing fairness for tourists** | | | | | | | | |
| Fairness of the accommodation price | 2.35 | 0.69 | 2.35 | 0.68 | 3.37 | 1.12 | 3.70 | 0.75 |
| Fairness of the gift/ souvenir price | 2.35 | 0.87 | 2.37 | 0.61 | 3.76 | 0.79 | 3.68 | 0.76 |
| Fairness of the price for beach / water activities | 2.61 | 0.84 | 2.40 | 0.67 | 3.41 | 0.87 | 3.34 | 0.98 |
| **Market Promotions** | | | | | | | | |
| Advertisements on newspapers , radio , and TV | 3.84 | 1.00 | 2.83 | 0.91 | 2.48 | 0.95 | 2.53 | 1.00 |
| Online communication throughout websites , or email | 4.11 | 0.97 | 2.85 | 0.78 | 2.57 | 0.87 | 2.88 | 0.85 |
| Hotel's attractiveness of sale promotions | 3.85 | 1.03 | 2.92 | 0.93 | 2.15 | 0.74 | 2.81 | 0.82 |
| Touring agencies attractiveness of sale promotions | 2.85 | 1.12 | 2.60 | 0.85 | 1.87 | 0.66 | 2.26 | 0.80 |
| **Tourism servicing process** | | | | | | | | |
| Process of announcing tourists for the safety rules and regulations for swimming and other beach activities | 2.94 | 1.10 | 2.68 | 0.63 | 3.64 | 0.86 | 3.61 | 0.93 |
| Availability of tourists service centers | 2.63 | 0.84 | 2.70 | 0.78 | 3.12 | 1.02 | 3.61 | 0.92 |
| Service providers' politeness to customers | 2.50 | 0.82 | 2.53 | 0.78 | 2.86 | 0.88 | 3.50 | 0.89 |
| **Physical Evidence** | | | | | | | | |
| Scenery of the beach and Clearness of sea water | 2.42 | 0.96 | 3.52 | 1.06 | 4.13 | 0.90 | 4.47 | 0.94 |
| Peacefulness of the beach | 2.05 | 0.94 | 2.55 | 0.97 | 3.86 | 0.88 | 4.48 | 0.94 |
| Convenient and lead to the location | 3.39 | 1.10 | 3.03 | 1.00 | 2.61 | 0.99 | 4.03 | 1.23 |
| Accreted and enough direction sign | 3.60 | 1.02 | 2.95 | 0.91 | 2.50 | 0.96 | 3.39 | 0.99 |
| **Tourism effectiveness and Quality** | | | | | | | | |
| Variety of activities | 3.96 | 0.88 | 2.99 | 1.06 | 3.04 | 1.05 | 3.64 | 0.91 |
| Popularity of the location | 3.94 | 0.90 | 2.88 | 0.96 | 2.57 | 1.00 | 3.65 | 0.89 |

number of tourists to the area.

In tourism servicing process factor, process of announcing tourists for the safety rules and regulations for swimming and other beach activities are the most important in Pattaya, Sor, and Toey Ngam to keep safety standard and to ensure safeness for the tourists. Availability of tourist service centers is the most important in Bang Saen and Toey Ngam to give correct information to the tourists.

In physical evidence factor, accreted and enough direction sign is the most important in Pattaya because the roads in Pattaya are complex. Scenery of the beach and clearness of sea water is the most important in Bang Saen and Sor because it is first priority of being a tourist attraction. Peacefulness of the beach is the most important of Toey Ngam because many of people on working ages in the area.

In tourism effectiveness and quality factor, variety of activities is the most important of Pattaya, Bang Saen and Sor to increase the number of tourists in the area. Popularity of the location is the most important of Toey Ngam for creditable. (See Table 5.)

### *5.1.4 Influence factors on beach improvement*

Regression analysis is used to estimate the possibility to improve the undeveloped beach to become tourist attractions by comparing with developed beach. To find transfer function, the dependent variable is developed beach and the independent variable is undeveloped beach. Substituted into equation y = Ax + B. If the result of independent closed to the dependent means that the undeveloped beach can be improved. This equation helps forecast the unknown variable and it is practical.

From the Table 6, substitutes the variable into the equation y = Ax + B.

Therefore, $y = (0.085)X_1 + (0.151)X_2 + (-0.070)X_3 + (-0.079) X_4 + (0.088)X_5 + (-0.061) X_6 + (-0.054)X_7 + (0.195)X_9 + (0.100) X_{10} + (0.187)X_{11} + (0.146)X_{12} + (0.056)X_{13} + (0.068)X_{14} + (0.152)X_{15} + (0.056)X_{16} + (0.056)X_{16} + (-0.039)X_{18} + (-0.029)X_{19} + 0.941$

After substituting values of Sor and Toey Ngam beach into x variable, y = 0.68 and 0.82, respectively. This means that Sor and Toey Ngam can be improved like Pattaya and Bang Sean.

Table 6: Regression analysis

| Independent variable | Coefficient | t-value | p-value |
|---|---|---|---|
| Constant | 0.941 | 17.325 | 0 |
| Peacefulness of the beach | 0.085 | -5.459 | 0 |
| Touring agencies attractiveness of sale promotions | 0.151 | 9.529 | 0 |
| Safety standard of the service providers on marine activities | -0.070 | -4.321 | 0 |
| Fairness of the gift/ souvenir price | -0.079 | -5.377 | 0 |
| Hotel's attractiveness of sale promotions | 0.088 | 5.316 | 0 |
| Cleanness and security of accommodation | -0.061 | -4.307 | 0 |
| Cleanness and condition of equipment on marine activities | -0.054 | -3.658 | 0 |
| Food / Restaurant outstanding | 0.195 | 5.111 | 0 |
| Education visiting purpose | 0.100 | 2.771 | 0.006 |
| Accommodations outstanding | 0.187 | 4.616 | 0 |
| Other factor outstanding | 0.146 | 3.66 | 0 |
| Holiday visiting purpose | 0.056 | -2.651 | 0.008 |
| Beach's peacefulness outstanding | 0.068 | 2.652 | 0.008 |
| Gift / Souvenir shops outstanding | 0.152 | 2.574 | 0.01 |
| 3 days spent in the area | 0.056 | 2.354 | 0.019 |
| Advertisements on newspapers, radio, and TV | 0.056 | 3.217 | 0.001 |
| Process of announcing tourists for the safety rules and regulations for swimming and other beach activities | -0.039 | -2.784 | 0.006 |
| Online communication throughout website, or email | -0.029 | -2.09 | 0.037 |

R-square = .726, F = 24.744 (p = 0.000)
Dependent variable: Pattaya and Bang Saen
Method: Stepwise

### *5.3 Environment*

Environmental beach quality was calculated according to Simple Weighting Score Equation follow by Morgan (1999) on equation, the beach score was 610 maximum and 134 minimum. These scores representing the environmental quality of beach, were classified into four classes: A, B, C and D shown in Table 1 already.

Table 7: Environmental analysis at Bang Saen beach

| | Physical | Social | Pollution | Biological | Total |
|---|---|---|---|---|---|
| **Summer** | (29*4) = 116 | (83*3) = 249 | (23*2) = 46 | (9*1) = 9 | 420 |
| **Rainy** | (29*4) = 116 | (83*3) = 249 | (21*2) = 42 | (8*1) = 9 | 415 |
| **Winter** | (29*4) = 116 | (83*3) = 249 | (22*2) = 44 | (9*1) = 9 | 418 |

Bang Saen beach's score in summer, rainy and winter season were 420, 415 and 418 points, respectively. This score was compared with the classification of beach quality in Table 1 above. As a result, it was classified in Class B (very good environmental beach quality) in every season. Those result reflect that, this beach is very good in physical factor and social factor, which are 116 (65%) and 249 (75%), respectively. So as you can see in the survey quality of sand is pretty in case of sand color, sand size, beach length, and sand softness. Bang Saen has a lot of restaurants, hotels, and souvenir shops enough to satisfy the tourists. And in case of emergency and safety, Bang Saen also has enough life-guard and standard equipment provided to the customer (e.g. jet-skiing, banana-boat, rental car, rental motorcycle). For the pollution they gain the highest score (46 out of 84) in the summer which is the high season in Thailand that has long weekend. Bang Saen as the high rate of tourist, it will cause a bit problem of noise and unpleasant odors from vehicles and other road vehicles which allowed pass along the beach, poor of car parking area. In the biological achieved 9 out of 20 (only 45%) because it difficult to find many trees along the beach. And also has jelly-fish, the abundance species was poor, can found only a few and general species such as ghost crab maybe because of many tourist disturbed by tourist activities and local community (See Table 7.)

Pattaya Beach's score in summer, rainy and winter season were 400, 396 and 392 points, respectively. This score was compared with the classification of beach quality in Table 1. As a result, it was classified in Class B (very good environmental beach quality) in every season. Especially, in social factor achieved the score of 237 out of 330 (about 72%); this is the highest score for every factor. Because as you can see from the survey, Pattaya has a lot of restaurants, hotels, souvenir shops enough to satisfy the tourists. And in case of emergency and safety, Pattaya also have enough life-guard and standard equipment provide to the customer (e.g. jet-skiing, banana-boat, rental car, Rent motorcycle). But also has a low score in traffic because there are many tourists and most of the streets in the city are two-lane road, so no matter the season is, the traffic is still very bad. The biological achieved the average of 9 out of

Table 8: Environmental analysis at Pattaya beach

| | Physical | Social | Pollution | Biological | Total |
|---|---|---|---|---|---|
| **Summer** | (27*4) = 108 | (79*3) = 237 | (23*2) = 46 | (9*1) = 9 | 400 |
| **Rainy** | (27*4) = 108 | (79*3) = 237 | (21*2) = 42 | (9*1) = 9 | 396 |
| **Winter** | (27*4) = 108 | (79*3) = 237 | (19*2) = 38 | (9*1) = 9 | 392 |

20 (only 45%) because it is hard to find many trees along the beach. And the beach also has jelly-fish. The abundance species was poor; only a few and general species were found such as ghost crab.

Pollution factor, in the summer, achieved the highest score 46 out of 84 (about 55 %) because in the summer, which has the long holiday, there are many tourists. So it will result in a problem of trash on the beach, water quality, noise and odor from motor vehicle. For the physical factor the score is very low 108 out of 176 (about 61%) because this beach is very popular and well-known so there is a lot of tourists all around the world visited here. So this beach also has activities that will make the tourist enjoy to came here and attractive to the tourists. Some activities, however, will destroy the quality of sand, sea and maybe produce the pollution as much as it should.

Sor beach's score in summer, rainy and winter season were 378, 374 and 376 points, respectively. This score was compared with the classification of beach quality. As a result, it was classified in Class B (very good environmental beach quality) in every season. This beach very good at biological factor 13 out of 20 (about 65%) better than the rest of the beach in this research because has some trees along the beach and less harmful animals (such as jelly-fish) and less cockroaches than other beaches. Also in the physical factor, it achieved very high score, which is 120 out of 176 (about 68%) because this beach has not much tourists so the environment in this area will not be destroyed by people and still maintain their quality. But it also has serious problem same as Toey-Ngam beach. The problem that we should consider is social factor (171 out of 330 (about 55%)) because this beach does not have a lot of hotels and accommodations including shops so this will not satisfy the tourists coming to the beach. And also the sea activity, this beach does not have enough activity to attract the tourists. Even though the environment is very good and beautiful, but tourists need more. So this criteria should also be considered.

Table 9: Environmental analysis at Sor beach

| | Physical | Social | Pollution | Biological | Total |
|---|---|---|---|---|---|
| **Summer** | (30*4) = 120 | (57*3) = 171 | (37*2) = 74 | (13*1) = 13 | 378 |
| **Rainy** | (30*4) = 120 | (57*3) = 171 | (35*2) = 70 | (13*1) = 13 | 374 |
| **Winter** | (30*4) = 120 | (57*3) = 171 | (36*2) = 72 | (13*1) = 13 | 380 |

Table 10: Environmental analysis at Toey Ngam beach

| | Physical | Social | Pollution | Biological | Total |
|---|---|---|---|---|---|
| **Summer** | (30*4) = 120 | (58*3) = 174 | (37*2) = 74 | (12*1) = 12 | 380 |
| **Rainy** | (30*4) = 120 | (58*3) = 174 | (36*2) = 72 | (12*1) = 12 | 378 |
| **Winter** | (30*4) = 120 | (58*3) = 174 | (37*2) = 74 | (12*1) = 12 | 380 |

Toey-Ngam beach's score in summer, rainy and winter season were 380, 378 and 380 points, respectively. As a result, it was classified in Class B (very good environmental beach quality) in every season. The advantages of this beach, as you can see from the score are physical factor and pollution factor. For the physical factor achieved the score of 120 out of 176 (about 68%). This beach is not a popular beach; the quality of sand, the color of water and sea are in good condition. Because this beach is not well-known so there are not a lot of tourists and the environment is still maintained at good quality as it should be. Pollution factor is 74 out of 84 (about 88%); water is very clear and water odor is absent. It also has less pollution than other beaches; there are a lot of parking lots and very good traffic flow. The biological score is 12 out of 20 (about 60%). But has the serious problem with the social factor 174 out of 330 (only 58 %) because this beach does not have enough accommodations and hotels to satisfy the tourists in the high season.

## 6. Conclusion

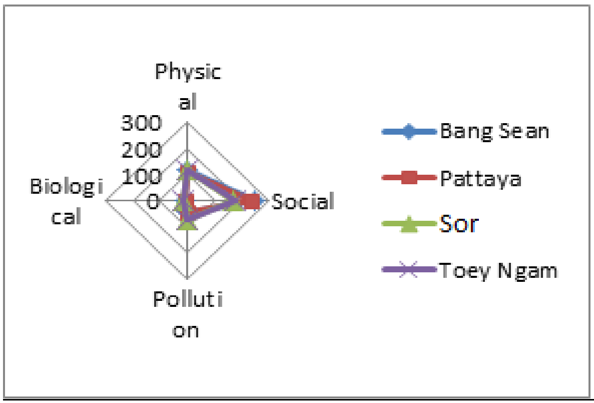
Conclusion is divided into two parts as those topics show the score of Business & Tourism, and Environment expected. The score can indicate the direction in which way they need to improve in order to be a successful tourism in term of sustainable environment and business preparation of local community in Chonburi area where the survey was taken. Not only in this province, can this method also be applied to many different areas that has potential to be promoted for tour destination.

For business and tourism parts, the survey are collected in three parts of information including general information, behavior of visitors, and satisfaction of visitors from 385 people at 4 beaches. For the basic information, the most respondents in four beaches are females, age between 21 – 40, employee and public officer, monthly income not more than 20000, come from central region. For the behavior part, most tourists stayed for two days; the main purpose is holiday/vacation/leisure. The most outstanding feature is water activities and tourists want to come back to the area again.

The marketing mix with highest rate of satisfaction in Pattaya, Bang Saen, Sor and Toey Ngam are tourism effectiveness and quality, physical evidence, pricing fairness for tourists and pricing fairness for tourists, respectively.

All in all, by the results form SPSS programs by regression analysis shown that Sor and Toey Ngam beaches can be developed to be tourism attract like Pattaya and Bang Saen beaches.

For the environment part, the survey shows that Sor and Toey-Ngam beaches which are not as well known as Pattaya and Bang Saen beach, the beach scores are almost in the same rate. But let's look in detail.



| | Physical | Social | Pollution | Biological |
|---|---|---|---|---|
| **Bang Saen** | 116 | 249 | 44 | 9 |
| **Pattaya** | 108 | 237 | 42 | 9 |
| **Sor** | 120 | 171 | 72 | 13 |
| **Toey Ngam** | 120 | 174 | 73.33 | 12 |

Figure 1: Comparison in environment score of four beaches

According to the charts, Toey Ngam and Sor beach have better score than Bang Sang and Pattaya beach in physical criteria. This means that beach and environment around these beaches are better. However, these beaches do not famous as they should be. When comparing to the famous beaches (which are Pattaya and Bang Saen beach), the big difference is in social criteria score. Thus, this is the point that we should concern about, in order to develop those beaches to become famous and gain the customer satisfaction. Additionally, the criteria that already good enough should be maintained in good quality when develop other criteria. For example, if built more of restaurants and accommodations to satisfy customer, which will gain a better score of social, the quality of physical, biological and pollution should be well taken care of.

## 6 Reference

[1] Anthony W. Dixon, Chi-Ok Oh and Jason Draper (2012) “Access to the Beach: Comparing the Economic Values of Coastal Residents and Tourists” http://jtr.sagepub.com/content/51/6/742 *Journal of travel research, 2012 51:742*

[2] Duk-Byeong, Kyung-Rok Doh, Kyung Hee Kim (2014) “Successful managerial behavior for farm-base tourism: A functional approach” *ScienceDirect, Tourism Management 45 (2014) 201-210*

[3] Göktug Morçöl, Triparna Vasavada and Sohee Kim (2013) “Business Improvement Districts in Urban Governance; A Longitudinal Case Study” http://aas.sagepub.com/content/46/7/796, *Administration & Social 2014 46:796*

[4] Harng Luh Sin, Claudio Minca (2014), “Touring responsibility: The trouble with ‘going local’ in community-based tourism in Thailand” *Geoforum 51 (2014) 96-106*

[5] Katanyu Hiransomboon (2012) “Marketing Mix Affecting Accommodation Service Buying Decision of Backpacker Tourist Traveling at Inner Rattanakosin Island in Bangkok, Thailand” *SciVerse ScienceDirect, Procedia Economics and Finance 3 (2012) 276-283*

[6] Sawanya Sealeum, Kasem Chankaew, Wasin Inkapattanakul, Charlie Navanugraha (2007) “Environment quality assessment criteria for recreational beach: a case study of cha-am beach” *Environment and Natural Resource Journal Vol. 5, No.1,June 2007*

# Probability Plot of the Length-biased Exponentiated Inverted Weibull Distribution

Palakorn Seenoi[1]* and Winai Bodhisuwan[2]

[1]*Department of Statistics, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand, palakorns@kku.ac.th*

[2]*Department of Statistics, Faculty of Science, Kasetsart University, Bangkok, Thailand, fsciwnb@ku.ac.th*

**Abstract**

Probability plotting is a technique that helps us decide whether data follows the pattern of a particular probability distribution. This paper develops a probability plot of the length-biased exponentiated inverted Weibull (LBEIW) distribution by using R program. In addition, the probability plot of the LBEIW distribution is used to illustrate by fitting the LBEIW distribution to the distance between cracks in a pipe dataset.

*Keywords*: probability distribution, probability plot, length-biased exponentiated inverted Weibull distribution

*Corresponding Author
E-mail Address: palakorns@kku.ac.th

## 1. Introduction

The Weibull distribution was introduced by Wallodi Weibull, Swedish scientist, in 1951 [1]. It is perhaps the most widely used distribution to analyze the lifetime data. This distribution provides vast impact of reliability and quality control. In some cases the Weibull distribution may not be fit to some datasets, however, alternatively, the inverse Gaussian, gamma, inverse gamma, log-normal distributions can be good representative. Recently the length-biased exponentiated inverted Weibulll (LBEIW) distribution has been proposed by Seenoi, Supapakorn, and Bodhisuwan [2] in 2014.

The LBEIW distribution has the following density function

$$g(x) = \frac{\beta\theta^{1-\frac{1}{\beta}}}{\Gamma(1-\frac{1}{\beta})} x^{-\beta}\{\exp(-x^{-\beta})\}^{\theta}; \quad x > 0, \beta > 1, \theta > 0, \tag{1}$$

and the distribution function

$$G(x) = \frac{\Gamma(1-\frac{1}{\beta}, \frac{\theta}{x^{\beta}})}{\Gamma(1-\frac{1}{\beta})}, \tag{2}$$

where $\Gamma(s,x) = \int_x^{\infty} t^{s-1} e^{-t}\,\mathrm{d}t$ is an upper incomplete gamma function.

Also, the survival and hazard functions of the LBEIW distribution with two shape parameters $\beta$ and $\theta$ are given by

$$S(x) = \frac{\gamma(1-\frac{1}{\beta}, \frac{\theta}{x^{\beta}})}{\Gamma(1-\frac{1}{\beta})}, \tag{3}$$

and

$$h(x) = \frac{\beta\theta^{1-\frac{1}{\beta}} x^{-\beta}\{\exp(-x^{-\beta})\}^{\theta}}{\gamma(1-\frac{1}{\beta}, \frac{\theta}{t^{\beta}})}, \tag{4}$$

where $\gamma(s,x) = \int_0^{x} t^{s-1} \exp(-t)\,\mathrm{d}t$ is a lower incomplete gamma function.

Some plots of the LBEIW density function with specific parameter values are shown in Figure 1.

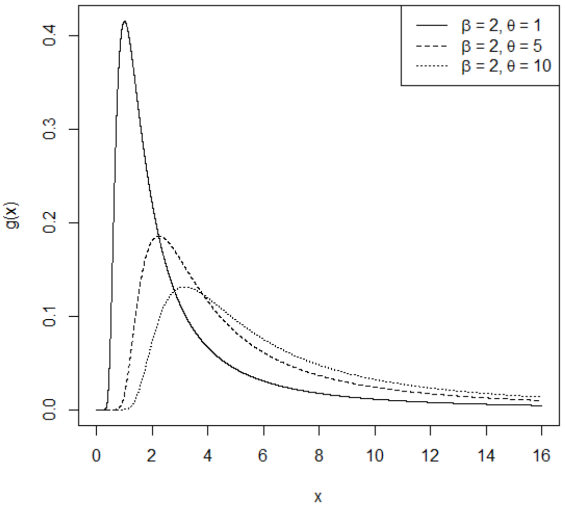


Figure 1: The density function of the LBEIW distribution for selected values of $\beta$ and $\theta$.

Probability plotting is a technique that helps us decide whether data follows the pattern of a particular probability distribution. Usually we wish to compare data to a Normal distribution – hence the name "Normal Probability Plot". This paper develops a probability plot of the LBEIW distribution [3].

The rest of the paper is organized as follows. In the next section, the basic idea of the probability plotting is presented. In Section 3, we calculate the points for a LBEIW probability plot. The graph of a LBEIW probability plot provided in Section 4. Finally conclusions appear in Section 5.

## 2. The basic idea

As usual we assume that the data values $x_1$, $x_2$, ..., $x_n$ are $n$ realizations of the LBEIW random variable $X$. We compare these observed values (observations) with theoretical values arising from the probability distribution of $X$.

The basic idea of probability plotting is that if these realized values of $X$, placed in ascending order, match up fairly well with the theoretical values of $X$ (what we would expect to get for $n$ observations) then when we plot these values against each other on a graph, we should get a straight line. If this is the case we conclude that the data are LBEIW distributed. On the other hand, if we get a plot that doesn't look linear, then we will conclude the data does not come from a LBEIW distribution. The only real complication will be to decide what to use for the theoretical values of $X$. The approach we will adopt is to consider percentiles, both of data and of the theoretical distribution. However, before we get to this we need some background ideas.

### *2.1 Ordering the data*

First we need to sort the data into increasing order. Let

$X_{(1)}$ = smallest observation,

$X_{(2)}$ = second smallest,

$\dots$

$X_{(n)}$ = largrst.

Thus $X_{(1)} \le X_{(2)} \le \dots \le X_{(n)}$.

The number in parentheses is called the rank of the data value.

If two data points are tied (i.e. they both have the same value) then in this primer we will use the convention that they can have different ranks $i$ and $i+1$ (it doesn't matter which one you choose to have rank $i$!). Note that in other contexts we would use a different convention, where both points would be given the same average rank $i+\frac{1}{2}$. This isn't so good for our purposes since it means that two points would be plotted one on top of the other. Our rule makes the computation easier and both points are visible on the plot. We can also justify our convention by saying that our plot wouldn't change if one data value was minutely smaller than the other, whereas the average-rank plot would change.

We will also make use of a concept known as depth of a data value. The depth refers to how far the particular point is in from the closest end, once the data have been sorted.

Then the smallest and largest values of a data set both have depth 1, the second-smallest and second-largest both the depth 2, and so on.

If $n$, the sample size, is old, then the median is the deepest points, and has depth $\frac{n+1}{2}$.

This also equals its rank. If n is even then there are two deepest points, both with depth $\frac{n}{2}$.

The median is the average of these values, and we still say its depth is $\frac{n+1}{2}$ even though this is no longer an integer.

### *2.2 Theoretical percentiles*

Consider a graph of the LBEIW density function; the area under the probability curve equals 1, or 100%. Now we consider cutting the $X$ axis at 99 points. We denote the *i*th point by $x_{i/100}$, since it is the point for which $P\left[X < x_{i/100}\right] = i/100 = i\%$. We can find these points from the LBEIW quantile function (inverse of the distribution function)

$$Q = G^{-1}(x).$$

### *2.3 Sample percentiles*

The *P*th sample percentile is a value such that at least *P*% of the measurements are less than or equal to that value, and at least (100-*P*)% of the measurements are greater than or equal to the value. Unfortunately, this definition has a serious drawback; it is one-directional. That is you can go from a percentage value to a percentile, but not back again unless you make further assumptions. For example, you can talk about the 45th, 50th or 55th sample percentile, and identify the corresponding data point, but unless you have a lot of data you may find that all these percentiles refer to the same data value.

### *2.4 The rule we use*

We will adopt the convention that $x_{(i)}$ is the $100\left(i-\frac{1}{2}\right)/n$ sample percentile. That is, the theoretical percentile corresponding to the *i*th ordered value is $x_{(i-\frac{1}{2})/n}$. A LBEIW probability plot would be a graph of the points $\left(x_{(1)}, x_{1/2n}\right)$, $\left(x_{(2)}, x_{3/2n}\right)$, $\left(x_{(3)}, x_{5/2n}\right)$, ..., $\left(x_{(n)}, x_{1-1/2n}\right)$.

Note that this definition will always put the median in the right place. That is, if $x_{(i)}$ is the sample median then $i = (n+1)/2 = n/2 + 1/2$, so that $\left(i-\frac{1}{2}\right)/n = \frac{1}{2}$. That is, it is the $x_{0.5}$ point, the theoretical median.

## 3. Calculating and Plotting Points

We consider uncensored data on distance between cracks in a pipe dataset [4]. The recent study found that the LBEIW MLE estimates of the parameters $\beta$ and $\theta$ are 3.3891 and 9508.9505, respectively. Moreover, the p-value of Anderson-Darling test [5] for the goodness of fit test of the LBEIW distribution is 0.9447 that the LBEIW distribution fits to the data is well.

We are now able to calculate the point for a LBEIW probability plot which corresponding to the distance between cracks in a pipe dataset. In this section it is assumed that we only have ordinary graph paper, with a linear-by-linear scale.

We shall draw up a table with four columns.

The first has $i$, the rank of each data point, going from 1 to $n$.

The second has the corresponding data values.

The third has the percent value, which from now on we will call the percentage $100\left(i-\frac{1}{2}\right)/n\%$.

The fourth has the percentile $x_{100\left(i-\frac{1}{2}\right)/n}$.

We will plot the second and fourth columns against each other. The process is illustrated by an example, given in Table 1.

Table 1: LBEIW probability plot calculations for distance between cracks in a pipe data

| Rank | Data | Percentage | Percentile |
|---|---|---|---|
| 1 | 10.24 | 2.08 | 10.59 |
| 2 | 10.67 | 6.25 | 11.84 |
| 3 | 12.26 | 10.42 | 12.69 |
| 4 | 14.10 | 14.58 | 13.42 |
| 5 | 14.70 | 18.75 | 14.09 |
| 6 | 14.97 | 22.92 | 14.74 |
| 7 | 16.62 | 27.08 | 15.38 |
| 8 | 17.12 | 31.25 | 16.04 |
| 9 | 18.14 | 35.42 | 16.71 |
| 10 | 18.51 | 39.58 | 17.41 |
| 11 | 18.84 | 43.75 | 18.16 |
| 12 | 19.08 | 47.92 | 18.96 |
| 13 | 19.40 | 52.08 | 19.83 |
| 14 | 22.38 | 56.25 | 20.78 |
| 15 | 22.57 | 60.42 | 21.85 |
| 16 | 22.85 | 64.58 | 23.07 |
| 17 | 25.43 | 68.75 | 24.47 |
| 18 | 27.47 | 72.92 | 26.15 |
| 19 | 27.98 | 77.08 | 28.20 |
| 20 | 29.93 | 81.25 | 30.83 |
| 21 | 30.94 | 85.42 | 34.40 |
| 22 | 36.02 | 89.58 | 39.76 |
| 23 | 49.56 | 93.75 | 49.40 |
| 24 | 51.56 | 97.92 | 78.43 |

## 4. The graph of a LBEIW probability plot

We are now in a position to plot the points. We will adopt the convention that the ranked data values $x_{(i)}$ will be plotted on the horizontal axis and the percentiles $x_{(i-\frac{1}{2})/n}$ plotted on the vertical axis. This is because the data vary, but the percentiles are fixed--they depend only on the sample size $n$. By using R program [6] **(See in Appendix A**), the LBEIW probability plot for distance between cracks in a pipe dataset is given in Figure 1. Note the axis labels. The graphs are all to the same scale, but this is not necessary. When plotting your graphs, make sure you allow a big enough scale to let you spread the data out, so that you can see any potential departures from a straight-line relationship.

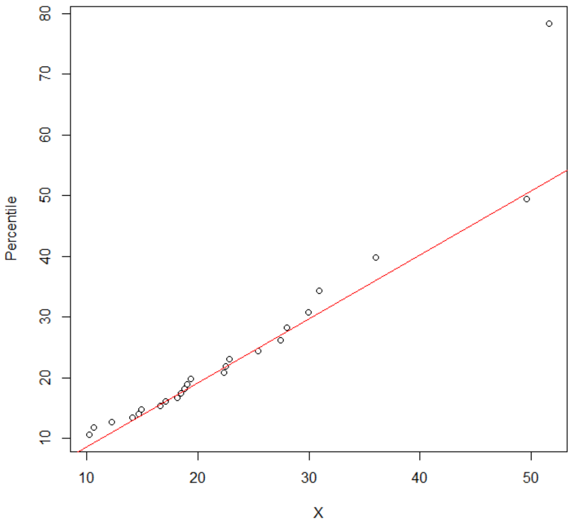


Figure 1: LBEIW probability plot for distance between cracks in a pipe dataset

## 3. Conclusion

The results in Figure 1 shown that a plot between ascending order and the theoretical values of $X$ is linear look, then we can conclude that the dataset is LBEIW distributed which corresponding to the p-value of the Anderson-Darling test.

## References

[1] Weibull W. A statistical distribution function of wide applicability. Applied Mechanics. 1951; 18(3): 293-297.

[2] Seenoi P, Supapakorn T, Bodhisuwan W. The length-biased exponentiated inverted weibull distribution. International Journal of Pure and Applied Mathematics. 2014; 92(2): 191-206.

[3] McDonald BW. A primer on normal probability plotting. Australia: RMIT Lecture Notes in Mathematics; 1993.

[4] Lawless JF. Statistical models and methods for lifetime data. 2nd ed. Canada: Weiley; 2003.

[5] Chen G, Balakrishnan N. A general purpose approximate goodness-of-fit test. Quality Technology. 1999; 27(2): 154-161.

[6] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2015. [Online]. Available: http://www.R-project.org/

## Appendix A. R Code for LBEIW Probability Plot

```
#Quantile function of LBEIW distribution
qLBEIW <- function(p,beta,theta)
{
   y <- p * gamma(1-(1/beta))
   library("zipfR")
   a <- Igamma.inv(1-(1/beta), y, lower=FALSE)
   x <- (theta/a)^(1/beta)
   return(x)
}
# LBEIW probability plot
x <- sort(x)
percentage <- 100*(i-0.5)/n
percentile <- qLBEIW(percentage/100,
   beta=beta,theta=theta)
plot(x, percentile)
yl <- quantile(x, c(0.25, 0.75))
xl <- qLBEIW(c(0.25, 0.75),beta=beta,
   theta=theta)
   slope <- diff(xl)/diff(yl)
int <- xl[1] - slope * yl[1]
abline(int, slope, col = "red")
```

# Assessing and Benchmarking the Efficiency of Local and City Governments in Building Resilience to Natural Disasters in Pathum Thani Province

Techanithi Sirikhum, Passakorn Pongsangiam, Ponanan Chonglikit*, Puvit Shujan, and Morrakot Raweewan
*School of Management Technology, Sirindhorn International Institute of Technology, Thammasat University, Muang, Pathum Thani, 12000, Thailand, ponanan.c@gmail.com*

## Abstract

Nowadays, the number of natural disasters and its economic losses has increased around the world including PathumThani Province of Thailand. Building resilience to natural disasters and major economic crisis is important policy issue in the field of public safety and security. Competences and operative responsibilities are mostly given to local and city governments. On one hand, local and city governments should be able to understand the gaps and challenges in disaster risk reduction. To serve such objective, UNISDR developed Local Government Self Assessment Tool (LGSAT), a self-assessment tool and feedback mechanism offered to local and city governments. On the other hand, local and city governments should be able to identify specific target for improvement if they are inefficient. A purpose of our study is to assess and benchmark the efficiency of local and city governments in PathumThani in building resilience to natural disasters. LGSAT is employed to identify gaps and challenges. The result from this step is used as outputs of Decision Making Units (DMU), i.e. a local or city government, in Data Envelopment Analysis (DEA). The input-oriented DEA is utilized to identify the best practice DMU and targets of improvement for inefficient DMUs when the given inputs are the budget, number of staff n the DMU.

*Keywords*: Resilience to disasters, local government, Data Envelopment Analysis

*Corresponding Author
E-mail Address: ponanan.c@gmail.com

## 1. Introduction

Since the number of natural disasters and its economics losses has increased around the world, building resilience to such disasters and economic crisis is important policy issue in the field of public safety and security. Building resilience engages multi-stakeholders within governments and whole community [1, 3]. How can local and city governments assess the achievements and challenges in building resilience? How can they assess efficiency of building resilience and set target for improvement? These two questions are studied here.

How can local and city governments assess the achievements and challenges in building resilience? Local and city governments should be able to understand the gaps and challenges in disaster risk reduction [8]. To serve this objective, United Nations International Strategy of Disaster Reduction (UNISDR) developed Local Government Self Assessment Tool (LGSAT) [8], a self-assessment tool and feedback mechanism offered to local and city governments. This framework was designed to assess progress against disaster risk reduction at the local level using as reference the Hyogo Framework for Action (HFA).

To build resilience to disasters, local and city governments use resources such as annual budget and staff. Are these resources utilized effectively? Resources (inputs) are transferred to build resilience to disasters (outputs). How can the governments assess efficiency of building resilience? How can best practice(s) be identified? How much target should be set to improve inefficient unit(s)? Data Envelopment Analysis (DEA) can offer answer to these questions.

In summary, stakeholders from local and city governments are identified and interviewed following LGSAT guideline. The result from this step is used as outputs of Decision Making Units (DMU), i.e. a local or city government, in DEA.

## 2. Methodology

### *2.1 Local Government Self Assessment Tool (LGSAT)*

LGSAT is self assessment tool for local government. Stakeholders, who should involve in the review process, include local government authorities, civil society organizations, local academia, the business community and community-based organizations. However, due to limit time of study, only government authorities were interviewed in our study [8].

LGSAT has 22 core indicators in 5 priorities for action. “*Priority for action 1*: Governance: ensure that disaster risk reduction is a national and local priority with strong institutional basis for implementation; *Priority for action 2*: Risk identification: identify, assess and monitor disaster risks and enhance early warning; *Priority for action 3*: Knowledge: use knowledge, innovation and education to build a culture of safety and resilience at all levels; *Priority for action 4*: Reducing the underlying risk factors in various sectors (environment, health, construction, etc.); and *Priority for action 5*: Strengthen disaster preparedness for effective response" [www.unisdr.org]. The score from five priorities is then used as five outputs in DEA.

Table 1: Scoring Scale - Level of Progress
The status and level of progress in the self-assessment shall be measured on a scale of 1-5 [8]

| Level of Progress | Description of level of progress for overall ranking for each question |
|---|---|
| 5 | Comprehensive achievement has been attained, with the commitment and capacities to sustain efforts at all levels. |
| 4 | Substantial achievement has been attained, but with some recognized deficiencies in commitment, financial resources or operational capacities. |
| 3 | There is some institutional commitment and capacities to achieving DRR, but progress is not comprehensive or substantial. |
| 2 | Achievements have been made but are incomplete, and while improvements are planned, the commitment and capacities are limited. |
| 1 | Achievements are minor and there are few signs of planning or forward action to improve the situation. |

In detail, 22 indicators in the five priorities of the Hyogo Framework are designed as a check list and the building block for disaster risk reduction. They are 10 essentials for making cities resilient.

Table 2: Description of 10 essentials for making cities resilient [8]

| Essential | Description of essential |
|---|---|
| 1 | Put in place organization and coordination to understand and reduce disaster risk, based on participation of citizen groups and civil society. Build local alliances. Ensure that all departments understand their role to disaster risk reduction and preparedness. |
| 2 | Assign a budget for disaster risk reduction and provide incentives for homeowners, low-income families, communities, businesses and public sector to invest in reducing the risks they face. |
| 3 | Maintain up-to-date data on hazards and vulnerabilities, prepare risk assessments and use these as the basis for urban development plans and decisions. Ensure that this information and the plans for your city's resilience are readily available to the public and fully discussed with them. |
| 4 | Invest in and maintain critical infrastructure that reduces risk, such as flood drainage, adjusted where needed to cope with climate change. |
| 5 | Assess the safety of all schools and health facilities and upgrade these as necessary. |
| 6 | Apply and enforce realistic, risk compliant building regulations and land use planning principles. Identify safe land for low-income citizens and develop upgrading of informal settlements, wherever feasible. |
| 7 | Ensure education programs and training on disaster risk reduction are in place in schools and local communities. |
| 8 | Protect ecosystems and natural buffers to mitigate floods, storm surges and other hazards to which your city may be vulnerable. Adapt to climate change by building on good risk reduction practices. |
| 9 | Install early warning systems and emergency management capacities in your city and hold regular public preparedness drills. |
| 10 | After any disaster, ensure that the needs of the survivors are placed at the centre of reconstruction with support for them and their community organizations to design and help implement responses, including rebuilding homes and livelihoods. |

### *2.2 Data Envelopment Analysis (DEA)*

DEA is a linear programming for measuring the relative performance of Decision Making Unit (DMU) where multiple inputs and outputs make comparisons difficult. DEA shows relative efficiency and determines targets to improve inefficiency units.

Let $\theta^*$be the efficiency:

$$\theta^* = min\ \theta$$

Subject to the following constraints:

$$\Sigma_{j=1}^{n} \lambda_j x_{ij} \leq \theta x_{io} \qquad \text{i = 1, 2,..., m;} \qquad (1)$$

$$\Sigma_{j=1}^{n} \lambda_j y_{rj} \geq y_{ro} \qquad \text{r = 1, 2,…, s;} \qquad (2)$$

$$\Sigma_{j=1}^{n} \lambda_j = 1 \qquad (3)$$

$$\lambda_j \geq 0 \qquad \text{j = 1, 2,…, n.} \qquad (4)$$

DMU$_o$ stands for one out of n DMUs that will be assessed and $x_{io}$ and $y_{ro}$ represent the *i*th input and *r*th output for DMU$_o$. If $\theta^*$ - which indicates the efficiency score - equals 1 then the input can no further be reduced, as DMU$_o$ is already on the frontier. Thus, if $\theta^* < 1$ then the efficiency frontier is above DMU$_o$ and DMU$_o$ can be pushed towards the frontier [10].

For inefficient DMU, it needs DEA slack to reduce input or output in order to become efficient. The first stage of DEA is computed for efficiency score ignoring the slacks. Then slacks are computed in the second stage of DEA. The linear programming of slack model is following:

$$max\ \Sigma_{i=1}^{m} s_i^- + \Sigma_{r=1}^{s} s_r^+ \qquad (5)$$

Subject to:

$$\Sigma_{j=1}^{n} \lambda_j x_{ij} + s_i^- = x_{io} \qquad \text{i = 1, 2,…,m} \qquad (6)$$

$$\Sigma_{j=1}^{n} \lambda_j y_{rj} - s_i^+ = \theta^* y_{ro} \qquad \text{r = 1, 2,…,s} \qquad (7)$$

$\Sigma_{j=1}^{n}\lambda_j = 1$ and (8)

$\lambda_j \geq 0$ j = 1, 2,…, n (9)

Here, $s_t^-$ and $s_r^+$ stand for input and output slack, as input needs to be reduced (-) and output be increased (+) to become more efficient. It should be noted that a DMU is fully efficient if and only if $\theta^* = 1$ $and$ $s_t^- = s_r^+ = 0$ [10].

*2.3 Data Collection*

Staff taking care of disaster management program and working at city and local government were interview based on LGSAT questions guideline. Data was collected from 21 city and local governments in PathumThani. They are Krong Three, Krong Four, Krong Five, Krong Six, Krong Seven, Ban Phatum, Ban Geaw, Ban Kra Chang, Chaing Rak Noy, Lak Six Municipal, Chaing Rak Yai, Saun Prig Thai, Ban Krang Municipal, Kra Chang, Bang Kra Beur, Taey Kor, Bang Pood, Ban Shang, Ku Bang Laung, Bor Guen, and Bang Ka Di Municipal For confidential reasons, we rearranged the order and name DUM i, where i = 1,…, 21. Data is calculated based on LGSAT's five priority actions, used as outputs. A complete data for DEA analysis is shown in Table 3.

Table 3: Inputs and outputs for DEA analysis

| Local Government | Inputs | | | Outputs (result from LGSAT) | | | | |
|---|---|---|---|---|---|---|---|---|
| | Annual budget | Number of Staff | | Priority for Action 1 | Priority for Action 2 | Priority for Action 3 | Priority for Action 4 | Priority for Action 5 |
| DMU 1 | 400,000 | 6 | | 39 | 32 | 19 | 43 | 40 |
| DMU 2 | 3,500,000 | 21 | | 37 | 40 | 21 | 46 | 45 |
| DMU 3 | 300,000 | 11 | | 43 | 43 | 23 | 55 | 46 |
| DUM 4 | 2,000,000 | 8 | | 40 | 32 | 17 | 44 | 42 |
| DMU 5 | 100,000 | 7 | | 40 | 30 | 14 | 38 | 36 |
| DMU 6 | 590,000 | 10 | | 33 | 27 | 12 | 22 | 25 |
| DMU 7 | 150,000 | 6 | | 28 | 21 | 14 | 30 | 31 |
| DMU 8 | 50,000 | 4 | | 36 | 32 | 16 | 43 | 35 |
| DUM 9 | 200,000 | 5 | | 36 | 40 | 21 | 42 | 43 |
| DUM 10 | 300,000 | 13 | | 37 | 30 | 13 | 48 | 38 |
| DMU 11 | 800,000 | 5 | | 31 | 30 | 18 | 41 | 34 |
| DMU 12 | 50,000 | 8 | | 24 | 20 | 13 | 28 | 33 |
| DMU 13 | 1,500,000 | 15 | | 38 | 32 | 15 | 47 | 42 |
| DMU 14 | 240,000 | 5 | | 34 | 24 | 14 | 42 | 31 |
| DMU 15 | 500,000 | 5 | | 42 | 29 | 17 | 44 | 37 |
| DMU 16 | 900,000 | 5 | | 43 | 38 | 23 | 55 | 46 |
| DMU 17 | 300,000 | 15 | | 33 | 30 | 15 | 44 | 36 |
| DMU 18 | 722,000 | 3 | | 39 | 33 | 16 | 46 | 39 |
| DMU 19 | 200,000 | 9 | | 32 | 29 | 12 | 19 | 37 |
| DMU 20 | 300,000 | 10 | | 44 | 39 | 19 | 51 | 37 |
| DMU 21 | 1,000,000 | 27 | | 46 | 40 | 23 | 38 | 42 |

## 3. Result and discussion

Efficiency score and slack of each local government is analyzed. The result shows that nine DMUs are efficient, and one (DMU 12) is weakly efficient (efficiency score = 1 but slacks are not zero). Looking at the input slack, DMU 12 should reduce the number of staff by 4 (8 - 4 = 4 persons). Likewise, to reach the efficient frontier, DUM 2 and 4 should decrease annual budget 593,333.30 and 166,000 baht, respectively. That is, they should control the budget within 2,906,666.70 (3,500,000- 593,333.30) and 1,834,000 (2,000,000-166,000), respectively. Similar reading can be done for other local governments listed in Table 4.

Table 4: Result of DEA analysis

| | | Slacks | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Local Government | Efficiency | Annual budget | Staff | Priority for Action 1 | Priority for Action 2 | Priority for Action 3 | Priority for Action 4 | Priority for Action 5 |
| DMU 1 | 0.8507 | 0 | 0 | 0 | 2.62 | 0 | 0.22 | 0 |
| DMU 2 | 0.3143 | 593,333.3 | 0 | 3.67 | 0 | 1.33 | 4.67 | 0 |
| DMU 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DUM 4 | 0.4821 | 166,000 | 0 | 0.71 | 3.14 | 2 | 5.86 | 0 |
| DMU 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DMU 6 | 0.3746 | 0 | 0 | 3.76 | 5.25 | 4 | 21.76 | 11.02 |
| DMU 7 | 0.6547 | 0 | 0 | 8.22 | 11.07 | 2 | 13.22 | 4.29 |
| DMU 8 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DUM 9 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DUM 10 | 0.5286 | 0 | 0 | 1.92 | 6.55 | 5.92 | 0 | 1.58 |
| DMU 11 | 0.7454 | 0 | 0 | 7.55 | 5.36 | 0 | 4.73 | 6.73 |
| DMU 12 | 1* | 0 | 4 | 12 | 12 | 3 | 15 | 2 |
| DMU 13 | 0.3064 | 0 | 0 | 0.67 | 5.01 | 5.16 | 0 | 0 |
| DMU 14 | 0.7606 | 0 | 0 | 2.59 | 8.2 | 2 | 1.59 | 4.79 |
| DMU 15 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DMU 16 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DMU 17 | 0.2957 | 0 | 0 | 3.59 | 2.86 | 1.64 | 0 | 0 |
| DMU 18 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DMU 19 | 0.47039 | 0 | 0 | 4.03 | 4.97 | 5.22 | 23.79 | 0 |
| DMU 20 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| DMU 21 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

1* = weak efficient

## 4. Conclusion

This paper used a guideline of LGSAT and input-oriented DEA to assess and benchmark the efficiency of local and city governments in PathumThani in building resilience to natural disasters. Nine DMUs are efficient and one DMU is weakly efficient. DEA slacks are target for weakly efficient and inefficient DMUs.

**Acknowledgements**

We would like to thank the staff of city and local governments in PrathumThani for their useful information and precious time for the interview.

**References**

[1] Fakhruddin S.H.M., Chivakidakarn Y. A case study for early warning and disaster management in Thailand. International Journal of Disaster Risk Reduction. 2014; 9: 159–180.

[2] Rahmayanti H., Suganda E., The Analysis of Community Adaptation Process in Constructing Disaster-Prone City (A Study on West Padang), Proceeding of the 4th International Conference on Sustainable Future for Human Security; 2013 October 19-21; Kyoto, Japan. 2013. p. 104-111.

[3] I-soon Raungratanaamporn, Penpathu Pakdeeburee, Akio Kamiko, Chaweewan Denpaiboon. Government-communities Collaboration in Disaster Management Activity: Investigation in the Current Flood Disaster Management Policy in Thailand. Procedia Environmental Sciences. 2014; 20: 658-667.

[4] Hiwasaki L., Luna E., Syamsidik, and Shawd R. Process for integrating local and indigenous knowledge with science for hydro-meteorological disaster risk reduction and climate change adaptation in coastal and small island communities. International Journal of Disaster Risk Reduction. 2014; 10: 15–27.

[5] Overseas Development Institute (ODI). The future framework for disaster risk reduction, a guild for decision-makers [Internet]. 2014 [updated 2014 Nov; cited 2015 May]. Available from: http://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/9230.pdf

[6] Pires S.M., Fidélis T., Ramos T.B.. Measuring and comparing local sustainable development through common indicators: Constraints and achievements in practice. Cities. 2014; 39: 1-9.

[7] Tingsanchali T., Urban flood disaster management. Procedia Engineering. 2012; 32: 25–37.

[8] United Nations Office for Disaster Risk Reduction - Regional Office for Asia and Pacific (UNISDR AP), Local self-assessment leads to resilience – Report [Internet]. 2014 [updated 2014 Jan 10; cited 2015 May]. Available from: http://search.preventionweb.net/?q=LGSAT&search.x=17&search.y=9

[9] WMO, UNEP. IPCC Fourth Assessment Report: Climate Change 2007 [Internet]. 2007 [cited 2015 May]. Available from http://www.ipcc.ch/publications_and_data/ar4/syr/en/main.html

[10] Zhu, J. Quantitative Models for Performance Evaluation and Benchmarking. New York: Springer; 2009.

# Assessing Greening Activities for Small and Medium Enterprises in the Thai Automotive Sector by DEA Approach

Niyat Tesfai[1*], Morrakot Raweewan[2]
[1]*Management Technology, Sirindhorn Institute of Technology, Thammasat University, Pathum Thani, Thailand,*
*niyat.tesfai@yahoo.com*
[2] *Management Technology, Sirindhorn Institute of Technology, Thammasat University, Pathum Thani, Thailand*
*morrakot@siit.tu.ac.th*

## Abstract

Small and medium enterprises (SMEs) are contributing more and more to Thailand's economy. As a result, the Thai industry is increasing in energy intensity, mainly in the manufacturing sector due to outdated machines. Furthermore, studies have revealed that countries that do not belong to the Organization for Economic Corporation and Development (OECD) will continue to increase its energy consumption, as a consequence of their growing numbers of inhabitants as well as economy [2]. The purpose of this study is to assess energy efficiency of SMEs in the Thai automotive sector. Data has been obtained from GIZ Bangkok, which are implementing a Green Auto Project in which they consult SMEs in terms of greening activities. The contribution of this proposed study is the application of Data Envelopment Analysis (DEA) to the Thai automotive sector for SMEs to assess energy efficiency. The DEA model is input-oriented and variable-return to scale (VRS), i.e. that the amount of input should be reduced (energy consumption) and the output-level maintained. Given the data that includes the consumption of gross energy and electrical energy as well as the production volume, the efficient Decision Making Units (DMUs) will be reported. Furthermore, the results will reveal a slack analysis, indicating which input factor to reduce in order to maximize efficiency.

*Keywords*: Automotive Sector, Data Envelopment Analysis, Energy Efficiency, Small and Medium Enterprises

*Corresponding Author
E-mail Address: niyat.tesfai@yahoo.com

## 1. Introduction

The importance of green manufacturing is raising awareness in today's global economy. Businesses are taking sustainable production and environmental activities as part of their corporate social responsibilities (CSR), as they have a huge impact on environmental changes  [9]. Not only are companies modifying single processes with respect to their CSR activities, but are involving the whole supply chain as well. The social and environmental aspects can be integrated into a company's supply chain through inter-organizational management [3]. This study focuses on small and medium enterprises (SMEs) in Thailand and aims to support the Green Auto Project of GIZ[1] (Deutsche Gesellschaft für internationale Zusammenarbeit) in Bangkok. The project aims to enhance sustainable production into the Thai automotive sector. Thailand's SMEs have contributed 36.6% of the overall gross domestic product (GDP) in 2011 and thus, have a great impact on the economy. When looking at the size of enterprises it shall be demonstrated that small enterprises (SE) account more to the countries GDP than medium enterprises (ME). While the value of SEs was 24.5% of Thailand's total GDP in 2011, MEs only contributed 12.1%. Another important finding that the Office of Small and Medium Enterprises Promotion (OSMEP) in Thailand made is that when considering the GDP structure, the agricultural sector contributed 13.3% while non-agricultural sector constituted 86.7% of Thailand's total GDP in 2011. When breaking down the non-agricultural sector the manufacturing sector stands out with the most influential contribution to the economy with 34% of Thailand's total GDP [6]. The purpose of this study is to assess and identify efficient SMEs, whose strategic approach can be implemented by those who did not perform well. It aims to motivate SMEs to participate in the Green Auto Project and inform them about the energy savings they can possibly achieve. The obtained results will provide information of SMEs relative performance within a peer group, which also aims to be a foundation for establishing policy recommendations in order to increase energy efficiency. Because Thailand is an emerging manufacturing country and has high-energy consumption, it has a huge potential of decreasing energy consumption by adapting new strategies and regulations. The Thai Government has been trying to reduce energy consumption by introducing laws and regulations. The under the Thai Energy Conservation and Promotion Act 1992 developed Energy Conservation Fund (ENCON Fund) has been a supportive financial act for developing energy efficiency (EE) and renewable energy (RE). This aid is controlled by the baseline of the governments conservation plan [11]. However, although the Thai government has established several EE and RE projects, the energy consumption has not changed much from the year 2000 onwards. As a result, the Thai government has developed a 20-Year National Energy Efficiency Development Plan that ensures to reduce energy

[1] Engl: German International Corporation

intensity by 25% likened with levels of 2005 by 2030 [5].

## 2. Methodology

### *2.1 Data Preparation*

A total number of 16 SMEs have been provided and grouped according to commonality of data. Thus, the SMEs have been grouped into 3 peer groups with respect to the same units of outputs. The data set includes 2 inputs (electrical and gross energy consumption) and one output (the end products). The SMEs in a peer group produce various and different types of automotive parts, as one product might require more energy than another, thus the data is rather heterogeneous. This is a problem because DEA assumes homogeneity of data. Even though DEA is capable of handling multiple inputs and outputs, the standardization of data is recommended for the purpose of this study. The statistical approach of standardization, namely z-standardization will be applied and the transformed data known as z-scores used for the DEA application [7]. Prior to the explanation of the standardization process, the following variables shall be defined:

$X_{ij}$: *the value of an attribute j for DMU i*

$\bar{X}_j$: *the mean for an attribute j across all i DMUs*

$\sigma_j$: *the standard deviation of an indicated value for attribute j*

$Z_{ij}$: *the z scores (the number of standard deviatons DMU i is above or below the mean for an attribute j)*

The z-scores can be computed by applying the following formula:

$$Z_{ij}: \frac{(X_{ij}-\bar{X}_j)}{\sigma_j} \qquad (1)$$

A z-score is a standardized value and indicates how many standard deviation units and in which direction (sign) a value deviates from the mean of its sample. Through the transformation the values of a distribution with different means and variations are made comparable with regard to its relative deviation from its mean. Note that the z-standardization does not change the form of a distribution. The sum of all z-scores of an attribute j equals 0, implying that there are negative z-scores as well. The standard deviation of a sample is always 1 and the mean 0 [10]. However, one must bear in mind that the DEA software is not capable of assessing negative data. Thus, the application of negative z-scores is not feasible. To circumvent this problem, a random variable R will be added to the z-scores of an attribute j, so that all the z-scores become positive. This procedure is based on the study of Powers and McMullen (2002), where the absolute value of the minimum z-score has been added to each z-score of an attribute. The mean of an attribute is then relocated from 0 to the random variable R, however, the standard deviation remains 1.

The table below illustrates the z-scores for all the peer groups that will be used for DEA application. It shall again be mentioned that the indicated random x-values have been added to the computed z-scores in order to overcome negativity.

Table 1: Standardized data

Peer Group 1

| DMUs | Electricity (kw/year) | Gross Energy (MJ/year) | Products in Kg |
|---|---|---|---|
| DMU1 | 1 | 1 | 0,22 |
| DMU2 | 1 | 1 | 0,37 |
| DMU3 | 3 | 3 | 2,41 |
| X-Value | 2 | 2 | 1 |

Peer Group 2

| DMUs | Electricity (kw/year) | Gross Energy (MJ/year) | Products in pieces |
|---|---|---|---|
| DMU1 | 3 | 1 | 2 |
| DMU2 | 3 | 4 | 2 |
| DMU3 | 2 | 2 | 1 |
| DMU4 | 2 | 2 | 1 |
| DMU5 | 1 | 1 | 5 |
| DMU6 | 4 | 4 | 2 |
| DMU7 | 2 | 2 | 1 |
| DMU8 | 1 | 1 | 1 |
| DMU9 | 1 | 1 | 2 |
| X-Value | 2 | 2 | 2 |

Peer Group 3

| DMUs | Electricity (kw/year) | Gross Energy (MJ/year) | Products in tons |
|---|---|---|---|
| DMU1 | 1,064 | 0,752 | 1,22 |
| DMU2 | 0,079 | 0,433 | 3,532 |
| DMU3 | 2,245 | 2,981 | 1,223 |
| DMU4 | 2,612 | 1,834 | 1,222 |
| X-Value | 1,5 | 1,5 | 1,8 |

The numbers in the table imply that for example DMU1 in peer group 1 needs one unit of electrical energy and one unit of gross energy to produce 0.22 products. The same interpretation applies for the remaining DMUs in the above table.

### *2.2 Data Envelopment Analysis*

DEA is a data-oriented analysis tool that assesses the performance of a particular number of peer entities, also known as Decision Making Units (DMUs), which transfer multiple inputs into multiple outputs. The DMUs used for DEA are diverse and flexible. Examples for DMUs are hospitals, banks, and organizations or simply processes, regions etc. One crucial function of DEA is that it does not require much internal information and assumptions; it simply considers internal information as a "black box". Solely the multiple inputs and multiple outputs of DMUs are of relevance for the analysis tool [12].

Farrell first developed DEA in 1957 and has developed the efficiency frontier. Subsequently, Charnes, Cooper and Rhodes (1978) enhanced this method by developing a nonparametric analysis in using efficiency assessment. DEA functions by using a linear programming methodology that converts inputs into outputs produced. DEA identifies a frontier (best practice) and compares the relative performance of organizations or units to the best practice. The range of efficiency is between 0 and 1, in which 1 is the most efficient unit/organization and 0 the least [4]. There are two major types of DEA; which can be either input- or output-oriented. The input-oriented model refers to the minimization of the objective function, i.e. the objective is to minimize the amount of input, while maintaining the same amount of output. Similarly, the output-oriented model considers the maximization of the problem. The objective is to maximize the amount of output with the same level of input [1]. DEA can also be differentiated according to its condition, which can be a constant or variable-return-to-scale, (CRS) and (VRS). The CRS model implies that the output level is proportional to the input level, while the VRS model supposes that the output level is higher or less than the increased input [4]. The method applied for this research is an input-oriented VRS envelopment model. Thus, the linear programming formulation of the model is as follows:

Let $\theta^*$be the efficiency:

$\theta^* = min\theta$

Subject to the following constraints:

$\Sigma_{j=1}^{n} \lambda_j x_{ij} \leq \theta x_{io}$ i=1,2,...m; (2)

$\Sigma_{j=1}^{n} \lambda_j y_{rj} \geq y_{ro}$ r=1,2,…s; (3)

$\Sigma_{j=1}^{n} \lambda_j = 1$ (4)

$\lambda_j \geq 0$ j=1,2,…n. (5)

Here, DMU$_o$ stands for one out of n DMUs that will be assessed and $x_{io}$ and $y_{ro}$ represent the *i*th input and *r*th output for DMU$_o$. If $\theta^*$ - which indicates the efficiency score - equals 1 then the input can no further be reduced, as DMU$_o$ is already on the frontier. Thus, if $\theta^* < 1$ then the efficiency frontier is above DMU$_o$ and DMU$_o$ can be pushed towards the frontier [12].

The individual reduction of input or output (depending on the model orientation) in order to reach the efficient frontier is known as DEA slack. Apart from the basic DEA envelopment model, DEA is mostly used as a 2-stage procedure as it first analyzes the efficiency scores and the slacks in the second stage. Hereby, the efficiency is first computed by ignoring the slacks and in the second stage the efficiency rate is fixed and the slacks calculated. The slack models can be designed on the basis or extended to the efficiency model, which are the following in particular:

$max\Sigma_{i=1}^{m} s_i^- + \Sigma_{r=1}^{s} s_r^+$ (6)

Subject to:

$\Sigma_{j=1}^{n} \lambda_j x_{ij} + s_i^- = x_{io}$ i= 1, 2,…,m (7)

$\Sigma_{j=1}^{n} \lambda_j y_{rj} - s_i^+ = \theta^* y_{ro}$ r= 1, 2,…,s (8)

$\Sigma_{j=1}^{n} \lambda_j = 1$ and (9)

$\lambda_j \geq 0$ j= 1, 2,…, n (10)

Here, $s_t^-$ and $s_r^+$ stand for input and output slack, as input needs to be reduced (-) and output be increased (+) to become more efficient. It should be noted that a DMU is fully efficient if and only if $\theta^* = 1\ and\ s_t^- = s_r^+ = 0$ [12]. Should the efficiency score be 1 and the slacks $\neq 0$, the DMUs are considered as weakly efficient. Weakly efficient units will still be considered as efficient, however, when calculating the slacks, i.e. in stage two, weakly DMUs will not be included in the reference set as a benchmarking unit to other DMUs [8].

## 3. Results

The table below indicates the results obtained with the DEA approach.

Table 2: DEA results

| Peer Group 1 | | | |
|---|---|---|---|
| DMUs | Efficiency Score | | Reference Set |
| DMU1 | 1 | Weakly efficient | 1 (DMU2) |
| DMU2 | 1 | Efficient | - |
| DMU3 | 1 | Efficient | - |
| **Peer Group 2** | | | |
| DMUs | Efficiency Score | | Reference Set |
| DMU1 | 1 | Weakly efficient | 1 (DMU5) |
| DMU2 | 0,33 | Inefficient | 1 (DMU5) |
| DMU3 | 0,5 | Inefficient | 1 (DMU5) |
| DMU4 | 0,5 | Inefficient | 1 (DMU5) |
| DMU5 | 1 | Efficient | - |
| DMU6 | 0,25 | Inefficient | 1 (DMU5) |
| DMU7 | 0,5 | Inefficient | 1 (DMU5) |
| DMU8 | 1 | Weakly efficient | 1 (DMU5) |
| DMU9 | 1 | Weakly efficient | 1 (DMU5) |
| **Peer Group 3** | | | |
| DMUs | Efficiency Score | | Reference Set |
| DMU1 | 0,57 | Inefficient | 1 (DMU2) |
| DMU2 | 1 | Efficient | - |
| DMU3 | 0,14 | Inefficient | 1 (DMU2) |
| DMU4 | 0,23 | Inefficient | 1 (DMU2) |

Out of the 3 DMUs in peer group 1, DEA identified 2 DMUs as efficient. DMU1 has an efficiency score of 1 and is weakly efficient. Recalling from the previous chapter a DMU is fully efficient, only and only if the efficiency score equals 1 and slacks equal zero. Because DMU2 uses less inputs and produces more output than DMU1, DEA indicates that DMU1 has the potential to either decrease its input by producing the same output or remain its input level and producing more output. In other words, DMU1 has found to be weakly efficient in comparison to DMU2, which is why it is being referred to in the reference set by 100%. The second peer group has also 1 efficient DMU, namely DMU5. DMU1, 8 and 9 are only weakly efficient, due to the fact that the slacks are non-zero. The remaining DMUs are relatively inefficient and are referred to DMU5 for improvement. As DMU5 is identified as relatively efficient, it will not be compared to any other DMU within the peer group and is its own reference. The computed weights in the column of the reference set indicated in the above table are assigned to each reference set member in order to calculate the efficiency rating. In other words, it reveals information for weakly efficient or inefficient DMUs about how much weight of which reference set member to apply in order to move towards the efficient frontier. In this case, only one efficient DMU has been identified per peer group, which is why a full weight of 100% is assigned to the best practice in the reference set. If there would be more than one efficient DMU, then DEA would compute every possible combination of weights to make the weakly efficient or inefficient DMU move towards the efficient frontier. Hence, the weakly and inefficient DMU are suggested to apply 100% of the input and output combination the best-practice DMU is using.

With regards to the slacks, DEA provides recommendation with respect to the computed weights in the reference set. Looking into peer group 1, DMU1 has the potential to increase its products because it uses more input to produce less output compared to DMU2. But because the objective is to minimize input, in this case energy consumption, it is advised to decrease the input and maintain the level of output. As for peer group 2 the weakly efficient and inefficient DMUs are fully benchmarked to DMU5. DEA computed slack adaption for DMU1, which should decrease the electrical energy consumption and for DMU2 to reduce the gross energy level towards the level of DMU5. The remaining DMUs have the potential to increase their products, but the reduction of input should preferably be targeted. For peer group 3, DMU1, 3 and 4 are fully benchmarked to DMU2. Thus, the reduction of electrical energy to the amount of DMU2 is recommended.

## 3. Conclusion

This paper assessed the energy efficiency of SMEs in the Thai automotive sector. The relative performance of SMEs in terms of energy efficiency has been analysed. 4 DMUs out of 16 DMUs are identified as energy efficient. Furthermore, 4 DMUs are identified as weakly efficient and 8 as inefficient. The findings highlight the value of benchmarking as a foundation for improvement opportunities in energy savings and the demonstration of energy consumption of SMEs in the Thai automotive sector. Additionally, results have indicated that some SMEs are able to use less energy consumption to produce more goods than others. Hence, results suggest that with respect to the GIZ project the reason of inefficiency should be further investigated. The value added to the SMEs is the target-oriented slack recommendation that gives flexibility on selecting actions to reduce energy consumption.

The results are of high importance because Thailand as one of the manufacturing countries in the world needs to further address the issue of energy efficiency as it has a huge potential in improving it. It is important that the energy-intensive sectors consider the high influence an efficient energy consumption can make and conduct arrangements that lead to a productive way of managing energy efficiency. The information obtained from the benchmarking results can be used to improve the relative performance of individual SMEs. Energy has a great impact on the economical performance of an organization. Improving energy efficiency reflects a great opportunity to save economic resources, improve production processes while also having a positive impact on the environment. Especially SMEs in Thailand that contribute a lot into the Thai GDP as pointed out in the introduction have a huge potential in enhancing energy efficiency, as it is often the case that out-dated machinery is used, awareness is missing or financial measures are unavailable.

For this to be implemented, awareness of the latest technology should be demonstrated. Other actions include the substitution of inefficient lightings and air conditioners to more efficient ones, the application of efficient appliances such as Minimum Energy Performance Standards or with efficiency labelling's. Lights and equipment's should be turned off when not in use and if financial resources allow it, investments in photovoltaic solar systems are recommended. The National 20-year Energy Efficiency Development Plan (EEDP) developed by the Thai Ministry of Energy targets the transportation and industrial sector. The goal is to reduce energy intensity by 25% in 2030, compared to the level of 2005. The EEDP has developed rules and regulations as well as provided subsidies that will help achieve energy efficiency. The rules and regulations involve the establishment of Minimum Energy Performance Standards (MEPS) and the labelling's of energy efficiency on appliances. The subsidies include a Standard Offer Program (SOP) or the approval of funding's for the achieved energy savings [5]. The policy recommendations address the issue of improving energy efficiency. The study has provided information on the performance and identification of a particular peer group within the Thai automotive sector that need to adapt the above-mentioned policy recommendations for efficiency maximization.

## References

[1] Atici, K. B., & Podinovski, V. V. *A REVIEW OF DATA ENVELOPMENT ANALYSIS STUDIES IN AGRICULTURE* . University of Warwick, Warwick Business School, Coventry, United Kingdom .

[2] EIA. (2013, December 3). *eia.gov*. (U. E. Administration, Producer) Retrieved October 12, 2014, from http://www.eia.gov/todayinenergy/detail.cfm?id=14011

[3] Kogg, B. (2012, November). Environmental and social responsibility in supply chains: The practise of choice and inter-organisational management. *Ecological Economics* .

[4] Liu, W.-B., Wongchai, A., & Peng, K.-C. *Adopting Super- efficiency and Tobit Model on Analyzing the Efficiency of Teacher's Colleges in Thailand.* National Pingtung University of Science and Technology, Taiwan R.O.C, Department of Tropical Agriculture and International Cooperation.

[5] Ministry of Energy. (n.d.). Retrieved April 4, 2015, from enconfund.go.th: http://www.enconfund.go.th/pdf/index/EEDP_Eng.pdf

[6] OSMEP. (2011). *Office of Small and Medium Enterprises Promotion.* Retrieved October 10, 2014, from http://eng.sme.go.th/Pages/home.aspx

[7] Powers, J., & McMullen, P. R. (2000). USING DATA ENVELOPMENT ANALYSIS TO SELECT EFFICIENT LARGE MARKET CAP SECURITIES. *Journal of Business and Management* , *7* (2), 31-42.

[8] Sherman, H. D., & Zhu, J. (2006). *Service Productivity Management: Improving Service Performance using DATA ENVELOPMENT ANALYSIS (DEA).* New York: Springer.

[9] Tikul, N. (2014, January). Assessing environmental impact of small and medium ceramic tile manufacturing enterprises in Thailand. *Journal of Manufacturing Systems* .

[10] Uni Hamburg. (n.d.). *uni-hamburg.de.* Retrieved March 28, 2015, from http://www2.jura.uni-hamburg.de/instkrim/kriminologie/Mitarbeiter/Enzmann/Lehre/StatIKrim/Z-Werte.PDF

[11] Wang, X., Stern, R., Limaye, D., Mostert, W., & Zhang, Y. (2013, September). *worldbank.org.* Retrieved October 11, 2014, from http://elibrary.worldbank.org/doi/book/10.1596/978-1-4648-0020-7

[12] Zhu, J. (2009). *Quantitative Models for Performance Evaluation and Benchmarking.* New York: Springer.

# TRACER STUDY OF THE BS STATISTICS PROGRAM AT EASTERN VISAYAS STATE UNIVERSITY: 1997 - 2014

Dennis C. De Paz[1] and Thea C. Galos[2*]
[1]Mathematics and Statistics Department, Eastern Visayas State University, Tacloban City, Leyte 6500, Philippines, depaz_dennis@yahoo.com
[2] Mathematics and Statistics Department, Eastern Visayas State University, Tacloban City, Leyte 6500, Philippines, thea_galos@yahoo.com

## Abstract

In today's increasing competitive environment and pursuant to the envisioned reform in statistics education, this study assessed the status of the Bachelor of Science in Statistics graduates' at Eastern Visayas State University, Tacloban City, Philippines for appropriate policy redirection and curriculum improvement aligned with the regional, national and global developmental goals. Based on the principles of evaluation, it describes the employment profile and other factors of the graduates. It has been more than 20-years now since its first implementation that the university has produced 281 graduates and it is expected that these graduates must have brought with them the zeal of excellence and competence as statisticians imbued with positive values that could be an instrument of change towards human development and progress. However, some of the graduates are employed, under-employed and there are still who are unemployed. Graduates were employed as statistical analyst, enumerators, researchers, and in office works. The authors find that there is a need to heighten the employability of the of the graduates in line with their specialization immediately after graduation as there are more opportunities in the international and local arena of public and private statistical agencies.

*Keywords*: BS Statistics alumni, eastern visayas state university, statisticians, tracer study

*Corresponding Author
E-mail Address: thea_galos@yahoo.com

## 1. Introduction

Tracer study serves an important tool to measure the relevance of educational programs. As such many schools all over the world mandate such study for closely monitoring employment status of their graduates.

The importance of statistics education is highly recognized as a means towards empowering citizens. In the Philippines, among the disciplines highly recognized as a program is statistics. Many institutions have developed major programs in statistics and so more universities in the Philippines are now offering a degree program in statistics.

Bersales [1,2] reported that by 2006, there were 19 colleges and universities nationwide offering statistics degree program in the tertiary level. With the proliferation of state universities and colleges, the presence of academic institutions offering statistics is already nationwide.

The memorandum dated June 10, 2011 issued by the Commission on Higher education (CHED) on the "Positive Lists of Higher Education Institutions (HEIs) Offering Programs/Courses with Recognition/Permits, disclosed that about 20 state universities and colleges have offer the BS Statistics Program throughout the country. One of these state universities and colleges offering the BS Statistics Program is the Eastern Visayas State University in Tacloban City, Leyte, Philippines, the second university in Leyte to offer a curriculum leading to the B.S. degree in statistics.

The BS Statistics program began with the establishments of the College of Arts and Sciences under the Mathematics and Statistics Department, effective first semester of school year 1993 - 1994 under the Board Resolution No. 34 s. 1993 [3]. Its mission endeavor to prepares the students for a more rewarding career in statistics, technically equipped and imbued with intellectual honestly considered vital in decision making and in other fields of endeavor.

Equally important, the program goals primarily are the following: (a) satisfy the special demand of research, government & private offices for highly trained graduates in the field of applied statistics, (b) to produce graduates who are technically qualified to select and gather data and make an in-depth statistical analysis, (c) establish the Eastern Visayas State University as a center of excellence of BS Statistics program in the region. Basically, career prospect for BS Statistics graduates ranging from statistician, teaching, data analysis in industries, government agencies, research assistants, enumerators/encoders and office works

Along this vein, it is surmised that one way to assess the status of the program is by asking the graduates' perceptions and experiences on the value of their education. As such, this paper discusses the current status of the program's graduates as inputs in assessing and enhancing program quality.

## 2. Research Methodology

This tracer study utilized descriptive method of research using the standard survey questionnaires as the basic instrument in gathering the data.

### *2.1 The Study Population*

Table I presents the total number of graduates and study population. During the 18 years of existence as a program from 1997 – 2014, 281 BS Statistics graduates have been produced already. In the study, 208 graduates, representing 74% of the population were considered as respondents due to accessibility.

Most of the respondents in this study belong to class 2010 and 2006 where the biggest number of graduates was pegged at 30 and 24, respectively. It could be implied that the total number of graduates of this program has been low. Simply put, only few can make to senior years and graduate in this program. However, this is a common observation among BS Statistics programs nationwide [4,5,6]. Few graduates in the program may create an impression that the program produces very good students since only few can survive.

Table 1. Total Graduates and Study Population

| Alumni Class | No. of Graduates by Sex | | Total |
|---|---|---|---|
| | MALE | FEMALE | |
| 1997 | 3 | 8 | 11 |
| 1998 | 3 | 12 | 15 |
| 1999 | 5 | 12 | 17 |
| 2000 | 5 | 5 | 10 |
| 2001 | 6 | 8 | 14 |
| 2002 | 3 | 4 | 7 |
| 2003 | 1 | 7 | 8 |
| 2004 | 8 | 13 | 21 |
| 2005 | 5 | 9 | 14 |
| 2006 | 10 | 14 | 24 |
| 2007 | 5 | 11 | 16 |
| 2008 | 4 | 10 | 14 |
| 2009 | 3 | 7 | 10 |
| 2010 | 6 | 24 | 30 |
| 2011 | 3 | 14 | 17 |
| 2012 | 2 | 16 | 18 |
| 2013 | 4 | 12 | 16 |
| 2014 | 5 | 14 | 19 |
| **Total Graduates** | **81** | **200** | **281** |
| **Study Population** | **49** | **159** | **208** |

### *2.2 The Research Instrument*

The standard CHED questionnaire adopted in gathering the needed data. The research instrument focused on the graduates' profile, present employment and the relevance of general and professional subjects.

### *2.3 Data Collection*

Data on graduates were taken from the EVSU Registrar's Office. Based on the graduate addresses, the researchers sent questionnaires for them to fill out. However, with the aid of electronic devices such as internet connection the researchers were able to reach the respondents especially those employed and living outside the city. Answered questionnaires were retrieved through facebook and emails. Respondents that were accessible within the Region were reached through personal meeting, mails and phone interview.

### *2.4 Data Processing and Analysis*

Since the study is based on the survey questionnaire, quantitative data such as the numerical descriptive measures such as frequency and percentages were used in the study.

## 3. Research Results and Discussion

### *3.1 Graduates Profile Variables*

The survey elicited the graduates' age, sex, civil status, and highest educational attainment comprises the respondents' personal background. Their reasons for enrolling BS Statistics and professional development were also surveyed in order to get a broader picture of the graduates' profile variables.

#### *3.1.1 Personal Background of Graduates*

*Age.* As seen in Table 2, most of the respondents engaged in this study are very young (below 25 years old) comprising 65 or 31.3% of the graduates, young (25-29 years old) comprising 61 or 29.3%, the early middle age of 30-34 years old comprising 53 or 25.5%, and 35 and above regarded as middle age comprising 29 or 13.9% of the graduates. It can be implied from the study that most of the graduates who responded are very young. Further, these are the respondents who are reachable as some of the respondents pursue graduate studies while some are working in Region VIII, Samar and Leyte, Philippines.

*Sex.* It can be gleaned that most of the graduates who responded were females (159 or 76.4%) while less than one-third of the female totality were male (49 or 23.6%). This evidence suggests that the number of graduates was dominated by females.

*Civil Status.* It can be noted that 96 (71.10%) graduates were still singles and 39 (28.89%) were already married.

*Educational Attainment.* Majority of the graduates were bachelor's degree holders without masters units comprising 77.4% or 161 graduates, 15.4% or 32 graduates earned units towards a master's degree while 7.2% or 15 graduates were master's degree holders.

Table 2. Distribution of Graduates Profile Variables

| PERSONAL BACKGROUND (n = 208) | | f | % |
|---|---|---|---|
| **Age** | | | |
| | 35 and above (middle age) | 29 | 13.9% |
| | 30-34 (early middle age) | 53 | 25.5% |
| | 25-29 (young) | 61 | 29.3% |
| | Below 25 (very young) | 65 | 31.3% |
| **Sex** | | | |
| | Male | 49 | 23.6% |
| | Female | 159 | 76.4% |
| **Civil Status** | | | |
| | Single | 131 | 63% |
| | Married | 77 | 37% |
| **Educational Attainment** | | | |
| | Masters Degree | 15 | 7.2% |
| | BS with MS / MA units | 32 | 15.4% |
| | BS Statistics | 161 | 77.4% |

### *3.1.2 Reasons for Enrolling BS Statistics*

While it is true that there is a high demand for statistics [7] statistics is not a popular course with many students [8], the results as presented in Table 3 using a list of 14 possible reasons, the graduates were asked to select multiple reasons that best explained their pursuit of BS Statistics. The top three main reasons of the graduates for pursuing the degree was "Affordable for the Family", second with tied ranks were "Influence of Parents / Relatives" and Prospect for Employment". On the one hand, these are the respondents' most ticked reasons in the survey and on the other hand, the least frequently cited reasons are peer influence and prospect for career advancement.

Table 3. Reasons for Enrolling BS Statistics

| Reasons for Enrolling BS Statistics | % | Rank |
|---|---|---|
| High Grades Related to the Course | 5.7% | 11 |
| Good Grades in High School | 7.4% | 6 |
| Influence of Parents / Relatives | 9.9% | 2.5 |
| Peer Influence | 4.6% | 13.5 |
| Inspired by Role Model | 8.2% | 4 |
| Strong Passion for the Profession | 6.7% | 8 |
| Prospect for Immediate Employment | 9.9% | 2.5 |
| Status or Prestige of the Profession | 5.3% | 12 |
| Availability of Course Offering in Chosen Institution | 7.6% | 5 |
| Prospect of Career Advancement | 4.6% | 13.5 |
| Affordable for the Family | 10.7% | 1 |
| Prospect of Attractive Compensation | 6.1% | 9.5 |
| Opportunity for Employment Abroad | 6.1% | 9.5 |
| No Particular Choice of No Better Idea | 7.2% | 7 |

### *3.1.3 Professional Development of Graduates*

Figure 1 presents the professional development of graduates in terms of advanced studies (27.9%) and training programs (30.3%). However, 41.8% of the graduates did not pursue for professional development. The findings serve as an input that having a baccalaureate degree alone is not enough, one must earn a higher degree and update oneself of the most recent developments in order to be competitive and at par with others, and hence several challenges on teaching statistics are faced in the educational system, not only in the Philippines but also in other country [9,10].

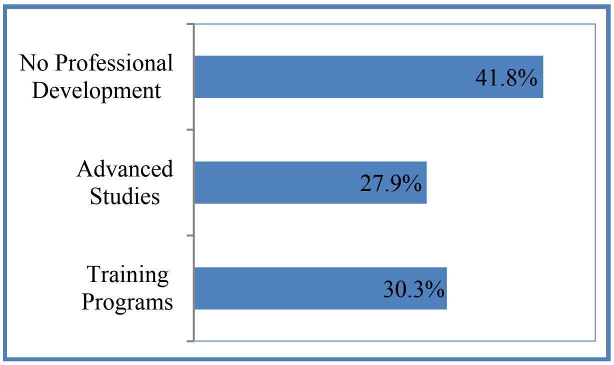


Figure 1. Professional Development of Graduates

### *3.2 Employment Profile of Graduates*

In this section, the status of the graduates in terms of employment is presented.

### *3.2.1 Reasons for Unemployment*

Figure 2 presents the reasons for unemployment of the graduates. The most numbered reasons for unemployment was lack of experience (32.2%), second was no job opportunity (25.0%), next in rank was did not look for a job (22.6%) and the last ranked was family concern (20.2%).

This implies that those graduates did not look for a job, instead pursue their graduate studies. Moreover, unemployed graduates specifically for female decided to prioritize their family concern rather than pursue their career.

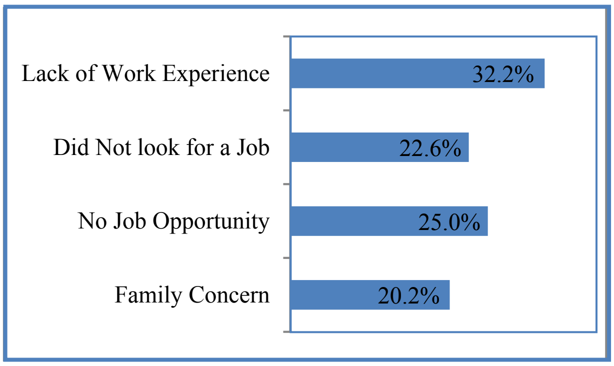


Figure 2. Reasons for Unemployment

### *3.2.2 Present Occupation*

It could be gleaned from Figure 3 that out of 208 graduates, 41% graduates were employed as statisticians and 59% graduates as non-statisticians. Further, statistically related jobs at present for graduates dwell along statistician aide, statistical analysts, research assistants and enumerators.

On the other hand, non-statisticians' job ranges from being clerk, businessman, police officer and secondary teacher. Some respondents completed a second course in law, nursing, and theology. For the time being, these graduates are already full-fledged attorney, nurse, and priest.

It could be implied from this study that there is a need to increase the employability of the graduates as statisticians immediately after graduation as there are more opportunities that exists within public and private sectors.

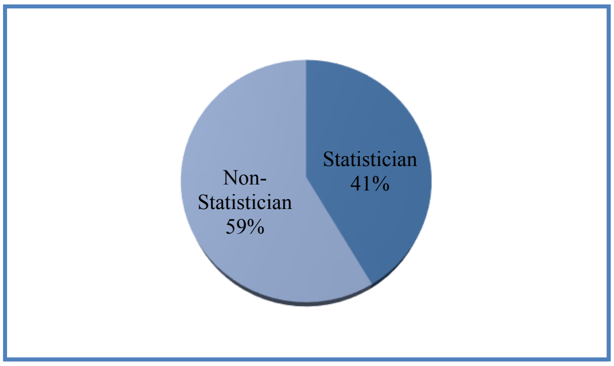


Figure 3. Distribution of Present Occupation by Type of Work

### *3.2.3 Present Employment Status*

It could be gleaned from Figure 4 that most of the respondents were employed as permanent (36.1%). This implies that the graduates were employed right after graduation. Further, it has been the observation of the department that most students before they graduate have already eligibility because of the career sub-professional and professional examination they pass; employed with temporary such reached 26.4% and respondents with contractual and other status have a percentage of 22.6% and others 14.9%, respectively.

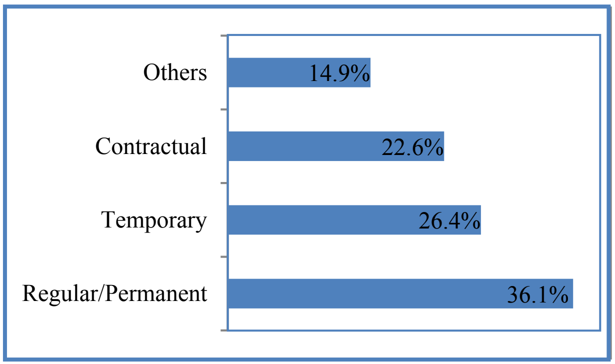


Figure 4. Present Employment Status

### *3.2.4 Place of Work*

Gleaned from Figure 5, there are more opportunities for working locally which comprises 84% of the graduates while 16% of the graduates were working abroad. The findings revealed that it is more convenient to find work here that abroad in the sense that the processing fees for working abroad is very expensive aside from being far from family and relatives.

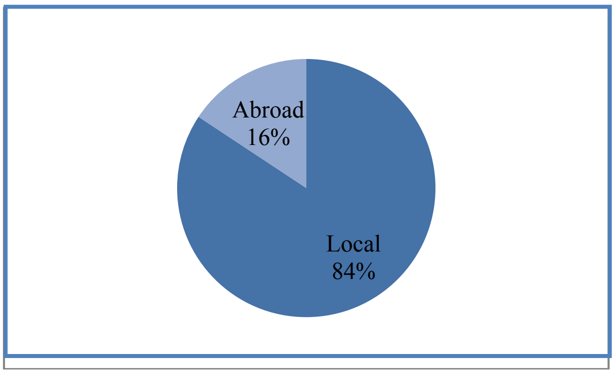


Figure 5. Place of Work of the Graduates

### *3.2.5 Duration of First Employment*

Figure 6 shows that majority of the graduates obtained their first job after completing the degree is “years” (50%). These graduates are most likely to be employed outside the region of Tacloban City. The next highest length of time up to the first job is “months” (35%). This finding coincides with their own self-categorization in number of months in the company.

It is over emphasized in this study that there are few graduates who obtained their first job in “days” with (15%). This observation in the Mathematics and Statistics department is manifested by the fourth-year students who are already working in the statistical agency while in the process of completing their degree.

As deduced, the main factor that facilitated in the short-period of getting the job in the company is the college preparation in the sense that almost every government agency uses statistically related works for various projects and research. Secondary to this is the eligibility. This implies that right after graduation, these graduates were able to get an employment.

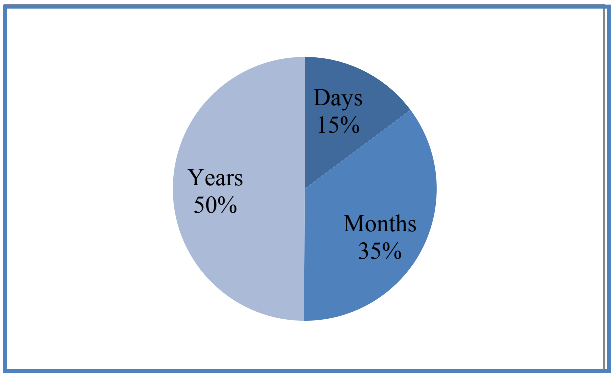


Figure 6. Duration of Employment by Length of Time

### *3.2.9 Salary*

Majority (28.2%) of respondents who work as temporary were given salary ranging between Php5,000 and Php10,999. Graduates who have been employed as regular received the highest salary range (23.7%) those with Php21,000 above while the lowest salary range below Php5,000 were earned by those working as casual and self-employed. This distribution of salaries appears in Figure 7.

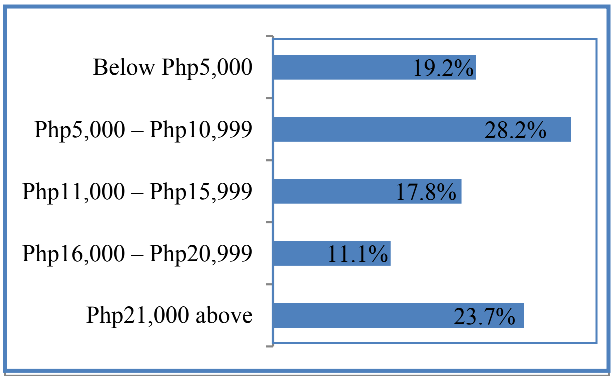


Figure 7. Distribution of Salary Ranges

### *3.3 Factors in Decision Making of the Graduates in Relation to their Acceptance of the Job*

Figure 8 shows the factors in decision making of graduates related to the acceptance of their job. It reflects the factors in terms of salary, geographic location, opportunity for further training, interpersonal relation at work, physical conditions of workplace, work-suitability to trainings, and opportunity for advancements. Based on the findings, all factors are indispensable major decisions of the graduates. Of the utmost importance, it can be inferred in the results that about 73.3% graduates considered "salary" as very important or the top priority in the acceptability of the job. High salaries and compensation are most likely the reasons that attract graduates as their skills are specialized, conversely compensation package is one of the extrinsic motivation. Equally very important factor is "opportunity for advancements" with 60.8%. It could be implied from this finding that graduates feel that they are highly motivated about their own professional development and promotional opportunities.

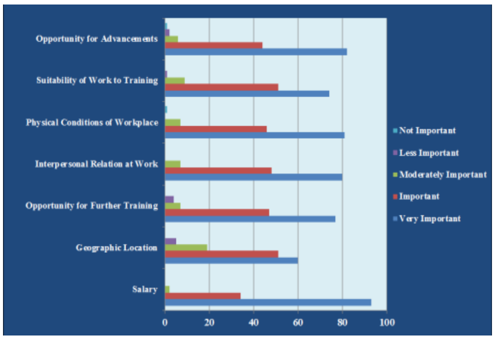


Figure 8. *Factors in Decision Making of the Graduates in Relation to their Acceptance of the Job*

Deemed necessary, the third factor is the physical conditions of workplace with 60%. The workplace environment plays a crucial role for graduates for accepting and/or keeping the jobs. Further, workplace satisfaction has been associated with job satisfaction. Beaconed on a widely accepted assumption the interpersonal relation at work motivates graduates to become productive in their workplace. Furthermore, interpersonal role is very important to encourage positive relations and increase self-confidence of the graduates. Thus, 59.3% graduates strongly agree that relations with peers and superiors at the workplace affect their performance.

Also important and essential factor is the opportunity for further training when it comes to accepting the job as reflected by 57.0% graduates. It cannot be denied that graduates also contemplate on how to broaden their skills. Necessitous is the suitability of work to trainings with a general agreement among 54.8% graduates. Many graduates feel that the acceptance of the job is suitable for training related to their work. Lastly, in determining the job there was a common agreement among 44.4% graduates that geographic location is the least priority.

### *3.4 Job Satisfaction*

Figure 9 shows the graduates job satisfaction in their present job. Many experts believe that job satisfaction can influence work productivity and is considered as a strong predictor of individual well-being. Every job is an instance of employment relationship, a substantive or implied ability to work for rewards both for material and recognition.

It could be further revealed that there is consensus among graduates that they are contented in their present job in terms of the following: opportunity for advancements with 43.7%; physical conditions of workplace with 43.0%; the salary and geographic location have an equal percentage of 42.2%. The finding strongly suggests that graduates experience a high level of satisfaction about their jobs and different aspects of their jobs. It could goes to show that graduates' jobs are fulfilled in their current situation.

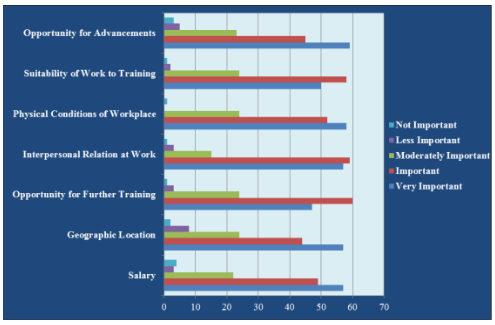


Figure 9. *Job Satisfaction in their Present Job*

On the other hand, secondary to the aforecited perceptions, graduates were satisfied in their present job in terms of opportunity for further training with 44.4%, interpersonal relation at work with 43.7%, and work-suitability to trainings with a frequency of 43.0%. It can be said also that these factors are composite measure of overall job satisfaction.

### ***3.5 Extent of Relevance of Statistics Curriculum***

Table 3 presents the mean rating on the extent of relevance of the curriculum. Graduates felt that the curriculum is "relevant". The highest mean score was the on-the-job training with $\bar{x} = 3.75$. This further implies that OJT is considered relevant as first-hand experience in preparation to their future work.

Secondary to OJT being considered also as "relevant" is the professional education with $\bar{x} = 3.72$. The result serves as an input to revisit the curriculum. Lastly, the general education subjects with $\bar{x} = 3.13$ interpreted as "fairly relevant".

On the basis of the results, there is a need to come up with a development improvement plan for the program in order to achieve the quality of statistics education.

*Table 3. Extent of Relevance of Statistics Curriculum*

| Learning Areas | Mean | Rank | Description |
|---|---|---|---|
| General Education | 3.13 | 3 | Fairly Relevant |
| Professional Education | 3.72 | 2 | Relevant |
| OJT | 3.75 | 1 | Relevant |
| **Overall Mean** | **3.53** | | **Relevant** |

## 4. Conclusions

This tracer study utilizes 208 out of 281 graduates from 1997 – 2014. It could be deemed that majority of the students are young, female, single, and with regards to the educational attainment, most of the graduates are still Bachelor's degree holder, thus, further implies that based on the findings most graduates have no professional development and lack of work experience.

The number of graduates in the program has been low and majority of the respondents refer the course BS Statistics because it was affordable for the family. The present employment statuses of the graduates for the non-statisticians are temporary, contractual, and others having salary ranges from Php5,000 – Php10,999. Most of the graduates' statisticians are regular employee employed locally. It is overemphasized in this study that there is a need to conduct a continuous comprehensive reviews and assessment for program improvement.

Consequently, as corroborated, BS Statistics is one of the fully recognized programs in the pure science stream which immense contribution in propelling sustainable development is acknowledged. In today's rapid changes and emergent reforms in statistics education, it is quintessential for any higher education institution to do a tracer study. Hence, the results of the study would serve as an input to the researcher of formulating a development improvement plan for the program. The findings of the study will also serve as a useful reference for policy makers, practitioners, and professionals in planning appropriate actions and decisions.

## 5. References

[1] Bersales, L. G. S. (2006) Academic Programs in Statistics in the Philippines. The Philippines Statistician, 55(3-4), 29-41.

[2] Bersales, L. G. S. (2004). Academic Programs in Statistics in the Philippines. 9th National Convention on Statistics. Convention Papers, 1G-4.

[3] Board Resolution No. 34, S. 1993. Approving the Recommendation of the Curriculum for the Implementation of the following New Curricular Offering Effective School Year 1993-1994.

[4] Nebres, Bienvinedo F. (1998). Assessment of the Mathematics Sciences in the Philippines and their Applications in Industry.

[5] Thomas, Dana L. and Barry, Ronald P. (n.d.) Program Review Report of University of Alaska Fairbanks for BS in Statistics, S.Y. 2005-2006. Retrieved October 22, 2014 from https://www.uaf.edu/files/dms/assessment/DMS%20PR%2005-06%20Report%20Undergrad%20Stats.pdf.

[6] Reston, E. D. and Bersales, L. G. S. (2008). Reform Efforts in Training Statistics in the Philippines: Challenges and prospects. In C. Batanero, G. Burril, C. Reading and A. Rossman (Eds.), Joint International Committee on Mathematics Instruction and International Association on Statistics Education Study: Teaching Statistics in School Mathematics, Challenges for Teaching and Teacher Education. Proceedings of the ICMI Study 18 and 2008 IASE Roundtable Conference.

[7] The Florida State University (October, 2012). Undergraduate Statistics Program. Retrieved December 8, 2014 from www.careercast.com/jobs-rated/best-jobs-2014.

[8] Verhoeven, Pieternel (2006). Statistics Education in the Netherlands and Flanders: an Outline of Introductory Courses at Universities and Colleges, ICOTS-7.

[9] Opolut-Okurut, Charles (2010). Statistics in Ugandan Schools: Challenges on Instruction and Assessments, ICOTS-8.

[10] Tabunda, A. L. (2006). Some Aspects of the Teaching of Statistics in the Philippines. The Philippine Statistician, 55(3-4), 43-54.

[11] David, I.P. and Maligalig, D.S. (2006). Are We Teaching Correctly to Our Youth? The Philippine Statistician, 55(3-4), 1-28.

# Mathematical Competence of Pre-service Teacher Education Students of State Universities and Colleges in Region 8

Saladaga, Thelma Agamon*

*Department of Arts and Sciences, Eastern Visayas State University, Tacloban City, Leyte 6500,Philippines,*
*thelma.saladaga@yahoo.com*

## Abstract

The study assessed the mathematical competence of the Pre-Service Teacher Education Students in SUCs in Region 8 during the first semester of the school year 2012-2013, in terms of PSTE's conceptual and procedural skills in mathematics. The study used the descriptive-comparative-correlational method utilizing a researcher-structured inventory test to elicit the pertinent data needed. The study identified the level of conceptual and procedural skills of the PSTE students, the difference in the conceptual skills and the difference in procedural skills among the PSTE students in the different SUCs in Region 8. It also determined whether a significant relationship exists between conceptual and procedural skills of the PSTE students. An intervention scheme is designed to enhance and produce highly competent education graduates. Conclusions: The level of conceptual skills of the PSTE students is satisfactory, the procedural skills is fair. The level of conceptual skills among the PSTE students is significantly different; likewise, the level of procedural skills. A significant relationship exists between the conceptual and procedural skills of the PSTE students. Recommendations offered: A rigid screening for admission to the College of Education for the mathematics major must be observed. The teacher should implement a variety of learning styles throughout the course to maximize the students' potentials. The DOST, CHED, and the university should offer scholarships to attract highly intelligent students to enroll. The PSTE students should be given opportunities to attend pre-service workshops and seminars on mathematical competencies in all levels. The inventory test should be given before the semester starts.

*Keywords*: mathematical competence, pre-service teacher education students, conceptual skills, procedural skills

*Corresponding Author
E-mail Address: thelma.saladaga@yahoo.com

## 1. Introduction

Quality pre-service teacher education is a key factor in quality education. In the Philippines, the pre-service preparation of teachersfor the primary and secondary educational sectors is a very important function and responsibility that has been assigned to higher education institutions. All efforts to improve the quality of education in the Philippines are dependent on the services of teachers who are adequately prepared to undertake the various important roles and functions of teachers. It is therefore of utmost importance that the highest standards be set in defining the objectives, components, and processes of the pre-service teacher education curriculum (Art.1, Sec. 1 CMO No. 30 s.2004).

The curriculum of the teacher education program emphasizes the interweaving of foundational general education knowledge and skills, theoretical knowledge about teaching and learning, methodological skills, experiential knowledge in the various learningexperiences in the curriculum.

### *1.1 Background of the Study*

The results of the of the Trends in Mathematics and Science Study (TIMSS, 2003) show that the National Mathematics and Science mean scores in1999 and 2003 are very much lower than the international mean score (United Nations Population Division and Statistics Division, 2010).

Another study conducted by Razona (2006) as cited by Pagaduan that pre-service preparation of teachers should be upgraded to include mastery of subject matter, content, methodology, teacher's competence, communication skills, values and professional skills. The bold step taken for the improvement of teaching profession came in the form of Republic Act No. 7836, an Act to Strengthen the Regulation and Supervision of the Practice Teaching in the Philippines and Prescribing a Licensure Examination for Teachers. This Act became a low onDec, 16, 1994 and was fully implemented on August 1996.

The first licensure examination was in August 1996. Applicants for the first Licensure Examination for Teachers (LET) reached 96,000 and the national passing percentage was a low 29.92%. Ten years later, the trend in the results of the LET for the entire seven years, from 2006 to 2012 did not vary so much on the increased or decreased on the national passing rates (see appendix G). On the average, the national passing rate is only 25.55%.

Basil (1996) claimed that developing good and effective teachers who possess the pedagogical skills requires long and tedious work. He further pointed out that the honing of teaching craft starts from the first encounter with the actual teaching of the student teacher, thus, the process of pre-service training and the learning that goes with it play pivotal roles in teacher preparation. Therefore, it is necessary to assess the mathematical competence of the pre-service teacher education students in term of conceptual and procedural skills as they are applied in actual learning situations through teaching.

### *1.2 Statement of the Problem*

The study aimed to assess the mathematical competence of the Pre-Service Teacher Education (PSTE) students in State Universities and Colleges (SUCs) in Region 8, during the first semester of SY 2012-2013.

Specifically, the study sought answers to the following questions:

1. What is the level of conceptual skills of the pre-service teacher education students?
2. What is the level of procedural skills of the pre-service teacher education students?
3. Is there a difference in the conceptual skills of the PSTE students in the different SUCs?
4. Is there a difference in the procedural skills of the PSTE students in the different SUCs?
5. Is there a relationship between conceptual and procedural skills of the PSTE students?
6. What intervention scheme can be develop to improve the conceptual and procedural skills of the PSTE students?

### *1.3 Null Hypothesis*

The null hypotheses tested were the following:

1. There is no significant difference in the conceptual skills of the PSTE students in the different SUCs in Region 8.
2. There is no significant difference in the procedural skills of the PSTE students in the different SUCs in Region 8.
3. There is no significant relationship between the conceptual and procedural skills of the PSTE students.

### *1.4 Theoretical/Conceptual Framework*

This study is anchored on three theories of learning that work together on the process of learning. These are the constructivist learning theory by Jerome Bruner, the cognitive learning theory by Jean Piaget, and the evaluation learning theory by Donald Kirkpatrick.

Constructivism is a learning theory that attempts to explain how learners learn by constructing and understanding concept and other phenomena for themselves. It gives teachers perspective to re-think how students learn and how to focus on process and provide ways of documenting change and transformation. It also reminds teachers to look for different ways to engage individual student develop rich environments for exploration, prepare coherent problem sets and challenges that focus the model building effort, elicit and communicate student perceptions and interpretations (Abdal-Haqq, 1998). For a constructivist classroom, learning is active, constructive, reflective, collaborative, inquiry- based, and evolving (Brooks & Brooks, 1993; Devries & Zan, 2003).

Another theory anchored in this study is Cognitive Learning Theory. It is the theory that humans generate knowledge and meaning through the sequential development of anindividual’s cognitive abilities, such as, the mental process incorporating the abilities to recognize, recall, analyze, reflect, apply, create, understand and evaluate (Piaget, 1964).

To Piaget, cognitive development was a progressive reorganization of mental process as a result of biological maturation and environmental experience. Children construct an understanding of the world around them, then experience discrepancies between what the already know and what they discover in their environment (Smith, 1993). In addition, cognition refers to the mental processes which involve in gaining knowledge and comprehension. These processes include thinking, knowing, remembering, judging and problem solving. These are higher level functions of the brain and encompasses language, imagination, perception and planning (Pritchard, 2005).

Another theory by Donald Kirkpatric is the Learning Evaluation Theory. This model can help us analyze objectively the effectiveness and impact of the training so that it can be improved in the future. This theory is structured into four levels of evaluation, which essentially measures reaction of students, learning, behavior and results. Reaction evaluation is how the delegates felt and their personal reaction to the training or learning experiences. Learning evaluation is the measurement of the increase of knowledge before and after the learning experiences. Behavior evaluation is the extent of applied learning back on the job-implementation. Observation and interview overtime are required to assess change, relevance of change, and sustainability of change. Result evaluation, is the effect on the business or the environment of the trainee. It is much challenging not because of the reliance on the line management and the frequency and scale of changing structure, but on the responsibilities and roles, which complicates the process of attributing clear accountability.

### 1.5 The Scheme of the Study

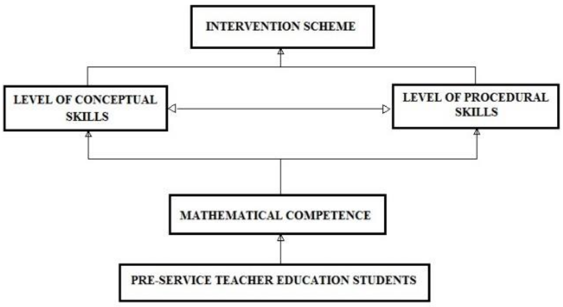

INTERVENTION SCHEME
LEVEL OF CONCEPTUAL SKILLS
LEVEL OF PROCEDURAL SKILLS
MATHEMATICAL COMPETENCE
PRE-SERVICE TEACHER EDUCATION STUDENTS

Figure 1. Schema showing the conceptual flow of the study

The 97 PSTE students were made to answer the inventory test wherein their mathematical competence in terms of conceptual and procedural skills are being tested and assessed through the strategies and procedures they illustrated at, in arriving at a particular answer.

### 1.6 Scope and Delimitation of the study

This study focuses on the assessment of the mathematical competence of the Pre-Service Teacher Education (PSTE) students in the seven State Universities and Colleges in Region 8. It looked into the conceptual and procedural skills demonstrated by the PSTE students on their answers and solutions to the problems in the Inventory Test. The Inventory Test is a researcher-made test which is taken from the secondary mathematics curriculum such as Elementary Algebra for first year, Intermediate Algebra for second year, Geometry for third year and Advanced Algebra, Trigonometry and Statistics for fourth year. There are 25 test items testing the conceptual skills and also 25 test items testing the procedural skills.

## 2. Research methodology

This study used the descriptive-comparative-correlational method of research. Best and Khan (1998) defined descriptive research as one that uses quantitative methods to describe “what is” describing, recording, analyzing, and interpreting conditions that exist. It involves some types of comparisons or contrasts and attempts to discover relationships between existing non-manipulative variables. This is the appropriate method for the present study because it gathers data, records, describes, analyzes, and interprets the level of mathematical competence of PSTE students in Region 8.

This study further determined the level of conceptual skills and procedural skills of the PSTE students, the difference in conceptual skills, and the difference in procedural skills of the PSTE students. It also established whether a significant relationship exists between the conceptual and procedural skills of the PSTE students, and provided an intervention scheme for the students in order to produce highly competent education graduates.

The statistical measures used were the percent, mean, One-Way-Analysis of Variance (One-Way ANOVA), Scheffe’s test, Pearson’s Product-Moment Coefficient of Correlation (Pearson r), and Kendall’s Coefficient of Concordance W.

### 2.1 Respondents of the Study

The respondents of the study were the PSTE students of the seven State Universities and Colleges in Region 8 who were enrolled during the first semester of SY 2012-2013. A total enumeration of 97 PSTE students was involved.

The Eastern Samar State University (ESSU) has 11 respondents; Leyte Normal University (LNU), 28; Naval State University (NSU), 8; Northwest Samar State University (NwSSU), 12; Samar State University (SSU), 14; University of Eastern Philippines (UEP), 11; and Visayas State University (VSU), 13.

### 2.2 Research Instrument

The research instrument used to gather data in the study was the researcher-made-inventory test. It consists of 50 items for which 25 items were testing the conceptual skills of the PSTE students, and the other 25 items were testing the procedural skills of the same test takers.

The types of problems included in testing the conceptual skills are: multiple choice, conversion, definition of certain concept and problems which can be answered directly. This means that the problem does not require a step-by-step solution. It can be answered directly through simple recall of concept or theory. All problems testing the conceptual skills of the students require only simple recall of the theories or cognition.

On the other hand, the problems testing the procedural skills of the students are problems which require a step by step solution. It also involves simple recall of concepts and theories needed in performing correct procedures in solving the problems.

### 2.3 Data Gathering Procedure

Before fielding the instrument to the PSTE students, a written permission from the seven SUCs was secured. Upon approval by the university presidents the researcher requested the respective deans of the College of Education for scheduling and other arrangements in administering the inventory test. The researcher personally administered the test to the PSTE students in their respective campuses. It lasted for 2 hours.

### 2.4 Method of Scoring

A rubric was prepared as guide in correcting the test papers. Each problem is given credit of 5 points. Points were assigned for each correct step by the checker for the students to get a cumulative sum of 5 points as credit for a perfect solution to the problem. Credit for an answer would be lesser than 5 if the solution is not perfect. For the problems which were left unsolved, a point 1 was credited so that nobody got zero in the test. This method was done for the entire test items testing the procedural skills of the PSTE students.

Some of the problems for testing the conceptual skills of the students are multiple-choice tests which do not need computations. So, if their answer is correct they earned 5 points for that particular problem. If their choice is wrong, a point 2 is given to his credit, and for those who did not answer a point 1 was given. The other problems testing the conceptual skills require a simple recall on the definition of the concept and derivation of formula. So, points were assigned for each step for a cumulative sum of 5points for a perfect answer, and less than 5 points if errors were made (see appendix L and J on scoring procedure).

After the tests were checked and scored by the same mathematics experts their scores were further tested using Kendall's Coefficient of Concordance, W. This is to find out if there was an agreement or concordance among the three checkers/raters of the inventory test. It was found out that W = 1 of the three checkers/raters, which means they have highly agreed on the checking/rating or giving of points to the solution and answers of the PSTE students on both conceptual and procedural.

The perfect raw score is 125 each for conceptual skills and for procedural skills and the lowest is 25 for both skills. Below is the range of raw scores for both conceptual and procedural skills and the corresponding description.

| Range of Raw Scores | Description |
|---|---|
| 105 – 125 | Outstanding |
| 85 – 104 | Very Satisfactory |
| 65 – 84 | Satisfactory |
| 45 – 64 | Fair |
| 25 – 44 | Poor |

*2.5 Statistical Treatment of Data*

The statistical tools used in this study were the following: Percent, Mean, One-Way Analysis of Variance (One-Way ANOVA), Scheffe's test, Pearsons Product-Moment Coefficient of Correlation (Pearson r), and the Kendall's Coefficient of Concordance W.

Percent is a statistical tool used to find the number of parts taken from the whole. This was used in finding what percent of the PSTE students are on a certain level of conceptual and procedural skills.

The Mean Scores were computed and were used to describe the levels of conceptual and procedural skills for each of the identified SUCs.

One-Way ANOVA was used to determine whether significant differences exist both in conceptual and procedural skills among the PSTE students.

Scheffe's Test was used as the post comparison test to find out which among the groups caused the significant difference among the SUCs.

The Pearson's Product-Moment Coefficient of Correlation (Pearson r) was used to establish whether significant relationship exist between the conceptual skills and procedural skills.

Kendall's Coefficient of Concordance was used to find out if there is an agreement or concordance among the three checkers/raters of the inventory test.

The level of significance was set at a = 0.05 level of two tailed test.

## 3. Research results and discussion

The findings and the results of the study is presented as follows:

*3.1 Level of Conceptual Skills and Percentage Distribution*

The frequency and the corresponding percentage of the PSTE students with particular level of conceptual skills were determined. Figure 1 is a graphical presentation of the percentage distributionof the level of conceptual skills of the PSTE students in Region 8.

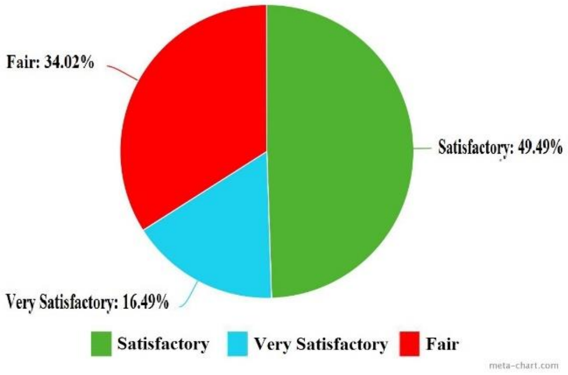


Figure 2. Level of conceptual skills and percentage distribution of PSTE students in Region 8

It presents the distribution of the PSTE by level of conceptual skills. It shows that 49.49% got satisfactory rating, followed by 34.02% with fair rating and 16.49% with very satisfactory rating. It may be noted that nobody got an outstanding rating but nobody got zero either. It implies that learning potentials of the students were not maximized.

Table 1: Level of conceptual skills of the pre-service teacher education students in region 8

| School | Mean Scores | Rank | Description |
|---|---|---|---|
| A | 69.69 | 5 | Satisfactory |
| B | 61.00 | 7 | Fair |
| C | 70.00 | 4 | Satisfactory |
| D | 71.29 | 3 | Satisfactory |
| E | 69.14 | 6 | Satisfactory |
| F | 74.00 | 2 | Satisfactory |
| G | 81.82 | 1 | Satisfactory |
| Overall Mean | 70.99 | | Satisfactory |

Table 1 shows the overall mean of 70.99 obtained by the PSTE on conceptual skill and is described as satisfactory. Six out of the seven SUCs obtained satisfactory rating, however, one out of seven SUCs obtained a mean score of 61.0 which is described as fair. This implies that the PSTE students in the region can be categorically described as almost having a homogeneous competence in conceptual skills at satisfactory level only. This result falls short of the government's aim to improve the quality of education and to keep pace with the demands for global competitiveness. There is much room for improvement in order to attain higher level in mathematical competence.

### *3.2 Level of Procedural Skills and Percentage Distribution*

Figure 3 shows the graphical presentation of the percentage distribution of the level of the procedural skills. It shows that 44.33% got a fair rating. This level consists the highest number of PSTE in the distribution.

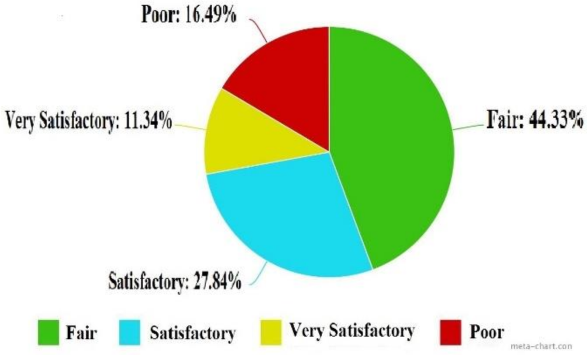


Figure 3. Level of Procedural Skills and Percentage Distribution of the PSTE students in Region 8

It is followed by 27.84% with satisfactory rating, then 16.49% with poor rating and the least number of students 11.34% got very satisfactory rating. It is distressing to note that as much as 16 students or 16.49% got poor ratings which implies that these students are poor performers in mathematics and did not know anything on what to do or how to solve the problem.

Table 2: Level of procedural skills of the pre-service teacher education students in region 8

| School | Mean Scores | Rank | Description |
|---|---|---|---|
| A | 59.85 | 5 | Fair |
| B | 49.17 | 7 | Fair |
| C | 67.63 | 2 | Satisfactory |
| D | 60.29 | 4 | Fair |
| E | 64.79 | 3 | Satisfactory |
| F | 59.00 | 6 | Fair |
| G | 76.36 | 1 | Satisfactory |
| Overall Mean | 62.44 | | Fair |

Table 2 shows the overall mean score of 62.44 obtained by the PSTE on procedural skills and is described as fair. Four out of seven SUCs obtained fair ratings while three out of seven SUCs obtained satisfactory ratings. It implies that these students are far from being well-prepared for a teaching job. They lacked the requisite skills to become effective teachers. It precludes the probability of good teaching because "one cannot teach what one does not know."

The rating of conceptual skills (satisfactory) is one level higher than the rating for procedural skills (fair). As explained by Heibert and Lefevre (1986) this difference shows that students are not fully competent if they are deficient in either kind of knowledge. Students may have understanding of the concept but cannot solve the problem or they may be able to perform some tasks yet without understanding of what they are doing.

### *3.3 Difference in the Level of Conceptual skills and Procedural Skills of the PSTE students in the Different SUCs in Region 8.*

The difference in the level of conceptual skills among the PSTE students in the seven SUCs are computed using the One-Way Analysis of Variance. It shows the significant difference between/among the means of variables. Since a significant difference exists using One-Way ANOVA, it was tested further to determine which among the pairwise comparison caused the significance.

Pairwise comparison was done by Scheffe's test (F') on conceptual skills to determine further which among the group differences caused the significance. It was found out that the difference between SUCs B and G was significant which obtained the p-level of 0.007. This was the one and only one pair which turn out to be significantly different, SUC B being at the lowest rank with mean value of 61.00 while G being the highest in rank with mean value of 81.82. It shows that the PSTE students in SUC G were better performer than theother PSTE students and those in SUC B were the lowest performers.

Similarly, the One-Way ANOVA was used in performing a statistical teston the difference in the level of procedural skills among the PSTE students among the SUCs identified. It was found out that a significant difference in the procedural skills occurred among the PSTE students in the different SUCs. It shows a very big difference between the extreme scores. It was tested further to verify which among the group difference led the significance. Scheffe's test was used to establish the post comparison. It was found out that SUCs G, with mean value of 76.36 and B, with mean value of 49.17 appeared to be extremely ranked pairs in the procedural skills. This validates the findings that SUC G and B were of opposite extreme and were identified as significantly different pairs. Hence, it can be gleaned that the mean scores obtained from these SUCs B and G were the reason behind why significant difference occurred when the F-test was performed. Same as in the conceptual skills, the PSTE students in SUC G were better performers than the other PSTE and that those of SUC B were the lowest performers.

### *3.4 Relationship Between the Conceptual and Procedural skills of the PSTE students.*

The test on relationship between conceptual and procedural skills of the PSTE students among the SUCs identified in this study was conducted using the Pearson's Product-Moment Coefficient of Correlation (Pearson r). It was found out that a significant correlation exists between the PSTEs' level of conceptual skills and their level of procedural skills in mathematics. It indicates that the lower the level ofconceptual skills of the students, the lower the level of their procedural skills. The connection between the conceptual and procedural knowledge is very important to students in solving mathematics problems. When students could not appropriately connect between conceptual and procedural knowledge they may have some understanding of mathematical concepts but not solve the problem, or they may be able to perform some task but could not understand what they are doing (Gallager, 1981). A variety of learning styles may be implemented throughout the course to allow the students to learn in at least one way that matches their learning styles.

### *3.5 Intervention scheme*

An implementable scheme was conceived as a result of salient findings of the study. The main purpose of which is to enhance the competency level of the Pre-Service Teachers in mathematics of the schools under the study.

#### *3.5.1 Objectives*

1. Enhance the mathematical competencies of the PSTE students in Region 8 in preparation for their actual teaching profession.
2. Provide learning alternatives, employing the constructivist theory, and instructional activities for the conceptual and procedural skills development of the PSTE students based on the concepts that need emphasis and mastery.

#### *3.5.2 Intervention Scheme*

Intervention scheme is composed of academic-related activities such as: seminars and forum in mathematics; Interschool competition in search for "Math Wizard"; Joining academic organization, Junior Philippine Mathematical Society(JPMS) and the Future Teachers Organization (FTO).

## 4. Conclusion

Based on the findings of the study, the following conclusions are arrived at:

1. The level of conceptual skills of the PSTE students is satisfactory.
2. The level of procedural skills of the PSTE students is fair.
3. There is a significant difference in the conceptual skills of the PSTE students in the different SUCs in Region 8.
4. There is a significant difference in the procedural skills of the PSTE students in the different SUCs in Region 8.
5. There is a significant relationship between the conceptual skills and procedural skills among the PSTE students in Region 8.

## References

[1] Best, J.W., & Kahn, J.V. Research in education. 8th ed. Singapore: Prentice Hall, Inc. 1998

[2] Gallager, J.M. & Reid, D.K. The learning theory of Piaget and Inhelder. Austin, Texas; 1981

[3] Heibert, J. & Lefevre, P. Conceptual and procedural knowledge in mathematics: an introductory analysis. Hillsdale, New Jersey: Lawrence Erlbaum and Associates.

# Teachers' Perceptions and Decision on the Inclusion of History of Mathematics to Instruction and their Students' Mathematical Competence and Attitude

Grace D. Severo, MSME
*University of Perpetual Help System DALTA*
*Las-Piñas City, 1740, Philippines, grace.severo2012@gmail.com*

## Abstract

This study aimed to determine the factors that affect teachers' decision whether to include History of Mathematics (HOM) or not in their instructions; teachers' perceptions on HOM; and to determine significant differences on students' attitude and competence toward mathematics between teachers who include HOM and teachers who do not include HOM in their instructions. The study was conducted at one of the biggest public high schools in the city of Las-Piñas, located at the South of Metro Manila during the academic year 2013-2014. There were 247 fourth year high school students randomly selected and 41 mathematics teachers who participated in the study. Significant differences were found on the perceptions of HOM between teachers who include HOM and those who do not; and the students' attitude and competence toward mathematics between the two groups of teachers. Results show that those teachers who enjoyed teaching HOM indicated that their students enjoyed learning HOM; HOM helps students to see the development of the connections among mathematical concepts; HOM provides the conditions to capture students' interests; and improve attitude toward mathematics.

*Keywords*: Teachers' Perception; History of Mathematics; Students' Mathematical Attitude & Competence

## 1. Introduction

Learning mathematics comes through the introduction of concepts, then techniques of computation, and then drills. If time permits, the application of the concepts follows. Teachers tend to look at the results of drills to see whether the majority of the students grasped the concepts and the techniques. If not, a remedial period is allocated. After some time, when the majority understands it, another topic is brought in through the introduction of concepts, techniques, and drills ad infinitum.

This kind of routine has been the kind of introduction to the formal learning of mathematics, which students come to dislike [11]. Reference [7] claimed that majority of the students look at mathematics as a set of rules and formulas to be memorized devoid of what makes it interesting. They were just taught how is it done, and how to apply without knowing how it came to be. Sadly, the final goal is just to pass the subject.

However, when students are given awareness that every mathematician around the world are like them, struggling, and was in constant struggle to find the answers, students' interests toward the concepts is piqued [1-2]. It was the researcher's experience through teaching that if students were taught how the solutions were discovered by the mathematicians who toiled for so many years to find it and how it came into being, attitude towards mathematics will change over time.

Mathematics has a long, rich history of which teachers and students are often unaware [1]. As a result, they fail to appreciate that mathematics is constantly developing, that it is a human endeavor, and that it is a subject that has varied applications in different times and cultures [7].

In looking for the best way of improving students' attitude and interests in mathematics, the researcher found out that for decades, other researchers have been suggesting a number of benefits to the integration of the history of mathematics into mathematics education

"An understanding of the history of mathematics and the context in which problems arose can lead to greater understanding of the content and nature of mathematics" [2]. Likewise, [3] the history of the subject can humanize mathematics and thus can be motivating to students who conceive mathematics as "dead," "boring," or "too abstract."

At present, however, there is little empirical research on the effectiveness of the integration of history of mathematics into mathematics education, local and international.

If students perceive mathematics as a set of discrete topics with no historical background or discussion of historical significance, it is likely that they will fail to see the relationship and relevance of mathematics to other related sciences. When students are exposed to varied experiences related to the cultural and historical aspects of evolution of mathematics, they are likely to develop an appreciation of mathematics and its role in the development of society [4]. The researcher, therefore, postulates that a historical background provides a perspective that lays a foundation for learning. In particular, it may be beneficial to the student whose learning of mathematics is a struggle.

Considering the entire aforementioned context, this paper provides a modest contribution to that endeavor.

The researcher came up with this study to look at the following variables: (1) factors that affect teachers' decision to include and not to include HOM into their instructions; (2) teachers' perceptions of history of mathematics; and (3) the differences on students' mathematical attitude and competency between teachers who include HOM and teachers who do not.

## 2. Research Methodology

This study made use of the descriptive and inferential methods of research. A survey scale questionnaires was used and distributed to collect the data needed.

### *2.1 Respondents*

The respondents of these study were teachers and students in the 4th year high school (grade 10) from one of the largest public high schools in the City of Las-Piñas, located at the South of Metro Manila. The study was conducted during the second term of academic year 2013-2014. During this academic year, the 4th year high school had a total of 23 sections with 1,344 students enrolled. A total of 41 teachers and 247 students who were under the supervision of these teachers responded to the survey. The teacher-respondents were classified into two namely: 24 teachers who indicated that they include HOM in their instruction; and 17 teachers who do not. Student-respondents under these teachers were randomly selected. One hundred twenty three (123) students were under the teachers who include HOM, while the remaining 124 were from teachers who do not include HOM.

### *2.2 Instruments*

The researcher utilized three survey questionnaires. The first one was a researcher-made test for Mathematics Competency; second was an adopted questionnaire on Teachers' Perceptions of History of Mathematics and Teachers' Reasons for including and not including HOM into their instructions [5]; and third was a standardized questionnaire on Attitude Toward Mathematics Inventory (ATMI) [6]. These questionnaires were validated by some mathematics professors, pilot–tested, and analyzed for reliability and validity purposes.

## 3. Research Results and Discussion

The following presents discussions on teachers' perceptions of including HOM into instruction and the factors that affect their decision whether to include HOM or not; and the student-respondents' attitude and competency toward mathematics.

### *3.1 Teachers' Perception on History of Mathematics.*

(Table 1 and 2) presents the frequency and percentage distribution of the five components in the survey questionnaire of Teachers' Perception on History of Mathematics.

The first highest number (15, 62.5%) of teachers who responded that they include HOM in their instruction, "agree" that "knowing the history of mathematics may assist students in the learning mathematical concepts." The second highest number (14, 58%) of the teachers "agree" that "all students of mathematics should be taught some history of mathematics." It is significant to note that the importance of HOM in mathematics instruction was positively perceived by the teachers who responded that they include HOM in their classroom instruction.

Teachers (8, 49.6%) who responded that they do not include HOM in their instruction, nonetheless "agree" that "knowing the history of mathematics may assist students in learning mathematical concepts." Some of them, however, were "neutral" on their stand on the other four statements.

### *3.2 Factors that Affect Teachers' Decision Whether to Include HOM or Not in their Instructions.*

Finding show that those teachers who enjoyed teaching HOM believe that their students enjoyed learning HOM. This is made evident by 13 teachers who "agree" on the statement "I enjoy teaching the facts and episodes from history of mathematics" (Table 3). They also perceived that HOM helps students to see the development of the connections among mathematical concepts and provides the conditions to capture students' interests and improve their attitude toward mathematics, which implies that, HOM plays a critical role in the construction of knowledge and contributes to the development of "affective domain" of learning.

The statement, "I consider myself lacking expertise on the history of mathematics" received a mean rating of 4 – "Agree" (Table 4). This implies that the teachers who have had relatively little or no exposure to HOM, may perceive themselves lacking expertise and consequently not expose students to HOM.

The availability of resources is another deciding factor influencing a teachers' decision. Two of the most highly-ranked reasons for not including HOM were, "History of mathematics is not in the textbook that I use" and "There are not enough appropriate resource materials" (Table 4). Seemingly, the lack of readily available resources may be a factor for teachers who do not include HOM. There was no overwhelming majority of teachers who felt HOM does not belong in the curriculum, while there was a large majority who would be interested in knowing how to include HOM.

Lastly, teachers who do not include HOM are intrinsically motivated to integrate some topics from the HOM and agree that it will benefit the students by capturing their interest, and, perhaps assist, in the learning of mathematics. It is unfortunate that HOM is not a part of the required curriculum.

.

### *3.3 Students' Attitude toward Mathematics*

Findings show that the highest mean (3.98) score of the four domains in the Attitude toward Mathematics Inventory (ATMI) of the students under the teachers who include HOM into instructions was "Value" (Table 5). This implies that integrating HOM in the classroom could improve students' attitude toward mathematics because they perceived to value learning of HOM. Many researchers [7-10], have argued that the use of history in mathematics lessons can lead to better attitudes and achievement in mathematics. In addition, the use of anecdotes and biographies of mathematicians make lessons more interesting and dynamic [10]. Hence, students should find their mathematics lessons more enjoyable when history is used. The use of history in lessons can improve students' perception regarding the value of mathematics to mankind and motivate them to learn mathematics [10]. Moreover, learning takes place more effectively when a learner retraces the key steps in the historical development of the subject [8-9].

On the other hand, the students under the group of teachers who do not want to consider HOM of this study obtained a lower mean value of 3.10 in "Motivation" and 3.00 in "Enjoyment" domains of ATMI. This findings supports [8] the proposal of the five reasons for using the history of mathematics in instructions, to wit: (1) history can help increase motivation and can help develop a positive attitude toward learning; (2) past obstacles in the development of mathematics can help explain today's students difficulties; (3) historical problems can help develop students' mathematical thinking; (4) history reveals the humanistic facets of mathematical knowledge; and (5) history gives teachers a guide for teaching

### *3.4 Students' Mathematical Competence*

Surprisingly, there were only 12 out of 123 students passed the 40-item mathematical competency test under the teachers who include HOM in their instruction. Ten of these obtained a "Minimum" level, whereas two got the highest score of 28 and 29 which belongs to the "Intermediate" level. On the other hand, only 8 out of 124 students got a minimum level of competency under the teachers who do not include HOM in their instruction. These students were also the ones who spent the highest time studying at home.

The students who obtained the highest score of 29 preferred to take science course in college, but those 8 students who got the lowest score preferred to enroll in BSED Math. The rest of the students performed very poor in the mathematics competency test, but surprisingly, there were 68, 43, and 22 of these students responded that they preferred to take Engineering, BSED Math, and BS Math courses in college.

### *3.5 Differences on perceptions of HOM between teachers who include HOM and who do not.*

Results found that there is a significant difference on the perceptions of HOM between teachers who include HOM into their instruction and those who do not. This indicates that teachers who view HOM as valuable, necessary, or worthwhile may seemingly tend to include HOM in their classroom lessons. The contributing factors on the reason of teachers who do not include HOM were "Understanding mathematics would be easier if the history of mathematics was taught," but they still perceived that "Knowing the history of mathematics may assist students in learning mathematics concepts," because this was rated at a much higher degree by this group of teachers. Teachers that include HOM highly ranked "Understanding mathematics would be easier if the history of mathematics was taught."

### *3.6 Differences on students' attitude and competency toward mathematics*

There is a significant differences on students' attitude and competency toward mathematics between the two groups of teachers. This implies that students who belong to the teachers that include HOM in their instructions had a positive attitude, and may tend to get a higher level of competency toward mathematics than to those students that belong to teachers who do not include HOM. To support this findings, similar researches [7-11] concluded that the role of history for the teaching and learning of mathematics with reference to students' mathematical competence is the main purpose of mathematics education; and that awareness of mathematics history promote active learning, create motivation, alleviate boredom and improve students' understanding of mathematics.

## 4. Conclusions

The results imply that HOM may not only benefits a student but also the teacher. The impact of the findings about the factors mentioned above is far reaching. Teachers indicated a lack of confidence in teaching HOM. In line with this, the HOM may serve as the foundation upon which the teacher can construct strong mathematical connections which would, in turn, help to strengthen the teacher's mathematical content knowledge and confidence [4-5, 13].

The students under the teachers who include HOM in their instruction had a positive attitude toward mathematics than the students under the teachers who do not. There is a significant difference found at the lowest p-value of 0.0001 of the four factors of ATMI, hence, the use of history of mathematics in instruction may most likely improve students' attitude toward mathematics.

Normally, a student working conscientiously on a mathematics program would reach minimum competence after two years of high school study; intermediate after three; and advanced after four [16]. These results as to be expected have not been realized by the 4th year high school students being tested for this research. This high gap between the ideal and the actual may result from high school courses that do not cover the mathematical material central to college preparation. Also, [16] incoming college level students are expected to bring hands-on skills in computation and algebraic manipulation, as well as conceptual knowledge rooted in a deep and profound understanding of numbers and basic geometry. They are expected to know basic mathematical concepts in computation, algebra and geometry because these are readiness benchmarks on college success. Because of these results in the competency test, the need for improvement of mathematics learning and achievement seemingly critical for students' numeracy must be the main focus of teacher-respondents of this study. And since HOM is a humanizing factor because it piques students' interests, this may likely a good instructional tool in improving students' mathematical skills and attitude towards numbers and numeracy [3-5; 13-15].

Table 1. Perceptions of HOM for teachers who include HOM

| | All students of mathematics should be taught some history of mathematics. | | The history of mathematics is worthwhile and necessary to the understanding mathematics. | | Knowledge of the history of mathematics is valuable to non-scientists or non-mathematicians. | | Knowing the history of mathematics may assist students in learning mathematical concepts. | | Understanding mathematics would be easier if the history of mathematics was taught. | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Freq. | % | Freq. | % | Freq. | % | Freq. | % | Freq. | % |
| Strongly Disagree | 1 | 4.2 | 2 | 8.3 | 0 | 0.0 | 1 | 4.2 | 0 | 0.0 |
| Disagree | 1 | 4.2 | 2 | 8.3 | 2 | 8.3 | 2 | 8.3 | 2 | 8.3 |
| Neutral | 3 | 12.5 | 5 | 20.8 | 5 | 20.8 | 1 | 4.2 | 11 | 45.8 |
| Agree | 14 | 58.3 | 11 | 45.8 | 13 | 54.2 | 15 | 62.5 | 8 | 33.3 |
| Strongly Agree | 5 | 20.8 | 4 | 16.7 | 4 | 16.7 | 5 | 20.8 | 3 | 12.5 |
| Total | 24 | 100.0 | 24 | 100.0 | 24 | 100.0 | 24 | 100.0 | 24 | 100.0 |

Table 2. Perceptions of HOM for those teachers who do not include HOM

| | All students of mathematics should be taught some history of mathematics. | | The history of mathematics is worthwhile and necessary to the understanding mathematics. | | Knowledge of the history of mathematics is valuable to non-scientists or non-mathematicians. | | Knowing the history of mathematics may assist students in learning mathematical concepts. | | Understanding mathematics would be easier if the history of mathematics was taught. | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Freq. | % | Freq. | % | Freq. | % | Freq. | % | Freq. | % |
| | 0 | 0.0 | 1 | 5.9 | 1 | .8 | 1 | .8 | 1 | .8 |
| Strongly Disagree | 1 | 5.9 | 0 | 0.0 | 0 | 4.5 | 1 | 5.9 | 1 | 5.9 |
| Disagree | 3 | 17.6 | 3 | 17.6 | 3 | 17.6 | 2 | 11.8 | 5 | 29.4 |
| Neutral | 7 | 41.2 | 8 | 47.1 | 7 | 41.2 | 5 | 27.8 | 8 | 47.1 |
| Agree | 4 | 23.5 | 5 | 29.4 | 5 | 29.4 | 8 | 49.6 | 3 | 17.6 |
| Strongly Agree | 2 | 11.8 | 1 | 5.9 | 2 | 11.8 | 1 | 6.8 | 0 | 0.0 |
| Total | 17 | 100.0 | 17 | 100.0 | 17 | 100.0 | 17 | 100.0 | 17 | 100.0 |

Table 3. Factors that Affect Teachers' Decision to Include History of Mathematics into Instructions

| **Factors** | **Frequency (%)** | | | **Mean** |
|---|---|---|---|---|
| | **Disagree** | **Neutral** | **Agree** | |
| 1. The history of mathematics may capture students' interests and improve their attitude toward mathematics. | 1<br>(8.3%) | 1<br>(12.5%) | 22<br>(79.2%) | 3.88 |
| 2. Students may enjoy learning facts about the history of mathematical concepts. | 2<br>(12.5%) | 2<br>(16.7%) | 20<br>(70.8%) | 3.75 |
| 3. It will help students to see the development of the connections among mathematical concepts. | 1<br>(16.7%) | 3<br>(25%) | 20<br>(58.3%) | 3.79 |
| 4. I enjoy teaching the facts and episodes from history of mathematics. | 0 | 11<br>(41.7%) | 13<br>(58.3%) | 3.54 |
| 5. I consider the history of mathematics as important as the mathematics itself even though it takes away time from the curriculum. | 1<br>(25%) | 23<br>(75%) | 0 | 2.95 |

Table 4. Factors that Affect Teachers' Decision to Not Include History of Mathematics into Instructions

| **Factors** | **Frequency (%)**<br>**Disagree** | <br>**Neutral** | <br>**Agree** | <br>**Strongly Agree** | **Mean** |
|---|---|---|---|---|---|
| 1. I do not know how to teach the history of mathematics. | 2<br>(11.8%) | 2 | 8<br>(47.1%) | 5<br>(29.4%) | 3.94 |
| 2. If students are not tested on the history of mathematics, they will not pay attention to the discussion about the history of mathematics. | 0 | 4<br>(23.5%) | 6<br>(35.3%) | 7<br>(41.2%) | 4.18 |
| 3. Since there are no questions pertaining to the history of mathematics in our school, the history of mathematics is not on my priority list of mathematical topics to teach in class. | 0 | 3<br>(17.6%) | 9<br>(52.9%) | 5<br>(29.4%) | 4.12 |
| 4. I consider myself lacking expertise on the history of mathematics. | 0 | 5<br>(29.4%) | 7<br>(41.2%) | 5<br>(29.4%) | 4.00 |
| 5. There are not enough appropriate resource materials. | 0 | 5<br>(29.4%) | 7<br>(41.2%) | 5<br>(29.4%) | 4.00 |
| 6. History of mathematics is not in the textbook I use. | 3<br>(17.6%) | 4<br>(23.5%) | 5<br>(29.4%) | 5<br>(29.4%) | 3.71 |
| 7. There is not enough time to teach history of mathematics along with the regular curriculum. | 2<br>(11.8%) | 4<br>(23.5%) | 5<br>(29.4%) | 6<br>(35.3%) | 3.88 |
| 8. I examined the resources about the history of math and found little or none. | 2<br>(11.8%) | 9<br>(52.9%) | 6<br>(35.3%) | 0 | 3.24 |

Table 5. Comparisons of Students' Attitude toward Mathematics

| Factors | Grouping | n | Mean | SD | F-value | p-value | Conclusion |
|---|---|---|---|---|---|---|---|
| Enjoyment | Group 1 | 123 | 3.40 | 0.75 | 4.07 | 0.003 | Significant |
| | Group 2 | 124 | 3.00 | 0.80 | | | |
| Motivation | Group 1 | 123 | 3.43 | 0.75 | 2.54 | 0.040 | Significant |
| | Group 2 | 124 | 3.10 | 0.86 | | | |
| Self-confidence | Group 1 | 123 | 3.71 | 0.84 | 5.57 | 0.000 | Significant |
| | Group 2 | 124 | 3.24 | 0.71 | | | |
| Value | Group 1 | 123 | 3.98 | 0.45 | 21.97 | 0.0001 | Significant |
| | Group 2 | 124 | 3.35 | 0.59 | | | |

Group 1 – students under the teachers who include HOM
Group 2 – students under the teachers who do not include HOM

## Acknowledgements

The researcher wishes to extend her sincerest thanks and cordial appreciation to all the special people who have extended their valuable time and assistance towards the completion and realization of this piece of work. To all of them, the researcher is and will be grateful forever.

## References

[1] Jankvist, U. A Categorization of the "Whys" and "Hows" of Using History in Mathematics education. Educational Studies in Mathematics [cited: 2009 Jan 21]. Available from: http://link.springer.com/article/10.1007/s10649-008-9174-9

[2] Alpaslan, M. Prospective Elementary Mathematics Teachers' Knowledge of History of Mathematics and their Attitudes and Beliefs towards the Use of History of Mathematics in Mathematics Education. Unpublished Master's Thesis, ODTÜ, Ankara; 2011.

[3] Byers, V. Why Study the History of Mathematics? International Journal of Mathematical Education in Science and Technology. 2010; 13(1), 59-66.

[4] Gürsoy, K. A survey of Prospective Mathematics Teachers' Beliefs and Attitudes Towards Using the History of Mathematics in Mathematics Teaching. Unpublished Master's Thesis, Karadeniz Tecnical University, Institute of Pure and Applied Sciences, Trabzon; 2010.

[5] Panasuk, R. & Leslie Bolinger Horton. Integrating History of Mathematics into the Classroom: Was Aristotle Wrong? Journal of Curriculum and Teaching. Vol. 2, No. 2; 2013. Available from: www.sciedu.ca/jct

[6] Tapia, M. Attitudes Toward Mathematics Inventory (ATMI). 1996. Available from: http://www.pearweb.org/atis/tools/48

[7] Fauvel, J. History in Mathematics Education: The ICMI Study. Dordrecht: Kluwer Academic Publishers; 2009.

[8] Blom, K. "A historical angle". A Survey of Recent Literature on the Use and Value of History in Geometrical Education. Educational Studies in Mathematics; 2011. Available from: http://connection.ebscohost.com/academic

[9] Liu, P. Do Teachers Need to Incorporate the History of Mathematics in their Teaching? The Mathematics Teacher. 2008; 96(6), 416. Available from: http://www.nctm.org

[10] Ho, W. Using history of Mathematics in the Teaching and Learning of Mathematics in Singapore. Paper Presented at the 1st RICE, Singapore: Raffles Junior College. 2008. Available from: www.researchgate.net

[11] Tzanakis, C & Arcavi, A. Integrating History of Mathematics in the Classroom: An Analytic Survey. In History in Mathematical Education: The ICMI Study, ed. J. Fauvel and J. Van Maanen. Dordrecht: Kluwer Academic Publishers. 2008. Available from: http://link.springer.com/chapter/10.1007

[12] Lim Siew Yee. Mathematics Attitudes and Achievement of Junior College Students in Singapore. Mathematics Education Research Group of Australasia, Paper presented at the Annual Meeting of the Mathematics Education Research Group of Australasia (33rd, Freemantle, Western Australia, Jul 3-7, 2010)

[13] Lim Siew Yee. Effects of Using History of Mathematics on Junior College Students' Attitudes and Achievement. Mathematics: Traditions and [New] Practices. Aamt & Merga. 2013. Available from: http://www.merga.net.au/documents/rp_siew.yee.lim_merga34-aamt.pdf

[14] Ho, W. K. Using history of mathematics in teaching and learning of mathematics in Singapore. Paper presented at Raffles International Conference on Education; 2008 March; Singapore. p. 1-38.

[15] ZUYA, H. E. The Need for the Inclusion of History of Mathematics into Secondary School Curriculum: Perceptions of Mathematics Teachers. Department of Science Education. Modibbo Adama University of Technology, Yola, Adamawa State. International Journal of Innovative Education Research 2 (2):46-51, June 2014. Available from: http://seahipub.org/wpcontent/uploads

[16] Department of Education (DEPED) [Internet]. Available from: http://www.deped.gov.ph/curriculum-guides

### Appendix A. Students' Profile

| Gender | Frequency | Percent |
|---|---|---|
| Female | 145 | 58.70 |
| Male | 102 | 41.30 |
| **Number of hours spent studying mathematics at home** | **Frequency** | **Percent** |
| Less than 1 hour/day | 140 | 56.68 |
| 1 - 2 hours/day | 102 | 41.30 |
| 3 - 4 hours/day | 4 | 1.62 |
| more than 4 hours/day | 1 | .40 |
| **Most Preferred Course to take in College** | **Frequency** | **Percent** |
| BSED Mathematics | 43 | 17.41 |
| BS Mathematics | 22 | 8.91 |
| BS Engineering | 69 | 27.94 |
| Science Course | 77 | 31.17 |
| Other Course | 36 | 14.57 |

### Appendix B. Independent t-test

| | | n | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| Perceptions of History of Mathematics | Teachers who do not include HOM | 17 | 3.222 | .244 | 0.059 |
| | Teachers who include HOM but not often | 24 | 3.792 | .303 | 0.062 |

t-test for Equality of Means

| | | t | Df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Lower | Upper |
| Perceptions of history of mathematics | Equal Variances assumed | -6.782 | 39 | .000 | -.570 | .039 | -.714 | -.431 |
| | Equal variances not assumed | -6.659 | 30 | .000 | -.570 | .042 | -.720 | -.425 |

# Two-echelon inventory model with return items and lead time

Jarumas Sangsavang[1*] and Wuttichai Srisodaphol[2]
[1] Department of Statistics, Khon Kaen University, Khon Kaen, 40002, Thailand, Jarumas1705@gmail.com
[2]Department of Statistics, Khon Kaen University, Khon Kaen, 40002 Thailand, wuttsr@kku.ac.th

## Abstract

Inventory management with return items has drawn the attention from many researchers. Since the problem increasing of the items after used a lot. Return items to recovery is one of the most efficient to protect environment. The objective of this study is to consider a two-echelon inventory model with return items. Demand and returned items are dependent and lead time is constant. The optimal of re-order point, order quantity, and number of cycle of order quantity that minimize the total cost of the system are obtained. The sensitivity analysis is also considered and showed that when the fraction between demand and returned and lead time at Stage 1 (distributor) are increasing; the total cost of system will be increased. Moreover, lead time at State 2 (warehouse) is increasing; the total cost of system will be the same.

*Keywords*: Two-echelon inventory, product returns, continuous review, lead time

*Corresponding Author
E-mail Address: Jarumas1705@gmail.com

## 1. Introduction

Currently, inventory management with return items has drawn the attention from many researchers. Since the items after used increase, damaged and become garbage a lot. The disposal garbage by incineration or landfill enormous amounts causes environmental pollution. This is a problem of people in many countries. Until now, people interested to deal with these problems. Whether it is a government law or human subconscious is aware of the environmental pollution caused by disposal garbage. The organization is involved in manufacturing products that will be responsible for dealing with the items or their items after they have been used or discarded by the customer. This makes many researchers are interested in study about the inventory system with return items. To bring items after used that can be recycled or recovery to become new items that can be put back on sale. The quality and price of the items is the same value as the procurement items such as electronic equipments. In fact, the return items are uncertain more than demand customers. Whether it is in terms of quantity, quality and timing makes inventory control are more difficult and more complex than that inventory system without return items. The inventory system with return items consist: returned items, recovered items and procurement items.

Korugan and Gupta [3] developed model for two-echelon inventory system with return items. The demand items and return items are independent and distributed as Poisson distribution with mean demand $\lambda$ and mean return $\gamma_{c_i}$ where $c_i$ is retailers $i$ $(i=1,...,N)$. This describes the problem by using an open queuing network with finite buffers to minimize the total costs without the set-up cost. Mitra [5] developed deterministic and stochastic models for a two-echelon inventory system with return items. The demand items and return items are independent. For a stochastic model, demand and returns are Normal distribution under continuous review policy. Wang and Di [6] developed a model for inventory control under policy $(s,S)$ with return items. Offered in 2 Cases, demand the return items are independent and distributed as Poisson distribution, and demand the return items are dependent and distributed as Poisson distribution. They also found the optimal re-order point and optimal order up to level.

Therefore, in this study we are interested to consider a two-echelon inventory system model with a single warehouse inventory and a single distributor with return items under continuous review policy. We assume that the recovered items will be the same value as the procurement items. This study extended Mitra's model [5] which demand and return items are dependent and distributed as Poisson distribution with the constant lead time.

## 2. Research Methodology

In this study, we consider the two-echelon inventory system under continuous review policy. Demand and return items are dependent and distributed as Poisson distribution with the constant lead time. The optimal of re-order point, order quantity, and number of cycle of order quantity with minimizing the total cost of the system will be obtained. This system is separated into 3 Stages: Stage 1 is the inventory at the distributor, Stage 2 is inventory at the warehouse for replenishment to Stage 1 and Stage 3 is the return of the defective items to recovered at the warehouse which the recovered items are as good as new and have the same value as the procurement items. This system is shown in Figure 1.

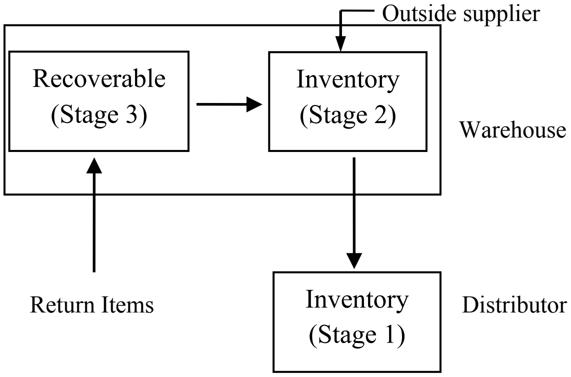


Figure 1: A one depot one distributor inventory system with return items

From Figure 1, the two-echelon inventory system consist warehouse and distributor. Customer can return items directly to the warehouse. Return items will return to Stage 3 and recover as good as a new one and also have the same value as the procurement items. The products at Stage 2 include the new items purchased from outside supplier and recovered items from Stage 3. The products at Stage 2 are intended to replenish the items to the inventory at Stage 1 and available for the demand of customers.

### *2.1 Assumptions*

We assume the assumptions for formulating the model as follows.

1. Demand and return items are random variables with Poisson distribution. Dependence is assumed between demand and return.
2. The return rate is less than the demand rate.
3. The time to recover of defective items is much less than the lead time.
4. The lead time between Stage 1 and 2 is constant.
5. The inventory system is not shortage items.
6. The time to return items is zero and no charged price to return.
7. The defective items are directly returned to the warehouse and recovered items are as good as new and have the same value as the procurement items.
8. The distributor knows the customer demand and order with the depot for replenishment.

### *2.2 Notations*

We introduce the following notations.

$A_i$ : Set-up cost at Stage $i\ (i=1,2,3)$

$h_i$ : Holding cost per unit time at Stage $i\ (i=1,2,3)$

$c$ : Item cost per unit items

$l_i$ : Lead time at Stage $i\ (i=1,2,3)$

$SS_i$ : Safety stock at Stage $i\ (i=1,2,3)$

$Z_\alpha$ : Safety factor at Stage $i\ (i=1,2,3)$

$\alpha$ : Error of lead time

$\lambda_D$ : Mean demand per unit time

$D$ : Quantity demand items per unit time; random variable

$R$ : Quantity return items per unit time; random variable

$\beta$ : Fraction of demand returned per unit time $0<\beta<1$

$Q$ : Order quantity

$Q^*$ : Optimal order quantity

$n^*$ : Optimal number of cycle at Stage 1

### *2.3 Mathematical model*

Stage 1 follows a stationary ordering policy. When inventory level falls at or below $s_1$, an order quantity of $Q$ units will order from Stage 2. Inventory level at Stage 2 must be enough to replenishment number of cycle $n$ cycles of Stage 1 .When inventory level at Stage 2 falls at or below $s_2$, an order quantity of $nQ$

units will order from two sources- recovered items at Stage 3 $\beta\lambda_D$ units and outside supplier $nQ-\beta\lambda_D$ units. The cycle length of Stage 2 is $nQ/\lambda_D$. The lead time at Stage 1 and 2 are constant. Inventory level is shown in Figure 2.

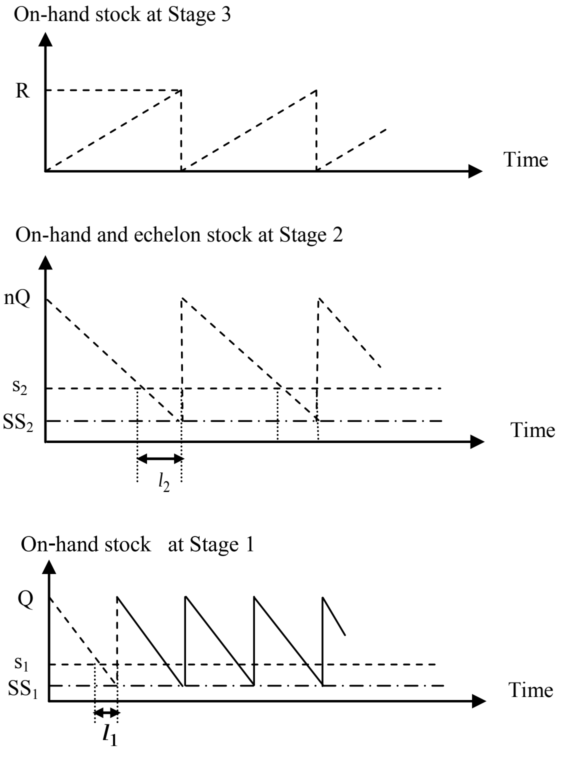


Figure 2: Inventory level at Stage 1, 2 and 3

From Figure 2, there are three parts of cost in inventory model such that item cost, set-up cost and holding cost. Each cost is calculated as follows.

1. The item cost ($TC_1$) includes the item cost per unit items $(c)$ and quantity demand items per unit time $(D)$. Therefore, the item cost is given by

$$TC_1 = cD \tag{1}$$

2. The set-up cost per unit time ($TC_2$) for 3 stages includes the set-up cost in each Stage ($A_i$ : $i=1,2,3$). Therefore, the set-up cost per unit time for 3 stages is given by

$$TC_2 = \frac{A_1 D}{Q} + \frac{A_2 D}{nQ} + \frac{A_3 D}{nQ}. \tag{2}$$

3. The holding cost per unit time ($TC_3$) at Stage 1 and 2 includes the order quantity $(Q)$ and safety stock ($SS_i$ : $i=1,2$) and number of return items $(R)$ at Stage 3. Therefore, the holding cost per unit time for 3 stages is given by

$$TC_3 = h_1(\frac{Q}{2} + SS_1) + h_2(\frac{(n-1)Q}{2} + SS_2) + h_3(\frac{RnQ}{2D}) \tag{3}$$

where $SS_1 = E(s_1 - D_{l_1}) = s_1 - \lambda_D l_1$.

The safety stock at Stage 2 $(SS_2)$ is a protection uncertainty in returns. The safety stock consist safety factor and variance of replenishment items in Stage 3 to Stage 2. So we have to know the distribution of replenishment items in Stage 3 to Stage 2. We assume that Poisson distribution with a mean and variance is $\beta\lambda_D$, and time to replenishment in each cycle from Stage 3 and Outside supplier to Stage 2 is $nQ/\lambda_D$. Thus, the distribution of replenishment items from Stage 3 to Stage 2 between $nQ/\lambda_D$ is Poisson with mean and variance $\beta\lambda_D\left(nQ/\lambda_D\right)$. To make easier to find the safety factor, we approximate Poisson distribution by Normal distribution [2, p. 390]. Hence, the safety stock at Stage 2 can be written as

$$SS_2 = Z_\alpha \sigma = Z_\alpha \sqrt{\beta n Q}.$$

Therefore, the total cost $(TC)$ can be written as

$$TC = cD + \frac{A_1 D}{Q} + \frac{A_2 D}{nQ} + \frac{A_3 D}{nQ} + h_1(\frac{Q}{2} + SS_1) + h_2(\frac{(n-1)Q}{2} + SS_2) + h_3(\frac{RnQ}{2D}). \tag{4}$$

Since, the quantity of demand items $(D)$ and the quantity of return items $(R)$ are random variables. So, we take expectation of equation (4) and obtain the average total cost $E(TC)$ as equation (5).

$$E(TC) = c\lambda_D + \frac{A_1\lambda_D}{Q} + \frac{A_2\lambda_D}{nQ} + \frac{A_3\lambda_D}{nQ} + h_1(\frac{Q}{2} + SS_1) + h_2(\frac{(n-1)Q}{2} + SS_2) + h_3(\frac{n\beta Q}{2}). \tag{5}$$

The optimal order quantity $Q^*$ and the optimal number of cycle $n^*$ can be calculated from equation (5). Eugene et al. [1] has studied the distribution of demand during lead time. Thus, we can calculate the re-order point at Stage 1 and 2 from equation (6),

$$\sum_{D=0}^{s_1} \frac{e^{-\lambda_D l_1}(\lambda_D l_1)^D}{D!} = 1-\alpha \text{ and}$$

$$\sum_{D=0}^{s_2} \frac{e^{-\lambda_D (l_1+l_2)}(\lambda_D (l_1+l_2))^D}{D!} = 1-\alpha \qquad (6)$$

where the lead time at Stage 1 and 2 are constant.

## 3. Numerical examples

In this section, we use the data from Mitra [5] and Wang and Di [6] to illustrate the results as follows.

$\lambda_D = 100$, $c = 100$, $\beta = 0.5$, $A_1 = 25$, $A_2 = 100$, $A_3 = 50$, $h_1 = 2$, $h_2 = 1$, $h_3 = 0.3$, $l_1 = 0.25$, $l_2 = 0.5$, $\alpha = 0.05$.

We calculate the optimal values of average total cost, order quantity $Q^*$, number of cycle $n^*$ and re-order point $s_i^*$. The results are showed in Table 1 and Figure 3.

Table 1: The optimal values of average total cost, order quantity and number of cycle

| $n$ | $Q$ | $E(TC)$ |
|---|---|---|
| 1 | 122 | 10,310 |
| 2 | 74 | 10,295 |
| 3 | 55 | 10,298 |
| 4 | 44 | 10,306 |
| 5 | 38 | 10,315 |

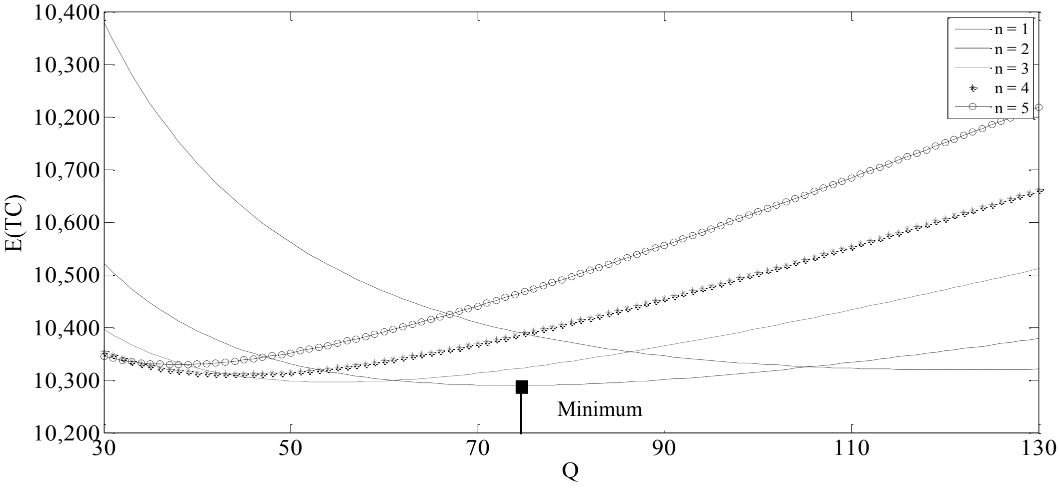


Figure 3: $E(TC)$ for cycle $n = 1,2,...,5$

From Table 1 and Figure 3, the result shows that minimum average total cost is 10,295 when the number of cycle as $n^* = 2$ and the order quantity as $Q^* = 74$. From equation (6), the optimal re-order point of Stage 1 and Stage 2 are 33 and 90, respectively. Thus, at Stage 1 when the inventory level falls at or below 33 units, an order quantity of 74 units will be ordered from Stage 2. At Stage 2, when the inventory level falls at or below 90 units, an order quantity of 148 units will be ordered from two sources, which are 1) Outside supplier $nQ - \beta\lambda_D = 98$ units and 2) Recovery items at Stage 3 $\beta\lambda_D = 50$ units.

We consider the sensitivity analysis of the average total cost based on the various values of the fraction of demand returned per unit time $(\beta)$, lead time at Stage 1 $(l_1)$ and Stage 2 $(l_2)$. The results are presented in Tables 2-4 and Figures 4-6.

1. Table 2 and Figure 4, the average total cost $E(TC)$ increases when the fraction of demand returned per unit time $(\beta)$ increases.

Table 2: Effect on $E(TC)$ for various values of $\beta$

| $\beta$ | $n$ | $Q$ | $E(TC)$ |
|---|---|---|---|
| 0.1 | 2 | 79 | 10,272 |
| 0.3 | 2 | 76 | 10,284 |
| 0.5 | 2 | 74 | 10,295 |
| 0.7 | 2 | 71 | 10,305 |
| 0.9 | 2 | 69 | 10,314 |

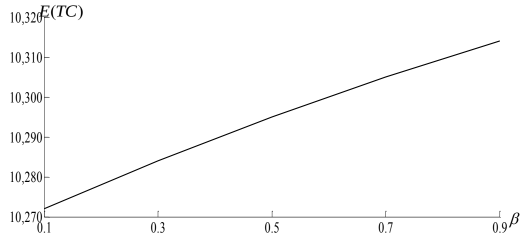


Figure 4: Effect on $E(TC)$ for various values of $\beta$

2. Table 3 and Figure 5, the average total cost increases when lead time at Stage 1 $(l_1)$ increases.

Table 3: Effect on $E(TC)$ for various values of $l_1$

| $l_1$ | $n$ | $Q$ | E(TC) |
|---|---|---|---|
| 0.1 | 2 | 74 | 10,289 |
| 0.25 | 2 | 74 | 10,295 |
| 0.5 | 2 | 74 | 10,303 |
| 0.75 | 2 | 74 | 10,309 |
| 0.9 | 2 | 74 | 10,311 |
| 1 | 2 | 74 | 10,313 |

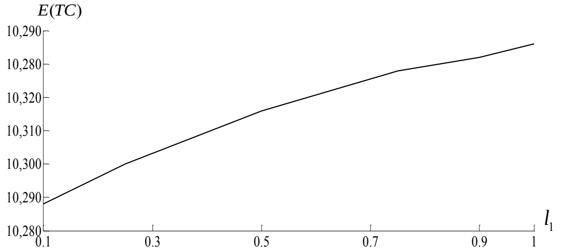


Figure 5: Effect on $E(TC)$ for various values of $l_1$

3. Table 4 and Figure 6, the average total cost is the same when lead time at Stage 2 ( $l_2$ ) increases.

Table 4: Effect on $E(TC)$ for various values of $l_2$

| $l_2$ | $n$ | $Q$ | E(TC) |
|---|---|---|---|
| 0.1 | 2 | 74 | 10,295 |
| 0.25 | 2 | 74 | 10,295 |
| 0.5 | 2 | 74 | 10,295 |
| 0.75 | 2 | 74 | 10,295 |
| 0.9 | 2 | 74 | 10,295 |
| 1 | 2 | 74 | 10,295 |

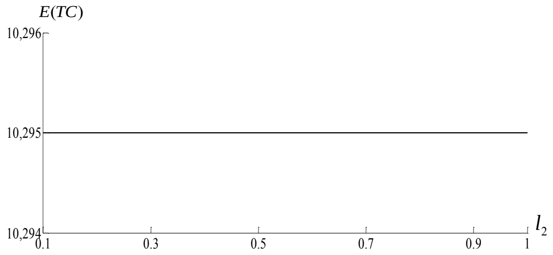


Figure 6: Effect on $E(TC)$ for various values of $l_2$

## 4. Conclusions

In this study, we propose a two-echelon inventory system with a single warehouse and a single distributor with return items under continuous review policy. The demand items and return items are dependent and distributed as Poisson distribution with constant lead time. The optimal of re-order point, order quantity, and number of cycle of order quantity that minimize the total cost of the system are obtained. Numerical examples are illustrated the sensitivity analysis of the average total cost that when the fraction between demand and returned and lead time at Stage 1 (distributor) are increasing; the average total cost of system will be increased. Moreover, lead time at State 2 (warehouse) is increasing; the average total cost of system is the same.

## References

[1] Kopytov E, Greenglaz L, Muravyov A, Puzinkevich E. Modelling of two strategies in inventory control system with random lead time and demand. Computer modelling and New Technologies. 2007; 21-30.

[2] Crawshaw J, Chambers J. A concise course in advanced level statistics with worked examples. 4th ed. Nelson Thornes; 2014.

[3] Korugan A, Gupta SM. A multi-echelon inventory system with returns. Computers and Industrial. 1998; 145-148.

[4] Silver EA, Pyke DF, Peterson R. Inventory management and production planning and scheduling. 3rd ed. New York: Wiley; 1998.

[5] Mitra S, Analysis of Two-echelon inventory system with returns. Omega 37. 2009; 106–115.

[6] Wang, D. An (s,S) Inventory control model with return flows. Business Management and Electronic Information. 2011; 328-331.

# A Comparison of Efficiency between Multivariate Shewhart and Multivariate CUSUM Control Chart for Bivariate Copula

Sasigarn Kuvattana[1*], Saowanit Sukparungsee[1] , Piyapatr Busababodhin[2] and Yupaporn Areepong[1]

[1]Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, 10800, Thailand, e-mail: sasigarn2010@gmail.com, swns@kmutnb.ac.th, yupaporna@kmutnb.ac.th
[2]Department of Mathematics, Faculty of Science, Mahasarakham University, Mahasarakham, 41150, Thailand, e-mail: piyapatr99@gmail.com

**Abstract**

The objective of this paper is to compare efficiency between multivariate Shewhart and multivariate CUSUM control chart for bivariate copula when observations are exponential distribution. Monte Carlo simulation is used to investigate the value of Average Run Length (*ARL*) and measure the observation dependencies by using Kendall's tau for each copula. The positive and negative dependence was proposed in the case of one and two parameter shifts. The numerical results indicate that multivariate CUSUM control chart is an effective alternative to the multivariate Shewhart control chart in term of sensitive to the detection of small to moderate shifts.

*Keywords*: *ARL*, copula, multivariate CUSUM control chart, multivariate Shewhart control chart

*Corresponding Author
E-mail Address: sasigarn2010@gmail.com

## 1. Introduction

Traditionally, Statistical Process Control (SPC) charts are widely implemented in monitoring and improving quality of production in industry manufacturing [1]. Control charts are the simplest type of statistical process control procedure. Quality control problems in industry may involve more than a single quality characteristic. Generally, multivariate detection procedures are based on a multi-normality assumption and independence but many processes are often non-normality and correlated. Most of these multivariate control charts are generalizations of their univariate counterparts [2], such as the multivariate Shewhart control chart, multivariate exponentially weighted moving average (MEWMA) control chart proposed by Lowry et al. [3] and the multivariate cumulative sum (MCUSUM) control chart. Multivariate Shewhart control chart is used to detect large shifts in the mean vectors. The MEWMA and MCUSUM control charts are commonly used to detect small or moderate shifts in the mean vectors [4]. Many multivariate control charts have the lack of the related joint distribution and copula can specify this property. The copula approach is a popular tool for modeling nonlinearity, asymmetricality and tail dependence in several fields; it can be used in the study of dependence or association between random variables. Copula modeling is based on a representation from Sklar's theorem [5] and it can estimate joint distribution of nonlinear outcomes and explain the dependence structure among variables through the joint distribution by eliminating the effect of univariate marginals. A bivariate copula is the simplest case for the description of dependent random variables and it can apply to control chart [6]. Recently, several papers use copula in control chart such as, copula based on bivariate ZIP control chart [7-8], copula Markov CUSUM chart [9], Shewhart control charts for autocorrelated and normal data [10], non-normal multivariate cases for the Hotelling $T^2$ control chart [11] and bivariate copula on the Shewhart control chart [12].

This paper presents comparison of efficiency between multivariate Shewhart and multivariate CUSUM control chart when observations are exponential distribution with the means shifts and use a bivariate copula function for specifying dependence between random variables.

## 2. Research Methodology

This section consisted of the following

### *2.1 The multivariate Shewhart control chart*

For multivariate random variables, suppose that $\bar{\mathbf{x}}$ and $\mathbf{S}$ are the sample mean vector and covariance of matrix, respectively. There are $m$ samples and $p$ is the number of quality characteristics observed in each sample. The Hotelling $T^2$ statistic is

$$T^2 = (\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) \qquad (1)$$

where $\bar{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i$ and $\mathbf{S} = \frac{1}{m-1}\sum_{i=1}^{m}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$.

SPC is usually split into two phases, namely phase I and phase II, as explained in Montgomery [13]. For the phase I, control limits are $UCL = \frac{(m-1)^2}{m} \; \beta_{\alpha, p/2,(m-p-1)/2}$ and $LCL = 0$ where $\beta_{\alpha, p/2,(m-p-1)/2}$ is the upper $\alpha$ percentage point of a beta distribution with parameters $p/2$ and $(m-p-1)/2$ (see Tracy et al. [14] and

Bersimis et al. [15] ). For the phase II, control limits for this statistic are $UCL = \frac{p(m+1)(m-1)}{m^2 - mp} F_{\alpha,p,(m-p)}$ and $LCL = 0$ where $F_{\alpha,p,(m-p)}$ is $F$ distribution with parameters $p$ and $(m-p)$.

Note that this article will focus on phase II control charts and their performance.

### *2.2 The multivariate cumulative sum control chart*

The multivariate cumulative sum (MCUSUM) control chart is the multivariate extension of the univariate cumulative sum (CUSUM) control chart. The MCUSUM control chart was initially proposed by Crosier [16]. The MCUSUM control chart may be expressed as follows:

$$C_t = [(\mathbf{S}_{t-1} + \mathbf{x}_t - \mathbf{a})' {\textstyle\sum}^{-1} (\mathbf{S}_{t-1} + \mathbf{x}_t - \mathbf{a})]^{1/2} \qquad (2)$$

where covariance$\left(\sum\right)$ and $\mathbf{S}_t$ are the cumulative sums; $t = 1,\ 2,\ 3,\ldots$

$$\mathbf{S}_t = \begin{cases} 0, & \text{if } C_t \le k \\ (\mathbf{S}_{t-1} + \mathbf{x}_t - \mathbf{a})\left(1 - \dfrac{k}{C_t}\right), & \text{if } C_t > k \end{cases} \qquad (3)$$

the reference value $k > 0$ and $\mathbf{a}$ is the aim point or target value for the mean vector [17]. The control chart statistics for MCUSUM chart is

$$Y_t = [\mathbf{S}_t' {\textstyle\sum}^{-1} \mathbf{S}_t]^{1/2};\ t = 1,\ 2,\ 3,\ldots \qquad (4)$$

The signal gives an out-of-control if $Y_t > h$ where $h$ is the control limit [18].

### *2.3 Copula function*

The Sklar's theorem for a bivariate case, let $X$ and $Y$ be continuous random variables with joint distribution function $H$ and marginal cumulative distribution $F(x)$ and $F$(y), respectively. Then $H(x,y) = C\left(F(x), F(y);\theta\right)$ with a copula $C:[0,1]^2 \to [0,1]$ where $\theta$ is a parameter of the copula called the dependence parameter, which measures dependence between the marginals. For the purposes of statistical method it is desirable to parameterize the copula function. Let $\theta$ denote the association parameter of the bivariate distribution and there exists a copula $C$. Then $F(x) = u$, $F(y) = v$ where $u$ and $v$ are uniformly distributed variates [19]. This paper focuses on Normal copula and one type of Archimedean copulas which is Frank copula [20].

### *2.3.1 Normal copula*

$$C(u,v;\ \theta) = \Phi_N(\Phi^{-1}(u), \Phi^{-1}(v);\ \theta)\,; -1 \le \theta \le 1 \qquad (5)$$

where $\Phi_N(u,v)$ is the cumulative probability distribution function of the bivariate normal distribution, $\Phi^{-1}(u)$ and $\Phi^{-1}(v)$ are the inverse of the cumulative probability function of the univariate normal distribution.

### *2.3.2 Frank copula*

For Archimedean copulas, let a class $\Phi$ of functions $\phi:[0,1] \to [0,\infty]$ with continuous, strictly decreasing, such that $\phi(1) = 0$, $\phi'(t) < 0$ and $\phi''(t) > 0$ for all $0 < t < 1$ [20-22]. An Archimedean copula of Frank copula is generated as follow:

Frank copula

$$C(u,v;\theta) = -\frac{1}{\theta} ln(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1}) \qquad (6)$$

where $\phi(t) = -ln(\frac{e^{-\theta t} - 1}{e^{-\theta} - 1})$ ; $\theta \in (-\infty,\infty) \setminus 0$.

### *2.4 Dependence measures for data*

Theoretically, a parametric measure of the linear dependence between random variables is correlation coefficient and nonparametric measures of dependence are Spearman's rho and Kendall's tau. According to the earlier literature, the copulas can be used in the study of dependence or association between random variables and the values of Kendall's tau are easy to calculate so this measure is used for observation dependencies.

Let $X$ and $Y$ be continuous random variables whose copula is $C$ then Kendall's tau for $X$ and $Y$ is given by $\tau_c = 4\iint_{\mathbf{I}^2} C(u,v)\ dC(u,v)$ - 1 where $\tau_c$ is Kendall's tau of copula $C$ and the unit square $\mathrm{I}^2$ is the product $\mathrm{I} \times \mathrm{I}$ where $\mathrm{I} = [0,1]$ and the expected value of the function $C(u,v)$ of uniform (0,1) random variables U and V whose joint distribution function is $C$, i.e., $\tau_c = 4E[C(U,V)] - 1$ [21].

Genest and McKay [20] considered Archimedean copula $C$ generated by $\phi$, then $\tau_{Arch} = 4\int_0^1 \frac{\phi(t)}{\phi'(t)}\ dt$ +1 where $\tau_{Arch}$ is Kendall's tau of Archimedean copula $C$.

### *2.4.1 Normal copula*

$$\tau = arcsin(\theta)/(\pi/2)\ ;\ \theta \in [-1,1] \qquad (7)$$

### *2.4.2 Frank copula*

$$\tau = 1 + 4\left(\frac{1}{\theta}\int_0^\theta \frac{\mathrm{t}}{\mathrm{e}^{\mathrm{t}} - 1}\,\mathrm{dt}\ \text{-1}\right)/\theta\ ;\ \theta \in (-\infty,\infty) \setminus \{0\} \quad (8)$$

### *2.5 Average Run Length*

The performance of a control chart is measured by the Average Run Length (*ARL*). The ARL is classified into *ARL₀* and *ARL₁*. *ARL₀* is defined as the expected of false alarm time before an in-control process is taken to signal to be out of control. A sufficient large in-control *ARL₀* is desired. When the process is out-of-control, the performance of a control chart is usually used as $ARL_1$. It is the expected number of observations taken from an out-of-control process until the control chart signals that the process is out-of-control. Ideally, $ARL_1$ should be small.

### *2.6 Monte Carlo simulation*

In this paper, we use Monte Carlo simulation for ARL calculation by using R statistical software [23-25] with the number of simulation runs 50,000. Observations were from exponential distribution with parameter ( $\alpha$ ) equal to 1 for in control process ( $\mu_0 = 1$ ) and the shifts of the process level ( $\delta$ ) by $\mu = \mu_0 + \delta$. The process mean are equal to 1.02, 1.04, 1.05, 1.10, 1.25 and 1.50 with sample size is 1,000.

The simulation experiments carried out to assess the performance of control chart. Copula estimations are restricted to the cases of dependence (positive and negative dependence). For all copula models, setting $\theta$ corresponds with Kendall's tau. The level of dependence is measured by Kendall's tau values ( $-1 \le \tau \le 1$ ). For moderate and strong dependence, Kendall's tau values are defined to 0.5 and -0.8, respectively.

## 3. Research Results

The results are presented in Table 1-4 for the different values of Kendall's tau and denoted $\mu_1$ for the variables $X$ and $\mu_2$ for the variables $Y$. The control chart was chosen by setting the desired $ARL_0 = 370$ for each copulas. Table 1 and 2 show positive dependence $(\tau > 0)$ and Table 3 and 4 show negative dependence $(\tau < 0)$. For example, Table 1 shows moderate positive dependence when the shifts in one of exponential parameters. In the case of moderate dependence $(\tau = 0.5)$, for small and moderate shifts $(\mu_1 = 1,\ 1.02 \le \mu_2 \le 1.10)$, the $ARL_1$ values of Frank copula on MCUSUM control chart are less than multivariate Shewhart control chart. For large shifts $(\mu_1 = 1,\ 1.25 \le \mu_2 \le 1.5)$, the $ARL_1$ values of Normal copula on multivariate Shewhart control chart are less than MCUSUM control chart. Table 2 shows moderate positive dependence when the shifts in both of exponential parameters. In the case of moderate dependence $(\tau = 0.5)$, for small and moderate shifts $(1.02 \le \mu_1 \le 1.10,\ 1.02 \le \mu_2 \le 1.10)$, the $ARL_1$ values of Frank copula on MCUSUM control chart are less than multivariate Shewhart control chart. For large shifts $(1.25 \le \mu_1 \le 1.5,\ 1.25 \le \mu_2 \le 1.5)$, the $ARL_1$ values of Normal copula on multivariate Shewhart control chart are less than MCUSUM control chart. Table 3 shows strong negative dependence when the shifts in one of exponential parameters. In the case of strong dependence $(\tau = -0.8)$, for small shifts $(\mu_1 = 1,\ 1.02 \le \mu_2 \le 1.04)$, the $ARL_1$ values of Normal copula on MCUSUM control chart are less than multivariate Shewhart control chart. For large shifts $(\mu_1 = 1,\ 1.25 \le \mu_2 \le 1.5)$, the $ARL_1$ values of Normal copula on multivariate Shewhart control chart are less than MCUSUM control chart. Table 4 shows strong negative dependence when the shifts in both of exponential parameters. In the case of strong dependence $(\tau = -0.8)$, for small shifts $(\mu_1 = 1.02,\ \mu_2 = 1.02)$, the $ARL_1$ values of Frank copula on MCUSUM control chart are less than multivariate Shewhart control chart. For small and moderate shifts $(1.04 \le \mu_1 \le 1.10,\ 1.04 \le \mu_2 \le 1.10)$,the $ARL_1$ values of Normal copula on multivariate Shewhart control chart are less than MCUSUM control chart. For $1.25 \le \mu_1 \le 1.5,\ 1.25 \le \mu_2 \le 1.5$, the $ARL_1$ values of Frank copula on multivariate Shewhart control chart are less than MCUSUM control chart.

Table 1: *ARL* of control chart with Kendall's tau values equal to 0.5 when the shifts in one of exponential parameters

| Parameters | | *ARL₀* and *ARL₁* | | | |
|---|---|---|---|---|---|
| | | Multivariate Shewhart | | MCUSUM | |
| $\mu_1$ | $\mu_2$ | Normal | Frank | Normal | Frank |
| **1** | **1** | **369.909** | **369.952** | **370.017** | **370.016** |
| 1 | 1.02 | 354.200 | 359.144 | 355.204 | **351.755** |
| 1 | 1.04 | 339.385 | 343.251 | 338.425 | **336.598** |
| 1 | 1.05 | 330.687 | 334.279 | 329.892 | **328.320** |
| 1 | 1.10 | 291.963 | 296.346 | 291.047 | **288.525** |
| 1 | 1.25 | **193.047** | 197.586 | 199.409 | 195.186 |
| 1 | 1.5 | **100.576** | 105.701 | 111.201 | 109.707 |

Table 2: *ARL* of control chart with Kendall's tau values equal to 0.5 when the shifts in both of exponential parameters

| Parameters | | *ARL₀* and *ARL₁* | | | |
|---|---|---|---|---|---|
| | | Multivariate Shewhart | | MCUSUM | |
| $\mu_1$ | $\mu_2$ | Normal | Frank | Normal | Frank |
| **1** | **1** | **369.909** | **369.952** | **370.017** | **370.016** |
| 1.02 | 1.02 | 340.598 | 343.388 | 337.617 | **337.262** |
| 1.04 | 1.04 | 309.692 | 315.924 | 311.937 | **308.447** |
| 1.05 | 1.05 | 297.258 | 299.862 | 297.390 | **295.694** |
| 1.10 | 1.10 | 238.900 | 241.024 | 240.551 | **236.408** |
| 1.25 | 1.25 | **131.193** | 132.817 | 138.760 | 133.198 |
| 1.5 | 1.5 | **61.371** | 61.971 | 66.428 | 64.077 |

Table 3: *ARL* of control chart with Kendall's tau values equal to -0.8 when the shifts in one of exponential parameters

| Parameters | | *ARL₀* and *ARL₁* | | | |
|---|---|---|---|---|---|
| | | Multivariate Shewhart | | MCUSUM | |
| $\mu_1$ | $\mu_2$ | Normal | Frank | Normal | Frank |
| **1** | **1** | **370.012** | **369.959** | **369.840** | **369.898** |
| 1 | 1.02 | 358.275 | 357.948 | **357.675** | 360.032 |
| 1 | 1.04 | 343.146 | 342.560 | **340.720** | 342.948 |
| 1 | 1.05 | **331.405** | 333.608 | 334.221 | 335.469 |
| 1 | 1.10 | **297.878** | 298.801 | 300.079 | 302.912 |
| 1 | 1.25 | **200.245** | 200.871 | 212.436 | 213.616 |
| 1 | 1.5 | **106.630** | 107.303 | 121.119 | 122.585 |

Table 4: *ARL* of control chart with Kendall's tau values equal to -0.8 when the shifts in both of exponential parameters

| Parameters | | *ARL₀* and *ARL₁* | | | |
|---|---|---|---|---|---|
| | | Multivariate Shewhart | | MCUSUM | |
| $\mu_1$ | $\mu_2$ | Normal | Frank | Normal | Frank |
| **1** | **1** | **370.012** | **369.959** | **369.840** | **369.898** |
| 1.02 | 1.02 | 343.512 | 344.320 | 343.811 | **341.284** |
| 1.04 | 1.04 | **316.280** | 317.664 | 319.293 | 318.649 |
| 1.05 | 1.05 | **302.046** | 304.128 | 308.907 | 306.131 |
| 1.10 | 1.10 | **252.182** | 252.369 | 254.984 | 257.884 |
| 1.25 | 1.25 | 147.097 | **146.783** | 158.015 | 159.226 |
| 1.5 | 1.5 | 74.906 | **74.165** | 80.594 | 81.915 |

## 4. Conclusion

We compare efficiency between multivariate Shewhart and multivariate CUSUM control chart for bivariate copula when observations are exponential distribution by using Monte Carlo simulation approach. The results found that multivariate CUSUM control chart performs better than multivariate Shewhart control chart for small to moderate values of changes and the performance of multivariate Shewhart chart is superior to multivariate CUSUM chart for large changes.

## Acknowledgements

The authors would like to thank for the King Mongkut's University of Technology, North Bangkok and Thailand Ministry of Science and Technology for the financial support during this research.

## References

[1] Sukparungsee S. Combining martingale and integral equation approaches for finding optimal parameters of EWMA. Applied Mathematical Sciences. 2012; 90( 6): 4471-4482.

[2] Mahmoud MA, Maravelakis PE. The performance of multivariate CUSUM control charts with estimated parameters. Journal of Statistical Computation and Simulation. 2013; 83(4): 721-738.

[3] Lowry CA, WH Woodall, Champ CW, Rigdon SE. A multivariate exponentially weighted moving average control chart. Technometrics. 1992; 34: 46-53.

[4] Midi HM and Shabbak A. Robust multivariate control charts to detect small shifts in mean. Mathematical Problems in Engineering. 2011; Special section: 1-19.

[5] Sklar A. Random variables, joint distributions, and copulas, Kybernetica. 1973; 9: 449-460.

[6] Kuvattana S, Sukparungsee S, Busababodhin P, Areepong Y. Efficiency of bivariate copula on the CUSUM chart, Proceedings of the International MultiConference of Engineers and Computer Scientists 2015 Vol II, IMECS 2015, March 18-20, Hong Kong. 2015. p.829-832.

[7] Fatahi AA, Dokouhaki P, Moghaddam BF. A bivariate control chart based on copula function. IEEE International Conference on Quality and Reliability (ICQR), 2011 Sept 14-17; Bangkok, Thailand. 2011. p. 292-296.

[8] Fatahi AA, Noorossana R, Dokouhaki P, Moghaddam BF. Copula-based bivariate ZIP control chart for monitoring rare events. Communications in statistics theory and methods. 2012; 41: 2699-2716.

[9] Dokouhaki P, Noorossana R. A copula markov CUSUM chart for monitoring the bivariate auto-correlated binary observation. Quality and Reliability Engineering International. 2013; 29: 911-919.

[10] Hryniewicz O. On the robustness of the Shewhart control chart to different types of dependencies in data. Frontiers in Statistical Quality Control. 2012; 10: 20-33.

[11] Verdier G. Application of copulas to multivariate control charts. Journal of Statistical Planning and Inference. 2013; 143: 2151-2159.

[12] Kuvattana S, Sukparungsee S, Busababodhin P, Areepong Y. Efficiency of bivariate copula on the Shewhart control chart. The 21st International Conference on Computational Statistics (COMPSTAT 2014), 2014 August 19-22; Geneva, Switzerland.

[13] Montgomery DC. Statistical quality control: a modern introduction. 6th ed. New York: Wiley; c2009.

[14] Tracy ND, Young JC, Mason RL. Multivariate control charts for individual observations, Journal of Quality Technology. 1992; 24(2): 88-95.

[15] Bersimis S, Psarakis S, Panaretos J. Multivariate statistical process control chart: an overview. Quality and Reliability Engineering International. 2007; 23: 517-543.

[16] Crosier RB. Multivariate generalizations of cumulative sum quality-control schemes. Technometrics. 1988; 30: 291–303.

[17] Khoo BC, Atta MA, Phua HN. A Study on the performances of MEWMA and MCUSUM charts for skewed distributions. In Proc.10th Islamic Countries Conference on Statistical Science; Cairo, Egypt. 2009. p. 817-822.

[18] Alves CC, Samohyl RW, Henning E. Application of multivariate cumulative sum control charts(MCUSUM) for monitoring a machining process. 16th Int. Conf. on Industrial Engineering and Operations Management; 2009 Oct 12-15; Sao Carlos, Brazil. 2010. p. 1-7.

[19] Trivedi PK, Zimmer DM. Copula Modeling: an introduction for practitioners. Foundations and Trends in Econometrics. 2005.

[20] Genest C, McKay RJ. The joy of copulas: bivariate distributions with uniform marginals. American Statistician. 1986; 40: 280–283.

[21] Nelsen RB. An introduction to copulas. 2nd ed. New York: Springer; 2006.

[22] Genest C and Rivest L-P. Statistical inference procedures for bivariate Archimedean copulas. Journal of the American Statistical Association. 1993; 88: 1034–1043.

[23] Yan J. Enjoy the joy of copulas: with a package copula. Journal of Statistical Software. 2007; 21: 1-21.

[24] Hofert M, M̈achler M, McNeil AJ. Likelihood inference for Archimedean copulas in high dimensions under known margins. Journal of Multivariate Analysis. 2012; 110: 133–150.

[25] M̈achler M, Zurich E. Numerically stable Frank copula functions via multiprecision: R Package Rmpfr. [Internet]. 2013 [cited 2013 March 15]. Available from: http://cran.r-project.org/web/packages/copula/vignettes/Frank-Rmpfr.pdf

# Economic Design for Fuzzy Variable Parameters $\overline{X}$ control charts using $\alpha$-cut under Non- normality

Chaowalit Panthong[1] and Adisak Pongpullponsak[2*]

[1]Department of Mathematics, King Mongkut's University of Technology Thonburi, Bangkok, 10140, Thailand,
Chao-kmutt@hotmail.com

[2]Department of Mathematics, King Mongkut's University of Technology Thonburi, Bangkok, 10140, Thailand,
adisak.pon@kmutt.ac.th

**Abstract**

The purposes of research were to present the economic model of fuzzy variable parameters $\overline{X}$ control charts(FVP) and fuzzy variable parameters $\overline{X}$ control charts by weighted variance method (FVP-WV) for non - normality data,used Weibull distribution.The performances of the control charts are the expected value of all expernses per one single unit(ECTU) of time as standard,the results indicated that the production small process mean shift the fuzzy variable parameters $\overline{X}$ control charts by weighted variance method (FVP-WV) have more efficocied of the fuzzy variable parameters $\overline{X}$ control charts (the coefficient of skewness at 0.1, 0.5 and 1.0).

*Keywords*: Economic design ; Fuzzy variable parameters $\overline{X}$ control charts ; $\alpha$ - cut; non-normality distribution

* Adisak Pongpullponsak
*adisak.pon@kmutt.ac.th*

## 1. Introduction

The fuzzy method were considered and developed in the variable parameters $\overline{X}$ control charts so more efficiency. Zadeh[1] was the first person to use fuzzy theory with type of data which consisted of human error ,masuring device and environmental conditions. Gullay[2] suggested the $\alpha$ - cut for fuzzy control charts for linguistic data. Senturk and Erginel[3] presented the fuzzy $\overline{X}-R$ and $\overline{X}-S$ control charts using $\alpha$ - cut. The methods used in the transformation of fuzzy sets into scalars used the $\alpha$ -level fuzzy midrange,see Wang[4]. Pongpullponsak and Intaramo[5] apply fuzzy theory and control chart to upgrade control charts in non-normallity.The topic of the research was development of fuzzy extreme value theory control chart using $\alpha$ - cut for skewed used data distributions for skewed are Weibull , lognormal and Burr's.The effective of control charts are investigated by ARL. Pongpullponsak and Panthong[6] studied $\overline{X}$ control charts for variable parameters by weighted variance used the fuzzy method are $\alpha$ - cut and $\alpha$ - level fuzzy midrange.Used the triangular fuzzy number(a,b,c) for create the limit control charts, data distribution Weibull and lognormal are used and the average number of observations to signal(ANOS), the adjusted average time to signal (AATS) and the average time to signal(ATS) to determine the efficiency of the control charts.

Actually, the investigation of manufacturing process must have been a cost involved such as cost of sampling data and cost per unit atc.Yan-Kwang Chen[7] studied economic design of control chart for non-normal data using variable sampling policy , De Magalh$a$ s et al.[8] proposed an economic model for variable parameters $\overline{X}$ control charts used the expected cost per time unit(ECTU) for was used for assessing the efficiency of the control charts.For the skewed data, Pongpullponsak and Panthong[9] presented the economic model of $\overline{X}$ control charts using Shewhart method for skewed distributions, Weibull , lognormal and Burr's are used for data distribution.

The objective of this research is to introduce the economic model of the fuzzy variable parameters $\overline{X}$ control charts by weighted variance method using $\alpha$ - cut under non-normality. We can use the expected cost per time unit(ECTU) investigate the efficiency of the $\overline{X}$ control charts.

## 2. Research Methodology

### 2.1 Data distribution

Weibull distribution is continuous distribution that is used widely. Let $X$ equal continuous random variables that are weibull distribution with $\beta > 0$ and $\theta > 0$.

Density function

$$f(x;\theta,\beta) = \frac{\beta}{\theta^{\beta}} x^{\beta-1} e^{-(x/\theta)^{\beta}} \; ; \; x > 0 \qquad (1)$$

cumulative distribution function

$$F(x;\theta,\beta) = 1 - e^{-(x/\theta)^{\beta}} \quad ; \quad x > 0 \qquad (2)$$

mean

$$\mu = E(X) = \frac{\theta}{\beta}\Gamma(\frac{1}{\beta}) \tag{3}$$

Variance

$$\sigma^2 = V(X) = \frac{\theta^2}{\beta}\left\{2\Gamma(\frac{2}{\beta}) - \frac{1}{\beta}\left[\Gamma(\frac{1}{\beta})\right]^2\right\} \tag{4}$$

where

$\theta$ is scale parameter.

$\beta$ is shape parameter.

In this study using $\theta = 0.1, 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9$ and $\beta$ are relevant with a coefficient of skewness at $.\alpha_3 \in \{0.1, 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ shown in table 1.

Table 1: A coefficient of skewness and shape parameters of Weibull distribution

| Coefficient of skewness | shape parameters |
|---|---|
| 0.1 | 3.2219 |
| 0.5 | 2.2110 |
| 1 | 1.5630 |
| 2 | 1.0000 |
| 3 | 0.7686 |
| 4 | 0.6478 |
| 5 | 0.5737 |
| 6 | 0.5237 |
| 7 | 0.4873 |
| 8 | 0.4596 |
| 9 | 0.4376 |

## 2.2 Control charts

Pongpullponsak and Panthong[6] has developed the variable parameters $\overline{X}$ control charts from fuzzy theor. Firstly,A triangular fuzzy number(a,b,c) is used for constructing control charts and then the data that are necessary are cut by considering $\alpha$ - cut theory. Then the data are transformed by $\alpha$ -level fuzzy midrange. The control charts can be writlen as [6] and some of sample are as the following :

### 2.2.1 The $\alpha$ - cut of fuzzy variable parameters $\overline{X}$ control charts: VP - WV

$$\left.\begin{aligned}
U\tilde{C}L_i^{\alpha} &= \left(\overline{\overline{X}}{}^{\alpha}_{a,i} + \frac{k_i\overline{R}{}^{\alpha}_{a,i}}{3}, \overline{\overline{X}}_{b,i} + \frac{k_i\overline{R}_{b,i}}{3}, \overline{\overline{X}}{}^{\alpha}_{c,i} + \frac{k_i\overline{R}{}^{\alpha}_{c,i}}{3}\right) \\
U\tilde{W}L_i^{\alpha} &= \left(\overline{\overline{X}}{}^{\alpha}_{a,i} + \frac{w_i\overline{R}{}^{\alpha}_{a,i}}{3}, \overline{\overline{X}}_{b,i} + \frac{w_i\overline{R}_{b,i}}{3}, \overline{\overline{X}}{}^{\alpha}_{c,i} + \frac{w_i\overline{R}{}^{\alpha}_{c,i}}{3}\right) \\
C\tilde{L}_i^{\alpha} &= \left(\overline{\overline{x}}{}^{\alpha}_{a,i}, \overline{\overline{x}}_{b,i}, \overline{\overline{x}}{}^{\alpha}_{c,i}\right) \\
L\tilde{W}L_i^{\alpha} &= \left(\overline{\overline{X}}{}^{\alpha}_{a,i} - \frac{w_i\overline{R}{}^{\alpha}_{a,i}}{3}, \overline{\overline{X}}_{b,i} - \frac{w_i\overline{R}_{b,i}}{3}, \overline{\overline{X}}{}^{\alpha}_{c,i} - \frac{w_i\overline{R}{}^{\alpha}_{c,i}}{3}\right) \\
L\tilde{C}L_i^{\alpha} &= \left(\overline{\overline{X}}{}^{\alpha}_{a,i} - \frac{k_i\overline{R}{}^{\alpha}_{a,i}}{3}, \overline{\overline{X}}_{b,i} - \frac{k_i\overline{R}_{b,i}}{3}, \overline{\overline{X}}{}^{\alpha}_{c,i} - \frac{k_i\overline{R}{}^{\alpha}_{c,i}}{3}\right)
\end{aligned}\right\} \tag{5}$$

where i = 1,2.

### 2.2.2 The $\alpha$ -cut of fuzzy Variable parameters $\overline{X}$ control charts by weighted variance method : FVP – WV

The $\alpha$ -cut fuzzy mean control limits by weighted variance method are

$$\left.\begin{aligned}
U\tilde{C}L_i^{\alpha} &= \left(\overline{\overline{X}}{}^{\alpha}_{a,i} + \frac{k_iW_{U_{a,i}}\overline{R}{}^{\alpha}_{a,i}}{3}, \overline{\overline{X}}_{b,i} + \frac{k_iW_{U_{b,i}}\overline{R}_{b,i}}{3}, \overline{\overline{X}}{}^{\alpha}_{c,i} + \frac{k_iW_{U_{c,i}}\overline{R}{}^{\alpha}_{c,i}}{3}\right) \\
U\tilde{W}L_i^{\alpha} &= \left(\overline{\overline{X}}{}^{\alpha}_{a,i} + \frac{w_iW_{U_{a,i}}\overline{R}{}^{\alpha}_{a,i}}{3}, \overline{\overline{X}}_{b,i} + \frac{w_iW_{U_{b,i}}\overline{R}_{b,i}}{3}, \overline{\overline{X}}{}^{\alpha}_{c,i} + \frac{w_iW_{U_{c,i}}\overline{R}{}^{\alpha}_{c,i}}{3}\right) \\
C\tilde{L}_i^{\alpha} &= \left(\overline{\overline{x}}{}^{\alpha}_{a,i}, \overline{\overline{x}}_{b,i}, \overline{\overline{x}}{}^{\alpha}_{c,i}\right) \\
L\tilde{W}L_i^{\alpha} &= \left(\overline{\overline{X}}{}^{\alpha}_{a,i} - \frac{w_iW_{L_{a,i}}\overline{R}{}^{\alpha}_{a,i}}{3}, \overline{\overline{X}}_{b,i} - \frac{w_iW_{L_{b,i}}\overline{R}_{b,i}}{3}, \overline{\overline{X}}{}^{\alpha}_{c,i} - \frac{w_iW_{L_{c,i}}\overline{R}{}^{\alpha}_{c,i}}{3}\right) \\
L\tilde{C}L_i^{\alpha} &= \left(\overline{\overline{X}}{}^{\alpha}_{a,i} - \frac{k_iW_{L_{a,i}}\overline{R}{}^{\alpha}_{a,i}}{3}, \overline{\overline{X}}_{b,i} - \frac{k_iW_{L_{b,i}}\overline{R}_{b,i}}{3}, \overline{\overline{X}}{}^{\alpha}_{c,i} - \frac{k_iW_{L_{c,i}}\overline{R}{}^{\alpha}_{c,i}}{3}\right)
\end{aligned}\right\} \tag{6}$$

where i = 1,2

Then, the condition of process control for each sample can be defined as

Process control =

$$\begin{cases}
warning\ control\,; UCL \geq S^{\alpha}_{mr-\overline{X},j} > UWL and\ LWL < S^{\alpha}_{mr-\overline{X},j} \leq LWL \\
in\ \ control\,;\ LWL \leq\ S^{\alpha}_{mr-\overline{X},j}\ \leq\ UWL \\
out\ of\ control\,;\ S^{\alpha}_{mr-X,j}\ <\ LCL\ \ and\ \ S^{\alpha}_{mr-\overline{X},j} > UCL
\end{cases} \tag{7}$$

where

$$S^{\alpha}_{mr-\overline{X},j} = \frac{(\overline{X}_{a_j} + \overline{X}_{c_j}) + \alpha\left[(\overline{X}_{b_j} - \overline{X}_{a_j}) - (\overline{X}_{c_j} - \overline{X}_{b_j})\right]}{2} \qquad (8)$$

## 2.3 The cost model

The expected cost per time unit $(ECTU)$ is the ratio of the expected cost per cycle $(E(C))$ to the expected cycle time $(E(T))$, that is

$$ECTU = \frac{E(C)}{E(T)} . \qquad (9)$$

**ECTU calculation (in case of fuzzy theory)**

$$E(C) = (1/\lambda)C_0 + C_1\left[AATS + n'G + \delta_1 T^* + \delta_2 T^{**}\right] + YE(F) + W + (a+bn)\tilde{s} + (a+bn')s' \qquad (10)$$

the value of used following Pongpullponsak[6] $C_0$ , $C_1$ , $G$ , $T_*$ , $T_{**}$ , $T_0$ , $W$ , $Y$ , $\delta_1$, $\delta_2$ , $a$ and $b$ .

$AATS$ is the mean time interval since a process shift until an alarm occurs.

Determination of $n$ is

$$n = \overline{n}_a P_{0,i} + \overline{n}_c (1 - P_{0,i}) , \qquad (11)$$

determination of $n'$ is

$$n' = \overline{n}_a P_{0,i}(\delta) + \overline{n}_c (1 - P_{0,i}(\delta)), \qquad (12)$$

determination of $h'$ is

$$h' = h_1 P_{0,i}(\delta) + h_2 (1 - P_{0,i}(\delta)). \qquad (13)$$

The parameters calculation based the probability $p_{0,i}$ as follows,

$$p_{0,i} = P(|M| < w_i \mid |M| < k_i), i = 1,2 \qquad (14)$$

where M is non-normal random variable

$$E(T) = \frac{1}{\lambda} + (1-\delta_1)E(T_{fa}) + AATS + n'G + T^* + T^{**} . \qquad (15)$$

Determination of $E(\overline{R})$ is

$$E(\overline{R}) = \left\{h_1 - \frac{1 - e^{-\lambda h_1}(1+\lambda h_1)}{\lambda(1 - e^{-\lambda h_1})}\right\} P(A = h_1) + \left\{h_2 - \frac{1 - e^{-\lambda h_2}(1+\lambda h_2)}{\lambda(1 - e^{-\lambda h_2})}\right\} P(A = h_2), \qquad (16)$$

determination of $E(\overline{S})$ is

$$E(\overline{S}) = E(T_1)P(B = B_1) + E(T_2)P(B = B_2). \qquad (17)$$

Therefore $ECTU$ is

$$ECTU = \frac{(1/\lambda)C_0 + C_1\left[AATS + n'G + \delta_1 T^* + \delta_2 T^{**}\right] + YE(F) + W + (a+bn)\tilde{s} + (a+bn')s'}{\frac{1}{\lambda} + (1-\delta_1)E(T_{fa}) + AATS + n'G + T^* + T^{**}} . \qquad (18)$$

$$AATS = E(R) + E(S), \qquad (19)$$

determination of $E(\overline{R})$ is

$$E(\overline{R}) = \left\{h_1 - \frac{1 - e^{-\lambda h_1}(1+\lambda h_1)}{\lambda(1 - e^{-\lambda h_1})}\right\} P(A = h_1) + \left\{h_2 - \frac{1 - e^{-\lambda h_2}(1+\lambda h_2)}{\lambda(1 - e^{-\lambda h_2})}\right\} P(A = h_2), \qquad (20)$$

determination of $E(\overline{S})$ is

$$E(\overline{S}) = E(T_1)P(B = B_1) + E(T_2)P(B = B_2), \qquad (21)$$

determination of $E(\overline{F})$ is

$$E(\overline{F}) = \left(\alpha_1 P_{0,i} + \alpha_2\left(1 - P_{0,i}\right)\right). \qquad (22)$$

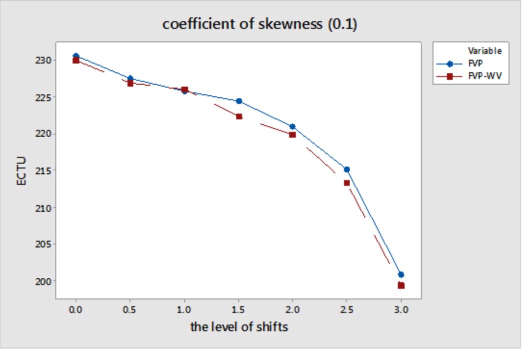
## 3. Research Results and Discussion

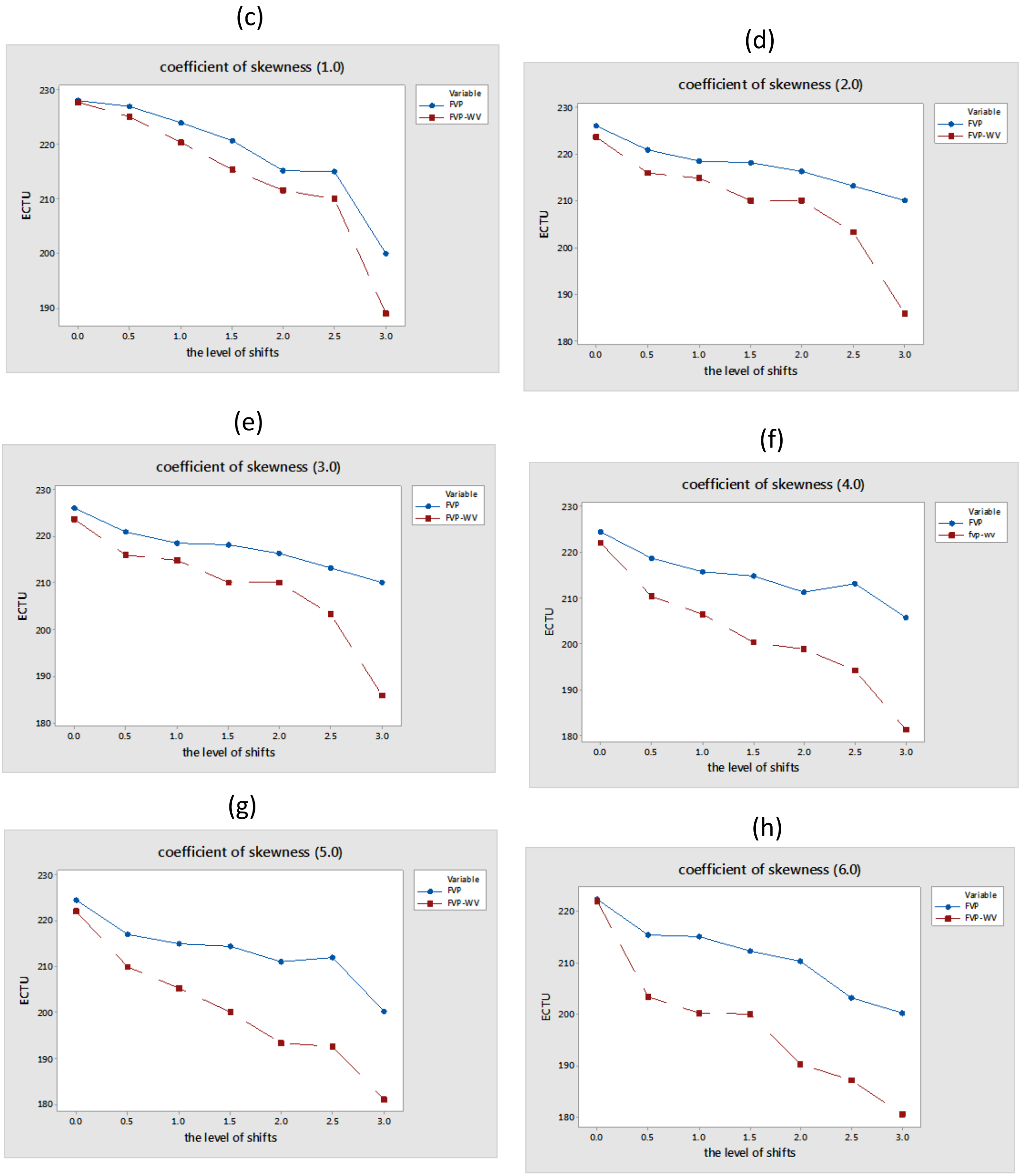
The objective of this research is to compare the efficiency of FVP and FVP-WV control charts using $\alpha - cut$ under non-normality is Weibull distributions used coefficient of skewness and $\theta = 0.1, 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9$

$\beta = 3.2219, 2.211, 1.563, 1.0, 0.7686, 0.6478, 0.5737,$
$0.5237, 0.4873, 0.4596$ and $0.4376$.

The procedure is repeated 10,000 times. From this study, it showed values of parameters ECTU to compare coefficient of control chart when mean shifts occurred in the process $\delta$ = 0.0 , 0.5 , 1.0 , 1.5 , 2.0 , 2.5 and 3.0. the result are as :

(a)





(c)

(d)

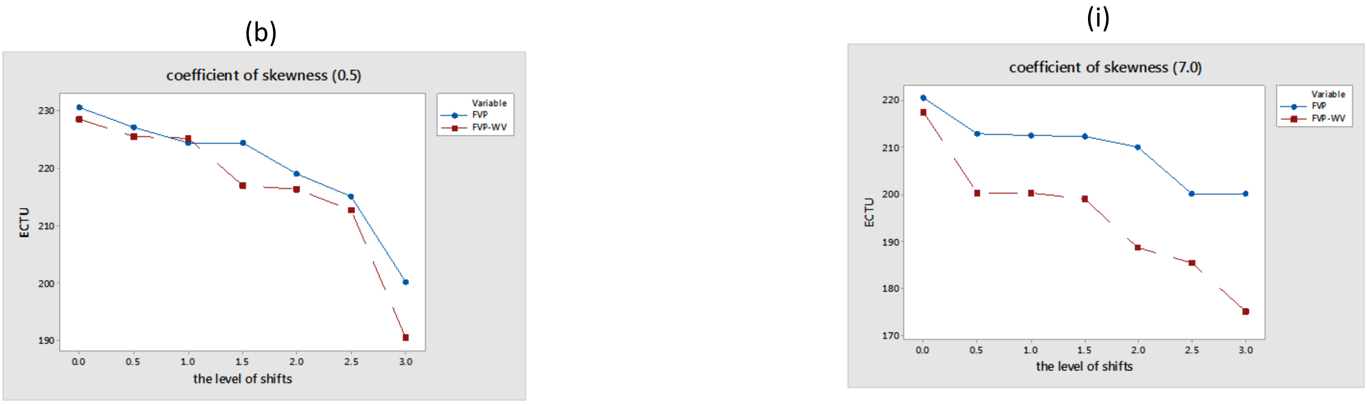
(e)

(f)

(g)

(h)

Figure 1: The economic performance of FVP and FVP – WV under Weibull distribution



(b)

(i)

(k)

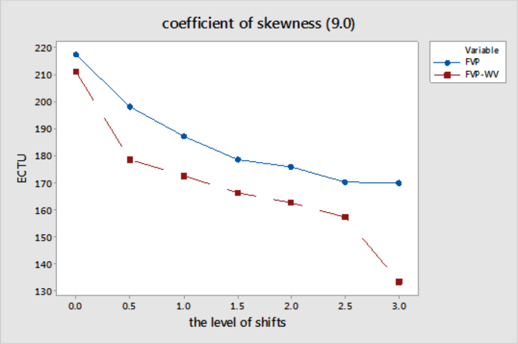


Figure 1(continous) : The economic performance of FVP and FVP – WV under Weibull distribution

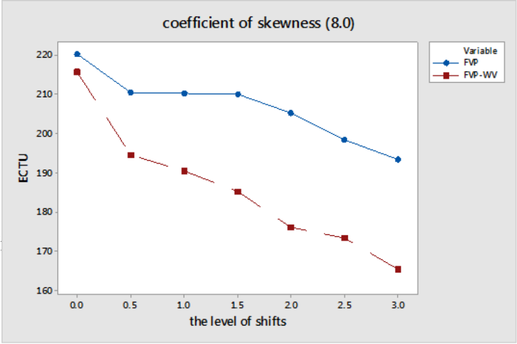
From fig 1. a , b and c the process mean shift (0 , 0.5 and 1.0 ) have a similar values.The part of process mean shift (1.5-3.0) have a more different ECTU of FVP-WV method less than FVP method.

## 4.Conclusions

The aim of this research was to introduce the economic model of fuzzy variable $\overline{X}$ control charts(FVP) and the fuzzy variable $\overline{X}$ control charts by weighted variance method (FVP-WV) using fuzzy theory ; $\alpha$ - cut and $\alpha$ - level fuzzy midrange.

Form results,If the process mean shift less than or equal to 1 then the ECTU will receive similar value but ECTU have a more different value then the process mean shift greater than 1 same as coefficient of skewness. In further study, we may study from other non-normal distributions such as Lognormal distribution, Burr's distribution and Gramma distribution, using other control charts.

(j)



## References

[1] Zadeh, L.A. “Fuzzy Sets.” Information and Control 8(1965): 338-353.

[2] Gullbay M., Kahraman C., Ruan D. “α-cuts Fuzzy Control Charts for Linguistic Data .” International Journal of Intelligent Systems 19(2004): 1173-1196.

[3] Senturk, S., Erginel N. “Development of fuzzy $\overline{X} - R$ and $\overline{X} - S$ Control Charts Using α-cuts.” Information Science 179(2009): 1542-1551.

[4] Wang, J.H., Raz T. “On The Construction of Control Charts Using Linguistic Variables.” Intelligent Journal of Production Research 28(1990): 477-487.

[5] Pongpullponsak,A., Intaramo R.Development of fuzzy extreme value theory control charts using α-cuts for skewed populations,Applied Mathematical Sciences, 6(2012), 5811 – 5834 .

[6] Yan-Kwang Chen , Economic design of control chart for non-normal data using variable sampling policy , International Journal of Production Economics, 92(2004): 61-74

[7] Pongpullponsak,A., Panthong C.Development of fuzzy variable parameters $\overline{X}$ control charts by weighted variance method using α-cuts under non-normality, International Conference on Applied Statistics. (2014) :118-125.

[8] M.S. De Magalh$a$s., E.K. Epprecht, Antonio, F.B. Costa.Economic design of a Vp $\overline{X}$ Chartss, International Journal of Production Economics 74 (2001) 191-200.

[9] Pongpullponsak,A., Suracherkeiti W. and Panthong C,The economic model of $\overline{X}$ control chart using shewhart method for skewed distributins, journal of Thailand statistician.7(2009) :81-99.

# Economic Design of the New Variable Parameters Hodges-Lehmann Estimator Control Chart

Vadhana Jayathavaj[1*] and Adisak Pongpullponsak[2*]
[1]*College of Oriental Medicine, Rangsit University, Pathumthani, 12000, Thailand, vadhana.j@rsu.ac.th*
[2]*Faculty of Science, King Mongkut's University of Technology Thonburi, Bangkok, 10140, Thailand, adisak.pon@kmutt.ac.th*

## Abstract

In the distribution free process control, the original control limits of the Hodges-Lehmann estimator (HLE) control chart (HLC) are the median of Walsh averages at predetermined locations of the Wilcoxon signed rank statistics (WSR). This study uses the new proposed control limits derive from the distribution of the HLE which are the valid Type I and Type II probabilities for the known process distributions. The 7 shifts of 0, 0.5,1, 1.5, 2, 2.5 , and 3 times of standard deviation for the symmetric standard normal distribution are included in this study. The economic cost per time unit (ECTU) of the variable parameters HLC results from the new HLC charting technique is also in the same range as in the Shewhart $\bar{X}$ chart and ECTUs are decreasing while the mean shifts are increasing. The Hodges-Lehmann estimator control chart is an alternative control chart for the process that requires the robust to outliers in measure of central tendency, but the Walsh average needs more computation steps.

*Keywords*: Economic design, Hodges-Lehmann estimator, Variable parameters control chart, Walsh average, Wilcoxon signed rank

*Corresponding Author
E-mail Address: vadhana.j@rsu.ac.th

## 1. Introduction

The parametric traditional Shewhart type $\bar{X}$ chart with the normality assumption had been developed by Dr. Walter A. Shewhart since 1920s [1]. But normality is the exception rather than the rule and was often not justified in the statistical practice especially in the field of health sciences [2] and in quality control operations which encounter many quality characteristics, the process data may not always follow the normal distribution [3]. Besides distribution-based procedures, a branch of statistic known as nonparametric statistics or distribution-free statistics is used when the population from which selected samples are not normally distributed or normality cannot be met. A nonparametric model is the one in which no assumption is made about the process distribution, the only assumption made about the observations is that they are independent identically distributed (i.i.d.) from an arbitrary continuous distribution [4].

In a distribution free inference, whether for testing or estimation, the methods are based on functions of the observation, which does not depend on the specific distribution function of the population from which the sample was drawn [5]. By definition of the distribution-free control chart in statistical process control, the in control average run length ($\text{ARL}_0$) is the same for every underlying process distribution [6].

The Hodges–Lehmann estimator (HLE) was proposed originally for estimating the location parameter of one-dimensional populations in 1963 [7]. It has been used to estimate the differences between the members of two populations, and also has been generalized from univariate populations to multivariate populations. HLE is based on the Wilcoxon signed-rank statistic. In statistical theory, it was an early example of a rank-based estimator, an important class of estimators both in nonparametric statistics and in robust statistics [8].

The "Hodges–Lehmann" statistic estimates the location parameter for a univariate population. For a dataset with $n$ measurements, its Cartesian product with itself has $n(n + 1)/2$ pairs, including the pair of each item taken twice. For each such pair, the mean is computed; finally, the median of these $n(n + 1)/2$ averages is defined to be the Hodges–Lehmann estimator of location.

The HLE is the nonparametric statistic that has very robust to outliers [9] and is widely used in clinical trial medical research [10-12], pharmacy research, drug simulation [13] and bioequivalent study [14], even in a drug annotation also has the statement "the Placebo Corrected Median Treatment Effect (Hodges-Lehmann estimate with 95% CI)", and water resources quality survey [15]. The processes that need robust to the outliers feature are also in biochemical to banking industries.

The Hodges-Lehmann estimator control chart (HLC) is different from another nonparametric statistics in hypothesis testing by associating HLE to the Wilcoxon signed rank statistic (assuming that the locations of the Walsh average have the Wilcoxon signed rank probability mass function) while another use the distribution of the test statistics directly. For example; the Sign test (the Mann-Whitney) control chart uses the binomial (the Mann-Whitney) distribution.

For a known process data distribution, the performance of the HLC will not be conformed to their associated Wilcoxon signed rank probability mass function [16].

In this study, Section 2 demonstrates the new Hodges-Lehmann estimator control limits and the economic design for the new variable parameters Hodges-Lehmann estimator control chart (VpHLC). Section 3 explains simulation steps in computing the economic cost per time unit (ECTU) of the VpHLC, and the process distributions which are represented by the standard normal with the process mean shifts in $\delta$ times of standard deviation ($\delta$ = 0.25, 0.50, 0.75, 1, 1.5, 2, 2.5, and 3) and their computational results. Section 4 presents conclusion and suggested further study.

## 2. Research Methodology

### *2.1 The Hodges-Lehmann estimator Control Charts*

### *2.1.1 The original Hodges-Lehmann estimator control charts*

The Shewhart-type chart for the Hodges–Lehmann estimator based on the Hodges–Lehmann estimator and the associated confidence interval are based on the order statistics instead of traditional measures of mean and dispersion [17].

Let $x_1, x_2, ..., x_n$ be a random sample of sample size = $n$,

$\mathrm{M} = n(n+1)/2$,

The Walsh averages $\overline{W_r}$ for $r = 1, 2, ..., \mathrm{M}$,

$$\overline{W_r} = \left(\frac{x_i + x_j}{2}\right) \qquad (1)$$

for $i \le j$, $i = 1, 2, ..., n$, $j = 1, 2, ..., n$.

The Hodges-Lehmann estimator $\overline{H}$ is defined as the median of the Walsh averages for the sample.

$$\overline{H} = \begin{cases} \overline{W}_{l+1} & \text{if M is odd} \\ (\overline{W}_l + \overline{W}_{l+1})/2 & \text{if M is even} \end{cases} \qquad (2)$$

where

$$l = \begin{cases} \text{M/2} & \text{if M is odd} \\ \text{(M-1)/2} & \text{if M is even} \end{cases}.$$

Using the connection with Wilcoxon signed-rank statistic (WSR) [18], the positions of the two ordered Walsh averages determine the control limit values. The 100(1$-\alpha$)% confidence interval for $\overline{H}$ is given by the following order statistics of the Walsh averages

$$\left(\overline{W}\left(WSR_{(\alpha/2,n)}\right), \overline{W}\left(WSR_{(1-\alpha/2,n)}\right)\right)$$

where $\overline{W}\left(WSR_{(\alpha/2,n)}\right)$ is the Walsh average at the Wilcoxon signed rank location that corresponds to the sample sizes $n$ at left-tail probability $\alpha$/2.

### *2.1.2 The new Hodges-Lehmann estimator control charts using the Hodges-Lehmann estimator probability distribution*

The new proposed Hodges-Lehmann estimator control chart is using probability distribution of the Hodges-Lehmann estimator itself in constructing the control limits. In the case of unknown process variable distribution, the control limits from Wilcoxon signed rank are the Hodges-Lehmann estimator control chart in subsection 2.1.1 which is the original implementation that the performance of the control chart also assumed to be fixed with the Wilcoxon signed rank distribution. But if the process variable has its own distribution, the Hodges-Lehmann estimator control chart should test hypothesis against its distribution vis-à-vis the Shewhart $\overline{X}$ chart tests the sample $\overline{X}$ statistic against the probability distribution of $\overline{X}$.

- Determine the Type I error probability and the probabilities in the central region, warning regions, and the action regions.
- Determine the sample size = $n$.
- Compute 100,000 Hodges-Lehmann estimators from the random numbers of sample size *n* generate by the given distribution and sort them in ascending order.
- Identify the control limits, the two action limits (the lower action limit ($-k$) and the upper action limit ($+k$) and the two warning limits (the lower warning limit ($-w$) and the upper warning limit ($+w$) from the locations correspond to the given probabilities of the 100,000 sorted Hodges-Lehmann estimators. For example, given the probability α=0.0027 (as at ±3σ in the standard normal distribution), the locations at α/2 and 1-α/2 for the lower control limits $-k$ and the upper control limits $+k$ and are 136 and 99,865 (99,865 = 100,000 - 135) [16].

### *2.2 Economic Design of the new Variable Parameters Control Charts*

The economic design of a variable parameters $\overline{X}$ chart proposed by De Magalhães, Epprecht and Costa [19] is applied to this new variable parameters Hodges-Lehmann estimator control chart. The probability distribution of Hodges-Lehmann estimator within the control limits ($\pm k$ or $\pm w$) (and also the transition matrix) can be achieved by simulation, then the economic cost per unit time (ECTU) can be solved by determining the policy-dependent input scenarios: the probability by regions (central, warning, and action), sample sizes, and sampling intervals of the regular and tight control schemes to generate target dual scheme variable parameters solutions.

**Nomenclature**

$1/\lambda$ the mean time the process stays in control which is exponentially distributed with this mean

$\delta$ shift in the process mean $\mu_0$ to $\mu_0+\delta\sigma$

$\delta_1$ indicator variable (1, if production continues during searches, 0 otherwise)

$\delta_2$ indicator variable (1, if production continues during repair, 0 otherwise)

$n_1, n_2$ sample sizes of scheme 1 and 2 (smaller and larger sample sizes)

$h_1, h_2$ sampling intervals of scheme 1 and 2 (larger and smaller sampling intervals)

$\pm k_1, \pm k_2$ action limits of scheme 1and 2 (regular and tight control)

$\pm w_1, \pm w_2$ warning limits of scheme 1and 2 (regular and tight control)

$p_0$ probability that the Hodges-Lehmann estimator falls in the central region ±w$_1$ or ±w$_2$, the scheme 1 is used for the next sample

$p_0(\delta)$ probability that the Hodges-Lehmann estimator falls in the warning region $[-k_1, -w_1)$ or $(+w_1, +k_1]$ or $[-k_2, -w_2)$ or $(+w_2, +k_2]$ then the scheme 2 is used for the next sample

$p_{ij}$ an element of transition matrix from scheme i to scheme j

$C_0$ the costs per hour due to nonconformities produced while the process is in control.

$C_1$ the costs per hour due to nonconformities produced while the process is out of control.

$Y$ the cost per false alarm

$a$ fixed cost of sampling per sample

$b$ variable cost of sampling per unit sampled.

$W$ cost of finding and repairing an assignable cause

$A$ a random variable representing the length of the interval in which the shift occurs

$R$ the time from the process shift until the first sample after shift

the time since the first sample after the shift until an alarm

T$'$ represent the time from last sample before the shift until the process shift

$T_*$ expected time to discover the assignable cause

$T_{**}$ expected time to eliminate the assignable cause

$G$ mean time to take a sample and chart it

$N$ the number of samples taken before the process shift

$s$ ( $s'$ ) the average number of samples drawn while the process is in control (out of control)

$n$ ( $n'$ ) the average sample size while the process is in control (out of control).

$h'$ the average time between samples while the process is off target

$AATS$ the adjusted average time to signal

The expected cost and expected time are as follows,

$$\begin{aligned} E(C) = {} & C_0/\lambda \\ & + C_1\left[AATS + E(T_a) + \delta_1 T_* + \delta_1 T_{**}\right] \\ & + YE(F) + W + (a+bn)s + (a+bn')s' \end{aligned} \quad . \qquad (3)$$

$$\begin{aligned} E(T) = {} & 1/\lambda + (1-\delta_1)E(T_{fa}) \\ & + AATS + n'G + T_* + T_{**} \end{aligned} \quad . \qquad (4)$$

The computational steps are as follows:

(i) Determine the parameters (policy-dependent variables) of the dual scheme variable parameters control chart: $n_1, h_1$ and the probability by region (regular scheme : small sample size with large sampling interval and small warning region), and $n_2, h_2$ and probability by region (tight scheme : larger sample size with smaller sampling interval and larger warning region).

(ii) Compute the control limits (action and warning limits) correspond to the desired probability by region for each scheme from the given process data distribution (without shift ( $\delta$ =0) or the process is in normal operation) by

- Generate 100,000 Hodges-Lehmann estimators $\overline{H}$

- Choose the Hodges-Lehmann estimator at the location correspond to the given probability by region to be the Hodges-Lehmann control limits ( $-k$ , $-w$ , $+w$ , $+k$ ).

(iii) Generate 100,000 Hodges-Lehmann estimators from the given process data distribution including the desired shift ( $\delta$ ) with the sample size $n_1$ and $n_2$ .

Finding the transition matrix, the probability of scheme 1

$p_{11}$ – number of the Hodges-Lehmann estimator plot within action limits of scheme 1 divided by 100,000

$p_{12}$ – number of the Hodges-Lehmann estimator plot within warning region of scheme 1 divided by 100,000

and also for $p_{21}$ and $p_{22}$ of scheme 2

(iv) Compute $n'$ from $n_1$, $n_2$, and $p_0$

Compute $h'$ from $p_0(\delta)$

$p_0$ derives from $p_{11}$ and $p_{12}$ when $\delta = 0$ which

$$p_0 \begin{aligned} &= p\left(-w_1 < \overline{H} < w_1 \mid -k_1 < \overline{H} < k_1\right) \\ &= p\left(-w_2 < \overline{H} < w_2 \mid -k_2 < \overline{H} < k_2\right) \end{aligned} \quad \text{and}$$

$p_0(\delta)$ derives from $p_{11}$ and $p_{12}$ when $\delta > 0$ which

$$p_0(\delta) = p(-w_i < \overline{H}(x(\delta)) < w_i \\ \left| -k_i < \overline{H}(x(\delta)) < k_i \right)$$

$\overline{H}(x(\delta))$ is the HLE derived from the shifted process.

(v) Compute $ATTS = E(R) + E(S)$ by

Calculate $E(R)$ from $\lambda, h_1, h_2$

and $E(S)$ from $\lambda, h_1, h_2, p_0, p_{11}, p_{12}, p_{21}, p_{22}$

(vi) Compute $s'$ from $AATS, n', h', G, \delta_1, \delta_2, T_*, T_{**}$

(vii) Compute $ECTU$ by substituting the input variables into $E(C)$ and $E(T)$ in equation (3- 4).

All models in this study are designed in MATLAB using custom scripts [20] and MINITAB [21].

## 3. Research Results and Discussion

*3.1 The economic design scenarios*

*3.1.1 Control schemes*

The regular scheme (R) has the probabilities in central, warning, and action region follows the traditional parametric Shewhart $\overline{X}$ chart which derived from ±3σ and ±2σ for action limits and warning limits respectively. The probabilities for central, warning, and action region are 0.97725, 0.02005, and 0.00270 respectively. The three tight schemes T5, T10, and T15 are as follows,

T5 – Reduce in control region (central + warning) by 5% from the regular scheme, the action region increased by 18.10 times from 0.0027 to 0.0516.

T10 - Reduce in control region by 10%, increase warning by 5%, then central region decreased by -10%, from the regular scheme, the action region is increased by 36.94 times from 0.0027 to 0.1024.

T15 – Reduce in control region by 15% %, increase warning by 10%, then central region decreased by -16%, from the regular scheme, the action region is increased by 55.41 times from 0.0027 to 0.1523.

The detailed probabilities classified by regions, the location of Hodges-Lehmann estimator corresponding to the given probabilities are shown in Table 1.

Table 1 Probabilities classified by regions and locations for the new control limits

| | Region probability | | | Location at<br>100000 HL Estimators | | | |
|---|---|---|---|---|---|---|---|
| Scheme | Central | Warning | Action | -k | -w | +w | +k |
| R-Regular | 0.9773 | 0.0201 | 0.0027 | 135 | 1138 | 98863 | 99865 |
| T5-5% tight | 0.9294 | 0.0191 | 0.0516 | 2579 | 3532 | 96468 | 97421 |
| change from R | -5% | -5% | 1810% | | | | |
| T10-10% tight | 0.8765 | 0.0211 | 0.1024 | 5122 | 6174 | 93826 | 94879 |
| change from R | -10% | 5% | 3694% | | | | |
| T15-15% tight | 0.8257 | 0.0221 | 0.1523 | 7615 | 8718 | 91283 | 92385 |
| change from R | -16% | 10% | 5541% | | | | |

*3.1.2 The underlying process distribution*

The process data in this study are the standard normal distribution ( $N(0,1)$ ).

Assumes that $X$ to be the normal process variable with mean $\mu$ and standard deviation $\sigma$, $X \sim N(\mu, \sigma)$, the change variable $z = (x - \mu) / \sigma$ converts $X$ from $N(\mu, \sigma)$ random variable into $N(0,1)$ random variable.

*3.1.3 The process parameters*

The parameters are arranged as follows,

- 7 shifts in times of standard deviation,

- 6 combinations of probability by region or control limit (3 from $n_1$=5, $n_2$=5,10,15 and 3 from $n_1$=10, $n_2$=10,15,20) as shown in Table 2,

- 9 sampling intervals by fixed $h_1$ at 1.00 and vary $h_2$=0.75, 0.50, 0.25, and 0.10, and fixed $h_2$ at 0.10 and vary $h_1$=2.00,1.50,0.75,0.50, and 0.25 as shown in Table 3.

Table 2 Six dual scheme combinations of variable parameters Hodges-Lehmann estimator control charts

| Regular | $n_1$ | Tight | $n_2$ |
|---|---|---|---|
| R | 5 | T5 | 5 |
| R | 5 | T10 | 10 |
| R | 5 | T15 | 15 |
| R | 10 | T5 | 10 |
| R | 10 | T10 | 15 |
| R | 10 | T15 | 20 |

Total combination of variable parameters control charts to compute economic cost per time unit (ECTU) are 6 (Vp combinations) x 9 (dual sampling intervals) = 54 for each shift.

Table 3 Nine sampling intervals for variable parameters Hodges-Lehmann estimator control charts

| $h_1$ | $h_2$ |
|---|---|
| 1.00 | 0.75 |
| 1.00 | 0.50 |
| 1.00 | 0.25 |
| 1.00 | 0.10 |
| 1.50 | 0.10 |
| 2.00 | 0.10 |
| 0.75 | 0.10 |
| 0.50 | 0.10 |
| 0.25 | 0.10 |

### *3.1.4 Costs and Times*

The input data come from foundry operations, where sample of molten iron are taken to monitor the carbon-silicate content periodically from Lorenzen & Vance [22].

$G=T_*=T_0=5/60$ $T_{**}=45/60$

$1/\lambda=50$

$C_0=\$114.24/\text{hour}$ $C_1=\$949.20/\text{hour}$

$Y=W=\$997.40$ $a=0$; $b=\$4.22$

$\delta_1=1$; $\delta_2=0$.

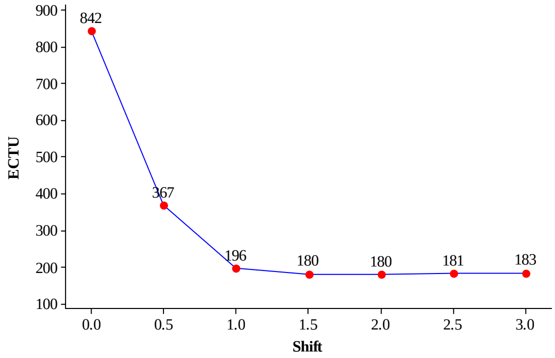


Figure 1: The economic cost per time unit of variable parameters Hodges-Lehmann estimator control chart with N(0,1) using (n1,n2,h1,h2)=(5,10,1,0.1) and control limits=Regular and 5% tight

### *3.2 The economic design results*

The numerical example for the economic cost per time unit ( *ECTU* ) of new variable parameters Hodges-Lehmann estimator control chart with $N(0,1)$ using $(n_1,n_2,h_1,h_2)=(5,10,1,0.1)$ and control limits=Regular (R) and 5% tight (T5) are shown in Table 4 and Figure 1. The ECTUs of R &T5 are decreasing while the shifts are increasing. The ECTU values are in the same range from the original study [19].

Table 4 Numerical examples for the economic design of variable parameters Hodges-Lehmann estimator control charts for Standard Normal Distribution N(0,1)

| | **Regular** | | | | | | **5% Tight** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Shift | $n_1$ | $h_1$ | $-k_1$ | $-w_1$ | $w_1$ | $k_1$ | $n_2$ | $h_2$ | $-k_2$ | $-w_2$ | $w_2$ | $k_2$ | **ECTU** |
| 0.0 | **5** | 1.00 | -1.3942 | -1.0606 | 1.0589 | 1.3976 | **10** | 0.10 | -0.6387 | -0.5940 | 0.5915 | 0.6392 | 841.81 |
| 0.5 | **5** | 1.00 | -1.4047 | -1.0718 | 1.0541 | 1.4028 | **10** | 0.10 | -0.6330 | -0.5888 | 0.5949 | 0.6404 | 366.70 |
| 1.0 | **5** | 1.00 | -1.3934 | -1.0625 | 1.0588 | 1.3976 | **10** | 0.10 | -0.6365 | -0.5896 | 0.5923 | 0.6360 | 196.04 |
| 1.5 | **5** | 1.00 | -1.3870 | -1.0421 | 1.0565 | 1.3901 | **10** | 0.10 | -0.6366 | -0.5901 | 0.5934 | 0.6390 | 180.45 |
| 2.0 | **5** | 1.00 | -1.4086 | -1.0610 | 1.0605 | 1.4044 | **10** | 0.10 | -0.6412 | -0.5946 | 0.5939 | 0.6388 | 180.07 |
| 2.5 | **5** | 1.00 | -1.4014 | -1.0604 | 1.0550 | 1.3722 | **10** | 0.10 | -0.6415 | -0.5936 | 0.5916 | 0.6380 | 181.05 |
| 3.0 | **5** | 1.00 | -1.4074 | -1.0631 | 1.0611 | 1.3823 | **10** | 0.10 | -0.6386 | -0.5919 | 0.5913 | 0.6367 | 183.18 |

### *3.3 The minimum economic design for the standard normal distribution*

The minimum economic cost per time unit (ECTU) from the total 54 combination of variable parameters control charts for each shift (δ) for the standard normal distribution is shown in Table 5 and Figure 2.

## 4. Conclusion

For the economic design, the use of Hodges-Lehmann estimator probability distribution in finding the control limits through simulation, this method is able to compute the valid process probability when the shift occurs (both Type I error and Type II error, and the transition matrix in variable parameters control chart), this lead to make economic cost per time unit (ECTU) computation possible.

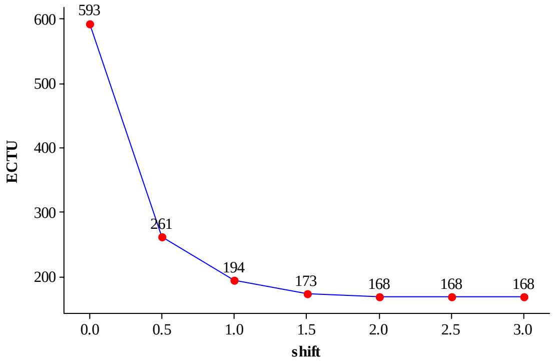


Figure 2 Minimum economic cost per time unit of variable parameters Hodges-Lehmann estimator control chart for each shift of N(0,1)

Table 5 Minimum economic cost per time unit of Variable Parameters Hodges-Lehmann estimator control charts for the standard normal distribution

| Shift | n1 | h1 | n2 | h2 | VP | ECTU |
|---|---|---|---|---|---|---|
| 0.00 | 5 | 0.25 | 5 | 0.10 | R&T15 | 593.50 |
| 0.50 | 10 | 1.00 | 20 | 0.10 | R&T15 | 261.86 |
| 1.00 | 5 | 1.00 | 10 | 0.10 | R&T15 | 194.10 |
| 1.50 | 5 | 1.50 | 5 | 0.10 | R&T10 | 173.10 |
| 2.00 | 5 | 2.00 | 5 | 0.10 | R&T15 | 168.80 |
| 2.50 | 5 | 2.00 | 5 | 0.10 | R&T5 | 168.60 |
| 3.00 | 5 | 1.50 | 5 | 0.10 | R&T15 | 168.90 |

The minimum ECTU of variable parameters Hodges-Lehmann estimator control chart from the predetermined deterministic policy-dependent inputs scenarios – the probability by control regions (the control limits – warning and action limits), the sample sizes, the sampling intervals, when the shift is increasing from 0.5, 1, 1.5, 2, 2.5 and 3 times of standard deviation, the ECTU's are decreasing.

Even though the Hodges-Lehmann estimator has more computations (the median of Walsh averages has many more computations than another tests), the robust to outliers feature will still persuade the quality controllers in implementing the Hodges-Lehmann estimator to their operations

The asymmetrical shape of underlying process distribution and the optimal economic design are the subjects for further study.

## Acknowledgements

This work was supported by Rangsit University.

## References

[1] Montgomery DC. Statistical Quality Control A Modern Introduction. 7th ed. John Wiley & Sons Singapore Pte, Ltd; 2013.

[2] Neuhäuser M. Nonparametric Statistical Tests : A computational Approach. Florida: CRC Press; 2012.

[3] Das N. A comparison study of three non-parametric control charts to detect shift in location parameters. The International Journal of Advanced Manufacturing Technology. 2009; 41: 799-807.

[4] Bluman AG. Elementary Statistics: a step by step approach. 3rd ed. McGraw Hill Companies, Inc; 1998.

[5] Gibbons JD. Nonparametric Statistical Inference. Tokyo: McGraw-Hill Kogakusha, ltd; 1971.

[6] Bakir ST. A distribution-free Shewhart quality control chart based on signed-ranks. Quality Engineering. 2004; 16(4):613-623.

[7] Hodges JL, Lehmann EL. Estimates of Location Based on Rank Tests. The Annals of Mathematical Statistics. 1963; 34(2): 598-611.

[8] Lehmann EL. Nonparametrics: Statistical methods based on ranks. New York: Springer; 2006.

[9] Duchnowski R. Hodges_Lehmann estimates in deformation analyses. Journal of Geodynamics. 2013; 87: 873-884.

[10] Arroliga AC, Pollard CL, Wilde CD, Pellizzari SJ, Chebbo A, Song J, Ordner J, Cormier S, Meyer T. Reduction in the incidence of ventilator-associated pneumonia: a multidisciplinary approach. Respiratory Care. 2012; 57(5): 688-696.

[11] Shuster JJ, Theriaque DW, Ilfeld BM. Applying Hodges-Lehmann scale parameter estimates to hospital discharge times. Clinical Trials. 2008; 5(6): 631-634.

[12] Mirko S, Reinmar K, Martin K, Abolmaali AN, Jennifer F,Thomas JV. Vascular guide wire navigation with a magnetic guidance system: eperimantal results in a Phantom. Radiology. 2004; 232(2): 475-78.

[13] Allen IE, Seanman CA. Simulation can forecast probable success in clinical trials. Quality Progress. February 2009; February: 60-63.

[14] Florin EM, Ion CG, Roxana S, Constantin M. Estimation of therapeutic equivalence statistical methods for ALGOPIRIN tablets versus EXCEDRIN analgesic formulations. Farmacia. 2012; 60(2): 227-239.

[15] Langman JB, Anderholm SK. Effects of reservior installation, san Juan-Chama project water and reservior operations on streamflow and water qaulity in the Rio Chama and rio Grande, northern and central New Mexico, 1938-2000. Scientific invetigation report 2004-5188 (U.S. Geological Survey, Reston, Virginia). 2004.

[16] Pongpullponsak A, Jayathavaj V. The new Hodges-Lehmann estimator control charting technique for the known process distributions, Proceeding of the The International Conference on Applied Statistics 2014; 2014 May 21-24, 2014; Khon Kaen, Thailand. 2014. p. 257-268.

[17] Alloway JA, Raghavachari M. Control Chart based on Hodges-Lehmann Estimator. Journal of Quality Technology. 1991; 23: 336-347.

[18] Beaumont GP, Knowles JD. Statistical tests An introduction with MINITAB commentary.

Hertfordshire: Prentice Hall international (UK) Limited; 1996.

[19] De Magalhães MS, Epprecht EK, Costa AF.B. Economic deisgn of a Vp $\overline{X}$ chart. International Journal of Production Economics. 2001; 74: 191-200.

[20] The Math Works. MATLAB 7.6.0 (R2009a). License Number 350306, February 12, 2009.

[21] MINITAB Thailand. "MINITAB 16 Order Number 100004968850, Single License." February 2, 2010.

[22] Lorenzen TJ, Vance LC. The economic design of control charts: A unified approach. Technometrics, 1986; 28: 3-10.

# The role of driver's attitude to trucking firm performance

Le Thi Ngoc Lan[1], Chawalit Jeenanunta[1*], Nattharika Rittippant[2], Pornpimol Chongphaisal[3], Tomohiro Machikita[4], Yasushi Ueki[5], Masatsugu Tsuji[6]
[1]*Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, Thailand*
[2]*Institute of Developing Economies, Chiba, Japan*
[3]*Economic Research Institute for ASEAN and East Asia, Jakarta, Indonesia*
[4]*Graduate School of Applied Informatics, University of Hyogo, Hyogo, Japan*

## Abstract

The third party logistics services provider (3PL) companies continuously develop and take important role in supply chain system. They need to keep improving their firm performance to stay competitive. Literature shows many evidences about the link of employee's attitude and firm performance. In this research paper, we investigate the impact of driver's attitude on trucking companies' performance. We aimed to answer two questions: Does driver's attitude influence to the firm operational performance in term of transportation efficiency? And does driver's attitude mediates the relationship between transportation efficiency and profitability? In this research, the survey is constructed consisting of three sections, the demographic of the firm, the driver's attitude and the firm operational performance and profitability. Survey data collected from top management of 86 trucking companies in Thailand during December 2013 to January 2014. The interesting results are consistent with literature when driver's attitude positive impact on transportation efficiency and it's also mediate the relationship between transportation efficiency and firm profitability.

*Keywords*: *driver's attitude*: *logistics services provider*: *trucking firm*: *performance*

*Corresponding Author
E-mail Address: chawalit@siit.tu.ac.th

## 1. Introduction

The continuous enhancing of competition in market nowadays is spurring organizations to make their performance better in order to keep distinction with competitors. This challenge effects on all kind of organization includes manufacturers, service suppliers or even outsourcing suppliers. To deal with that challenge, organization need to delivering greater efficiency, quality and more flexibility of services. In other words, company need to improve their operational performance which makes increasing of their financial performance.

There are one way that companies nowadays prefer to use is to outsource their extra work to others outsourcing firm like logistics service providers, to concentrates on their core competences. This trend open a new opportunity to logistics company to get more customer but they also faces with more competitive in order to get long term commitment or long term contract with customer. Hence, again logistics service providers also have to join the race of all other organizations, enhancing their performance.

Literature shows evidences that quality of work and employee engagement can influence on outcome of organizations, working quality leads to product improvement [1]. Employees' attitude, behaviors, and employees' justice perceptions connect to the work performance and hence outcome of firm [1].

This research aim to find whether driver's quality which is in term of driver's attitude about their work influences on trucking firm performance or not. So we collect related information from trucking company in logistics association in Thailand who have responsibility to supply trucking service from consignor to consignee for analyzing.

## 2. Literature review

### 2.1 Logistics organization performance

There are many different performance definition based on company activities area. In general, firm performance was classified into two categories: operational performance which measures the outcomes of organization's process such as reliability, speed of delivery and quality of service and financial performance which measures gross profit margin, return on sales, operating profit margin return on assets, return on equity, and accounts receivable turnover, current ratio, and debt ratio, return on investment [2].

For logistics services point of view, operational performance was measured by delivery efficiency (includes: expedited shipment, short delivery lead-time, on time and accurately delivery), quality (includes: customer satisfaction rate, lower customer complain, undamaged delivery), flexibility (includes: accommodate special or non-routine requests, handle unexpected events, provide quicker response to customers), cost (includes: low overall operating cost, the rate of utilization of facilities/equipment/ manpower), innovation (includes: aggressiveness in increasing the value-added content of services, aggressiveness in the reduction of order cycle time, provide new and better services) [2-4]. Financial performance was measured differently even though in logistics sector, it depend on purpose of company or research. In general, financial performance measured using dimensions above.

### *2.2 Employee's attitude, driver's attitude and firm performance*

This paper aims to examine the impact of driver's quality in term of their attitude and their adoption attitude in work to firm performance of trucking firm in Thailand.

Prior research found many evidences about the role of employee to their firm performance. The research of Ellinger, Ketchen [5] proposed that employee with high working skilled influence on both employee and organizational performance.

The research of Wang, Tsui [6] supposed that employees' attitude can influence on firm performance by increasing their own work performance.

A numerous research found that there are positive relationship between general workplace attitudes and service intentions, customer perceptions and individual performance outcomes. The way employees treat can positive effect on their action, then effect on their performance [7-10]. Positive relations between employee attitudes, organizational unit performance, customer satisfaction, and turnover was found by RYAN, Schmit [11].

In particularly with trucking activities, there are many finding argued that driver's attitude impacts on performance of drivers in their driving process. Mirzaei, Hafezi-Nejad [12] said that drivers' knowledge, attitude and practice regarding traffic regulations effect on road traffic crashes. Drivers with higher knowledge, safer attitude, and safer practice were associated with a decreased number of road traffic crashes. The increasing of education and awareness in relation to safe driving behavior, road rule can improve safety [13]. Driver with higher driver anger, sensation seeking, urgency, and with a lack of premeditation and perseverance in daily activities have riskier driving acts [14]. These evidences argued that drivers' attitude do have influence on driving risk, and traffic crashes, then influence on driving performance.

## 3. Methodology

### *3.1 Measurement*

Based on the literature review and results from interviewing 10 experts who are top management on logistics firms in Thailand and experts working on logistics research center in Thailand and Japan, we measure transportation efficiency for trucking firm which is the improvement of six dimensions: transportation volume per truck, load efficiency (actual loading capacity per maximum load capacity), laden miles (driving distance carrying cargo / total driving distance), usage efficiency (Total days trucks are actually utilized / Total days trucks are available), delivery in full on time, and number of damages or losses of cargos. For financial performance, we measure by using the improvement rate of profitability. For the driver's attitude measurement, we measure the level of driver's attitude on five following categories: Consciousness about safe driving, compliance with road traffic laws and other laws related to trucking industry, compliance with your firm's rules/manuals for driving, understanding about customer requirements, and implementation of "5S". All of these will be measured by using the 5-point Likert scale. The type of goods that trucks handle may effect on the firm transportation efficiency and firm profitability, hence we use "type of goods trucks handle" to control the model.

### *3.2 Sample*

The trucking company list is from Siam List Database Marketing Company which contains 13,418 logistics companies and the additional 9,607 logistics companies are from the ministry of transportation. We randomly selected 200 companies from the list and send the mail survey. In total we received 86 responses where 14 responses via mail, 25 responses via phone, and 47 responses via in-person survey

### *3.3 Model and hypothesis*

We assume that operational performance positive influences on financial performance in trucking firm, and driver's quality positive influences on operational performance. Hence, we set three hypothesis: H1+ Driver's attitude positive influences on transportation efficiency, H2+

driver's attitude positive influence on profitability, and H3+ driver's attitude mediates the positive relationship between transportation efficiency and profitability.

The figure 1 shows the total image of this paper.

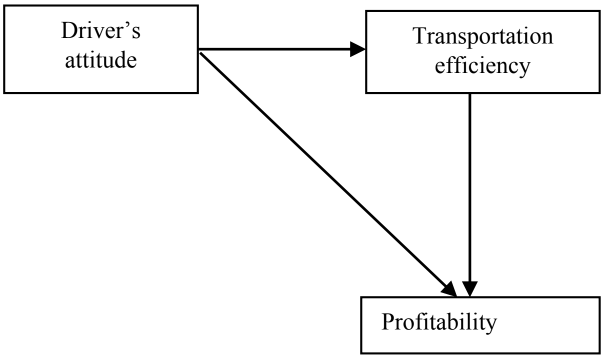


Figure 1: Research model

*3.4 Statistical method*

The first step, we use reliability test to warrant the appropriate of item, reliability is usually expressed on the basis of the Cronbach's alpha coefficient (reliability coefficients). Levels of 0.70 or more are generally accepted as representing good reliability [15]. After that, we use correlation and simple regression to find the relationship between factors.

The pairwise approach was used for addressing missing data in this study in order to maximize the use of valid data [2].

## 4. Results and analysis

*4.1 Data description*

Form of firms in our survey consist individual proprietor, juristic partnership and company limited form. Number of each type of firm is quite equal (35, 27 and 22, respectively). Most of firm in the survey have capital less than 25 million Thai Baht with 81.4%, only 5.9% companies have capital over 25 million Baht. Almost companies in this survey have number of employees less than 50 including driver with 83.7%, the rest has number of employees over 50 with 15.1%. Top management education of these companies was divided into two equal group with 48.8% of them have bachelor degree or less than it, and 46.5% of them have master degree or over. The table 1 below shows in detail the demographic of firms which are interviewed in our survey.

Table 1: Demographic of respondent

| | | Frequency | % |
|---|---|---|---|
| Form of legal organization | Individual proprietor | 35 | 40.7 |
| | Juristic partnership | 27 | 31.4 |
| | Company limited, Public company limited | 22 | 25.6 |
| Capital (THB) | | | |
| | 5 million or less | 37 | 43.0 |
| | 6-25 million | 33 | 38.4 |
| | Upper 25 million | 5 | 5.9 |
| Number of employee | | | |
| | Less than 25 | 41 | 47.7 |
| | From 25-50 | 31 | 36.0 |
| | upper 50 | 13 | 15.1 |
| Education of top manager | | | |
| | Bachelor degree and less | 42 | 48.8 |
| | Master degree and upper | 40 | 46.5 |

*4.2 Reliability test*

The instrument from driver's attitude group and transportation efficiency group have high level of reliability with Cronbach's alpha quite high 0.895 and 0.83, respectively. Hence, the instrument for these two group are appropriate for analyzing. The table 2 shows in detail results of reliability test.

Table 2: Reliability test result

| | Number of item | Cronbach's alpha |
|---|---|---|
| Driver's attitude | 5 | 0.895 |
| Transportation efficiency | 6 | 0.830 |

After doing factor analysis, 6 item of transportation efficiency was eliminated in one group. Factor loading of each item is over 0.5, then all of them can representative for the factor [16]. The Kaiser – Meyer – Olkin (KMO) measures 0.738, exceeding the recommended figure of 0.6, so the factor analysis can be used. Detailed result was showed in table 3 below.

Table 3: Factor analysis for transportation efficiency-factor loading and Kaiser-Meyer-Olkin test

| Transportation efficiency | Factor loading |
|---|---|
| | |
| 1.Transportation volume per truck | 0.683 |
| 2. Load efficiency | 0.712 |
| 3. Laden miles | 0.813 |
| 4. Usage efficiency | 0.848 |
| 5. Delivery in Full on Time (DIFOT) | 0.783 |
| 6. Number of damages or losses of cargos | 0.588 |
| **Kaiser-Meyer-Olkin Measure of Sampling Adequacy** | **0.738** |

Similarly, the table 4 shows in detail the factor analysis of driver's attitude. Five items were eliminated into one group, each item represent highly for driver's attitude factor with factor loading over 0.8. KMO measures 0.822 ensure that this factor analysis can be good for using.

Table 4: Factor analysis for driver's attitude measurement - factor loading and Kaiser-Meyer-Olkin test

| Driver's attitude | Factor loading |
|---|---|
| | |
| 1. Consciousness about safe driving | 0.808 |
| 2. Compliance with road traffic laws and other laws related to trucking industry | 0.912 |
| 3. Compliance with your firm's rules/manuals for driving | 0.870 |
| 4. Understanding about your consignees' businesses and needs for logistics | 0.833 |
| 5. Implementation of so called "5S" | 0.803 |
| **Kaiser-Meyer-Olkin Measure of Sampling Adequacy** | **0.822** |

*4.3 Statistics descriptive and inter-correlation*

The table 5 below (appendix) shows information about descriptive of data collected in driver's attitude group and operational performance with its mean and standard deviation, this table also shows the inter-correlation between factors in these two groups.

The first fifth factors belong to driver's attitude group which have mean fluctuate from 3.41 to 3.77. Its means lie between "average scales" and "above average scales" but near to the "above average scale". The after sixth factors have means fluctuate from 3.09 to 3.22, its means also lie between "average scales" and "above average scales" but near to the "average scale".

The correlation between driver's attitude and operational performance is showed in bold number. We can see that "transportation volume per truck" and "load efficiency" have high correlation with all factors of driver's quality, "usage efficiency" has moderate correlation with the first three factors of driver's quality, and "number of damages" or losses of cargos only correlates with the first factor of driver's attitude that is "consciousness about safe driving". Whereas the factors "laden mile" and "delivery in full on time" of operational performance do not correlate with driver's attitude.

*4.4 Test of hypothesis*

Hypothesis 1 states that driver's attitude positive influences on firm operational performance. Table 6 below shows the results of regression test. We use variable "type of goods truck handle" as control variable of this regression test. The result indicates that driver's attitude has positive influence on transportation efficiency (β = 0.377, p = 0.001). Hence, the hypothesis 1 is supported.

Table 6: Regression between driver's attitude and transportation efficiency

| Variables | Transportation efficiency | | |
|---|---|---|---|
| | B | SE | β |
| Driver's attitude | $0.377^{***}$ | 0.123 | $0.377^{***}$ |
| R2 | 0.45 | | |
| Change in R2 | 0.089 | | |
| F change | $9.410^{***}$ | | |
| Adjusted R square | 0.260 | | |
| ANOVA (F) | $2.369^{***}$ | | |

Note: *p<0.05; **p<0.01, ***p<0.001

Hypothesis 2 states that driver's attitude and profitability have positive relationship, but the result in table 7 below give us a very small confidence level (greater than 0.05). Hence, we don't have enough evidence to support this hypothesis.

Table 7: Regression between driver’s attitude and profitability

| Variables | Profitability | | |
|---|---|---|---|
| | B | SE | β |
| Driver's attitude | 0.92 | 0.071 | 0.138 |
| R2 | 0.591 | | |
| Change in R2 | 0.012 | | |
| F change | 1.681 | | |
| Adjusted R square | 0.449 | | |
| ANOVA (F) | 4.182 | | |

Note: *p<0.05; **p<0.01, ***p<0.001

Hypothesis 3 states that driver’s attitude mediates the positive relationship between transportation efficiency and profitability. The two stage regression test’s result is showed in the table 8 below. In the first stage, only driver’s attitude was put in the regression and the second stage, both of driver’s attitude and transportation efficiency was put in. Profitability was put in the regression test like a dependent variable. Based on the result, driver’s attitude do mediate the relationship between transportation efficiency and profitability (β = 0.336, p<0.001). Hence, hypothesis 3 was supported.

Table 8: Two stage regression of driver’s attitude, transportation efficiency and profitability

| | Profitability | | | | | |
|---|---|---|---|---|---|---|
| | Step 1 | | | Step 2 | | |
| | B | SE | β | B | SE | β |
| Driver's attitude | 0.920 | 0.071 | 0.138 | 0.920 | 0.071 | 0.138 |
| Transportation efficiency | | | | $0.226^{***}$ | 0.071 | $0.336^{***}$ |
| R2 | 0.591 | | | 0.653 | | |
| Change in R2 | 0.012 | | | 0.062 | | |
| F change | 1.681 | | | $10.196^{***}$ | | |
| Adjusted R square | 0.449 | | | 0.525 | | |
| ANOVA (F) | 4.182 | | | $5.100^{***}$ | | |

Note: *p<0.05; **p<0.01, ***p<0.001

## 5. Discussion

The primary purpose of this paper is to examine the role of driver’s attitude on firm performance in term of transportation efficiency and firm profitability. The results from analysis part indicates that driver’s attitude does have positive and directly impact on firm transportation efficiency and it also has indirectly impact on firm profitability for trucking companies. This result consistent with the findings of all authors we have mentioned in the literature review part. Our findings suggest that for trucking companies, driver’s attitude is an important thing that they have to consider in order to improve their performance. In the detail of the relationship between driver’s attitude and transportation efficiency, the driver’s attitude in term of safe driving, compliance with road traffic law and implementing “5S” have most impact on transportation efficiency especially on the number of damaged or loss of cargo, and usage efficiency of firm. In the other hand, these aspects of transportation efficiency also have the most impact on firm profitability. Hence, trucking firm need to consider more about these driver’s attitude.

This paper limit the driver’s attitude on five aspects and only use “the type of goods that truck handle” to control the effect of others variables on firm transportation efficiency and firm profitability. It may limit the result of analysis. Later papers can extend other driver’s attitude aspects, firm performance also, and put more control variables in order to get better results.

**Acknowledgements**

This paper was funded by Institute of Developing Economies and it was partially supported by Center of Excellence in Intelligent Informatics, Speech and Language Technology and Service Innovation (CILS), Thammasat University.

### References

[1]. Mathew J., Ogbonna E., Harris, Culture L.C. Employee work outcomes and performance: An empirical analysis of Indian software firms. Journal of World Business. 2012; 47(2): 194-203.

[2]. Liu C.-L., Lyons A.C. An analysis of third-party logistics performance and service provision. Transportation Research Part E: Logistics and Transportation Review. 2011; 47(4): 547-570.

[3]. Brooks M.R. Performance evaluation by North American carriers. Transport Reviews. 1999; 19(1): 1-11.

[4]. Van Hoek R.I. The contribution of performance measurement to the expansion of third party logistics alliances in the supply chain. International Journal of Operations & Production Management. 2001; 21(1/2): 15-29.

[5]. Ellinger A.E., et al. Market orientation, employee development practices, and performance in logistics service provider firms. Industrial Marketing Management. 2008; 37(4): 353-366.

[6]. Wang H., Tsui A.S., Xin K.R. CEO leadership behaviors, organizational performance, and employees' attitudes. The Leadership Quarterly. 2011; 22(1): 92-105.

[7]. Harter J.K., Schmidt F.L., Hayes T.L. Business-unit-level relationship between employee satisfaction, employee engagement, and business outcomes: a meta-analysis. Journal of applied psychology. 2002; 87(2): 268.

[8]. Denison D.R., Corporate culture and organizational effectiveness. John Wiley & Sons, 1990.

[9]. Ostroff C. The relationship between satisfaction, attitudes, and performance: An organizational level analysis. Journal of applied psychology. 1992; 77(6): p. 963.

[10]. Schneider B., et al. Which comes first: employee attitudes or organizational financial and market performance? Journal of Applied Psychology. 2003; 88(5): 836.

[11]. RYAN A., Schmit M.J., Johnson R., Attitudes and effectiveness: Examining relations at an organizational level. Personnel psychology. 1996; 49(4): 853-882.

[12]. Mirzaei, R., et al., Dominant role of drivers' attitude in prevention of road traffic crashes: A study on knowledge, attitude, and practice of drivers in Iran. Accident Analysis & Prevention. 2014; 66: 36-42.

[13]. Johnson, M., et al. Safety in numbers? Investigating Australian driver behaviour, knowledge and attitudes towards cyclists. Accident Analysis & Prevention. 2014; 70: 148-154.

[14]. Bachoo S., Bhagwanjee A., Govender K., The influence of anger, impulsivity, sensation seeking and driver attitudes on risky driving behaviour among post-graduate university students in Durban, South Africa. Accident Analysis & Prevention. 2013; 55: 67-76.

[15]. Hair, J.F. Multivariate data analysis. Upper Saddle River NJ [etc.]: Pearson Prentice Hall; 2006.

[16]. Field A., Discovering statistics using SPSS. Sage publications; 2009.

## Appendix

Table 5: Descriptive and inter-correlation between driver’s attitude, transportation efficiency and profitability

| | | Mean | SD | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **I.** | **Driver's attitude** | | | | | | | | | | | | | | |
| 1 | Consciousness about safe driving | 3.77 | 0.965 | 1 | | | | | | | | | | | |
| 2 | Compliance with road traffic laws and other laws related to trucking industry | 3.62 | 0.874 | $0.781^{**}$ | 1 | | | | | | | | | | |
| 3 | Compliance with your firm’s rules/manuals for driving | 3.67 | 0.908 | $0.623^{**}$ | $0.766^{**}$ | 1 | | | | | | | | | |
| 4 | Understanding about customer requirement | 3.43 | 0.999 | $0.508^{**}$ | $0.621^{**}$ | $0.670^{**}$ | 1 | | | | | | | | |
| 5 | Implementation of so called “5S” | 3.41 | 1.11 | $0.491^{**}$ | $0.648^{**}$ | $0.589^{**}$ | $0.687^{**}$ | 1 | | | | | | | |
| **II.** | **Transportation efficiency** | | | | | | | | | | | | | | |
| 6 | Transportation volume per truck | 3.16 | 0.614 | $0.413^{**}$ | $0.415^{**}$ | $0.415^{**}$ | $0.564^{**}$ | $0.445^{**}$ | 1 | | | | | | |
| 7 | Load efficiency | 3.14 | 0.675 | $0.298^{**}$ | $0.353^{**}$ | $0.249^{*}$ | $0.433^{**}$ | $0.347^{**}$ | $0.603^{**}$ | 1 | | | | | |
| 8 | Laden miles | 3.18 | 0.819 | 0.175 | 0.168 | 0.204 | 0.119 | 0.107 | $0.415^{**}$ | $0.428^{**}$ | 1 | | | | |
| 9 | Usage efficiency | 3.22 | 0.807 | $0.267^{*}$ | $0.259^{*}$ | $0.326^{**}$ | 0.177 | 0.103 | $0.573^{**}$ | $0.465^{**}$ | $0.660^{**}$ | 1 | | | |
| 10 | Delivery in Full on Time (DIFOT) | 3.15 | 0.824 | 0.067 | 0.059 | 0.099 | 0.061 | -0.031 | $0.232^{*}$ | $0.367^{**}$ | $0.665^{**}$ | $0.664^{**}$ | 1 | | |
| 11 | Number of damages or losses of cargos | 3.09 | 0.908 | $0.247^{*}$ | 0.073 | 0.067 | 0.217 | 0.092 | $0.249^{*}$ | $0.347^{**}$ | $0.362^{**}$ | $0.312^{**}$ | $0.490^{**}$ | 1 | |
| **III.** | **Pofitability** | | | | | | | | | | | | | | |
| 12 | Pofitability | 3.13 | 0.669 | $0.267^{*}$ | $0.312^{**}$ | $0.274^{*}$ | $0.278^{*}$ | 0.090 | $0.498^{**}$ | $0.275^{*}$ | $0.501^{**}$ | $0.497^{**}$ | $0.353^{**}$ | $0.254^{*}$ | 1 |

# Mathematical Analysis of SARS Epidemic Model

Siwaphorn  Kanchanarat, Settapat  Chinviriyasit*
*Department of Mathematics, King Mongkut's University of Technology ThonBuri, Bangkok, 10140, Thailand,*
*siwaphonkp@gmail.com*
*settapat.chi@kmutt.ac.th*

## Abstract

This paper presents an *SEIRS* epidemic model for SARS outbreak in 2003. The epidemiological threshold, known as the basic reproduction number, of the model is derived. The model exhibits two equilibriums namely: the disease-free equilibrium and the endemic equilibrium. Using global stability analysis of the model at disease free equilibrium, based on constructing a Lyapunov function, it is shown that the disease can be eradicated from the population if $R_0 < 1$. If $R_0 > 1$, a unique endemic equilibrium which is locally asymptotically stable showing that the disease will persist within the population so that the disease becomes endemic. Further, the *SEIRS* epidemic model is applied to the real data of SARS outbreak in 2003. The result shows good agreement with the real data of SARS outbreak in Hongkong 2003. This verifies that a model *SEIRS* can be used to predict the SARS transmission in Hongkong 2003.

*Keywords*: *SEIRS* epidemic model, Stability, Reproduction number, SARS transmission

*Settapat Chinviriyasit
E-mail Address: settapat.chi@kmutt.ac.th

## 1. Introduction

In recent years, the understanding of infectious-disease epidemiology and control has been greatly increased through mathematical modelling. Mathematical modelling, thus, plays a key role in policy making, including health-economic aspects; emergency planning and risk assessment; control-program evaluation; and monitoring of surveillance data. In research, it is essential in study design, analysis (including parameter estimation) and interpretation. Severe acute respiratory syndrome (SARS) spread rapidly in 2002-2003, revealing just how vulnerable we might be to emerging diseases and how global transportation is critical to the spread of an epidemic [14,16]. The transmission of SARS is thus ideally suited to the testing of mathematical models.

Epidemic models study the transmission dynamics of infectious diseases in host populations. It is assumed that a susceptible individual first goes through a latent period after infection before becoming infectious. The resulting models are of *SEI* , *SEIR* or *SEIRS* type, respectively. Greenhalgh [9] considered *SEIR* models that incorporate density dependence in the death rate. Cooke and van den Driessche [5] introduced and studied *SEIRS* models with delays. Greenhalgh [10] studied Hopf bifurcations in models of the *SEIRS* type with density dependent contact rate and death rate. However, they have not analyzed the global dynamics of the model and have not used the mode to applied the real data of outbreak in order to analysis and interpretation for designing strategies to control rapidly spreading infectious diseases.

The aim of this paper is to analyze the global dynamic of an *SEIRS* epidemic model. The model consists of the following equations:

$$\begin{aligned}\frac{dS}{dt} &= a - \frac{\beta SI}{N} - bS + \alpha_2 R,\\ \frac{dE}{dt} &= \frac{\beta SI}{N} - (b+c)E,\\ \frac{dI}{dt} &= cE - (e+d)I,\\ \frac{dR}{dt} &= dI - (b+\alpha_2)R.\end{aligned} \tag{1}$$

where $S(t)$, $E(t)$, $I(t)$, and $R(t)$ represent susceptible, exposed, infectious, and recovered, respectively. The other parameter in (1) and their values are given in Table I.

From biological considerations, the system (1) is studied in the closed set

$$D = \{(S,E,I,R) \in \mathbb{R}^4_+ \,|\, S \geq 0, E \geq 0, I \geq 0, R \geq 0, \\ S+E+I+R \leq a/b\}$$

where $\mathbb{R}^4_+$ denotes the non-negative cone of $\mathbb{R}^4$ including its lower dimensional faces. It can be verified that $D$ is positively invariant with respect to (1).

This paper is organized as follows. The model (1) is analyzed for the existence and stability of its equilibria in Section 2. By constructing a Lyapunov function, it will be shown that the disease-free equilibrium is globally asymptotically stable if the basic reproductive number is less than unity. When  this value is greater than unity, the model (1) has a unique endemic equilibrium which is locally asymptotically stable based on the use of the centre manifold theory. Simulation results are presented in Section 3 to illustrate the theoretical results. The *SEIRS* model is applied to predict the SARS outbreak in Hongkong.

## 2. Analysis of the Model

In this section, the model (1) will be qualitatively analyzed to investigate the existence and stability of its associated equilibria.

### *2.1 Disease-free equilibrium*

The disease-free equilibrium, obtained by setting the right-hand sides of equations in (1) to zero, is given by

$$P_0(\mathrm{S}^0,0,0,0)=(\frac{a}{b},0,0,0). \qquad (2)$$

According to the concept of next generation matrix [19] and reproduction number presented in Van Den Driessche and Watmough [20], two matrices are defined as

$$F=\begin{pmatrix}0 & \beta\\ 0 & 0\end{pmatrix} \text{ and } V=\begin{pmatrix}b+c & 0\\ -c & e+d\end{pmatrix}.$$

Hence the reproduction number for (1) is

$$R_0=\rho(FV^{-1})=\frac{\beta c}{(e+d)(b+c)}. \qquad (3)$$

Consequently, using Theorem 2 of [14], the following results is established.

**Lemma 3.1**: The disease-free equilibrium (DFE), $P_0$, of the system (1) is locally asymptotically stable (LAS) if $R_0<1$ and unstable if $R_0>1$.

The quantity $R_0$ in (3) is called the basic reproduction number [1]. It measures the average number of secondary cases that one infected case can generate if introduced into a completely susceptible population. The local stability result in Lemma 3.1 implies that for $R_0<1$, the total number of infective individual in the population can be reduced to zero if the initial sizes of the sub-populations of the model are in the basin of attraction of $P_0$. That is, a small influx of infected population into the community would not generate large outbreaks, and the disease dies out in time if $R_0<1$ (but will do so if $R_0>1$). To ensure that the effective control (or elimination) of the number of infected population in the community at steady-state is independent of the initial sizes of the sub-populations of the model, it is imperative to show that the DFE $P_0$ is globally-asymptotically stable (GAS). The following result is claimed (the proof is based on using the following Lyapunov function, see Appendix A).

Table 1: Description and parameter values for the model (1)

| Parameters | Descriptions | Values | References |
|---|---|---|---|
| $a$ | Recruitment rate(by birth and by immigration) | 1 | [15] |
| $b$ | Natural death rate | 0.2 | [15] |
| $c$ | Rate of exposed individuals become infected individuals | 0.3 | [15] |
| $d$ | Transfer rate from infected individuals to recovered individuals | 0.1 | [13] |
| $e$ | Mortality rate for infected individuals | 0.4 | [15] |
| $\alpha_2$ | Rate that recovered individuals become susceptible individuals | 0.03 | [13] |
| $\beta$ | Transmission rate | $0\le\beta\le1$ | assumed |

**Theorem 3.1** : The DFE of the model (1) is globally asymptotically stable in $D$ if $R_0\le1$.

### *2.2 Endemic equilibrium*

In the presence of infection ( $E\neq0$ and $I\neq0$ ),model (1) has an endemic equilibrium $P_*\left(S^*,E^*,I^*,R^*\right)$, where

$$\begin{aligned}
S^*&=\frac{a\{(b+\alpha_2)(c+d+e)+cd\}}{\Omega},\\
E^*&=\frac{a\{(R_0-1)(b+\alpha_2)+(e+d)\}}{\Omega},\\
I^*&=\frac{ac(R_0-1)(b+\alpha_2)}{\Omega},\\
R^*&=\frac{acd\,(R_0-1)(b+\alpha_2)}{\Omega},
\end{aligned} \qquad (4)$$

with $N^*=R_0S^*$ and

$$\Omega=R_0b(be+cd+d\alpha_2+e\alpha_2+bd)$$
$$+c(b+\alpha_2)((R_0-1)e+b).$$

Clearly, it is evident from the above four equations that if $R_0<1$, then the model (1) has no positive endemic equilibrium (since $E^*$ and $I^*$ will assume negative values which are biologically unrealistic).Therefore, to ensure the existence of a positive endemic equilibrium, $R_0$ must be greater than one.

The local stability of the unique endemic equilibrium is guaranteed by the following theorem (the proof, is given in Appendix C).

**Theorem 3.1**: The unique endemic equilibrium $P_*$ of the model (1) is LAS if $R_0>1$.

## 3. Numerical Experiments

The model (1) is solved by using fourth-order Runge kutta method with the parameter values/ranges in Table 1. The results are shown in two experiments as follows.

### 4.1 Experiment 1

To illustrate theoretical, the transmission rate is set to be $\beta = 0.6, 0.95$. This value and the other parameter values/ranges in Table 1, give $R_0 = 0.72 < 1$ and $R_0 = 1.14 > 1,$, respectively. Figure 1(a)-(d) verify that the numerical solutions of the model (1) converge to disease-free equilibrium, $P_0$, whenever $R_0 < 1$, and to endemic equilibrium, $P_*$, if $R_0 > 1$ as shown in Figure 1(e) - (h), respectively.

### 4.2 Experiment 2

The *SEIRS* model (1) is used to predict the cumulative number of SARS cases in Hongkong from 17 March 2003 to 26 April 2003 [18]. The cumulative number of SARS cases is predicted by solving

$$\frac{dC}{dt} = kI,$$

where $C$ denotes the cumulative number of SARS cases and $k$ is the rate of progression from infective to diagnosed. The model (1) is simulated with the parameter values:

$a = 3$ day$^{-1}$, $b = 0.000034$ day$^{-1}$, $c = \frac{1}{6.4}$ day$^{-1}$,

$d = \frac{1}{4}$ day$^{-1}$, $e = 0.007934$ day$^{-1}$, $\alpha_2 = 0.001$ day$^{-1}$,

$k = \frac{1}{3}$ day$^{-1}$ and $\beta = 0.679$ day$^{-1}$.

For numerical simulations, the initial conditions are assumed to be $S(0) = 1{,}100,\ E(0) = 95,\ R(0) = 0$ and $C(0) = 95.$ For $I(0) = 95,$ corresponds to number of infectious cases on 17 March 2003. The numerical results of model (1) are shown in Figure 2-3. Figure 2 shows that the number of susceptible individuals decrease whereas the number of exposed, infected and recovered individuals increase. This means that when the disease spread occurs, the number of susceptible individuals decrease since the susceptible individuals contact with infected individuals. Thus, susceptible individuals can require exposed individuals. After 2 - 10 days [7], the exposed individuals is progression to symptoms development, therefore, exposed individual is called infected individuals. After that infected individuals is hospitalized about 3-5 days [7] and then infected individuals is becomes recovered individuals. It can be concluded that SARS is highly infectious base on the gradient of the susceptible curve. Figure 3 shows the predicted total cases obtained by (5) The resulting curve for $C$ fits very well with the observed total cases from 17 March 2003 to 26 April 2003 (totally 54 days ).

$R_0 < 1$ $\qquad\qquad$ $R_0 > 1$

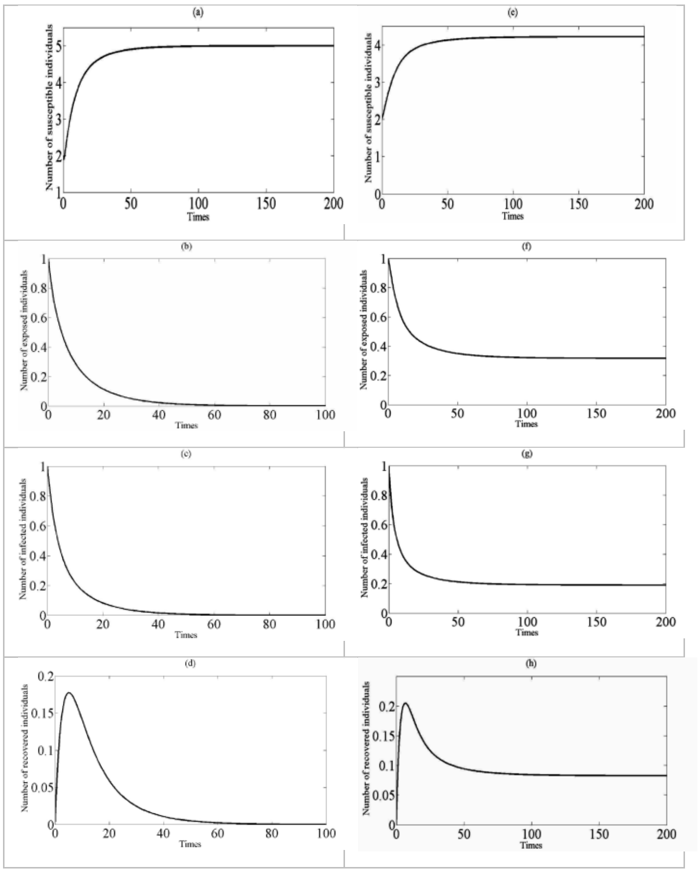


Figure 1. Time series plot of the model (1) with parameter values in Table 1 and initial conditions $S(0) = 2$, $E(0) = 1$, $I(0) = 1,$ $R(0) = 0$ : (a)-(d) profiles of all populations for $\beta = 0.6$, $R_0 = 0.72 < 1$; (e)-(h) profiles of all populations for $\beta = 0.95$, $R_0 = 1.14 > 1,$

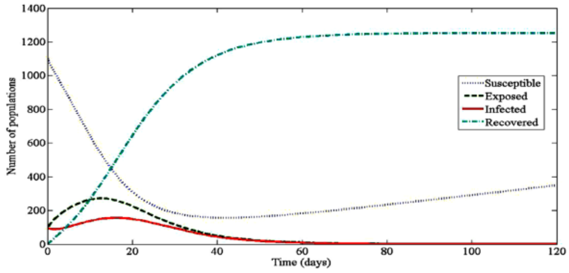


Figure 2. The number of all populations produced by the model (1). with parameter values in Table 1 and initial condition $S(0) = 1{,}100,\ E(0) = 95,\ \mathrm{I}(0) = 95$ and $\mathrm{R}(0) = 0.$

This implies that *SEIRS* model (1) can be used to predict the SARS transmission in Hongkong 2003.

## 4. Conclusions

In this paper, a *SEIRS* epidemic model is analyzed to study the transmission of SARS in Hongkong 2003. By analyzing the model, a threshold $R_0$ is derived.

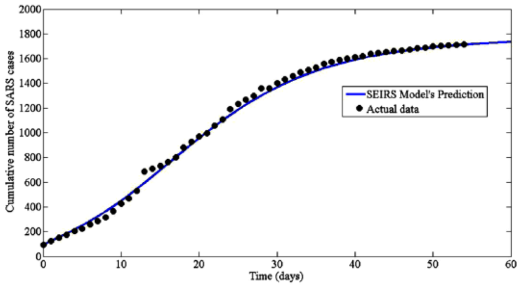


Figure 3. Comparison the cumulative numbers of SARS between actual data by WHO [18] (dotted lines) and predicted by *SEIRS* model(1) (solid lines).

It is noted that when $R_0 < 1$ then disease dies out and when $R_0 > 1$ the disease becomes endemic. The model has two non-negative equilibria namely $P_0(A/\mu, 0, 0, 0)$ , the disease-free equilibrium and the endemic equilibrium, $P_*\left(S^*, E^*, I^*, R^*\right)$

It is found that the equilibrium state $P_0$, corresponding to disappearance of disease is locally asymptotically stable if $R_0 < 1$, and $R_0 > 1$, for it is unstable which implies that the infection is maintained in the population. The endemic equilibrium $P^*$, which exists only when is $R_0 > 1$, always locally asymptotically stable. The result of the *SEIRS* model (1) is good agreement with the real data of SARS outbreak in Hongkong 2003. This implies that *SEIRS* model (1) can be used to predict the SARS transmission in Hongkong 2003.

## 5. References

[1] Anderson R.M. and May R.M. Infectious Diseases of Humans, Dynamics and Control, Oxford University Press, London, NewYork, 1991.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892; 68–73.

[2] Carr J. Applications Centre Manifold Theory, Springer-Verlag, New York, 1981.K. Elissa.

[3] Castillo-Chavez C. and Song B. "Dynamical models of tuberculosis and their applications", Math. Biosci. Eng., 2004; 1(2): 361–404.

[4] Chowella G. Fenimorea P.W. Castillo-Garsowc M.A. and Castillo-Chavez C. "SARS outbreaks in Ontario, Hong Kong and Singapore: the role of diagnosis and isolation as a control mechanism", J. Theor. Biol., Vol. 224, pp. 1--8, 2003.

[5] Cooke K. and van den P. Driessche, "Analysis of an SEIRS epidemic model with two delays", J. Math. Biol., 1996; 35: 240-260.

[6] Diekmann O., Metz J.A.J., and Heesterbeek J.A.P. "On the definition and the computation of the basic reproduction ratio $R_0$ in models for infectious diseases in heterogeneous populations", J. Math. Biol., 1990; 28: 365-382.

[7] Donnelly C.A., et al., Epidemiological determinants of spread of cusal agent of severe acute respiratory syndrome in Hong Kong[Internet]. 2003. Available from: http://image.thelancet.com/extras/03art4453-web.pdf.

[8] Fulford G.R., Roberts M.G., and Heesterbeek J.A.P., The metapopulation dynamics of an infectious disease: tuberculosis in possums, J. Theor. Biol., 2002; 61: 15-29.

[9] Greehalgh D., Some results for a *SEIR* epidemic model with density dependence in the death rate, IMAJ. Math. Appl. Med. Biol. 1992; 9: 67-106.

[10] Greehalgh D., Hopf bifurcation in epidemic models with a latent period and nonpermanent immunity, Math. Comput. Model., 1997; 25: 85-107.

[11] Gumel A.B., et al., "Modelling strategies for controlling SARS outbreaks", Proc. Roy. Soc. B, 2004; 271: 2223-2232.

[12] Lipsitch M., et al., Transmission dynamics and control of severe acute respiratory syndrome, Science, 2003; 300: 1966-1970.

[13] Liu J. and Zhou Y. Global stability of an *SIRS* epidemic model with transport-related infection, Chaos, Solitons & Fractals, 2009; 40: 145-158.

[14] Van den Driessche P. and Watmough J., "Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission", Math. Biosci., 2002; 180: 29-48,.

[15] H. Wan and J. Cui, An SEIS epidemic model with transport--related infection, J. Theor. Biol., 2007; 247:507-524.

[16] Wilder-Smith A., The severe acute respiraltory syndrome: impact on travel and tourism, Travel Med. Infect. Dis., 2006; 4, : 53-60.

[17] World Health Organization, Severe acute respiratory syndrome (SARS): status of the outbreak and lessons for the immediate future, Geneva, May 20, 2003.

[18] World Health Organization, "Cumulative number of reported probable cases of severe acute respiratory syndrome (SARS)", 2011, www.who.int/csr/sarscountry//en/. (Accessed August 2011).

## Appendix A.

Consider the following Lyapunov function:

$$V = cE + (b+c)I,$$

with Lyapunov derivative (where " ' " represents differentiation with respect to $t$ ):

$$V' = (e+d)(b+c)\left[\frac{c\beta S}{N(b+c)(e+d)} - 1\right]I$$
$$= (e+d)(b+c)\left[\frac{R_0 S}{N} - 1\right]I$$
$$\le (e+d)(b+c)(R_0 - 1)I$$

and $V' = 0$ if and only if $I = 0$. The largest compact invariant set in $\{(S,E,I,R) \in D : V' = 0\}$ is the singleton $\{P_0\}$ Therefore, byLasalle.Lyapunov theorem [7], every solution that starts in $D$ approaches $P_0$ as $t \to \infty$.

## Appendix B.

**Theorem A** (Castillo-Chavez and Song [3, (Theorem 4.1)]).

Consider the following general system of ordinary differential equations with a parameter $\phi$

$$\frac{dx}{dt} = f(\mathrm{x}, \phi),\ f : \mathbb{R}^n \times \mathbb{R} \text{ and } f \in \mathbb{C}^2(\mathbb{R} \times \mathbb{R}). \text{(B.1)}$$

where 0 is an equilibrium point of the system (B.1) (i.e., $f(0,\phi) \equiv 0$ for all $\phi$ and assume

A1. $A = D_x f(0,0) = (\frac{\partial f_i}{\partial x_i}, 0, 0)$ is the linearization matrix of the system (B.1) around the equilibrium 0 with $\phi$ evaluated at 0. Zero is a simple eigenvalues of $A$ and other eigenvalues of $A$ have negative real parts;

A2. Matrix $A$ has a right eigenvectors, $w$ , and a left eigenvector, $v$, corresponding to the zero eigenvalue.

Let $f_k$ be the *kth* component of $f$ and

$$a = \sum_{k,i,j=1}^{n} v_k w_i w_j \frac{\partial^2 f_k}{\partial x_i \partial x_j}(0,0),$$

$$b = \sum_{k,i=1}^{n} v_k w_i \frac{\partial^2 f_k}{\partial x_i \partial \phi}(0,0).$$

The local dynamics of the system around 0 is totally by $a$ and $b$.

i. $a > 0, b > 0.$ When $\phi < 0$ with $|\phi| \ll 1, 0$ is locally asymptotically stable and there exists a positive unstable equilibrium; when , 0 is locally asymptotically stable equilibrium;

ii. $a < 0, b < 0.$ When $\phi < 0$ with $|\phi| \ll 1$ is unstable; when $0 < \phi \ll 1, 0$ is locally asymptotically stable equilibrium, and there exists a positive unstable equilibrium;

iii. $a > 0, b < 0.$ When $\phi < 0$ with $|\phi| \ll 1$, 0 is unstable, and there exists a locally asymptotically stable negative equilibrium; when $0 < \phi \ll 1, 0$ is stable, and a positive unstable equilibrium appears;

iv. $a < 0, b > 0.$ When $\phi$ changes from negative to positive, 0 changes its stability from stable to unstable. Correspondingly a negative unstable equilibrium becomes positive and locally asymptotically stable

## Appendix C.

### Proof of theorem 3.1

Linearizing the model (1) around its unique endemic equilibrium is laborious (and not really tractable) owing to its high dimensionality. Consequently, an alternative approach is considered. It is based on the use of the centre manifold theory [2], as described in Appendix B, the following simplification and change of variables are made first of all. Let $S = x_1, E = x_2, I = x_3$ and $R = x_4$, so that $N = x_1 + x_2 + x_3 + x_4$. Further, byusing vector notation $X = (x_1, x_2, x_3, x_4)^T$ the model (1) can be written in the form $\frac{dX}{dt} = (f_1, f_2, f_3, f_4)^T$, , as follows:

$$\begin{aligned} \frac{dx_1}{dt} &= f_1 = a - \mu x_1 - \frac{\beta x_1 x_3}{N} + \alpha_2 x_4, \\ \frac{dx_2}{dt} &= f_2 = \frac{\beta x_1 x_3}{N} - (b+c)x_2, \\ \frac{dx_3}{dt} &= f_3 = cx_2 - (e+d)x_3, \\ \frac{dx_4}{dt} &= f_4 = dx_3 - (b+\alpha_2)x_4. \end{aligned} \tag{C.1}$$

The Jacobian of the system(C.1) at the DFE, $P_0$, is given by

$$J(P_0) = \begin{pmatrix} -b & 0 & -\beta & \alpha_2 \\ 0 & -(b+c) & \beta & 0 \\ 0 & c & -(e+d) & 0 \\ 0 & 0 & d & -(b-\alpha_2) \end{pmatrix}, \tag{C.2}$$

from which it can also be shown (as before) that

$$R_0 = \frac{\beta c}{(e+d)(b+c)}.$$

Consider the case when $R_0 = 1$. Suppose, further, that $\beta = \beta^*$ is chosen as a bifurcation parameter. Solving for $\beta$ from $R_0 = 1$ gives

$$\beta = \beta^* = \frac{(e+d)(b+c)}{(c)}.$$

The eigenvalues of Jacobian of the system (C.1) evaluated at $P_0$ with $\beta = \beta^*$ are given by

$\lambda_1 = -b,\ \lambda_2 = -(b+\alpha_2),\ \lambda_3 = -(b+c+e+d),\ \lambda_4 = 0.$

Thus $\lambda_4 = 0$ is a simple zero eigenvalue and the other eigenvalues are real and negative. Hence, when $\beta = \beta^*$ (or equivalently when $R_0 = 1$), the disease-free equilibrium $P^0$ is a nonhyperbolic equilibrium: the assumption (A1) of Theorem A is then verified.

Now, a right eigenvector associated with the zero eigenvalue

$\lambda_4 = 0$ is given by $w = (w_1, w_2, w_3, w_4)^T$, where

$$\begin{aligned} w_1 &= -\left[\frac{(b+c)e+bd}{b} + \frac{d}{b+\alpha_2}\right] w_3, \\ w_2 &= \frac{(e+d)}{c} w_3,\ w_3 = w_3 > 0,\ w_4 = \frac{d}{b+\alpha_2} w_3 \end{aligned} \tag{C.3}$$

Further, a left eigenvector associated with the zero eigenvalue $\lambda_4 = 0$

is given by $v = (v_1, v_2, v_3, v_4)^T$, where

$$v_1 = 0,\ v_2 = \frac{\varepsilon}{2\mu+\gamma+\varepsilon},\ v_3 = \frac{\mu+\varepsilon}{2\mu+\gamma+\varepsilon},\ v_4 = 0. \tag{C.4}$$

The coefficients $a$ and $b$ defined in Theorem A are computed as in the following.

For the system (C.1) the associated non-zero partial derivatives of the right hand side functions $\left(f_i\right)$ are given by

$$\frac{\partial^2 f_1}{\partial x_2 \partial x_3} = \frac{\partial^2 f_1}{\partial x_3 \partial x_2} = \frac{\partial^2 f_1}{\partial x_3 \partial x_4} = -\frac{b\beta^*}{a}, \frac{\partial^2 f_1}{\partial x_3^2} = \frac{2b\beta^*}{a},$$
$$\frac{\partial^2 f_2}{\partial x_2 \partial x_3} = \frac{\partial^2 f_2}{\partial x_3 \partial x_2} = \frac{\partial^2 f_2}{\partial x_3 \partial x_4} = -\frac{b\beta^*}{a}, \qquad \text{(C.5)}$$
$$\frac{\partial^2 f_2}{\partial x_3^2} = -\frac{2b\beta^*}{a}, \frac{\partial^2 f_1}{\partial x_1 \beta^*} = -1, \frac{\partial^2 f_2}{\partial x_3 \beta^*} = 1.$$

Using the expressions (C.3)-(C.5), it follows that

$$a = \sum_{k.i.j=1}^{4} v_k w_i w_j \frac{\partial^2 f_k}{\partial x_i x_j}(0,0)$$
$$= -\frac{2b\beta v_3 w_3^2 \left[\left(c+d+e\right)\left(b+\alpha_2\right)+bd\right]}{a\left(b+c\right)};$$

and

$$b = \sum_{k,i=1}^{4} v_k w_i \frac{\partial^2 f_k}{\partial x_i \partial \beta^*}(0,0) = \frac{\mathrm{c} v_3 w_3}{b+c}.$$

It is found that $a < 0$ and $b > 0$. Hence, by Theorem A(iv) in

Appendix A, the unique endemic equilibrium $P_*$ is LAS whenever

$R_0 > 1$ and $\beta^* < \beta$ with $\beta$ close to $\beta^*$ . The proof is complete.

# The meshless local Petrov-Galerkin method for solving the time fractional Burgers' equation

N. Thamareerat[1*], A. Luadsong[1] and N. Aschariyaphotha[2]
[1]Department of Mathematics, Faculty of Science,
King Mongkut's University of Technology Thonburi (KMUTT),
126 Pracha Uthit Rd., Bang Mod, Thung Khru, Bangkok 10400, Thailand,
nuntapontham@gmail.com, anirut.lua@kmutt.ac.th
[2]Ratchaburi Campus, King Mongkut's University of Technology Thonburi (KMUTT),
Rang Bua, Chom Bueng, Ratchaburi 70150, Bangkok, Thailand, nitima.asc@kmutt.ac.th

**Abstract**

In this paper, we present a numerical scheme to solve the nonlinear time fractional Burgers' equation in one dimension. In this method, we first apply the meshless local Petrov-Galerkin (MLPG) method based on a local weak formulation for the spatial discretization and then we will approximate the time fractional derivative interpreted in the sense of Caputo by quadrature formula. The moving Kriging (MK) interpolation which possesses the Kronecker delta property is employed to construct shape functions. The aim of this paper is to show that the truly MLPG method regarded as an alternative numerical approach can readily be extended to the viscous Burgers' equation of fractional order. The numerical results are compared with analytical solution to confirm the accuracy and efficiency of the proposed algorithm. Very good agreement can be observed in the verification. The present MLPG method has proved its accuracy and efficiency for solving the time fractional Burgers' equation arising in fluid dynamics as well as several other problems in science and engineering.

*Keywords*: Time Fractional Burgers' equation: Meshless Local Petrov-Galerkin Method: Moving Kriging Interpolation: Quadrature Formula

*Corresponding Author
E-mail Address: nuntapontham@gmail.com

## 1. Introduction

Over the last few decades, more and more attempts have been devoted to the fractional differential equations (FDE), also known as extraordinary differential equations, due to its applications in various fields of science and engineering. Many physical phenomena in fluid mechanics, viscoelasticity, control theory of dynamical systems, chemical physics, biology, stochastic processes, finance and other sciences can be described very successfully by fractional models. There have been a lot of achievements on the theoretical analysis, but exact solutions of most FDE cannot be derived explicitly. Only approximate analytical and numerical solutions can be obtained using procedures such as linearization, perturbation, or discretization. Hence, proposing a new algorithm to find the numerical solution of FDE is of practical significance. Differential equations involving fractional order derivative are generalization of classical differential equations of integer order. There are several definitions to the generalization of the notion of differentiation to fractional orders, including the Grünwald-Letnikov's definition, Riemann-Liouville's definition, Caputo's definition, Jumarie's definition and generalized function approach. The two most frequently used generalizations of the derivative are the Riemann–Liouville and Caputo. The Riemann-Liouville fractional derivative is not suitable to deal with the physical problems in the real world because it requires the definition of fractional order initial conditions, which have no physically meaningful interpretation yet. Caputo introduced an alternative definition, which is not necessary to define the initial conditions for the fractional order differential equations since the fractional order initial conditions take on the same form as for integer order differential equations with a known physical interpretation. Another main advantage of the Caputo fractional derivative over the Riemann–Liouville is that the Caputo derivative of a constant function is zero. The same does not hold for the Riemann-Liouville derivative. It is reasonable that the fractional derivative of a constant should be zero from a physical viewpoint. We begin by reviewing the essentials of the fractional calculus. The fractional calculus is a name for the theory of integrals and derivatives of arbitrary order, which unifies and generalizes the notions of integer-order differentiation and n-fold integration.

**Definition 1**. For $\mu \in \mathbf{R}$ and $x > 0$ a real function $f(x)$ is said to be in the space $C_\mu$ if there exists a real number $p > \mu$ such that $f(x) = x^p f_1(x)$, where $f_1(x) \in C(0, \infty)$, and for $m \in \mathbf{N}$ it is said to be in the space $C_\mu^m$ if $f^m \in C_\mu$.

**Definition 2**. The Riemann–Liouville fractional integral operator of order $\alpha \geq 0$ for a function $f \in C_\mu, \mu \geq -1$ is defined as

$$J_t^\alpha f(x,t) = \begin{cases} \frac{1}{\Gamma(\alpha)} \int_0^t (t-\tau)^{\alpha-1} f(x,\tau) d\tau, \alpha > 0 \\ f(x,t) \qquad\qquad\qquad\qquad , \alpha = 0, \end{cases} \quad (1.1)$$

where $\Gamma(\cdot)$ denotes the Gamma function.

**Definition 3**. If $m$ be the smallest integer that exceeds $\alpha$, the Caputo time fractional derivative operator of order $\alpha > 0$ is defined as

$$D_t^\alpha f(x,t) = \begin{cases} J_t^{n-\alpha}\left(\frac{\partial^m f(x,t)}{\partial t^m}\right), m-1<\alpha<m \\ \frac{\partial^m f(x,t)}{\partial t^m}, \alpha = m. \end{cases} \qquad (1.2)$$

The Burgers' equation is the fundamental equation of fluid mechanics for describing the mathematical model illustrating the theory of turbulence [1] and the approximate theory of flow through a shock wave propagating in a viscous fluid [2]. Mathematically, the Burgers' equation is the nonlinear parabolic partial differential equations containing both nonlinear advection term and dissipation effect, which can be derived by dropping the pressure term of the incompressible Navier–Stokes equations. The Burgers' equation can be generalized to the time fractional Burgers' equation by simply replacing the first-order time derivative term by a derivative of non-integer order but retaining the second-order space derivatives. We consider the one-dimensional time fractional Burgers' equation in the following form:

$$\frac{\partial^\alpha u}{\partial t^\alpha} + \varepsilon u \frac{\partial u}{\partial x} = \nu \frac{\partial^2 u}{\partial x^2} + f(x,t), x \in \Omega, t > 0 \qquad (1.3)$$

subject to the initial condition

$$u(x,0) = g(x), \quad x \in \Omega \qquad (1.4)$$

where $\varepsilon$ is a parameter, $\nu$ is the viscosity coefficient $(\nu > 0)$, $\Omega$ is a bounded domain, $\alpha$ is the parameter describing the order of the fractional, $f(x,t)$ is the source term and $\frac{\partial^\alpha u(x,t)}{\partial t^\alpha}$ is the Caputo fractional derivative of order $0 < \alpha < 1$ defined by

$$\frac{\partial^\alpha u(x,t)}{\partial t^\alpha} = \frac{1}{\Gamma(1-\alpha)} \int_0^t \frac{\partial u(x,\tau)}{\partial \tau} (t-\tau)^{-\alpha} d\tau, 0 < \alpha < 1. \qquad (1.5)$$

In the case of $\alpha = 1$, Eq. (1.5) reduces to the classical nonlinear Burgers' equation.

Some analytical and numerical methods such as the Adomian decomposition method (ADM) [3], the variational iteration method (VIM) [4], the differential transform method (DTM) [5], the generalized differential transformation and homotopy perturbation method (HPM) [6] have been proposed to derive the approximate and exact solutions of nonlinear time fractional Burgers' and coupled Burgers' equations. However, so far only few researchers have attempted to extend and develop the so-called meshless method to the models governed by fractional Burgers' equation. The meshless or meshfree methods regarded as an alternative numerical approach have been developed to overcome the difficulties and limitations of mesh generation. This method is used to establish the system of algebraic equations for the whole problem domain without the use of a predefined mesh or domain discretization. There are a number of meshfree methods that use local nodes for approximating the field variable, for example, the element free Galerkin method (EFG) [7], the meshless local Petrov-Galerkin (MLPG) method [8], the reproducing kernel particle method (RKPM) [9], the point interpolation method (PIM) [10], the radial point interpolation method (RPIM) [11,12,13] and so forth. Some meshless methods based on global weak forms such as the RPIM and EFG method, being “meshless” only in terms of the interpolation of the field variables, have to use background cells to evaluate integrals appearing in the local weak formulation. This is one reason why the above-mentioned methods are not truly meshless method. One of the most popular meshless methods is the meshless local Petrov-Galerkin (MLPG) method first proposed by Atluri and Zhu [8] for solving linear potential problems. The MLPG method is one of the truly meshless methods since it does not require any background integration cells so that all integrations are carried out locally over small sub-domains of regular shapes. This approach is also different from the truly meshless method based on the local boundary integral equation (LBIE) method [14,15] in the fact that there is no singular integral in the present MLPG method, while some kinds of singular integrals have to tackled in the meshless LBIE method [16]. Meshfree shape function construction is one of the central and most important issues that significantly effect on the performance of meshfree methods. A number of ways to efficiently create shape functions have been proposed. In this study, we employ the moving Kriging (MK) interpolation. The MK interpolation technique was first introduced in computational mechanics by Gu [17]. One notable feature of shape function constructed using the MK interpolation is that it possesses the Kronecker delta property in which special treatment is not required to impose the essential boundary condition.

To the best of our knowledge, the MLPG method has never been used to approximate the solution of the time fractional Burgers' equation. Thus, the objective of this work is to extend the application of the MLPG method to the viscous Burgers' equation of fractional order. This paper is organized as follows. In Section 2, we introduce the moving Kriging interpolation for constructing shape functions and then we describe how to formulate the weak form and discretization system. In Section 3, the numerical experiments are presented and discussed in details to demonstrate the accuracy and efficiency of the proposed scheme. We complete the paper with conclusion and recommendation for further work given in Section 4.

## 2. Research Methodology

### *2.1 The moving Kriging interpolation method*

Kriging or Gaussian process regression was originally applied in geostatistics for spatial interpolation. Subsequently, Kriging interpolation was employed to construct shape function for enhancement of the finite element method (FEM). The procedure of constructing shape functions for meshfree methods using the MK interpolation is detailed in this section. Similar to the moving least square (MLS) approximation, Kriging interpolation can be extended to any sub-domain, $\Omega_s \subset \Omega$. To approximate a distribution function, $u(\mathbf{x}_i)$, based on all nodal values $\{\mathbf{x}_i, i = 1,2,3,\dots N\}$ where $N$ is the total number of nodes in the sub-domain, the formulation of the MK interpolation is given by

$$u^h(\mathbf{x}) = \sum_{j=1}^{N} \phi_j(\mathbf{x})\hat{u}_j = \mathbf{\Phi}(\mathbf{x})\mathbf{U},\ \mathbf{x} \in \Omega_s, \tag{2.1}$$

where $\mathbf{U} = [\hat{u}(\mathbf{x}_1)\ \hat{u}(\mathbf{x}_2)\ \hat{u}(\mathbf{x}_3) \dots \hat{u}(\mathbf{x}_N)]^T$ and $\mathbf{\Phi}(\mathbf{x})$ is a $1 \times N$ vector of Kriging shape function defined by

$$\mathbf{\Phi}(\mathbf{x}) = \mathbf{p}^T(\mathbf{x})\mathbf{A} + \mathbf{r}^T(\mathbf{x})\mathbf{B}. \tag{2.2}$$

The matrices $\mathbf{A}$ and $\mathbf{B}$ are determined by

$$\begin{aligned}\mathbf{A} &= (\mathbf{P}^T\mathbf{R}^{-1}\mathbf{P})^{-1}\mathbf{P}^T\mathbf{R}^{-1},\\ \mathbf{B} &= \mathbf{R}^{-1}(\mathbf{I} - \mathbf{PA})\end{aligned} \tag{2.3}$$

where $\mathbf{I}$ is an unit matrix of size $N \times N$ and $\mathbf{p}(\mathbf{x})$ is a vector of the polynomial with $m$ basis functions given by

$$\mathbf{p}(\mathbf{x}) = [p_1(\mathbf{x})\ p_2(\mathbf{x})\ \ p_3(\mathbf{x}) \dots p_m(\mathbf{x})]^T. \tag{2.4}$$

A commonly used linear basis in one-dimensional space is given by

$$\mathbf{p}^T(x) = \{1, x\},\ m = 2,$$

the quadratic polynomial basis is

$$\mathbf{p}^T(x) = \{1, x, x^2\},\ m = 3,$$

and the cubic polynomial basis is

$$\mathbf{p}^T(x) = \{1, x, x^2, x^3\},\ m = 4.$$

The matrix $\mathbf{P}$ has a size $N \times m$ and represents the collected values of (2.4) as

$$\mathbf{P} = \begin{bmatrix} p_1(\mathbf{x}_1) & \cdots & p_m(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ p_1(\mathbf{x}_N) & \cdots & p_m(\mathbf{x}_N) \end{bmatrix},$$

and $\mathbf{r}(\mathbf{x})$ in Eq. (2.2) has the form of

$$\mathbf{r}(\mathbf{x}) = [\gamma(\mathbf{x}, \mathbf{x}_1)\ \gamma(\mathbf{x}, \mathbf{x}_2)\ \dots \gamma(\mathbf{x}, \mathbf{x}_N)]^T,$$

where $\gamma(\mathbf{x}, \mathbf{x}_j)$ is the correlation between any pair of nodes located at $\mathbf{x}$ and $\mathbf{x}_j$, and it belongs to the covariance of the field value $u(\mathbf{x})$ ,i.e., $\gamma(\mathbf{x}_i, \mathbf{x}_j) = cov[\mathbf{u}_i, \mathbf{u}_j]$. The correlation matrix $\mathbf{R}[\gamma(\mathbf{x}_i, \mathbf{x}_j)]_{N\times N}$ is given by

$$\mathbf{R} = \begin{bmatrix} \gamma(\mathbf{x}_1, \mathbf{x}_1) & \cdots & \gamma(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ \gamma(\mathbf{x}_N, \mathbf{x}_1) & \cdots & \gamma(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix}.$$

Many different correlation functions can be used for the correlation matrix. Gaussian function with a correlation parameter $\theta$ is often used to best fit the model due to its simplicity.

$$\gamma(\mathbf{x}_i, \mathbf{x}_j) = e^{-\theta r_{ij}^2}, \tag{2.5}$$

where $r_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$ and $\theta > 0$ is a correlation parameter.

Let $C^{k_1}(\Omega)$ be the space of $k_1$th continuously differentiable functions on $\Omega$. If $\gamma(\mathbf{x}, \mathbf{x}_i) \in C^{k_1}(\Omega)$ and $p_j(\mathbf{x}) \in C^{k_2}(\Omega), i = 1,2,\dots,N,\ j = 1,2,\dots,m$ then $\phi_i(\mathbf{x}) \in C^k(\Omega)$ with $k = \min(k_1, k_2)$. The partial derivatives of the shape function $\mathbf{\Phi}(\mathbf{x})$ with respect to $\mathbf{x}_i$ are obtained as

$$\mathbf{\Phi}_{,i}(\mathbf{x}) = \mathbf{p}_{,i}{}^T(\mathbf{x})\mathbf{A} + \mathbf{r}_{,i}{}^T(\mathbf{x})\mathbf{B}, \tag{2.6}$$

$$\mathbf{\Phi}_{,ii}(\mathbf{x}) = \mathbf{p}_{,ii}{}^T(\mathbf{x})\mathbf{A} + \mathbf{r}_{,ii}{}^T(\mathbf{x})\mathbf{B}, \tag{2.7}$$

where $(\cdot)_{,i}$ and $(\cdot)_{,ii}$ denote the first- and second-order spatial derivatives, respectively.

### *2.2 The MLPG formulation and numerical implementation*

#### *2.2.1 The local weak form*

Instead of giving the global weak form, the MLPG method constructs the weak form over local sub-domains, $\Omega_s$, which is a small region taken for each node in the global domain $\Omega$. The local sub-domains overlap each other and cover the whole global domain. The local sub-domains could be of any geometric shape and size, such as open or closed intervals in one dimension, circles or squares in two dimensions and spheres or cubes in three dimensions. For simplicity they are taken to be of open interval. The local weak form of Eq. (1.3) for $x_i \in \Omega_s^i$ can be written as

$$\int_{\Omega_s^i} \left(\frac{\partial^\alpha u}{\partial t^\alpha} + \varepsilon u \frac{\partial u}{\partial x}\right) v_i dx = \int_{\Omega_s^i} \left(\nu \frac{\partial^2 u}{\partial x^2} + f(x,t)\right) v_i dx, \tag{2.8}$$

where $\Omega_s^i$ is a local sub-domain associated with the point $i$ ,i.e., a bounded interval centered at $x_i$ of characteristic length regarding the nodal spacing close to the point of interest, $r_0$, and $v$ is a test function.

Substitution of $u^h(x,t)=\sum_{j=1}^{N}\phi_j(x)\hat{u}_j(t)$ into local integral Eq. (2.8) yields, excluding the nonlinear term,

$$\sum_{j=1}^{N}\left(\int_{\Omega_s^i}\phi_j(x)v_i dx\right)\frac{\partial^\alpha \hat{u}_j}{\partial t^\alpha}+\sum_{j=1}^{N}\left(\int_{\Omega_s^i}\left(\varepsilon u\phi_{j,x}(x)-\nu\phi_{j,xx}(x)\right)v_i dx\right)\hat{u}_j=\int_{\Omega_s^i} f(x,t)\,v_i dx. \qquad (2.9)$$

Eq. (2.9) can be written in the following matrix form:

$$\mathbf{A}\frac{\partial^\alpha \hat{\mathbf{U}}}{\partial t^\alpha}+\mathbf{B}\hat{\mathbf{U}}=\mathbf{C}, \qquad (2.10)$$

where $\mathbf{A}=\left[\mathrm{A}_{ij}\right]_{N\times N}; \mathrm{A}_{ij}=\int_{\Omega_s^i}\phi_j(x)v(x_i)dx,$

$\mathbf{B}=\left[\mathrm{B}_{ij}\right]_{N\times N};$

$\mathrm{B}_{ij}=\int_{\Omega_s^i}\left(\varepsilon u\phi_{j,x}(x)-\nu\phi_{j,xx}(x)\right)v(x_i)dx,$

$\mathbf{C}=[\mathrm{C}_i]_{N\times 1}^{T}; \mathrm{C}_i=\int_{\Omega_s^i} f(x_i,t)v(x_i)dx,$

and $\hat{\mathbf{U}}=[\hat{u}_1\ \hat{u}_2\ \ldots \hat{u}_N]^T$.

In order to avoid the evaluation of any numerical integration in the weak form, the Kronecker delta function is chosen as the test function in each sub-domain. The coefficient matrices $\mathbf{A},\mathbf{B}$ and vector $\mathbf{C}$ can be simplified to the following expression:

$\mathrm{A}_{ij}=\phi_j(x),$

$\mathrm{B}_{ij}=\varepsilon u\phi_{j,x}(x)-\nu\phi_{j,xx}(x),$

$\mathrm{C}_i=f(x_i,t).$

Taking the advantage of shape function constructed by MK interpolation that the Kriging shape function maintains a unit value at the node and vanishes at all other nodes, i.e., the delta function property. This gives

$$A_{ij}=\phi_j(x_i)=\begin{cases}1, i=j\\ 0, i\neq j\end{cases} \quad \text{or} \quad \mathbf{A}=\mathbf{I},$$

where $\mathbf{I}$ is the $N\times N$ identity matrix.

Therefore Eq. (2.10) reduces to

$$\frac{\partial^\alpha \hat{\mathbf{U}}}{\partial t^\alpha}+\mathbf{B}\hat{\mathbf{U}}=\mathbf{C}. \qquad (2.11)$$

### *2.2.2 The time discretization*

For some positive integers $M_1$ and $M_2$, let $\Delta t=\frac{T}{M_1}$ be the step size of time variable and $\Delta x=\frac{b-a}{M_2}$ denotes the step size of space variable. The grid points in the space interval $[a,b]$ and time interval $[0,T]$ are defined by

$x_i=a+ih, i=0,1,2,\ldots M_1,$

$t_n=n\Delta t, n=0,1,2,\ldots,M_2,$

respectively. The approximate solutions at the grid points are denoted $u_i^n=u(x_i,t_n)$ and $f_i=f(x_i)$.

The linearization method based on Taylor series expansion of a function is used to approximate the nonlinear term

$$u(x,t_n)=u(x,t_{n-1})+O(\Delta t) \qquad (2.12)$$

The discrete approximation to the time fractional derivative in Caputo's sense can be obtained using a simple quadrature formula [18] as

$$\frac{\partial^\alpha \hat{\mathbf{U}}(x,t_n)}{\partial t^\alpha}=\sigma_{\alpha,\Delta t}\sum_{k=1}^{n}\omega_k^{(\alpha)}(\hat{\mathbf{U}}^{n-k+1}-\hat{\mathbf{U}}^{n-k})+O(\Delta t) \qquad (2.13)$$

where $\omega_k^{(\alpha)}=k^{1-\alpha}-(k-1)^{1-\alpha}$

and $\sigma_{\alpha,\Delta t}=\frac{1}{\Gamma(1-\alpha)}\frac{1}{1-\alpha}\frac{1}{\Delta t^\alpha}$.

Substituting Eq. (2.13) into Eq. (2.11), regarding to Eq. (2.12), and omitting higher-order terms lead to the $N\times N$ discretized system of linear algebraic equations

$$\sigma_{\alpha,\Delta t}\sum_{k=1}^{n}\omega_k^{(\alpha)}(\hat{\mathbf{U}}^{n-k+1}-\hat{\mathbf{U}}^{n-k})+\mathbf{B}^{n-1}\hat{\mathbf{U}}^n=\mathbf{C}^n,$$

or equivalently

$$\sigma_{\alpha,\Delta t}\left(\hat{\mathbf{U}}^n-\hat{\mathbf{U}}^{n-1}\right)+\mathbf{B}^{n-1}\hat{\mathbf{U}}^n=-\sigma_{\alpha,\Delta t}\sum_{k=2}^{n}\omega_k^{(\alpha)}\left(\hat{\mathbf{U}}^{n-k+1}-\hat{\mathbf{U}}^{n-k}\right)+\mathbf{C}^n. \qquad (2.14)$$

For $n=1$, we get

$$\left(\sigma_{\alpha,\Delta t}\mathbf{I}+\mathbf{B}^0\right)\hat{\mathbf{U}}^1=\sigma_{\alpha,\Delta t}\hat{\mathbf{U}}^0+\mathbf{C}^1, \qquad (2.15)$$

and for $n\geq 2$,

$$\left(\sigma_{\alpha,\Delta t}\mathbf{I}+\mathbf{B}^{n-1}\right)\hat{\mathbf{U}}^n=\sigma_{\alpha,\Delta t}\left(\hat{\mathbf{U}}^{n-1}-\sum_{k=2}^{n}\omega_k^{(\alpha)}\left(\hat{\mathbf{U}}^{n-k+1}-\hat{\mathbf{U}}^{n-k}\right)\right)+\mathbf{C}^n. \qquad (2.16)$$

## 3. Research Results and Discussion

### *3.1 Numerical experiments*

In this section, we provide the numerical examples to demonstrate the accuracy and efficiency of the proposed method. The results of numerical experiments are compared with analytical solution by performing the infinity norm (maximum norm) and root mean square (RMS) error:

$$L_\infty = \max_{i=1}^{N}|U_i - u_i|, \tag{3.1}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(U_i - u_i)^2}, \tag{3.2}$$

where $U_i$ and $u_i$ are the analytical and approximate solutions, respectively. $N$ is the number of nodal points. In these examples the regular node distribution is used. Also to implement the meshless local weak form, the radius of the support domain (interval) is taken to be $r_0 = 2.1\Delta x$. The whole global domain of problem must be enclosed with the union of these sub-domains. To give an appropriate approximation, this size should be large enough to have sufficient number of nodes inside the influence domain. In the MK procedure, the cubic basis function $\mathbf{p}^T(x) = \{1, x, x^2, x^3\}$ is used for all numerical computations because in general cubic polynomial basis will yield a better result than quadratic and linear basis. As studied in the previous work, the correlation parameter has a significant effect on the solution. The correlation parameter is taken to be $\theta = 1/\Delta x^2$ which can be used to smooth out small features in the data.

**Test problem 1.**

**Example 1.1** Consider the time fractional Burgers' equation with the following initial value problem [3]:

$$\frac{\partial^\alpha u}{\partial t^\alpha} + u\frac{\partial u}{\partial x} = \nu\frac{\partial^2 u}{\partial x^2}, 0 < x < 1, 0 < t \le T \tag{3.3}$$

with the exact solution in a series form for $\alpha = 1$

$$u(x,t) = \frac{\mu+\sigma+(\sigma-\mu)\exp(\gamma)}{1+\exp(\gamma)} + \frac{2\mu\sigma^2\exp(\gamma)}{[1+\exp(\gamma)]^2\nu}t + \frac{(\mu^3\sigma^2\exp(\gamma))(\exp(\gamma)-1)}{[1+\exp(\gamma)]^3\nu^2}t^2 + \frac{(\mu^4\sigma^3\exp(\gamma))(1-4\exp(\gamma)+\exp(\gamma))}{3[1+\exp(\gamma)]^4\nu^3}t^3 + \cdots,$$

where $\gamma = \frac{\mu}{\nu}(x - \lambda)$ and the parameters $\mu, \sigma, \lambda$ and $\nu$ are arbitrary constant,

or in a closed form

$$u(x,t) = \frac{\mu+\sigma+(\sigma-\mu)\exp(\frac{\mu}{\nu}(x-\sigma t-\lambda))}{1+\exp(\frac{\mu}{\nu}(x-\sigma t-\lambda))}.$$

The initial and boundary conditions can be obtained from the exact solution.

**Example 1.2** Consider the inhomogeneous time fractional Burgers' equation with the following initial value problem [19]:

$$\frac{\partial^\alpha u}{\partial t^\alpha} + u\frac{\partial u}{\partial x} = \frac{\partial^2 u}{\partial x^2}, 0 < x < 1, 0 < t \le T \tag{3.4}$$

subject to the initial condition

$$u(x,0) = 2x,\ 0 \le x \le 1,$$

and the boundary conditions

$$u(0,t) = 0,$$

$$u(1,t) = \frac{2}{1+2t}.$$

We have the exact solution of Eq. (3.4) for $\alpha = 1$,

$$u(x,t) = \frac{2x}{1+2t}.$$

The results of numerical experiments are compared with analytical solutions that previously found by variational iteration method and Adomian decomposition method. The behavior of the solution by fractional model can be observed as the fractional derivative parameter is changed. In special case when $\alpha \to 1$, as expected, the results tend to the analytical solution of classical Burgers' equation. Table 1 and 2 give the difference between the numerical results and analytical solutions for different values of times in solving example 1.1 and 1.2, respectively. Also Fig. 1 and 2 show the graphs of approximate and exact solutions after 10 iterations with $\alpha = 0.99, \Delta x = \Delta t = 0.1$ at $t = 1$.

**Test problem 2.** Let's consider the following time fractional Burgers' equation

$$\frac{\partial^{1/2} u}{\partial t^{1/2}} + u\frac{\partial u}{\partial x} = \frac{\partial^2 u}{\partial x^2} = f(x,t), a < x < b, 0 < t \le T, \tag{3.5}$$

where $f(x,t) = (x-a)(b-x)(a+b-2x)t^2 + 1.13(x-a)(b-x)t^{1/2} + 2t.$

The exact solution of Eq. (3.5) is

$$u(x,t) = (x-a)(b-x)t.$$

The initial and boundary conditions can be obtained from the exact solution. The computational domain for this example is $\Omega = [0,1] = \{x | 0 \le x \le 1\}$. In Table 3 we show the $L_\infty$ and $RMS$ errors obtained in solving Test problem 2 for different values of times. The graphs of approximate and exact solutions after 10 iterations with $\Delta x = \Delta t = 0.05$ at $t = 1$ are depicted in Fig. 3.

### *3.2 Discussion*

The numerical results obtained by MLPG method are found to be in good agreement. As can be seen in Figs.1-3, the numerical results and analytical solutions appear to be congruent. The $L_\infty$ and $RMS$ errors are very small for the present choice of $\Delta x$ and $\Delta t$. The proposed algorithm is easy to implement and gives high accuracy.

Table 1: The $L_\infty$ and $RMSE$ for 10 iterations in solving example 1.1 when ν = 0 .1, μ = 1, σ = 0.9 and λ = 0.4

| $n$ | $L_\infty$ | $RMSE$ |
|---|---|---|
| 1 | 7.4698e-005 | 6.3333e-005 |
| 2 | 1.1112e-004 | 8.8541e-005 |
| 3 | 1.1816e-004 | 8.7748e-005 |
| 4 | 1.0316e-004 | 7.1784e-005 |
| 5 | 6.9997e-005 | 5.9166e-005 |
| 6 | 9.4608e-005 | 7.7668e-005 |
| 7 | 1.3745e-004 | 1.2310e-004 |
| 8 | 1.8733e-004 | 1.7904e-004 |
| 9 | 2.4260e-004 | 2.3822e-004 |
| 10 | 3.0232e-004 | 2.9708e-004 |

Table 2: The $L_\infty$ and $RMSE$ for 10 iterations in solving example 1.2

| $n$ | $L_\infty$ | $RMSE$ |
|---|---|---|
| 1 | 1.8988e-003 | 1.7653e-003 |
| 2 | 1.7074e-003 | 1.5967e-003 |
| 3 | 1.1247e-003 | 1.0571e-003 |
| 4 | 5.8602e-004 | 5.5101e-004 |
| 5 | 1.7681e-004 | 1.6357e-004 |
| 6 | 1.3393e-004 | 1.1810e-004 |
| 7 | 3.2804e-004 | 3.0529e-004 |
| 8 | 4.5922e-004 | 4.3067e-004 |
| 9 | 5.4370e-004 | 5.1192e-004 |
| 10 | 5.9738e-004 | 5.6263e-004 |

Table 3: The $L_\infty$ and $RMSE$ for 20 iterations in solving Test problem 2

| $n$ | $L_\infty$ | $RMSE$ |
|---|---|---|
| 1 | 5.5037e-006 | 5.3485e-006 |
| 2 | 4.2062e-004 | 4.0833e-004 |
| 3 | 7.3628e-004 | 7.1458e-004 |
| 4 | 9.5622e-004 | 9.2783e-004 |
| 5 | 1.1164e-003 | 1.0831e-003 |
| 6 | 1.2392e-003 | 1.2021e-003 |
| 7 | 1.3374e-003 | 1.2971e-003 |
| 8 | 1.4184e-003 | 1.3754e-003 |
| 9 | 1.4867e-003 | 1.4415e-003 |
| 10 | 1.5454e-003 | 1.4983e-003 |
| 11 | 1.5966e-003 | 1.5478e-003 |
| 12 | 1.6419e-003 | 1.5915e-003 |
| 13 | 1.6823e-003 | 1.6304e-003 |
| 14 | 1.7186e-003 | 1.6655e-003 |
| 15 | 1.7516e-003 | 1.6972e-003 |
| 16 | 1.7817e-003 | 1.7262e-003 |
| 17 | 1.8093e-003 | 1.7528e-003 |
| 18 | 1.8349e-003 | 1.7773e-003 |
| 19 | 1.8590e-003 | 1.8000e-003 |
| 20 | 1.8814e-003 | 1.8212e-003 |

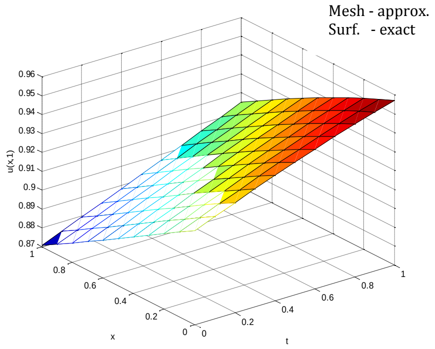


Figure 1: Graphs of approximate and exact solutions after 10 iterations with $\alpha = 0.99, \Delta x = \Delta t = 0.1$

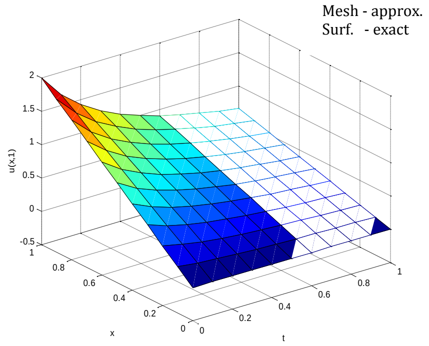


Figure 2: Graphs of approximate and exact solutions after 10 iterations with $\alpha = 0.99, \Delta x = \Delta t = 0.1$

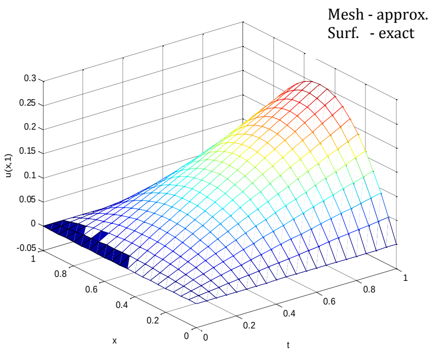


Figure 3: Graphs of approximate and exact solutions after 10 iterations with $\alpha = 0.5, \Delta x = \Delta t = 0.05$

## 4. Conclusion

In this paper, we have presented a scheme used to obtain the approximate solution of the time fractional Burgers' equation. The truly meshless local Petrov-Galerkin (MLPG) approach based on a local weak formulation is applied to the spatial discretization. We employ the moving Kriging (MK) interpolation for constructing shape functions at scattered points, and the Kronecker delta function is chosen as a test function in each sub-domain to avoid the evaluation of any numerical integration in weak form. The time fractional derivative interpreted in the sense of Caputo can be approximated by a simple quadrature formula. Also, the linearization method by Taylor series expansion is used to treat nonlinearity parts of Burgers' equation. The numerical examples are provided to illustrate the accuracy and efficiency of the proposed method. A very good agreement between the analytical and numerical results can be found. It is obvious that the MLPG method can readily be extended to solve the time fractional Burgers' equation as well as several other problems in science and engineering. In the future work, the present MLPG approach will be extended to the two dimensional coupled Burgers' equations.

## 5. Acknowledgements

This work was partially supported by Science Achievement Scholarship of Thailand (SAST). The authors are grateful for the support.

## 6. References

[1] Burger J.M. A Mathematical Model illustrating the Theory of Turbulence. Advances in Applied Mechanics. Academic Press. New York; 1948: 171-199.

[2] Cole J.D. On a quasi-linear parabolic equation occurring in aerodynamics. Quarterly of Applied Mathematics. 1951; 9: 225-236.

[3] Momani S. Non-perturbative analytical solutions of the space- and time-fractional Burgers equations, Chaos Solitons Fractals 28 (2006) 930-937.

[4] Inc M. The approximate and exact solutions of the space- and time-fractional Burgers equations with initial conditions by variational iteration method. Journal of Mathematical Analysis and Applications. 2008; 345: 476-484.

[5] Kurulay M. The approximate and exact solutions of the space and time-fractional Burgers equations. International Journal of Research and Reviews in Applied Sciences. 2010; 3(3): 257-263.

[6] Khan N.A., Ara A. and Mahmood A. Numerical solutions of time-fractional Burger equations: a comparison between generalized transformation technique with homotopy perturbation method. International Journal of Numerical Methods for Heat and Fluid Flow. 2012a; 22(2): 175-193.

[7] Belystchko T., Liu Y.Y. and Gu L. Element-Free Galerkin Methods. International Journal for Numerical Methods in Engineering. 1994; 37: 229-256.

[8] Atluri S.N., Zhu T. A new meshless local Petrov-Galerkin (MLPG) approach in computational mechanics. Computational Mechanics. 1994; 22(2): 117-127.

[9] Liu W.K., Jun S., Li S., Jonathan A. and Belytschko T. Reproducing kernel particle methods for structural dynamics. International Journal for Numerical Methods in Engineering. 1995; 38: 1655-1679.

[10] Liu G.R., Gu Y.T. A point interpolation method for two dimensional solids. International Journal for Numerical Methods in Engineering. 2001; 50: 937-951.

[11] Liu G.R., Gu Y.T. A local radial point interpolation method (LRPIM) for free vibration analyses of 2-D solids, Journal of Sound and Vibration. 2001; 246(1): 29-46.

[12] Wang J.G. and Liu G.R. Radial point interpolation method for elastoplastic problems, Proceedings of the 1st International Conference on Structural Stability and Dynamics; 2000 Dec. 7-9; Taipei, Taiwan. 2000. p.703-708.

[13] Wang J.G., Liu GR. A point interpolation meshless method based on radial basis functions. Issue International Journal for Numerical Methods in Engineering. International Journal for Numerical Methods in Engieering. 2002a; 54(11): 1623-1648.

[14] Atluri S.N., Zhu T.L. A new meshless local Petrov-Galerkin (MLPG) approach in computational mechanics. Computational Mechanics. 1998a; 22(2): 117-127

[15] Atluri S.N., Zhu T.L. A new meshless local Petrov- Galerkin (MLPG) approach to nonlinear problems in computer modeling and simulation.

Computer Modeling and Simulation in Engineering. 1998b; 3(3): 187-196.

[16] Atluri, S.N., Zhu T.L. The meshless local Petrov-Galerkin (MLPG) approach for solving problems in elasto-statics. Computational Mechanics. 2000; 25: 169-179.

[17] Gu L. Moving kriging interpolation and element-free Galerkin method. International Journal for Numerical Methods in Engineering. 2003; 56(1): 1-11

[18] Murio D.A. Implicit finite difference approximation for time fractional diffusion equations. Computers and Mathematics with Applications. 2008; 56: 1138-1145.

[19] Kaya D., Yokus A. A numerical comparison of partial solutions in the decomposition method for linear and nonlinear partial differential equations. Mathematics and Computers in Simulation. 2002; 60(6): 507-512.

# PROCESS CAPABILITY INDEX ($\tilde{\mathrm{C}}_{\mathrm{pk}}$) FOR A MIXED BIVARIATE DATA AS APPLIED TO STRATIFIED PRODUCT GROUPS

Dr. Dennis C. de Paz*

*Mathematics & Statistics Department, Eastern Visayas State University (EVSU-Main Campus), Tacloban City, Leyte/6500, Philippines, depaz_dennis@yahoo.com*

## Abstract

The very nature of process control in contract manufacturing is typically confounded by the magnitude of subcontractor's operations. As the number of the company's product groups increases, so does the number of input and quality parameters that need to be monitored and evaluated. Hence, to avoid proliferation of monitored parameters, stratification is done. Parametric monitoring is modeled using a mixture structure (Y, X) where X represents the product group and Y the monitored characteristic.

The study explored to establish a single characteristic of a bivariate mixture as given by Olkin and Tate (1961). Based on the given unconditional bivariate mixture, a new $\mathrm{C}_{\mathrm{pk}}$, say $\tilde{\mathrm{C}}_{\mathrm{pk}}$, was formulated.

It was observed, that $\hat{\tilde{\mathrm{C}}}_{\mathrm{pk}}$ was a biased estimator of $\tilde{\mathrm{C}}_{\mathrm{pk}}$. Similarly, $\hat{\tilde{\mathrm{C}}}^2_{\mathrm{pk}}$ was also biased estimator of $\tilde{\mathrm{C}}^2_{\mathrm{pk}}$. Since $\hat{\tilde{\mathrm{C}}}_{\mathrm{pk}}$ and $\hat{\tilde{\mathrm{C}}}^2_{\mathrm{pk}}$ are both estimators of their respective parameters, the asymptotic variance of $\hat{\tilde{\mathrm{C}}}_{\mathrm{pk}}$ is also a biased estimator. However, for large n the bias tends to zero.

In comparing the efficiency of $\hat{\tilde{\mathrm{C}}}_{\mathrm{pk}}$ and $\hat{\mathrm{C}}_{\mathrm{pk}}$, based on asymptotic variances, i.e., $\mathrm{Var}\left(\hat{\tilde{\mathrm{C}}}_{\mathrm{pk}}\right)$and $\mathrm{Var}(\hat{\mathrm{C}}_{\mathrm{pk}})$, it is noted that $\mathrm{Var}\left(\hat{\tilde{\mathrm{C}}}_{\mathrm{pk}}\right)$ is less than $\mathrm{Var}(\hat{\mathrm{C}}_{\mathrm{pk}})$. This implies that the derived variance is most appropriate for data coming from the unconditional bivariate mixture.

*Keywords*: Process Capability Index, mixed bivariate data, stratified product groups, biased estimator

*Corresponding Author
E-mail Address: depaz_dennis@yahoo.com

## 1. Introduction

A Process Capability Index (PCI) compares the inherent variability in the output of a process - often measured as $6\sigma$, $\sigma$ being the output standard deviation – with the specification range. One of the most popular indices is $\mathrm{C}_{\mathrm{pk}} = \min\left\{\frac{\mathrm{USL}-\mu}{3\sigma}, \frac{\mathrm{USL}-\mu}{3\sigma}\right\}$, where USL and LSL are the upper and lower specification limits for the output variable. The $C_{pk}$ index is useful as a tool for moving towards the "zero defects" approach (i.e., whether a process conforms to the specification) and towards "never ending reduction in variability" as criteria for excellence. In semiconductor industry, $\mathrm{C}_{\mathrm{pk}}$ is a standard part of manufacturing and has been applied to on-going measurement of manufacturing line process capability as well as to the process maturity evaluation during transfer of new technology from development to production. In addition, it also has been used for customer reports of manufacturing line process capability.

In contract manufacturing environment, process control is typically confounded by the magnitude of a subcontractor's operations, exerting a severe strain on a manufacturer's quality system. As the number of the company's product groups increases, so does the number of input and quality parameters that need to be monitored and evaluated. Hence, this brings to fore the need to obtain viable and economic monitoring system.

One way to avoid the proliferation of monitored parameters is through stratification. In stratification, product groups are clustered according to some criterion (e.g., die size in IC manufacturing) with the intention of establishing a monitoring system for each stratum instead of per product. In this system, product groups within a stratum can be assumed not to vary much with respect to the parameters being monitored, and monitoring is simplified by adopting a single system for the process characteristics of interest. Parametric monitoring in such a set-up can be modeled using a bivariate mixture structure (Y,X) where X represents the product group and Y the characteristics being monitored.

The goal of this paper is to explore how such a set-up can be established for a single characteristic. We do this in the context of a bivariate mixture as given by Olkin and Tate (1961). A new $\mathrm{C}_{\mathrm{pk}}$, say $\tilde{\mathrm{C}}_{\mathrm{pk}}$, is formulated and the efficiency of its asymptotic variance is compared to the asymptotic variance of the usual distribution, i.e., Y is normally distributed with mean $\mu$ and variance $\sigma^2$ with $\mathrm{C}_{\mathrm{pk}}$ as its corresponding index. In addition, interval estimates and hypothesis testing procedures are constructed via the asymptotic distribution of the unconditional bivariate mixture.

### *1.1 Objectives of the Study*

Using the bivariate mixture of Olkin and Tate (1961) as the theoretical underpinning, this study answered the following:

1. To formulate a new $\mathrm{C}_{\mathrm{pk}}$, say $\tilde{\mathrm{C}}_{\mathrm{pk}}$, based on the unconditional bivariate mixture;

2. To formulate an estimate of $\tilde{C}_{pk}$, say $\hat{\tilde{C}}_{pk}$, and find the asymptotic expectation and asymptotic variance of $\hat{\tilde{C}}_{pk}$, i.e., $AE[\hat{\tilde{C}}_{pk}]$ and $AV[\hat{\tilde{C}}_{pk}]$, respectively;
3. To compare the efficiency of $\hat{\tilde{C}}_{pk}$ and $\hat{C}_{pk}$, the estimate of $C_{pk}$ based on a normal unconditional distribution, based on asymptotic variance, i.e., $AV\left(\hat{\tilde{C}}_{pk}\right)$ and $AV(\hat{C}_{pk})$; and
4. To construct confidence interval and hypothesis testing procedure for $\tilde{C}_{pk}$.

### 1.2 Significance of the Study

The study has significant use in the formulation of an alternative $C_{pk}$ index for the specific case when the product set-up utilizes only a single characteristic from a mixed bivariate data. Based on readings, there is no available index of the said set-up in the literature; hence, the study was made.

The results of the study have significant use in semiconductor manufacturing industry: The existing index (i.e., coming from Y which is normally distributed with mean μ and variance $\sigma^2$) will only underestimate or overestimate the true index when used for stratified product groups. The study corrects this bias by proposing an alternative measure.

### 1.3 Scope and Delimitation of the Study

The study covers the formulation of a new $C_{pk}$ index, say $\tilde{C}_{pk}$, derived for a single characteristic from a mixed bivariate data. Asymptotic expectation and asymptotic variance of the said index are obtained. In addition, the asymptotic variance of the estimate of $\tilde{C}_{pk}$ will be compared to the usual estimate of $C_{pk}$ index, i.e., the case when Y is normally distributed with mean $\mu$ and variance $\sigma^2$. Interval estimates and hypothesis testing procedure will be made via the asymptotic distribution of the unconditional bivariate mixture. The paper will limit the discussion in the case when Y|X=i is distributed as normal with mean $\mu_i$ and variance $\sigma_i{}^2$ with P(X=i) = $p_i$ for i=1, 2,..., k such that $\sum_{i=1}^{k} p_i = 1$.

Also, asymptotic distributions was used in the study due to some difficulty in getting the exact distribution of the estimates.

## 2. Research Methodology

The study considered a subgroup of size n whose observations $(Y_i, X_i)$, i = 1,2,...,n, can be categorized into k product groups. These k product groups define a specific stratum in the stratification scheme.

For a stratum consisting of k product groups, the bivariate distribution of (Y, X=i) is given by f(y,i) = $N(\mu_{i,}\ {}_{\sigma_i{}^2})p_i$, i = 1,2,...,k, where $\sum_{i=1}^{k} p_i = 1$. This induces a mixture of k normal distributions as the unconditional distribution of Y.

The mean of this unconditional distribution is $\bar{\mu} = \sum_{i=1}^{k} \mu_i p_i$ and the variance is $\bar{\sigma}^2 = \sum_{i=1}^{k} p_{i[\sigma_i{}^2+(\mu_i-\bar{\mu})^2]}$.

### 2.1 $C_{pk}$ Formulation and Estimation

A likelihood based modification of the $C_{pk}$ will be used to take into account the statistical complication brought about by stratification. Given the parameters $\bar{\mu}$ and $\bar{\sigma}^2$ a simple modification of $C_{pk}$ is given by

$$\tilde{C}_{pk} = \min\left\{\frac{USL-\bar{\mu}}{3\bar{\sigma}}, \frac{LSL-\bar{\mu}}{3\bar{\sigma}}\right\} \tag{1}$$
$$= \frac{d-|\bar{\mu}-\underset{.}{a}|}{3\bar{\sigma}}$$

where USL and LSL are the upper and lower specification limits for the output variable, d = ½ |USL - LSL| and a = ½ (LSL + USL). Letting $\theta = (\bar{\mu}, \bar{\sigma}^2)$, we can look at $\tilde{C}_{pk}(\theta)$ as an estimand, continuous in θ. The estimation requires the use of $\hat{\theta}_n$ such that $\hat{\theta}_n \overset{a.s.}{\rightarrow} \theta$ and $\hat{\theta}_n \overset{a.s.}{\rightarrow} N(\theta, v(\theta))$ so that $\tilde{C}_{pk}(\hat{\theta}_n) \overset{a.s.}{\rightarrow} \tilde{C}_{pk}(\theta)$ and $\tilde{C}_{pk}(\hat{\theta}_n) \overset{d}{\rightarrow} N[\tilde{C}_{pk}(\theta), \bar{\omega}(\theta)]$. [1]

Now, we present a likelihood-based approach in estimating the parameter $\theta = (\bar{\mu}, \bar{\sigma}^2)$, utilizing the MLE's of the mixture distribution. We simplify the approach by using the normal mixture as reference distribution. Hence, the MLE for θ is given by

$$\hat{\theta} = (\hat{\bar{\mu}}, \hat{\bar{\sigma}}^2) \tag{2}$$

where:

$$\hat{\bar{\mu}} = \sum_{i=1}^{k} \hat{p}_i \qquad \hat{\bar{\sigma}}^2 = \sum_{i=1}^{k} \hat{p}_i \left[\hat{\bar{\sigma}}_i{}^2 + (\hat{\mu}_i - \hat{\mu}_i)^2\right]$$
$$\hat{p}_i = \frac{n_i}{n} \qquad \hat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$$
$$\hat{\sigma}_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i}\left(Y_{ij} - \hat{\mu}_i\right)^2$$

The estimate of $\tilde{C}_{pk}(\theta)$ now is defined as

$$\hat{\tilde{C}}_{pk} = \min\left\{\frac{USL-\hat{\bar{\mu}}}{3\hat{\bar{\sigma}}}, \frac{LSL-\hat{\bar{\mu}}}{3\hat{\bar{\sigma}}}\right\} \tag{3}$$
$$= \frac{d-|\hat{\bar{\mu}}-\underset{.}{a}|}{3\hat{\bar{\sigma}}}$$

From here, the asymptotic expectation and the asymptotic variance of $\hat{\tilde{C}}_{pk}$ will be determined. In addition, the efficiency of $\hat{\tilde{C}}_{pk}$ and $\hat{C}_{pk}$, the estimate of $C_{pk}$ based on a normal unconditional distribution, will be compared based on their asymptotic variances, i.e., $AV(\hat{\tilde{C}}_{pk})$ and $AV(\hat{C}_{pk})$. Lastly, confidence interval and hypothesis testing procedure will be made via the asymptotic distribution of the unconditional bivariate mixture.

## 3. Research Results and Discussion

The stratification considered in the previous chapter induces a mixture of k normal distribution as the unconditional distribution of Y with mean $\bar{\mu} = \sum_{i=1}^{k} \mu_i p_i$ and variance $\bar{\sigma}^2 = \sum_{i=1}^{k} p_i[\sigma_i^2 + (\mu_i - \bar{\mu})^2]$. Using the parameters $\bar{\mu}$ and $\bar{\sigma}^2$, a simple modification of $C_{pk}$ is given by

$$\tilde{C}_{pk} = \min\left\{\frac{USL-\bar{\mu}}{3\bar{\sigma}}, \frac{LSL-\bar{\mu}}{3\bar{\sigma}}\right\} \qquad (1)$$
$$= \frac{d - |\bar{\mu} - \underset{.}{a}|}{3\bar{\sigma}}$$

where USL and LSL are the upper and lower specification limits for a given output variable, d = ½ |USL − LSL|, and a = ½ (USL + LSL).

Utilizing the MLE's of the mixture distribution, an estimator of $\tilde{C}_{pk}$ is as follows:

$$\hat{\tilde{C}}_{pk} = \min\left\{\frac{USL-\hat{\bar{\mu}}}{3\hat{\bar{\sigma}}}, \frac{LSL-\hat{\bar{\mu}}}{3\hat{\bar{\sigma}}}\right\} \qquad (3)$$
$$= \frac{d-|\hat{\bar{\mu}}-\underset{.}{a}|}{3\hat{\bar{\sigma}}}$$

where: d = ½|USL − LSL|
a = ½ (LSL-USL)
$\hat{\bar{\mu}} = \sum_{i=1}^{k} \mu_i p_i$
$\hat{\bar{\sigma}}^2 = \sum_{i=1}^{k} p_i[\hat{\sigma}_i{}^2 + (\hat{\mu}_i - \hat{\bar{\mu}})^2]$.

Note that $\hat{p}_i = \frac{n_i}{n}$

$\hat{\mu}_i = \frac{1}{n_i}\sum_{i=1}^{n_i} Y_{ij}$

$\hat{\sigma}_i^2 = \frac{1}{n_i}\sum_{i=1}^{n_i}\left(Y_{ij} - \hat{\mu}_i\right)^2$

Reference [1] showed that:

a) $\hat{\bar{\mu}} \sim AN\left(\bar{\mu}, {}^{\bar{\sigma}^2}/_{n}\right)$ ; and

b) $\hat{\bar{\sigma}} \sim AN\left(\bar{\sigma}, \frac{\mu_4 - \bar{\sigma}^4}{4n\bar{\sigma}^2}\right)$.

The above results are essential in finding the asymptotic expectation and asymptotic variance of $\hat{\tilde{C}}_{pk}$. In what follows, we assume that $\hat{\tilde{C}}_{pk}$ is uniformly integrable. We start with the following lemma.

**Lemma 1**. Let Y be distributed as Mixture Normal with mean $\bar{\mu}$ and variance $\bar{\sigma}^2$. Then $\hat{\bar{\mu}}$ and $\hat{\bar{\sigma}}$ are independent.

**Lemma 2.** Let Y be distributed as Mixture Normal with mean $\bar{\mu}$ and variance $\bar{\sigma}^2$. Then the asymptotic expectation of $\hat{\bar{\sigma}}^{-1}$ is $AE[\hat{\bar{\sigma}}^{-1}] = \frac{1}{\bar{\sigma}} + \left[\frac{\bar{\mu}_4 - \bar{\sigma}^4}{4n\bar{\sigma}^5}\right]$.

**Theorem 1**. Given the assumptions (a) – (b) under the mixture set-up and assuming also that $\hat{\tilde{C}}_{pk}$ is uniformly integrable, the asymptotic expression of $\hat{\tilde{C}}_{pk}$ is

$$AE\left[\hat{\tilde{C}}_{pk}\right] = \frac{1}{3}\left(\frac{1}{\bar{\sigma}} + \frac{(\mu_4 - \bar{\sigma}^4)}{4n\bar{\sigma}^5}\right)\left\{d - \bar{\sigma}\sqrt{\frac{2}{n\pi}}\exp\left\{-\frac{n}{2\bar{\sigma}^2}(\bar{\mu} - a)^2\right\} - |\bar{\mu} - a|\left[1 - 2\Phi\left(-\frac{\sqrt{n}}{\bar{\sigma}}(\bar{\mu} - a)\right)\right]\right\} \qquad (4)$$

Note that $\hat{\tilde{C}}_{pk}$ is a biased estimator of $\tilde{C}_{pk}$. The bias arises from two sources:

a) $AE[\hat{\bar{\sigma}}^{-1}] = \frac{1}{\bar{\sigma}} + \left\{\frac{(\bar{\mu} - \bar{\sigma}^4)}{4n\bar{\sigma}^5}\right\} \neq \frac{1}{\bar{\sigma}}$

b) $AE[|\bar{\mu} - \underset{.}{a}|] \neq |\bar{\mu} - \underset{.}{a}|$.

However, as n→∞ the bias disappers.

**Lemma 3**. Let Y be distributed as Mixture Normal with mean $\bar{\mu}$ and variance $\bar{\sigma}^2$. Then, the asymptotic expectation of $\hat{\bar{\sigma}}^{-2}$ is

$$AE[\hat{\bar{\sigma}}^{-2}] = \frac{1}{\bar{\sigma}^2} + 3\left\{\frac{\bar{\mu} - \bar{\sigma}^4}{n\bar{\sigma}^4}\right\}. \qquad (5)$$

**Lemma 4**. Let Y be distributed as Mixture Normal with mean $\bar{\mu}$ and variance $\bar{\sigma}^2$. Then, the asymptotic expectation of $\hat{\bar{\mu}}^2$ is

$$AE[\hat{\bar{\mu}}^2\ ] = \bar{\mu}^2 + \left\{\frac{\bar{\sigma}^2}{n}\right\}. \qquad (6)$$

**Theorem 2.** Let $\hat{\tilde{C}}_{pk}^2$ be uniformly integrable, the asymptotic variance of $\hat{\tilde{C}}_{pk}$ of the mixture set-up is

$$AV\left[\hat{\tilde{C}}_{pk}\ \right] = \frac{1}{9}\left(\frac{1}{\bar{\sigma}^2} + \frac{3(\mu_4 - \bar{\sigma}^4)}{\bar{\sigma}^4 n}\right) \times$$
$$\left\{d^2 - 2d\left(\bar{\sigma}\sqrt{\frac{2}{n\pi}}\exp\left\{-\frac{n}{2\bar{\sigma}^2}(\bar{\mu} - \underset{.}{a})^2\right\} - |\bar{\mu} - \underset{.}{a}|\left[1 - 2\Phi\left(-\frac{\sqrt{n}}{\bar{\sigma}}(\bar{\mu} - \underset{.}{a})\right)\right]\right) + (\bar{\mu} - \underset{.}{a})^2\right\}$$
$$-\left\{AE\left[\hat{\tilde{C}}_{pk}\right]\right\}^2 \qquad (7)$$

where $AE\left[\hat{\tilde{C}}_{pk}\ \right]$ is given in equation (4).

Note that $\hat{\tilde{C}}_{pk}^2$ is a biased estimator of $\tilde{C}_{pk}^2$. The bias arises from two sources:

a) $AE[\hat{\bar{\sigma}}^{-2}] = \frac{1}{\bar{\sigma}^2} + 3\left\{\frac{(\mu_4 - \bar{\sigma}^4)}{\bar{\sigma}^4 n}\right\} \neq \frac{1}{\bar{\sigma}^2}$

b) $AE[(\hat{\bar{\mu}} - \underset{.}{a})^2] = (\bar{\mu} - \underset{.}{a})^2 + \frac{\bar{\sigma}^2}{n} \neq (\bar{\mu} - \underset{.}{a})^2$.

However, as n → ∞ the bias tends to zero.

Results that follows determine what will happen if we insist using the $\hat{C}_{pk}$ index where in fact we know that our data is coming from a mixture set-up.

**Lemma 5**. Let Y be distributed as MixN $(\bar{\mu}, \bar{\sigma}^2)$. The mean and the variance now becomes

$$\bar{\bar{Y}} = \sum_{i=1}^{k} \bar{Y}_i \hat{p}_i \qquad (9)$$

$$S^2 = \frac{n}{n-1}\hat{\bar{\sigma}}^2. \qquad (10)$$

**Theorem 3.** Let Y be distributed as MixN $(\bar{\mu}, \bar{\sigma}^2)$ and assume that $\hat{C}_{pk}$ is uniformly integrable. The asymptotic expectation of $\hat{C}_{pk}$ now becomes

$$AE[\hat{C}_{pk}] = \sqrt{\frac{n-1}{n}}\,AE\left[\hat{\tilde{C}}_{pk}\right]$$

where $AE\left[\hat{\tilde{C}}_{pk}\right]$ is given in equation (4)

Notice that if we insist on using the $\hat{C}_{pk}$ index even that we know that our data is coming from a mixture set-up it will only underestimate the true value of the $\hat{\tilde{C}}_{pk}$index in an asymptotic sense. Which means that the $\widehat{\mathrm{C}}_{\mathrm{pk}}$ is not applicable to be used when the set-up is mixture.

**Theorem 4.** Let Y be distributed as MixN $(\overline{\mu}, \overline{\sigma}^2)$ and assume that ${\widehat{\mathrm{C}}_{\mathrm{pk}}}^2$ is uniformly integrable. The asymptotic variance of $\widehat{\mathrm{C}}_{\mathrm{pk}}$ is equal to

$$\mathrm{AV}\left[\widehat{\mathrm{C}}_{\mathrm{pk}}\right] = \frac{\mathrm{n}-1}{\mathrm{n}}\mathrm{AV}\left[\widehat{\widetilde{\mathrm{C}}}_{\mathrm{pk}}\right] \qquad (11)$$

where $\mathrm{AV}\left[\widehat{\widetilde{\mathrm{C}}}_{\mathrm{pk}}\right]$ is given in equation (7)

Proof:

In solving for the variance of $\widehat{\mathrm{C}}_{\mathrm{pk}}$, we only need to solve for the asymptotic expectation of $\widehat{\mathrm{C}}^2_{\mathrm{pk}}$. That is,

$$\mathbf{AE}[\hat{\mathbf{C}}^2_{\mathbf{pk}}] = \frac{1}{9}\ \mathbf{AE}[\,\mathbf{S}^{-2}]\ \mathbf{AE}[\,(\mathbf{d} - \left|\hat{\mu} - \mathbf{a}\right|)^2\,]$$

Again, $AE[(\mathrm{d} - |\hat{\mu} - \underset{\cdot}{\mathrm{a}}|)^2]$ was already derived earlier and so we are left with finding the $\mathrm{AE}[\mathrm{S}^{-2}]$. Now,

$$\mathrm{AE}[\mathrm{S}^{-2}] = \frac{\mathrm{n}-1}{\mathrm{n}}\mathrm{AE}[\widehat{\overline{\sigma}}^2]$$

$$= \frac{\mathrm{n}-1}{\mathrm{n}}\left[\frac{1}{\overline{\sigma}^2} + \frac{3(\mu_4 - \overline{\sigma}^4)}{n\sigma^4}\right]$$

Substituting the results, we have

$$AE\left[\widehat{\mathrm{C}}^2_{\mathrm{pk}}\right] = \frac{1}{9}\frac{\mathrm{n}-1}{\mathrm{n}}\left[\frac{1}{\widehat{\overline{\sigma}}^2} + \frac{3(\mu_4 - \widehat{\overline{\sigma}}^4)}{n\widehat{\overline{\sigma}}^4}\right] \times$$
$$\left\{d^2 - 2d\left(\sqrt{\frac{2}{n\pi}}exp\left\{-\frac{n}{2\overline{\sigma}^2}(\overline{\mu} - \underset{\cdot}{\mathrm{a}})^2\right\} - \right.\right.$$
$$\left.\left.|\overline{\mu} - \underset{\cdot}{\mathrm{a}}|\left[1 - 2\Phi\left(-\frac{\sqrt{\mathrm{n}}}{\overline{\sigma}}(\overline{\mu} - \underset{\cdot}{\mathrm{a}})\right)\right]\right) + (\overline{\mu} - \underset{\cdot}{\mathrm{a}})^2\right\}$$

$$\mathrm{AE}\left[\widehat{\mathrm{C}}^2_{\mathrm{pk}}\right] = \frac{\mathrm{n}-1}{\mathrm{n}}\mathrm{AE}\left[\widehat{\widetilde{\mathrm{C}}}^2_{\mathrm{pk}}\right].$$

Thus,

$$\mathrm{AV}\left[\hat{\mathrm{C}}_{\mathrm{pk}}\right] = \frac{\mathrm{n}-1}{\mathrm{n}}\mathrm{AE}\left[\widehat{\widetilde{\mathrm{C}}}^2_{\mathrm{pk}}\right] - \left\{\frac{\mathrm{n}-1}{\mathrm{n}}\mathrm{AE}\left[\widehat{\widetilde{\mathrm{C}}}_{\mathrm{pk}}\right]\right\}^2$$
$$= \frac{\mathrm{n}-1}{\mathrm{n}}\left\{\mathrm{AE}\left[\widehat{\widetilde{\mathrm{C}}}^2_{\mathrm{pk}}\right] - \left(\mathrm{AE}\left[\widehat{\widetilde{\mathrm{C}}}_{\mathrm{pk}}\right]\right)^2\right\}$$
$$\mathrm{AV}\left[\hat{\mathrm{C}}_{\mathrm{pk}}\right] = \frac{\mathrm{n}-1}{\mathrm{n}}\mathrm{AV}\left[\widehat{\widetilde{\mathrm{C}}}_{\mathrm{pk}}\right]. \qquad \blacksquare$$

So, we already have derived the two asymptotic variances from two different set- ups. One, was in the case for the mixture set-up and the other one was in the case wherein we insisted in using the usual index (i.e., from a normally distributed random variable) where in fact we know that our data is coming from a mixture set-up. Also, the result of Kotz and Johnson (1993) was presented right after the asymptotic variance of the mixture set-up. These results were taken for comparison to the mixture set-up. Our next goal now is to show that the asymptotic variance of $\hat{\mathrm{C}}_{\mathrm{pk}}$ (as defined in 1.4) is greater compared to the asymptotic variance of $\hat{\tilde{\mathrm{C}}}_{\mathrm{pk}}$ (as defined in 1.2). This is presented in the theorem below.

**Theorem 5.** The asymptotic variance of $\hat{\mathrm{C}}_{\mathrm{pk}}$ is greater compared to the asymptotic variance of $\hat{\tilde{\mathrm{C}}}_{\mathrm{pk}}$, i.e.,

$$\mathrm{AV}\left[\hat{\mathrm{C}}_{\mathrm{pk}}\right] > \mathrm{AV}\left[\hat{\tilde{\mathrm{C}}}_{\mathrm{pk}}\right].$$

The result above showed that $\hat{\mathrm{C}}_{\mathrm{pk}}$ is as good as $\hat{\tilde{\mathrm{C}}}_{\mathrm{pk}}$ as n → ∞ but there is a difference when n is small.

In our previous results, we have come-up with estimate of the parameter $\tilde{\mathrm{C}}_{\mathrm{pk}}$, or more precisely, point estimation of value of a function of the $\tilde{\mathrm{C}}_{\mathrm{pk}}$. Our next goal now is to construct a confidence interval for $\tilde{\mathrm{C}}_{\mathrm{pk}}$.

**Theorem 6.** Let $\hat{\tilde{\mathrm{C}}}_{\mathrm{pk}}$ be distributed as asymptotic normal and that the asymptotic expectation and asymptotic variance of $\hat{\tilde{\mathrm{C}}}_{\mathrm{pk}}$ is given in equations 1.1 and 1.2 respectively. An approximate 100(1 – α)% confidence interval for $\tilde{\mathrm{C}}_{\mathrm{pk}}$ is

$$\hat{\tilde{\mathbf{C}}}_{\mathbf{pk}} \pm \mathbf{z}\sqrt{\mathbf{V}[\hat{\tilde{\mathbf{C}}}_{\mathbf{pk}}]}\ .$$

We have just come-up an approximate 100(1-α)% confidence interval for $\tilde{\mathrm{C}}_{\mathrm{pk}}$. This result can be extended to hypothesis testing. As the term suggests, one wishes to decide whether or not some hypothesis that has been formulated is correct. The choice here lies between only two decisions: accepting or rejecting the hypothesis. Two hypotheses are discussed here: First, the hypothesis being tested, is called the null hypothesis, denoted by Ho, and the second is called the alternative hypothesis, denoted by Ha. The thinking is that if the null hypothesis is false, then the alternative hypothesis is true, and vice versa. In our case, we want to test the hypothesis that the parameter $\tilde{\mathrm{C}}_{\mathrm{pk}}$ is equal to some specified value. This is presented on theorem form below.

**Theorem 7.** Let $\hat{\tilde{\mathrm{C}}}_{\mathrm{pk}}$ be distributed as asymptotic normal and that the asymptotic expectation and asymptotic variance of $\hat{\tilde{\mathrm{C}}}_{\mathrm{pk}}$ is given in equations 1.1 and 1.2, respectively. We test the hypothesis that $\tilde{\mathrm{C}}_{\mathrm{pk}} = \tilde{\mathrm{C}}_{\mathrm{pk}}(0)$ against the alternative hypothesis that $\tilde{\mathrm{C}}_{\mathrm{pk}} \neq \tilde{\mathrm{C}}_{\mathrm{pk}}(0)$ and the corresponding statistic is given below:

$$Z = \frac{\tilde{\mathrm{C}}_{\mathrm{pk}} - \tilde{\mathrm{C}}_{\mathrm{pk}}(0)}{\sqrt{\mathrm{V}[\hat{\tilde{\mathrm{C}}}_{\mathrm{pk}}]}}$$

The test rejects the null hypothesis if and only if: $\mathrm{Z} > \mathrm{Z}_{1-\alpha/2}$ or $\mathrm{Z} > -\mathrm{Z}_{1-\alpha/2}$.

## 4. Conclusion and Recommendation

Based on the given unconditional mixture, a new $\mathrm{C}_{\mathrm{pk}}$, say $\tilde{\mathrm{C}}_{\mathrm{pk}}$, was formulated. An estimator of $\tilde{\mathrm{C}}_{\mathrm{pk}}$, denoted as $\hat{\tilde{\mathrm{C}}}_{\mathrm{pk}}$, was formulated as:

$$\hat{\tilde{\mathbf{C}}}_{\mathbf{pk}} = \frac{\mathbf{d} - |\hat{\overline{\boldsymbol{\mu}}} - \underset{\cdot}{\mathbf{a}}|}{\mathbf{3}\hat{\overline{\boldsymbol{\sigma}}}}.$$

Taking the asymptotic expectation of the above equation resulted in:

$$\mathrm{AE}\left[\hat{\tilde{C}}_{pk}\right]=\frac{1}{3}\left(\frac{1}{\bar{\sigma}}+\frac{(\mu_4-\bar{\sigma}^4)}{4n\bar{\sigma}^5}\right)\left\{d-\bar{\sigma}\sqrt{\frac{2}{n\pi}}\exp\left\{-\frac{n}{2\bar{\sigma}^2}(\bar{\mu}-a)^2\right\}-|\bar{\mu}-a|\left[1-2\Phi\left(-\frac{\sqrt{n}}{\bar{\sigma}}(\bar{\mu}-a)\right)\right]\right\}.$$

However, it was observed, that $\hat{\tilde{C}}_{pk}$ was a biased estimator of $\tilde{C}_{pk}$. The bias arose from two sources, i.e.,

i) $\mathrm{AE}[\hat{\bar{\sigma}}^{-1}]=\frac{1}{\bar{\sigma}}+\left\{\frac{\mu_4-\bar{\sigma}^4}{4n\bar{\sigma}^5}\right\}\neq\frac{1}{\bar{\sigma}}$ and
ii)$\mathrm{AE}[|\hat{\bar{\mu}}-\underset{\cdot}{a}|]\neq|\hat{\bar{\mu}}-\underset{\cdot}{a}|$.

The derived asymptotic variance of the unconditional bivariate mixture follows:

$$\mathrm{AV}\left[\hat{\tilde{C}}_{pk}\right]=\mathrm{AE}\left[\hat{\tilde{C}}_{pk}^2\right]-\left\{\mathrm{AE}\left[\hat{\tilde{C}}_{pk}\right]\right\}^2$$

$$=\frac{1}{9}\left(\frac{1}{\bar{\sigma}^2}+\frac{3(\mu_4-\bar{\sigma}^4)}{\bar{\sigma}^4 n}\right)\times$$

$$\left\{d^2-2d\left(\bar{\sigma}\sqrt{\frac{2}{n\pi}}\exp\left\{-\frac{n}{2\bar{\sigma}^2}(\bar{\mu}-\underset{\cdot}{a})^2\right\}\right.\right.$$

$$\left.-|\bar{\mu}-\underset{\cdot}{a}|\left[1-2\Phi\left(-\frac{\sqrt{n}}{\bar{\sigma}}(\bar{\mu}-\underset{\cdot}{a})\right)\right]\right)$$

$$\left.+(\hat{\bar{\mu}}-\underset{\cdot}{a})^2+\frac{\bar{\sigma}^2}{n}\right\}-\left\{\mathrm{AE}\left[\hat{\tilde{C}}_{pk}\right]\right\}^2.$$

It can be observed from the above equation that $\hat{\tilde{C}}_{pk}^2$ was also a biased estimator of $\tilde{C}_{pk}^2$. The bias arose from two sources:

a) $\mathrm{E}[\hat{\bar{\sigma}}^{-2}]=\frac{1}{\bar{\sigma}^2}+3\left\{\frac{(\mu_4-\bar{\sigma}^4)}{\bar{\sigma}^4 n}\right\}\neq\frac{1}{\bar{\sigma}^2}$ ; and

b) $\mathrm{AE}[(\hat{\bar{\mu}}-\underset{\cdot}{a})^2]=(\bar{\mu}-\underset{\cdot}{a})^2+\frac{\bar{\sigma}^2}{n}\neq(\bar{\mu}-\underset{\cdot}{a})^2$.

Since $\hat{\tilde{C}}_{pk}$ and $\hat{\tilde{C}}_{pk}^2$ are both bias estimators of their respective parameters, the asymptotic variance of $\hat{\tilde{C}}_{pk}$ is also a biased estimator. However, for a large n the bias tends to zero.

In comparing the efficiency of $\hat{\tilde{C}}_{pk}$ and $\hat{C}_{pk}$, based on asymptotic variances, i.e., Var($\hat{\tilde{C}}_{pk}$) and Var($\hat{C}_{pk}$), it is noted that Var($\hat{\tilde{C}}_{pk}$) is less than Var($\hat{C}_{pk}$). This implies that the derived variance is most appropriate for data coming from the unconditional bivariate mixture.

In the case where it was insisted that the $\hat{C}_{pk}$ index be used, even if the data is coming from the unconditional bivariate mixture, the results showed that the asymptotic expectation of $\hat{C}_{pk}$ underestimates the asymptotic expectation of $\hat{\tilde{C}}_{pk}$. The same thing happened with regard to the asymptotic variance, which means that there is really a need for a new index intended specifically for the unconditional bivariate mixture.

The author then recommends that the derived asymptotic expectation and asymptotic variance of $\hat{\tilde{C}}_{pk}$ be used specifically if the data is coming from the unconditional bivariate mixture.

## Acknowledgment

Truly, our Almighty God continuously gives the author the courage, strength, and determination to overcome failing interests and moments of despair, has manifested His being through the generosity of the following whom I sincerely thankful:

Dr. Daniel C. Bonzo – UP Diliman Professor; Dr. Dominador O. Aguirre, Jr. - University President – EVSU; Dr. Segunda A. Lacaba, Vice President for Academic Affairs; and Dr. Felixberto E. Avestruz, Vice-President for Research Development and Extension.

Commision on Higher Education (CHED – RO8) for the financial assistance.

Dr. Jacqueline Saño-de Paz, my loving wife, Philippe Dominic Saño-de Paz, my dearest son and least to forget my lovely daughter, Miraj Claire Denine Saño-de Paz (in God's bossom).

## References

[1] Bonzo DC, De Leon, AR. Likelihood-Based Limits and $C_{pk}$ Measure for Stratified Product Groups. Proceedings of MASM, 2000.

[2] Olkin I, Tate RF. Multivariate Correlation Models with Mixed Discrete and Continuos Variables. The Annals of Mathematical Statistics. 1961; 32(2): 448-465.

# Forecasting early warning in Thai real estate crisis

Pichanun Boonpromgul[1], Atcharawan Ngarmyarn[2] and Yawaman Metapirak[3]
[1] *Program in Real Estate Business, Thammasat University, Bangkok, 10200 Thailand, pichanun.b@gmail.com*
[2] *Program in Real Estate Business, Thammasat University, Bangkok, 10200 Thailand, a_ngarmyarn@yahoo.com*
[3] *Program in Real Estate Business, Thammasat University, Bangkok, 10200 Thailand, yawaman@tbs.tu.ac.th*

## Abstract

Real estate business growth is important for the Thai economy as seen from 5.45% of GDP in 2014. This study attempts to find out leading indicator variables to forecast real estate crisis and cycle. The methodology is to use quantitative and qualitative information. Quantitative data is to use vector autoregressive (VAR) to figure out the variables with the most effective and suitable lagging periods. There are four selected variables, extracted from three case studies of the past real estate crisis, Sweden in 1992, Thailand in 1997, and USA in 2008. This study uses 74 quarterly data from 1995/Q3 to 2014/Q3. The dependent variable is a growth of Thai real estate value (RE-VALUE) with quarterly percent change of growth of the Thai stock market (SET), growth of real gross domestic product (GDP) and growth of minimum loan rate (MLR). The result is when the 1st lag of RE-VALUE increase 1 %, it will decrease the current quarter of RE-VALUE by 0.4391% at 90% significant. Qualitative data analysis is done by in-depth interview with selected six experts on four topics, definition of real estate business cycle, the upcoming real estate crisis characteristics in Thailand, opinion in speculation, opinion in macroeconomic variables which affect the real estate business. The result of VAR shows that Q4/2015 is the last quarter of growth in RE-VALUE. Subsequently, this ratio will decline in a linear movement. Forecasting shows growth of GDP is less than 0.5% from Q2/2017 which relates to qualitative analysis that GDP is significant to the growth of real estate business. In reality, there are some warning signals, for example extreme increase of sale promotion, price discount, and stopping price soar.

*Keywords*: Real estate crisis, oversupply, price bubble, real estate cycle, leading indicator, early warning

*Corresponding Author
E-mail Address: pichanun.b@gmail.com

## 1. Introduction

Currently, the growth of real estate business is important for the Thai economy. Noticeably, real estate and construction listed companies have high value, 5,173 MB which is 18.79% of overall stock market, retrieved on 10 March 2015. Moreover, real estate and related business is value at 276,709 MB annually, which is 2.27% of GDP. The construction sector is 387,249 MB, 3.18% of GDP. Thus, the two sections are 5.45% of GDP [1].

Real estate value transaction comparing with the real estate loan value for buyer from a bank is going in the same direction (Fig. 1). This implies that most of the higher transaction value is for a loan. Buyers buy properties by using loan as much as available. In 2014, 10.8 million of Thai households are indebted from total 20 million households. In household debt, there is home loan 36.8%, 60,016 baht per household [2]. Moreover, the shrink of GDP will be a catalyst for real estate business slow down (Fig 2). If the real estate business enter a recession period, it will affect overall economic system. The shrink of home buyer purchasing power will occur after the economic slowdown. However many institutes and private companies do the research for estimating growth of real estate they always refer to different variables. Thus, the question of this study is what variables are leading indicator to forecast real estate crisis and real estate cycle.

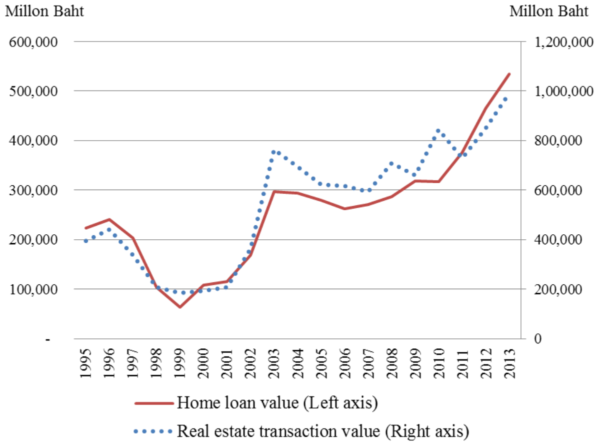


Figure 1: Comparison Home loan value and real estate transaction value 1995-2013 in Thailand

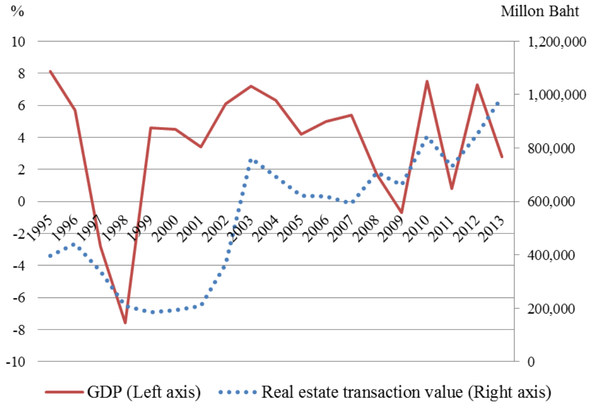


Figure 2: Comparison GDP and real estate transaction value 1995-2013 in Thailand

## 2. Main text

### *2.1 Literature review*

In this part, there are three parts, explanation of the real estate cycle definition, explanation the four past real estate crisis, and explanation selected variables in other studies.

### *2.1.1 Explanation of real estate cycle definition*

Vanichvatana [3] clarifies that real cycle has four periods.

1. The expansion phase - House and land price quickly rise because of developers and speculators aggressively buy. Supply is limited. Buyers buy properties for investment instead of living.

2. The contraction phase - After the peak point, supply is over. There are more sellers than buyers. House price stops increasing because the price is higher comparing with buyer's income. Developers stop buying new land and also new project development.

3. The recession phase - There are several negative signs, economic recession, decline of stock price, deflation, high interest rate, and high unemployment rate. Those signals lead to a real estate crisis. Buyers lose their confidence that this situation will be short term in V shape or long term in U shape of the graph.

4. The recovery phase - Economic starts recover. House inventory is decrease. It takes longer time to finish new houses. The seller has the power to negotiate prices. Economic investment is back.

### *2.1.2 Explanation of the four past real estate crisis*

All three selected cases have the same pattern which is rising land price and house price for 4-6 years for 2-3 times, then economic suddenly shrink and purchasing power is lost, due to increase of interest rate. Moreover, during to rising price, most of the buyers are disaster myopia (Appendix 2).

### *2.1.3 Explanation variables were employed in other fields*

Hussein [4] uses eight factors that cause real estate crisis which are vacancy rate, absorption rate, time lag, type of property, property present value, government intervention, gross domestic product (GDP).

Vanichvatana [3] uses factors in two types, macroeconomic and microeconomic. Macroeconomic factors are GDP, employment rate, financial factors, capital factors, and geographic factors. Microeconomic factor is housing start.

Kim&Lee [5] use by matching correlation coefficient of these factors, land price, nominal GDP, consumer price index, stock price index, wage, real GDP. The high correlation is land price and GDP, land price and CPI.

Pholphibul&Rukumnuaykit [6] use construction price index, money supply, property stock index, and post credit finance.

### *2.2 Methodology*

This study uses quantitative and qualitative information. Quantitative data use vector autoregressive (VAR) to figure out the variables with the most effective and suitable lagging periods. There are four selected variables, extracted from three case studies of the past real estate crisis, Sweden in 1992, Thailand in 1997, and USA in 2008. This study uses 74 quarterly data from 1995/Q3 to 2014/Q3. The pairwise correlation test is to test each pair to see how much it relates each other, the result is RE-VALUE has high relate with GDP in the same direction. On the other hand RE-VALUE relate to new house register in a negative direction. However, these variables are selected and considered together with the result from qualitative data, because the chosen statistic coincides with the period of crisis. So, it might contain abnormal information. The result is there are four variables selected. The dependent variable is a percent change in growth of Thai real estate value (RE-VALUE). During the 1997 real estate crisis, value of RE-VALUE was lost dramatically. Some buyers were unable to continue paying installments, transactions were thus turned into non-performing loan (NPL). NPL took 18 quarters for recovery and brought back the value of RE-VALUE to match with before the crisis. NPL is one of the main obstructions for economic acceleration. The remaining variables are quarterly percent change in growth of the Thai stock market (SET), quarterly percent change in growth of real gross domestic product (GDP) and quarterly percent change in growth of minimum loan rate (MLR).

*2.3 Result*

*2.3.1 Equations*

The first test is a unit root test, Augmented Dickey-fuller test, in order to confirm that all variables are stationary. Lag order, this study use the Akaike information criterion (AIC) method which is popular for VAR equation. The result of AIC is the 4th lag order. Cointegration test is to test in long term equilibrium. The hypothesis is no cointegration. The result is that Trace statistic is higher than the critical value, it means denying the hypothesis. All selected variables have cointegration. Vector autoregressive (VAR) is the main equation for forecasting early warning, coefficient of each lag variable normally use significant at 95%. For this study allow significant at 90% in order to obtain more variables.

$$Y_{1,t} = a^{0}_{11} + a^{1}_{11}y_{1,t-1} + a^{1}_{12}y_{2,t-2} + \ldots\ldots + e_{1,t} \quad (1)$$

$$RE\text{-}VALUE_{t} = -0.4391^{*}\ RE\text{-}VALUE_{t-1} + 0.2532^{*}\ SET_{t-3} + 0.0476^{*}\ MLR_{t-3} \quad (2)$$

$$SET_{t} = 0.2627^{*}\ RE\text{-}VALUE_{t-2} + 0.2740^{*}\ RE\text{-}VALUE_{t-3} - 0.0527^{*}\ MLR_{t-4} \quad (3)$$

$$GDP_{t} = 1.3098^{*} + 6.8209\ {}^{*}SET_{t-1} + 4.5532^{*}\ SET_{t-4}\ 0.3533^{*}\ GDP_{t-1} - 0.6039^{*}\ GDP_{t-2} - 0.3245\ {}^{*}GDP_{t-3} + 0.2698^{*}\ GDP_{t-4} + 0.9048^{*}\ MLR_{t-2} \quad (4)$$

$$MLR_{t} = 0.3591^{*}MLR_{t-1} - 0.2387^{*}\ MLR_{t-2} \quad (5)$$

*Where * is significant at 90%*

*2.3.2 Qualitative data analysis*

Qualitative data analysis does by in-depth interview. The interviewees are six experts from relevant subjects with four topic questions, definition of real estate business cycle, expectation of similarity of the upcoming real estate crisis in Thailand comparing with the last crisis, opinion in pseudo demand and speculation, and opinion in macroeconomic variables which affect to real estate business. To sum up, cycle of real estate in Thailand is not obvious. It potentially similar to the other business one cycle is around 7-9 years. The real estate cycle corresponds with GDP and interest rate. To compare Thai and US interest rate before and after crisis, the pattern before crisis interest rate climbs up slowly. Definitely, after crisis government need to reduce interest rate to stimulate economic drive. Most interviewees give the definition of crisis that the unbalancing of demand and supply. When the price bubble ended, over supply is more obvious. Developers cannot sale their assets. Buyers have higher risks. The risk might be buyers will not get properties as they had paid. Buying real estate will make long term debt, buyers must be confident that they can pay interest continually. If the economy starts to slow down, it will affect sales volume of real estate. In their opinion, speculative is normal in this type of business. Speculative always occur when interest rate is low. In order to keep the balance of three types of customer, real demand, investor, and speculator, developers must maintain a high percentage ratio of real demand.

## 3. Conclusion

The result of VAR shows Q4/2015 is the last quarter of growth in RE-VALUE. Subsequently, this ratio will decline in a linear movement (Fig. 3). RE-VALUE will have small cycle, but each peak will not be higher than 2015q4. This implies that Thai real estate business will enter a recession phase which might be in U shape because the forecast does not show any turning point. The forecast shows that the growth of GDP is less than 0.5% from Q2/2017. This relates to qualitative analysis that GDP is significant to the growth of real estate business. Before the real estate crisis in 1997, growth of real GDP q-o-q is less than 1.0 % to minus since Q2/1995 until Q3 / 1998, 14 quarters (Fig. 4). Impulse response function test shows change in RE-value in the past one unit affect 1st quarter in negative direction and then 2nd-6th quarter will affect in positive direction. 7 th quarter onward will be long term equilibrium (Fig. 5).

Nowadays, there are some warning signals, for example an extreme increase of sale promotion for condominium, price discount for low rise projects, stopping price soar. Those circumstances indicate the forecast result of the research. However, Thai developers can adjust their projects and have a better decision due to the past experience. Thus, the new real estate crisis might not be happen due to experience of developers**.**

The study limitation is another economic effect which suddenly happens such as new issue of an official development plan.

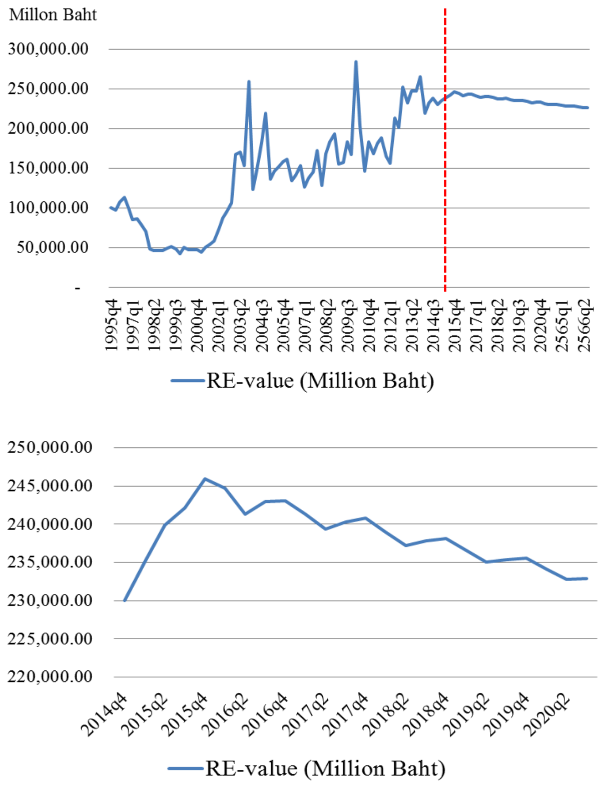


Figure 3: RE-VALUE from 1995q4 – 2020q2 (Top), Forecasting of RE-VALUE from 2014q4-2020q2 (Bottom)

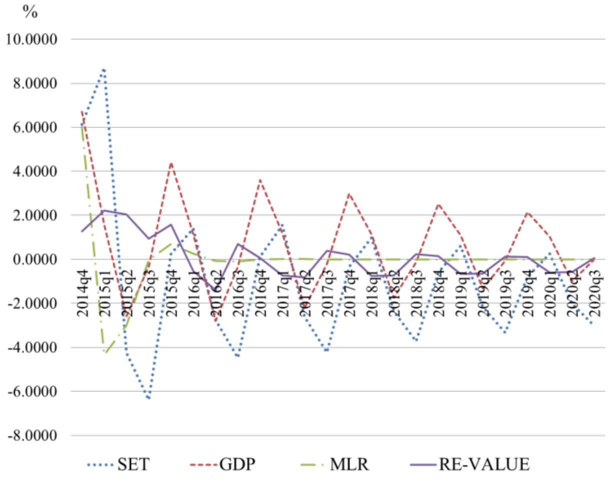


Figure 4: The result of forecasting percent change of SET, MLR, GDP and real estate transaction value 2014q4 -2020q3

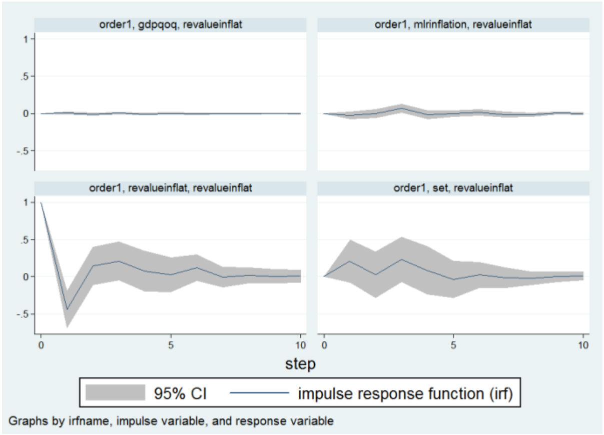


Figure 5: The result of impulse response function SET, MLR, GDP and real estate transaction value 2014q4 -2020q3

**Acknowledgements**

I would like to gratefully thank my advisor, Assoc. Prof. Dr. Atcharawan Ngarmyarn and Assoc. Prof. Yawaman Metapirak for their guidance, understanding and support.

**References**

[1] Gross Domestic Product :Q1/2015 [Internet]. 2015 [cited 26 May 2015] Available from: http://www.nesdb.go.th/Default.aspx?tabid=95

[2] Social statistic bureau. Summary : The survey of economic status and household debt 2014. National Statistical Office Ministry of Information and Communication Technology, 2014.

[3] Vanichvatana S. Thailand Real Estate Market Cycles: Case Study of 1997 Economic Crisis. GH Bank Housing Journal. 2007;1(1):38-47.

[4] Hussein SA. Business and Real Estate Cycles The Kuala Lumpur Office Market: KTH Royal Institute, 2011.

[5] Kim K-H, Lee HS. Real Estate Price Bubble and Price Forecasts in Korea. Department of Economics Sogang University, 2000.

[6] Pholphirul P, Rukumnuaykit P. The real estate cycle and real business cycle : Evidence from Thailand. Pacific Rim Property Research Journal, Vol 15, No 2, 2009.

## Appendix A. Result from VAR model forecasting

| Quarter | RE-value | SET | MLR | GDP |
|---|---|---|---|---|
| 2014q4 | 0.0127 | 0.0613 | 0.0603 | 0.0670 |
| 2015q1 | 0.0221 | 0.0870 | -0.0435 | 0.0157 |
| 2015q2 | 0.0203 | -0.0426 | -0.0300 | -0.0258 |
| 2015q3 | 0.0095 | -0.0638 | -0.0004 | -0.0033 |
| 2015q4 | 0.0158 | 0.0026 | 0.0070 | 0.0441 |
| 2016q1 | -0.0053 | 0.0140 | 0.0026 | 0.0113 |
| 2016q2 | -0.0139 | -0.0276 | -0.0007 | -0.0279 |
| 2016q3 | 0.0071 | -0.0446 | -0.0009 | -0.0033 |
| 2016q4 | 0.0006 | 0.0012 | -0.0001 | 0.0360 |
| 2017q1 | -0.0073 | 0.0155 | 0.0002 | 0.0118 |
| 2017q2 | -0.0082 | -0.0263 | 0.0001 | -0.0220 |
| Quarter | RE-value | SET | MLR | GDP |
| 2017q3 | 0.0039 | -0.0423 | 0.0000 | -0.0020 |
| 2017q4 | 0.0022 | -0.0033 | 0.0000 | 0.0299 |
| 2018q1 | -0.0076 | 0.0098 | 0.0000 | 0.0113 |
| 2018q2 | -0.0073 | -0.0237 | 0.0000 | -0.0174 |
| 2018q3 | 0.0024 | -0.0372 | 0.0000 | -0.0013 |
| 2018q4 | 0.0014 | -0.0061 | 0.0000 | 0.0253 |
| 2019q1 | -0.0066 | 0.0061 | 0.0000 | 0.0108 |
| 2019q2 | -0.0065 | -0.0217 | 0.0000 | -0.0133 |
| 2019q3 | 0.0013 | -0.0333 | 0.0000 | -0.0006 |
| 2019q4 | 0.0010 | -0.0081 | 0.0000 | 0.0215 |
| 2020q1 | -0.0059 | 0.0026 | 0.0000 | 0.0102 |
| 2020q2 | -0.0059 | -0.0200 | 0.0000 | -0.0100 |
| 2020q3 | -0.0005 | -0.0301 | 0.0000 | -0.0000 |

## Appendix B. Comparison of three past real estate crisis in the world

Table 1: Comparison of three past real estate crisis in the world

| | Sweden (end of 1992) | Thailand (mid 1997) | United State (2008) |
|---|---|---|---|
| Loan campaign | Encourage loaning, low interest rate, unrestricted loan. | Bank and non-bank offer easily loan. A lot foreign money supply. | Allow financial institute can give loan for home loan and also release bond for with has real estate to be collateral. |
| Price bubble | Real estate price rise quickly during 1986-1990 | Land price increase greatly and landlord resell to new owner repeatedly. | Real estate price increase quickly because how high cost of property that mean value of issued bond also increase. |
| Non-performing loan | Highly increase | After an issue of floating exchange rate, NPL increase highly due to developer lost their cash flow because foreign loan has cost 2-3 times more than before. | Highly increase during 2007-2010 |
| Exchange rate | Use floating exchange policy in November 1992 | Use floating exchange policy at 2 July 1997 | No change |
| Interest rate | Since 1985 real interest rate after tax was minus. Household saving decrease to minus 4% In 1992, the lost of Household wealth reduced household purchasing power. Saving increased to +8% The shrink of demand and GDP led to unemployed issue. | Even though high interest rate, most of developer loan foreign loan due to lowest cost and fixed exchange rate. Developers aggressively invest until oversupply coincident with change policy to floating exchange rate. Developer could not able to pay interest. | Low interest rate in 2001 lead to financial institute to try decrease loan standard in order to increase clients. Until June 2004 FED increase interest rate 1% in June 2006 interest rate is 5.25%, Price bubble was ended, house price was decrease, but interest was still high. |
| Duration rising of real estate price contentiously. | 10 years (1980-1989). House price index from 100 to 700 | 6 years (2002-2007). House price index from 120 to 190 | 11 years (1997-2007) US National index level from 80 to 190 |
| Interest rate 2 years before crisis | 6.59 % | 7.25 % | 4.74 % |

| | | | |
|---|---|---|---|
| Interest rate 1 year before crisis | 7.21 % | 9.02 % | 5.26 % |
| Interest rate during the crisis year | 14.04 % | 9.21 % | 3.07 % |
| Unemployment rate 1 year before crisis | 1.5 % | 1.10 % | 4.70 % |
| Unemployment rate during the year of crisis | 13% | 0.90% | 5.9 % |
| Unemployment rate 1 year after crisis | 9.50 % | 3.4 % | 9.30% |
| Duration from the last crisis | 20 years after recession of growth of GDP, GDP growth merely 1.1 % 1970-1991 | 14 years after trust company crisis and the 1st reduction of Baht exchange rate | 18 years |
| GDP recovery period | 3 years | 3 years | 3 years |

# POSTER PRESENTATIONS

# Population Size Estimation Based Upon Zero-Modified Poisson Models

Nupian Thepmong[1*] and Rattana Lerdsuwansri[2]
[1]*Department of Mathematics and Statistics, Faculty of Science and Technology Thammasat University, Pathumthani, Thailand, e-mail: nupian@grad.sci.tu.ac.th*
[2]*Department of Mathematics and Statistics, Faculty of Science and Technology Thammasat University, Pathumthani, Thailand, e-mail: rattana@mathstat.sci.tu.ac.th*

## Abstract

Capture-Recapture methodology has been used to estimate an unknown size $N$ of a population and the number of units in the population that are never observed. A counting distribution arises when we summarize how often a unit was identified. A widely used model for count data is Poisson Model having parameter $\lambda$ . Owing to the large number of missing units, the counting distribution can be approached with a zero-modified Poisson models. By fixing known zero-modification parameter to be [0,1], two new estimators are proposed. One is developed along the line of the Turing's estimator so-called Generalized Turing's estimator. The other is derived in the spirit of the Chao's estimator censoring the counts larger than two so-called Generalized Chao's estimator. As a result, we attach the estimators the indexes GTuring and GChao and call $\hat{N}_{GTuring}$ and $\hat{N}_{GChao}$, respectively. To investigate performance of the proposed estimators, the simulation study is done and compared with conventional estimators including the Chao's estimator, the Turing's estimator and the Lanumteang-Böhning's estimator. The results show that two new estimators perform the best with smallest bias and smallest mean square error for all population sizes. However, the $\hat{N}_{GTuring}$ shows benefits, in comparison with the $\hat{N}_{GChao}$, in terms of bias and mean square error. A study of count data in golf-tees experiment is undertaken to illustrate the potential of the two proposed estimators.

*Keywords*: Capture-Recapture, Zero-inflated Poisson Model, Turing's estimator, Chao's estimator

*Corresponding Author.
E-mail Address: nupian@grad.sci.tu.ac.th

## 1. Introduction

In an attempt to overcome the problem of estimating an unknown size $N$ of a population that is hidden and difficult to reach, capture-recapture methods have been widely used in enumerating the size $N$ of a population. We assume that a population is closed with size $N$. It means that the unknown $N$ remains constant during the period of the study (no birth, death or migration) [1]. To formulate an estimate of the population size $N$, an identifying mechanism, e.g. trapping, diagnostic device, register, is used to identify units having a characteristic of interest. According to the identifications, not only a count $y$ but also the number of units identified exactly $y$ times is obtained.

Let $Y_d$ be the number of times that unit $d$ is identified, $d = 1, 2, 3, \ldots, N$ . Thus, $Y_d = 0$ is unobserved. Let $n$ denotes the number of observed units. Without loss of generality, $Y_1, Y_2, \ldots, Y_n$ be the observed sample of non-zero counts and $Y_{n+1}, Y_{n+2}, \ldots, Y_N$ be the remaining which are unobserved. Also, let $f_i$ denotes the number of units identified $i$ times where $i = 0,1,2,\ldots,m$ and $m$ is the largest observed count. It is clear from the sample that $f_1, f_2, \ldots, f_m$ except $f_0$ can be obtained. As a result, $n = f_1 + f_2 + \ldots + f_m$ and $N = n + f_0$. Therefore, the unobserved zero count $f_0$ requires to be estimated in order to obtain an estimate for the population size $N$. Let $p_i$ denote probability of identifying a unit $i$ times. Accordingly, $p_0$ is the probability that a unit is not observed. The unknown $f_0$ may be replaced by the expected value $Np_0$. If $p_0$ were known then we can solve for $N = n + f_0 = n + N p_0$ . However, $p_0$ is unknown and is requires to be estimated. Solving for $N$ leads to the well-known Horvitz-Thompson estimator

$$\hat{N} = \frac{n}{1 - \hat{p}_0} \ . \qquad (1)$$

See [2] for more details.

A question arises how we can estimate the unknown $f_0$ or the unknown $p_0$ . Based upon repeated identifications in the period of the study, we postulate that $Y$ is modeled by Poisson distribution having mean parameter λ. It is assumed that all units in the population have the identical Poisson parameter λ. The corresponding model is called homogeneous Poisson model. However, count data modeled by identical λ are rare in practice. An alternative model incorporating heterogeneity might be more appropriate. There are numerous population size estimators which are built upon homogeneous and heterogeneous Poisson models. An example of the estimator based on homogeneity is

$$\hat{N}_{Turing} = \frac{n}{1 - \left(f_1/S\right)} \qquad (2)$$

where $S = 0f_0 + 1f_1 + \ldots + mf_m$ [see 3]. Furthermore, two estimators of Chao [4] and Lanumteang-Böhning [5] allow population heterogeneity. Chao's estimator and Lanumteang-Böhning estimator are given by

$$\hat{N}_{Chao} = n + \frac{f_1^2}{2f_2}$$

and

$$\hat{N}_{LB} = n + \frac{f_1^2}{2f_2}\left(\frac{3f_1f_3}{2f_2^2}\right)$$

, respectively.

For count data, the Poisson model does not fit for some situations because of too many zeros or too few zeros. For example, in industry, there is concern about finding the number of defects in total in manufacturing [6]. In public health and epidemiology, there is concern about demonstrating quantify dental status of children in Belo Horizonte, Brazil [7]. In social sciences, the interest is in estimating the number of people who is illegal gun ownership in Netherlands [8]. These might be the case that extra zeros are large relative to the other counts. In addition, the population size estimators which are developed under the homogeneity and heterogeneity Poisson might be often underestimating because of extra zero counts. To allow flexibility for extra zero counts including the Poisson model as a special form, the zero-inflated Poisson model is more appropriate [9]. As a result, it is interesting to develop alternative estimators based on the zero-inflated Poisson model. Two new estimators have been proposed in section 2. A simulation study is presented in section 3 to compare the proposed estimators with the conventional estimators including the Turing's estimator, Chao's estimator and Lanumteang-Böhning's estimator. Shown in section 4 is an application with the frequency of recovery counts in golf-tees experiment. Finally, the conclusion and discussion are shown in the last section.

## 2. The Proposed Estimators

Suppose that a population is closed with size *N*. The distribution of count *Y* is the zero-modified Poisson [9] which is of the form

$$f_{ZMP}(y;\lambda,\pi) = (1-\pi)Po(y;0) + \pi Po(y;\lambda)$$

where $\lambda$ denotes mean parameter, λ > 0 and $\pi$ denotes zero modification parameter, $0 \le \pi \le e^{\lambda}/(e^{\lambda}-1)$. Different values of $\pi$ lead to different modifications of the Poisson model. In the case that $\pi$ takes the values $[0,1]$, $f_{ZMP}(y;\lambda,\pi)$ becomes the zero-inflated Poisson model as follows:

$$f_{ZIP}(y;\lambda,\pi) = \begin{cases} (1-\pi)+\pi e^{-\lambda} & ; y = 0 \\ \dfrac{\pi e^{-\lambda}\lambda^y}{y!} & ; y > 0 \end{cases}$$

We restrict ourselves to the situation that the zero modification parameter $\pi$ is known. The contribution here is developing two new estimators as presented in following subsections.

### *2.1 The generalized Turing's estimator*

We consider the Horvitz-Thompson's estimator

$$\hat{N} = \frac{n}{1-\hat{p}_0}.$$

Under homogeneous Poisson model with parameter $\lambda$, the probability of identifying a unit *i* times has the form

$$p_i = \frac{e^{-\lambda}\lambda^y}{y!}.$$

We have that

$$p_0 = e^{-\lambda} = \frac{p_1}{E(Y)}$$

where $p_1$ is the probability of identifying a unit once. The estimate of $p_0$ is

$$\hat{p}_0 = \frac{f_1/\hat{N}}{S/\hat{N}} = \frac{f_1}{S}$$

where $S = \sum_{y=1}^{m} yf_y$ .

This leads to the Turing's estimator [see 3]

$$\hat{N}_{Turing} = \frac{n}{1-(f_1/S)}.$$

Along the line of the Turing's estimator, we consider the zero-inflated Poisson model. We know that

$$p_0 = (1-\pi) + \pi e^{-\lambda}$$

and also $\frac{P_1}{E(X)}$ for the zero-inflated Poisson and homogeneity Poisson is identical because

$$\frac{p_1}{E(Y)} = \frac{\pi e^{-\lambda}\lambda}{\pi\lambda} = e^{-\lambda}.$$

Hence,

$$p_0 = (1-\pi) + \pi\frac{p_1}{E(Y)}.$$

Estimating $p_1$ with $\hat{p}_1 = f_1/\hat{N}$ and estimating $E(Y)$ with $S/\hat{N}$, we arrive at

$$\hat{p}_0 = (1-\pi) + \pi\frac{f_1}{S}.$$

A new population size estimator is accomplished by

$$\hat{N}_{GTuring} = \frac{n}{\pi(1-\frac{f_1}{S})}.$$

We attach the new estimators the index GTuring and call the generalized Turing's estimator $\hat{N}_{GTuring}$. As can be seen, $\hat{N}_{GTuring}$ is closely associated with Turing's estimator. If $\pi = 1$ then $\hat{N}_{GTuring}$ becomes the Turing's estimator.

*2.2The generalized Chao's estimator*

Let $Y$ be given according to the zero-inflated Poisson model. In practice, units are likely to be captured only a few times. Consequently, most of counts are $f_1$ and $f_2$. If we constrain the counts to only two values, 1 and 2, and censor the counts larger than 2, the associated truncated zero-inflated Poisson are given as

$$f_{Y=1,2}(y;\lambda,\pi)=\frac{2\lambda^{y-1}}{y!(2+\lambda)} \quad . \tag{3}$$

The log- likelihood function takes the form

$$\log L(\lambda)= f_1 \log\left(\frac{2}{2+\lambda}\right)+ f_2 \log\left(\frac{\lambda}{2+\lambda}\right). \tag{4}$$

As can be seen, the log- likelihood function capturing only 1's counts and 2's counts is not affected by the zero modification parameter $\pi$. Maximizing (4) with respect to $\lambda$ produces the maximum likelihood estimator

$$\hat{\lambda}=\frac{2f_2}{f_1}.$$

To estimate the unobserved $f_0$, let $E(f_y \mid f_1, f_2;\lambda)$ is denoted by $e_y$.

$$e_y = Np_y = (e_0 + f_1 + f_2 + e_3^+)p_y \tag{5}$$

where $e_3^+ = \sum_{y=3}^{\infty} e_y$ .

Then,

$$e_0 = (e_0 + f_1 + f_2 + e_3^+)p_0 \tag{6}$$

$$e_3^+ = (e_0 + f_1 + f_2 + e_3^+)p_3^+ \tag{7}$$

where $p_3^+ = \sum_{y=3}^{\infty} p_y$ .

$$\begin{aligned} e_0 + e_3^+ &= (e_0 + f_1 + f_2 + e_3^+)(p_0 + p_3^+) \\ &= (e_0 + f_1 + f_2 + e_3^+)(1-p_1-p_2) \end{aligned}$$

Hence,

$$e_0 + e_3^+ = \frac{(f_1+f_2)(1-p_1-p_2)}{p_1+p_2} \quad . \tag{8}$$

Recall that we have

$$\begin{aligned} e_0 &= (e_0 + e_3^+ + f_1 + f_2)p_0 \\ &= (f_1+f_2)p_0 + \left[\frac{(f_1+f_2)(1-p_1-p_2)}{p_1+p_2}\right]p_0 \end{aligned}$$

$$e_0 = \frac{(f_1+f_2)p_0}{p_1+p_2}. \tag{9}$$

Estimating $p_0$ with $\hat{p}_0 = 1-\pi+\pi e^{-\hat{\lambda}}$, $p_1$ with $\hat{p}_1 = \pi\hat{\lambda}e^{-\hat{\lambda}}$ and $p_2$ with $\hat{p}_2 = \frac{\pi\hat{\lambda}^2 e^{-\hat{\lambda}}}{2!}$ whereas $\hat{\lambda}=\frac{2f_2}{f_1}$ in equation (9), the estimate of unobserved $f_0$ is given as

$$\hat{f}_0 = \frac{f_1^{\,2}}{2f_2}\left(\frac{1-\pi}{\pi e^{-\frac{2f_2}{f_1}}}+1\right).$$

Consequently, a new population size estimator based upon censoring the counts larger than two is

$$\hat{N}_{GChao} = n + \frac{f_1^{\,2}}{2f_2}\left(\frac{1-\pi}{\pi e^{-\frac{2f_2}{f_1}}}+1\right) \quad .$$

We attach the new estimators the index GChao and call the generalized Chao's estimator $\hat{N}_{GChao}$. As can be seen, $\hat{N}_{GChao}$ is closely associated with Chao's estimator. If $\pi = 1$ then $\hat{N}_{GChao}$ becomes the Chao's estimator which is of the form [4]

$$\hat{N}_{Chao} = n + \frac{f_1^{\,2}}{2f_2} \quad .$$

Recently, Lanumteang and Böhning (2011) have proposed a new estimator by extending Chao's estimator and using monotonicity of ratios of neighboring frequency counts under Poisson-Gamma mixture model or negative binomial. The Lanumteang-Böhning's estimator is given by [5]

$$\hat{N}_{LB} = n + \frac{f_1^{\,2}}{2f_2}\left(\frac{3f_1f_3}{2f_2^2}\right).$$

According to $\hat{N}_{GChao}$ and $\hat{N}_{LB}$ , it can be considered that both $\hat{N}_{GChao}$ and $\hat{N}_{LB}$ are adjusted estimators of the Chao estimator. Hence, we investigate the effect of the adjustment factors in the simulation study.

## 3. Simulation Study

The simulation study is done to investigate the performance of $\hat{N}_{GTuring}$ and $\hat{N}_{GChao}$ and to compare with conventional estimators such as the Turing's, Chao's and Lanumteang-Böhning's estimator.

*3.1 Designs of Simulation*

Data are generated from a number of models. Firstly, populations are simulated under zero-inflated Poisson distribution, $ZIP(\lambda,\pi)$ where λ = 1, 2, 3 and $\pi$ = 0.5, 0.7, 0.9, 1. These are the cases that $\hat{N}_{GTuring}$ and $\hat{N}_{GChao}$ are derived. Secondly, we consider the situations of contamination having equal weights under probabilities model $0.5ZIP(\lambda_1,\pi)+0.5ZIP(\lambda_2,\pi)$ ;

$(\lambda_1, \lambda_2)$ = (1,2), (1,3), (2,3) and $\pi$ = 0.5, 0.7, 0.9, 1. We also study the case of heterogeneity model; *ZINB*($k,\theta,\pi$), or the marginal of Poisson-Gamma mixture model, where ($k,\theta$) = (1,0.3), (3,0.5), (5,0.7) and $\pi$ = 0.5, 0.7, 0.9, 1. The population size to be estimated is $N$ = 100, 1,000 and 10,000. Each situation is repeated 10,000 times.

Performance of population size estimators is evaluated in terms of bias, variance and mean square error. Due to the fact that with increasing $N$ the expected values and the variance increase, we take Relative bias (*RBias*), Relative variance (*RVar*) and Relative mean square error (*RMSE*) to be the following:

$$RBias = \frac{E(\hat{N}) - N}{N} \qquad (10)$$

$$RVar = \frac{E(\hat{N} - E(\hat{N}))^2}{N^2} \qquad (11)$$

$$RMSE = \frac{E(\hat{N} - N)^2}{N^2} \;. \qquad (12)$$

*3.2 Results of Simulation*

Presented in Table 1 - Table 3 are results of the simulation study. Estimating the size $N$ of populations is consider under homogeneity and heterogeneity models. We summarize a few major results as following

Table 1: relative bias; *RBias*, relative variance; *RVar*, and relative mean square error; *RMSE* (homogeneity model)

| Estimator | *N*=100 | | | *N*=1,000 | | | *N*=10,000 | | |
|---|---|---|---|---|---|---|---|---|---|
| | *RBias* | *RVar* | *RMSE* | *RBias* | *RVar* | *RMSE* | *RBias* | *RVar* | *RMSE* |
| *Y*=*ZIP*(1,0.5) | | | | | | | | | |
| Turing | -0.4841 | **0.0127** | 0.2470 | -0.4984 | **0.0010** | 0.2495 | -0.4998 | **0.0001** | 0.2499 |
| Chao | -0.4635 | 0.0288 | 0.2436 | -0.4971 | 0.0016 | 0.2487 | -0.4996 | 0.0002 | 0.2498 |
| GTuring | **0.0318** | 0.0507 | **0.0517** | **0.0031** | 0.0041 | **0.0041** | **0.0005** | 0.0004 | **0.0004** |
| GChao | 0.0857 | 0.1019 | 0.1092 | 0.0070 | 0.0054 | 0.0054 | 0.0009 | 0.0005 | 0.0005 |
| LB | 0.0674 | 47.1473 | 47.1518 | -0.4849 | 0.0086 | 0.2437 | -0.4985 | 0.0007 | 0.2492 |
| *Y*=*ZIP*(1,0.7) | | | | | | | | | |
| Turing | -0.2858 | **0.0143** | 0.0960 | -0.2991 | **0.0013** | 0.0907 | -0.2999 | **0.0001** | 0.0900 |
| Chao | -0.2658 | 0.0300 | 0.1006 | -0.2974 | 0.0021 | 0.0905 | -0.2997 | 0.0002 | 0.0900 |
| GTuring | **0.0204** | 0.0293 | **0.0297** | **0.0013** | 0.0026 | **0.0026** | **0.0002** | 0.0003 | **0.0003** |
| GChao | 0.0542 | 0.0551 | 0.0581 | 0.0042 | 0.0037 | 0.0038 | 0.0004 | 0.0004 | 0.0004 |
| LB | 0.0284 | 13.1928 | 13.1936 | -0.2844 | 0.0111 | 0.0919 | -0.2986 | 0.0009 | 0.0900 |
| *Y*=*ZIP*(1,0.9) | | | | | | | | | |
| Turing | -0.0826 | **0.0167** | 0.0235 | -0.0986 | **0.0014** | 0.0015 | -0.0998 | **0.0001** | 0.0101 |
| Chao | -0.0654 | 0.0319 | 0.0362 | -0.0974 | 0.0024 | 0.0027 | -0.0997 | 0.0002 | 0.0102 |
| GTuring | **0.0194** | 0.0206 | **0.0210** | **0.0015** | 0.0018 | **0.0015** | **0.0002** | 0.0002 | **0.0002** |
| GChao | 0.0397 | 0.0379 | 0.0394 | 0.0030 | 0.0029 | 0.0027 | 0.0003 | 0.0003 | 0.0003 |
| LB | 0.1346 | 0.9801 | 0.9982 | -0.0857 | 0.0131 | 0.0150 | -0.0988 | 0.0012 | 0.0109 |
| *Y*=*ZIP*(1,1) | | | | | | | | | |
| Turing | **0.0159** | **0.0174** | **0.0177** | **0.0015** | **0.0015** | **0.0015** | **0.0002** | **0.0002** | **0.0002** |
| Chao | 0.0332 | 0.0361 | 0.0372 | 0.0031 | 0.0027 | 0.0027 | 0.0004 | 0.0003 | 0.0003 |
| GTuring | **0.0159** | **0.0174** | **0.0177** | **0.0015** | **0.0015** | **0.0015** | **0.0002** | **0.0002** | **0.0002** |
| GChao | 0.0332 | 0.0361 | 0.0372 | 0.0031 | 0.0027 | 0.0027 | 0.0004 | 0.0003 | 0.0003 |
| LB | 0.2556 | 11.1704 | 11.2357 | 0.0163 | 0.0150 | 0.0152 | 0.0017 | 0.0013 | 0.0013 |

Table 2: relative bias; *RBias*, relative variance; *RVar*, and relative mean square error; *RMSE* (heterogeneity model)

| Estimator | $N$=100 | | | $N$=1,000 | | | $N$=10,000 | | |
|---|---|---|---|---|---|---|---|---|---|
| | *RBias* | *RVar* | *RMSE* | *RBias* | *RVar* | *RMSE* | *RBias* | *RVar* | *RMSE* |
| $Y=0.5ZIP(1,0.7)+0.5ZIP(2,0.7)$ | | | | | | | | | |
| Turing | -0.3289 | **0.0060** | 0.1142 | -0.3342 | **0.0006** | 0.1122 | -0.3344 | **5.67E-05** | 0.1119 |
| Chao | -0.3004 | 0.0126 | 0.1029 | -0.3178 | 0.0010 | 0.1020 | -0.3190 | 9.66E-05 | 0.1018 |
| GTuring | -0.0413 | 0.0123 | **0.0140** | -0.0488 | 0.0012 | 0.0035 | -0.0491 | 0.0001 | 0.0025 |
| GChao | **0.0008** | 0.0204 | 0.0204 | **-0.0367** | 0.0015 | **0.0029** | **-0.0394** | 0.0001 | **0.0017** |
| LB | -0.1536 | 0.5625 | 0.5861 | -0.2915 | 0.0043 | 0.0893 | -0.2994 | 0.0004 | 0.0900 |
| $Y=0.5ZIP(1,0.9)+0.5ZIP(2,0.9)$ | | | | | | | | | |
| Turing | -0.1383 | **0.0060** | 0.0251 | -0.1438 | **0.0006** | 0.0212 | -0.1441 | **5.73E-05** | 0.0208 |
| Chao | -0.1061 | 0.0133 | 0.0245 | -0.1229 | 0.0011 | 0.0162 | -0.1243 | 0.0001 | 0.0156 |
| GTuring | -0.0425 | 0.0074 | **0.0092** | -0.0486 | 0.0007 | 0.0031 | -0.0490 | 7.08E-05 | 0.0025 |
| GChao | **-0.0075** | 0.0146 | 0.0146 | **-0.0293** | 0.0012 | **0.0020** | **-0.0311** | 0.0001 | **0.0011** |
| LB | 0.0231 | 0.2329 | 0.2334 | -0.0912 | 0.0052 | 0.0135 | -0.0992 | 0.0005 | 0.0103 |
| $Y=0.5ZIP(1,1)+0.5ZIP(2,1)$ | | | | | | | | | |
| Turing | -0.0449 | **0.0056** | **0.0077** | -0.0491 | **0.0005** | 0.0029 | -0.0492 | **5.43E-05** | 0.0025 |
| Chao | **-0.0111** | 0.0133 | 0.0134 | -0.0262 | 0.0011 | **0.0018** | -0.0272 | 0.0001 | 0.0009 |
| GTuring | -0.0449 | **0.0056** | **0.0077** | -0.0491 | **0.0005** | 0.0029 | -0.0492 | **5.43E-05** | 0.0025 |
| GChao | **-0.0111** | 0.0133 | 0.0134 | -0.0262 | 0.0011 | **0.0018** | -0.0272 | 0.0001 | 0.0009 |
| LB | 0.1178 | 0.2065 | 0.2204 | **0.0084** | 0.0057 | 0.0058 | **0.0004** | 0.0005 | **0.0005** |
| $Y=ZINB(5,0.7,0.7)$ | | | | | | | | | |
| Turing | -0.3374 | **0.0036** | 0.1175 | -0.3399 | **0.0004** | 0.1159 | -0.3400 | **3.62E-05** | 0.1156 |
| Chao | -0.3062 | 0.0072 | 0.1009 | -0.3187 | 0.0006 | 0.1021 | -0.3195 | 5.91E-05 | 0.1021 |
| GTuring | -0.0534 | 0.0074 | **0.0103** | **-0.0570** | 0.0007 | **0.0040** | **-0.0571** | 7.39E-05 | **0.0033** |
| GChao | **-0.0024** | 0.5759 | 0.5760 | -0.0619 | 0.0008 | 0.0047 | -0.0651 | 8.25E-05 | 0.0043 |
| LB | -0.1953 | 0.4169 | 0.4551 | -0.2972 | 0.0021 | 0.0905 | -0.3026 | 0.0002 | 0.0918 |
| $Y=ZINB(5,0.7,0.9)$ | | | | | | | | | |
| Turing | -0.1488 | **0.0032** | 0.0254 | -0.1514 | **0.0003** | 0.0232 | -0.1514 | **3.14E-05** | 0.0230 |
| Chao | -0.1113 | 0.0074 | 0.0198 | -0.1242 | 0.0006 | 0.0160 | -0.1251 | 6.03E-05 | 0.0157 |
| GTuring | -0.0542 | 0.0040 | **0.0069** | -0.0571 | 0.0004 | **0.0036** | -0.0572 | 3.87E-05 | 0.0033 |
| GChao | **-0.0170** | 0.0189 | 0.0192 | **-0.0389** | 0.0006 | 0.0021 | **-0.0403** | 5.95E-05 | **0.0017** |
| LB | -0.0056 | 0.1634 | 0.1634 | -0.0982 | 0.0024 | 0.0120 | -0.1037 | 0.0002 | 0.0110 |
| $Y=ZINB(5,0.7,1)$ | | | | | | | | | |
| Turing | -0.0558 | **0.0025** | **0.0057** | -0.0570 | **0.0003** | 0.0035 | -0.0571 | **2.50E-05** | 0.0033 |
| Chao | **-0.0171** | 0.0067 | 0.0070 | -0.0269 | 0.0006 | **0.0013** | -0.0279 | 5.61E-05 | 0.0008 |
| GTuring | -0.0558 | **0.0025** | **0.0057** | -0.0570 | **0.0003** | 0.0035 | -0.0571 | **2.50E-05** | 0.0033 |
| GChao | **-0.0171** | 0.0067 | 0.0070 | -0.0269 | 0.0006 | **0.0013** | -0.0279 | 5.61E-05 | 0.0008 |
| LB | 0.0786 | 0.1254 | 0.1316 | **0.0015** | 0.0025 | 0.0025 | **-0.0041** | 0.0002 | **0.0003** |

Table 3: The estimated population size of five estimators

| $N$ | $\hat{N}_{Turing}$ | $\hat{N}_{Chao}$ | $\hat{N}_{GTuring}$ | $\hat{N}_{GChao}$ | $\hat{N}_{LB}$ |
|---|---|---|---|---|---|
| $Y$=$ZIP$(1,0.5) | | | | | |
| 100 | 52 | 54 | 103 | 109 | 107 |
| 1,000 | 502 | 503 | 1003 | 1007 | 515 |
| 10,000 | 5002 | 5004 | 10005 | 10009 | 5015 |
| $Y$=$ZIP$(1,0.7) | | | | | |
| 100 | 71 | 73 | 102 | 105 | 103 |
| 1,000 | 701 | 703 | 1001 | 1004 | 716 |
| 10,000 | 7001 | 7003 | 10002 | 10004 | 7014 |
| $Y$=$ZIP$(1,0.9) | | | | | |
| 100 | 92 | 93 | 102 | 104 | 113 |
| 1,000 | 901 | 903 | 1002 | 1003 | 914 |
| 10,000 | 9002 | 9003 | 10002 | 10003 | 9012 |
| $Y$=$ZIP$(1,1) | | | | | |
| 100 | 102 | 103 | 102 | 103 | 126 |
| 1,000 | 1001 | 1003 | 1001 | 1003 | 1016 |
| 10,000 | 10002 | 10004 | 10002 | 10004 | 10017 |
| $Y$=0.5$ZIP$(1,0.7)+0.5$ZIP$(2,0.7) | | | | | |
| 100 | 67 | 70 | 96 | 100 | 85 |
| 1,000 | 666 | 682 | 951 | 963 | 708 |
| 10,000 | 6656 | 6810 | 9509 | 9606 | 7006 |
| $Y$=0.5$ZIP$(1,0.9)+0.5$ZIP$(2,0.9) | | | | | |
| 100 | 86 | 89 | 96 | 99 | 102 |
| 1,000 | 856 | 877 | 951 | 971 | 909 |
| 10,000 | 8559 | 8757 | 9510 | 9689 | 9008 |
| $Y$=0.5$ZIP$(1,1)+0.5$ZIP$(2,1) | | | | | |
| 100 | 96 | 99 | 96 | 99 | 112 |
| 1,000 | 951 | 974 | 951 | 974 | 1008 |
| 10,000 | 9508 | 9728 | 9508 | 9728 | 10004 |
| $Y$=$ZINB$(5,0.7,0.7) | | | | | |
| 100 | 66 | 69 | 95 | 100 | 80 |
| 1,000 | 660 | 681 | 943 | 938 | 703 |
| 10,000 | 6600 | 6805 | 9429 | 9349 | 6974 |
| $Y$=$ZINB$(5,0.7,0.9) | | | | | |
| 100 | 85 | 89 | 95 | 98 | 99 |
| 1,000 | 849 | 876 | 943 | 961 | 902 |
| 10,000 | 8486 | 8749 | 9428 | 9597 | 8963 |
| $Y$=$ZINB$(5,0.7,1) | | | | | |
| 100 | 94 | 98 | 94 | 98 | 108 |
| 1,000 | 943 | 973 | 943 | 973 | 1001 |
| 10,000 | 9429 | 9721 | 9429 | 9721 | 9959 |

### *3.2.1 Based on Homogeneity Model*

As can be seen from Table 1, $\hat{N}_{GTuring}$ and $\hat{N}_{GChao}$ are the best with the smallest relative bias among the others. In addition, $\hat{N}_{GTuring}$ and $\hat{N}_{GChao}$ are asymptotically unbiased with respect to *N*. However, $\hat{N}_{GTuring}$ performs better than $\hat{N}_{GChao}$ with smaller relative bias for $N$ = 100, 1,000 and 10,000. For situations that $\pi < 1$, $\hat{N}_{Turing}$, $\hat{N}_{Chao}$ and $\hat{N}_{LB}$ trend to be underestimating whereas $\hat{N}_{GTuring}$ and $\hat{N}_{GChao}$ trend to overestimate.

It is found that $\hat{N}_{Turing}$, $\hat{N}_{Chao}$, $\hat{N}_{GTuring}$ and $\hat{N}_{GChao}$ have rather low relative variance whereas $\hat{N}_{LB}$ has the worst performance with the largest variance for $N$ = 100, 1000 and 10000.

Achieving the smallest relative mean square error, $\hat{N}_{GTuring}$ performs best for $N$ = 100, 1,000 and 10,000 and for all values of $\pi$. $\hat{N}_{GChao}$ performs good as well. With increasing values of $\pi$, $\hat{N}_{GTuring}$ and $\hat{N}_{GChao}$ provide better performance with smaller *RMSE*. In contrast, $\hat{N}_{LB}$ performs the worst with the largest *RMSE* particularly for $N = 100$.

It is important to note in the case of $\pi = 1$ that results obtained from $\hat{N}_{Turing}$ are identical to those from $\hat{N}_{GTuring}$. Also, $\hat{N}_{Chao}$ gives similar results to $\hat{N}_{GChao}$.

### *3.2.2 Based on Heterogeneity Model*

According to simulation results in Table 2, all estimators are underestimating due to the heterogeneity of population. Nonetheless, two proposed estimators $\hat{N}_{GTuring}$ and $\hat{N}_{GChao}$ are doing good for all situations. $\hat{N}_{GChao}$ performs the best with the smallest relative bias under any cases of heterogeneity for $N$ = 100, 1,000 and 10,000.

Since $\hat{N}_{LB}$ is derived on the negative binomial distribution, it might be expected to be an appropriate choice under $ZINB(k,\theta,\pi)$ model in particular the case of $\pi = 1$. It is found that $\hat{N}_{LB}$ produces the smallest relative bias for $N$ = 1,000 and 10,000 under $ZINB(5,0.7,1)$ model. However, $\hat{N}_{GTuring}$ and $\hat{N}_{GChao}$ still behave well for the case of $\pi < 1$.

$\hat{N}_{Turing}$ provides the minimum *RVar* for $N$ = 100, 1,000 and 10,000 under any cases of heterogeneity. $\hat{N}_{GTuring}$ as well as $\hat{N}_{GChao}$ produce slightly larger of *RVar* than those of $\hat{N}_{Turing}$. Similar to the homogeneity model, $\hat{N}_{LB}$ is the worst with the largest variance for $N$ = 100, 1,000 and 10,000.

With respect to relative mean square error, It is shown that both $\hat{N}_{GTuring}$ and $\hat{N}_{GChao}$perform well when compared to the other estimators for $N$ = 100, 1,000 and 10,000 under any cases of heterogeneity. In addition, the higher the values of $\pi$ the smaller the *RMSE* of $\hat{N}_{GTuring}$ and $\hat{N}_{GChao}$. It is also found that *RMSE* of $\hat{N}_{GTuring}$ and $\hat{N}_{GChao}$ become smaller with increasing *N*.

It is interesting to look at the effect of the adjustment factors obtained from $\hat{N}_{GChao}$ and $\hat{N}_{LB}$. The results are presented in Table 3. Apparently, estimates provided by $\hat{N}_{GChao}$ are larger than those given by $\hat{N}_{LB}$ under heterogeneity and even under homogeneous model except the case of $\pi = 1$. $\hat{N}_{GChao}$ clearly adjusts for a likely underestimation bias of Chao estimator.

## 4. Application with Golf-tees Experiment

To illustrate the performance of the two proposed estimators, $\hat{N}_{GTuring}$ and $\hat{N}_{GChao}$, an application to real data set is done. Although the simulations have been done, it was generated from situations that support $\hat{N}_{GTuring}$ and $\hat{N}_{GChao}$.Hence it is of interest to investigate $\hat{N}_{GTuring}$ and $\hat{N}_{GChao}$ in data sets where the population size is known in advance. The data discussed relate to capture–recapture experiment in St. Andrews [see 10]. The 250 of golf-tees were placed on area of 1680 m$^2$ in St. Andrews. Then, eight difference students of the University of St. Andrews were again surveying the golf-tees. Shown in Table 4 is the frequencies of counts in golf-tees experiment.

Table 4: Frequency of recovery counts in golf-tees experiment (true $N = 250$)

| $y$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $f_y$ | 46 | 28 | 21 | 13 | 23 | 14 | 6 | 11 |

Here, population size is known (true $N$ = 250). Only zero-truncated counts are considered and $n$ = 162. Then 88 of the golf-tees were ignored and treated as missing datum. Consequently, the estimates are computed by means of Turing, Chao, Lanumteang-Böhning, GTuring and GChao. For the case of the proposed methods, the estimates of parameter $\lambda$ and $\pi$ require to be figured out. The maximum likelihood estimators of parameter $\lambda$ and $\pi$ are $\hat{\lambda}_{MLE} \approx 3.0$ and $\hat{\pi}_{MLE} \approx 0.7$, respectively. Then, the testing for goodness of fit is done to guarantee the fitting.

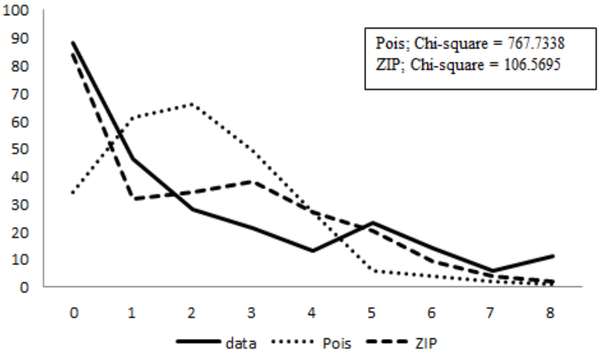


Figure 1: Show the fitting data by Poisson and ZIP models

As can be seen in Figure 1, the zero-inflated Poisson distribution provides a much better fit than the Poisson distribution although the fit of the zero-inflated Poisson is not perfect. One might argue that zero-inflated binomial model could be an althernative.

Table 5: Estimates of the total number of the golf-tees

| $N$ | $\hat{N}_{\text{Turing}}$ | $\hat{N}_{Chao}$ | $\hat{N}_{\text{GTuring}}$ | $\hat{N}_{GChao}$ | $\hat{N}_{\text{LB}}$ |
|---|---|---|---|---|---|
| 250 | 177 | 200 | 253 | 254 | 232 |

Table 5 presents the total number of golf-tees of five estimators. It is clearly seen that $\hat{N}_{GTuring}$ and $\hat{N}_{GChao}$ provide the best estimates achieving the smallest bias. These might be the case that extra zeros are large relative to the other counts. However, it is remarkable that $\hat{N}_{GTuring}$ and $\hat{N}_{GChao}$ significantly improve the underestimates of $\hat{N}_{Turing}$ and $\hat{N}_{Chao}$, respectively.

## 5. Conclusion and Discussion

Estimating the size $N$ of a close population that is affected by the large number of missing units is of interest. These situations are often seen in public health and social science. The conventional estimators such as the Turing's and the Chao's estimator are developed under the Poisson model. Consequently, the estimations are underestimating. We suggest alternative estimators, $\hat{N}_{GTuring}$ and $\hat{N}_{GChao}$, for such situations. It should be noted that the zero-modification parameter $\pi$ needs to be known. Otherwise, the model is not identifiable. The simulation studies provide the positive evidence that the new proposed estimators, $\hat{N}_{GTuring}$ and $\hat{N}_{GChao}$ show good performance and become a candidate for use. A study of count data in golf-tees experiment is undertaken to illustrate the potential of $\hat{N}_{GTuring}$ and $\hat{N}_{GChao}$. Nonetheless, it might be interest to consider an estimator based on zero-inflated binomial model.

According to the suggested approaches, we are assuming that the zero modification parameter $\pi$ is known. Determining an appropriate value of $\pi$ is an important concern to make the point estimate of $N$ trustworthy. We would suggest the testing for goodness of fit as well as using a graphical device to find the evidence for count distribution.

Not only the point estimate of $N$ but also confidence interval of $N$ is meaningful for inference. It is not easy to deal with variance estimates of the population size estimator $\hat{N}$ . The possibility to construct the confidence interval of unknown size $N$ of a population would be left for future research.

## Acknowledgements

We are grateful to the editor and anonymous reviewers for their helpful comments and suggestions.

## References

[1] White GC, Anderson DR, Burnham KP, Otis DL. Capture-Recapture and Removal Methods for Sampling Closed Populations. Mexico: Los Alamos National Laboratory; 1982.

[2] van der Heijden P, Bustami R, Cruyff M, Engbersan G, Houwelingen H. Point and interval estimation of population size using the truncated Poisson regression model. Statistical Modelling.2003; 3: 305–322.

[3] Lanumteang K. Estimation of the Size of a Target Population Using Capture-Recapture Methods Based upon Multiple Sources and Continuous Time Experiment [Dissertation]. UK: University of Reading; 2010.

[4] Chao A. Estimating the Population Size for Capture-Recapture Data with Unequal Catchability. Biometrics.1987; 43: 783-791.

[5] Lanumteang K, Böhning D. An extension of Chao's estimator of population size based on the first three capture frequency counts, Computational Statistics and Data Analysis.2011; 55: 2302-2311.

[6] Lambert D. Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. Technometrics.1992; 34: 1-14.

[7] Böhning D, Dietz E. The Zero-Inflated Poisson Model and the Decayed, Missing and Filled Teeth Index in Dental Epidemiology. Royal Statistical Society.1999; 162(2): 195-209.

[8] Baksh MF, Böhning D, Lerdsuwansri R. An extension of an over-dispersion test for count data. Computational Statistics and Data Analysis.2011; 55: 466-474.

[9] Dietz E, Böhning D. On estimation of the Poisson parameter in zero-modified Poisson models. Computational Statistics and Data Analysis. 2000; 34: 441-459.

[10] Niwitpong S, Böhning D, van der Heijden PGM, Holling H. Capture-recapture estimation based upon the geometric distribution allowing for heterogeneity . Metrika.2013; 76: 495-519.

# Confidence Interval for Population Size Using the Generalized Turing Estimator

Parawan Pijitrattana[1*] and Rattana Lerdsuwansri [2]

[1]Department of Mathematics and Statistics, Thammasat University, Pathumthani, 12121, THAILAND,
e-mail: parawan@grad.sci.tu.ac.th

[2]Department of Mathematics and Statistics, Thammasat University, Pathumthani, 12121, THAILAND,
e-mail: rattana@mathstat.sci.tu.ac.th

## Abstract

More estimators for estimating an unknown size $N$ of a closed population from capture-recapture data have been developed. Recently, a new estimator, the Generalized Turing, is presented using a simple weighted regression on the basis of the ratio of neighboring frequency counts from capture-recapture sampling which is valid under the Poisson-Gamma mixture or negative binomial. With regard to making inference about the unknown size $N$ of the population, confidence interval estimation based upon the Generalized Turing estimator is proposed. However, confidence interval estimation for $N$ is quite rare in developments. The usual approach known as a Wald method is based on the assumption of asymptotic normality and requires the standard error of the population size estimator ( $\hat{N}$ ). As a result, the variance estimation of the Generalized Turing estimator is discussed by means of the conditioning technique. A Monte Carlo simulation was provided to study the performance of the proposed variance estimation under the negative binomial. The results showed that the proposed variance estimation represents well the true variance of Generalized Turing estimator. Subsequently, we constructed confidence interval for population size $N$ based upon the Generalized Turing estimator by Normal approximation approach. Further simulation was conducted to compare confidence intervals obtained from the proposed one and several estimators including the Chao, the Zelterman and the Lanumteang&Böhning estimator. Performance of the confidence interval for $N$ was evaluated in terms of coverage probability and average length. Simulation results indicated that the proposed confidence interval tends to perform reasonably well. As an application, estimating the number of heroin users in Bangkok in the year 2002 is examined using the proposed method.

*Keywords:* Capture-recapture; Population size estimation; conditioning technique; negative binomial.

*Corresponding Author
E-mail Address: parawan@grad.sci.tu.ac.th

## 1. Introduction

Capture-Recapture methods aim to estimate the size of an elusive target population. A large number of application areas have adopted capture-recapture methods to estimate unobserved units as well as the total number in the population. Capture-Recapture methods has been widely used in many field, including biology, sociology, criminology, public health and epidemiology, such as estimating the number of female grizzly bears in the Yellowstone ecosystem [1], estimating the number of transport injuries in a developing country [2], estimating the size of a criminal population [3] and estimating the size of cancer patients [4] .

We assume that there are $N$ units in the closed population which is no birth, no death and no migration during the observational period. There is an identifying mechanism such as a register, trapping system, which is capable of repeatedly identifying observed units. The identifications provide a count $Y_i$ of how many times the individual $i^{th}$ was identified during the period of study, for $i = 1, 2, ..., N$ . We denote the frequency of units identified exactly $0, 1, 2, ..., m$ times by $f_0, f_1, f_2, ..., f_m$ respectively where $m$ is the largest observed count. It is interesting to note that $f_0$ , the frequency of units identified zero times, is unknown because units who were never identified did not appear. Consequently, the number of observed units, $n = f_1 + f_2 + .. + f_m$ and $N = n + f_0$ . The objective is to estimate $f_0$ which leads to the estimate of population size $N$.

A count distribution arises when we summarize how often a unit was identified. Let $p_i = P(Y = i)$ denote probability for identifying a unit $i$ times. Accordingly, $p_0$ is the probability of not identifying a unit. The unobserved $f_0$ might be replaced by the expected value $Np_0$ . If $p_0$ is known then we can solve for $N = n + f_0$ .

As $p_0$ is unknown and is required to estimate leading to the well-known Horvitz-Thompson estimator

$$\hat{N} = \frac{n}{1-\hat{p}_0} \qquad (1)$$

See [5] for more details. In general, the count data are often modeled by Poisson distribution having parameter $\lambda$. Due to the fact that identical $\lambda$ for all units in the population does not hold in practice. Incorporating heterogeneity of Poisson parameter into the model has been proposed in the literature [6-8]. As a consequence, the capture probability under heterogeneity Poisson with density $g(\lambda)$ is given by

$$p_i = \int_0^\infty \frac{e^{-\lambda}\lambda^i}{i!} g(\lambda)d\lambda \quad ,i = 0,1,2,... \quad . \qquad (2)$$

Suppose that $g(\lambda)$ is Gamma density with shape parameter $\kappa$ and scale parameter $\beta = \frac{1-\pi}{\pi}$. Then $p_i = \frac{\Gamma(\kappa+i)}{\Gamma(i+1)\Gamma(\kappa)}\pi^\kappa(1-\pi)^i$ is the negative binomial distribution.

In 2013,Böhning,Baksh,Lerdsuwansri,and Gallagher [9] defined the ratio $r_i = (i+1)\frac{p_{i+1}}{p_i}$ and pointed out that the negative binomial distribution was a power series density

$p_i = a_i t^i A(t)$ with $a_i = \frac{\Gamma(\kappa+i)}{\Gamma(i+1)\Gamma(\kappa)}$ and $A(t) = (1-t)^\kappa$.

We have

$\frac{a_i}{a_{i+1}} \cdot \frac{p_{i+1}}{p_i} = \frac{i+1}{i+\kappa} \cdot \frac{p_{i+1}}{p_i} = 1-\pi$ leading to the ratio

$r_i = (i+1)\frac{p_{i+1}}{p_i} = (i+\kappa)(1-\pi) = (1-\pi)\kappa + (1-\pi)i$ .

Given observed counts and observed frequencies, the estimated equation was given as

$$\hat{r}_i = \hat{\alpha} + \hat{\beta} i \, . \qquad (3)$$

The estimate of $\kappa$ could be obtained from $\frac{\hat{\alpha}}{\hat{\beta}}$ where the weighted regression on the basis of the ratio of neighboring frequency counts was exploited to estimate the regression coefficients. Since $\alpha = (1-\pi)\kappa$ and $\beta = 1-\pi$, the estimate of $\kappa$ was valid only if $\hat{\alpha} \in (0,\infty)$ and $\hat{\beta} \in (0,1)$. Furthermore, Böhning et. al. (2013) proposed a new estimator, the Generalized Turing estimator, based on Poisson-Gamma mixture model or negative binomial which was of the form

$$\hat{N}_{GT} = \frac{n}{1-\left(\frac{f_1}{S}\right)^{\frac{\kappa}{\kappa+1}}} \quad \text{where } S = \sum_{i=0}^{m} i f_i \quad . \qquad (4)$$

In addition, the Generalized Turing estimator $\hat{N}_{GT}$ was asymptotically unbiased with respect to *N*. See [9] for more details.

With regard to making inference about the unknown size *N* of a closed population, this paper proposes the confidence interval estimation based upon the Generalized Turing estimator $\hat{N}_{GT}$. In section 2, the variance estimation of $\hat{N}_{GT}$ is discussed and confidence interval for $\hat{N}_{GT}$ is constructed by Normal approximation approach. Performances of the proposed variance estimator as well as the confidence interval are evaluated by simulation studies in section 3. In section 4, the confidence interval for $\hat{N}_{GT}$ is applied to real data describing heroin user contacts in Bangkok, Thailand. Some points of work are remarked in section 5.

## 2. The proposed variance estimator and confidence interval

### *2.1 Variance estimator*

The variance estimation of the Generalized Turing estimator $\hat{N}_{GT}$ is discussed by means of the conditioning technique. It can be noted that the variation of the Generalized Turing estimator is arising from two sources: the random variation of sampling *n* units from *N* with *N* unknown, and the random variation due to the estimation of $\kappa$. Böhning [10] proposed a simple variance formula of population size estimators by conditioning which can be applied to derive the variance estimator of the Generalized Turing estimator as follows:

$$Var(\hat{N}) = Var_n\{E(\hat{N}|n)\} + E_n\{Var(\hat{N}|n)\} \qquad (5)$$

where $E_n$ and $Var_n$ are referred to the distribution of *n* which follows the binomial distribution with parameters *N* and $1-p_0$ .

#### *2.1.1* $Va\mathrm{r}_n\{E(\hat{N}|n)\}$

Assuming that $E(\hat{N}|n) \approx \frac{n}{1-\hat{p}_0}$ we have that

$$Var_n\{E(\hat{N}|n)\} = Var_n\{\frac{n}{1-\hat{p}_0}\} = \frac{1}{(1-\hat{p}_0)^2}Var_n\{n\}$$
$$= \frac{N\hat{p}_0(1-\hat{p}_0)}{(1-\hat{p}_0)^2} \cdot \qquad (6)$$

Since $\hat{p}_0 = \left(\frac{f_1}{S}\right)^{\frac{\hat{\kappa}}{\hat{\kappa}+1}}$ and $n = N(1-\hat{p}_0)$, eq.(6) can be estimated by

$$Var_n\{E(\hat{N}|n)\} = \frac{n\left(\frac{f_1}{S}\right)^{\frac{\hat{\kappa}}{\hat{\kappa}+1}}}{\left(1-\left(\frac{f_1}{S}\right)^{\frac{\hat{\kappa}}{\hat{\kappa}+1}}\right)^2}. \qquad (7)$$

### *2.1.2* $E_n\{Var(\hat{N}|n)\}$

Assuming that $E_n\{Var(\hat{N}|n)\} \approx Var(\hat{N}|n)$ we have that

$$Var(\hat{N}|n) = Var\left(\frac{n}{1-\left(\frac{f_1}{S}\right)^{\frac{\kappa}{\kappa+1}}}\right) = n^2 Var\left(\frac{1}{1-\left(\frac{f_1}{S}\right)^{\frac{\kappa}{\kappa+1}}}\right).$$

Using the delta method [11] deals with this reciprocal transformation: $g(x) = \frac{1}{(1-x)}$. This leads to $g'(x) = \frac{1}{(1-x)^2}$. Hence, we are able to achieve that

$$Var(\hat{N}|n) = \frac{n^2}{\left(1-\left(\frac{f_1}{S}\right)^{\frac{\kappa}{\kappa+1}}\right)^4} Var\left(\left(\frac{f_1}{S}\right)^{\frac{\kappa}{\kappa+1}}\right). \qquad (8)$$

To derive $Var\left(\left(\frac{f_1}{S}\right)^{\frac{\kappa}{\kappa+1}}\right)$ the conditioning technique is repeated again. We have that

$$Var\left(\left(\frac{f_1}{S}\right)^{\frac{\kappa}{\kappa+1}}\right) = \underbrace{Var_{f_1}\left\{E\left(\left(\frac{f_1}{S}\right)^{\frac{\kappa}{\kappa+1}} | f_1\right)\right\}}_{\{1\}}$$

$$+ \underbrace{E_{f_1}\left\{Var\left(\left(\frac{f_1}{S}\right)^{\frac{\kappa}{\kappa+1}} | f_1\right)\right\}}_{\{2\}}, \qquad (9)$$

where $E_{f_1}$ and $Var_{f_1}$ are referred to the distribution of $f_1$ which follows the binomial distribution with parameters $N$ and $p_1$.

Suppose that $Var_{f_1}\left\{E\left(\left(\frac{f_1}{S}\right)^{\frac{\kappa}{\kappa+1}} | f_1\right)\right\}$ in the first term of eq.(9) can be estimated by $Var_{f_1}\left\{\left(\frac{f_1}{S}\right)^{\frac{\kappa}{\kappa+1}}\right\}$. We have that

$$Var_{f_1}\left\{\left(\frac{f_1}{S}\right)^{\frac{\kappa}{\kappa+1}}\right\} \approx \frac{1}{S^{\frac{2\kappa}{\kappa+1}}} Var_{f_1}\left\{\left(f_1\right)^{\frac{\kappa}{\kappa+1}}\right\}.$$

Using the delta method [11] deals with this reciprocal transformation: $g(x) = x^{\frac{\kappa}{\kappa+1}}$. This leads to $g'(x) = \frac{\kappa}{\kappa+1}\left(x^{\frac{\kappa}{\kappa+1}-1}\right)$. Therefore, we are able to achieve that

$$\begin{aligned} Var_{f_1}\left\{\left(\frac{f_1}{S}\right)^{\frac{\kappa}{\kappa+1}}\right\} &\approx S^{-\frac{2\kappa}{\kappa+1}}\left(\frac{\kappa}{\kappa+1}\right)^2 f_1^{2\left(\frac{\kappa}{\kappa+1}-1\right)} Var_{f_1}\left(f_1\right) \\ &\approx S^{-\frac{2\kappa}{\kappa+1}}\left(\frac{\kappa}{\kappa+1}\right)^2 f_1^{2\left(\frac{\kappa}{\kappa+1}-1\right)} Np_1\left(1-p_1\right) \\ &\approx S^{-\frac{2\kappa}{\kappa+1}}\left(\frac{\kappa}{\kappa+1}\right)^2 f_1^{2\left(\frac{\kappa}{\kappa+1}-1\right)} f_1\left(1-\frac{f_1}{\hat{N}}\right) \\ &\approx S^{-\frac{2\kappa}{\kappa+1}}\left(\frac{\kappa}{\kappa+1}\right)^2 f_1^{\frac{\kappa-1}{\kappa+1}}\left(1-\frac{f_1}{\hat{N}}\right). \qquad (10) \end{aligned}$$

Now consider the second term of eq.(9), we assume that $E_{f_1}\left\{Var\left(\left(\frac{f_1}{S}\right)^{\frac{\kappa}{\kappa+1}} | f_1\right)\right\}$ can be estimated by $Var\left(\left(\frac{f_1}{S}\right)^{\frac{\kappa}{\kappa+1}} | f_1\right)$. We have that

$$Var\left(\left(\frac{f_1}{S}\right)^{\frac{\kappa}{\kappa+1}} | f_1\right) \approx f_1^{\frac{2\kappa}{\kappa+1}} Var\left(\left(\frac{1}{S}\right)^{\frac{\kappa}{\kappa+1}} | f_1\right).$$

Using the delta method [11] deals with this reciprocal transformation: $g(x) = x^{-\frac{\kappa}{\kappa+1}}$. This leads to $g'(x) = -\frac{\kappa}{\kappa+1}\left(x^{-\frac{\kappa}{\kappa+1}-1}\right)$. Consequently, we are able to achieve that

$$Var\left(\left(\frac{f_1}{S}\right)^{\frac{\kappa}{\kappa+1}} | f_1\right) \approx f_1^{\frac{2\kappa}{\kappa+1}}\left(-\frac{\kappa}{\kappa+1}\right)^2 S^{2\left(-\frac{\kappa}{\kappa+1}-1\right)} Var\left(S\right).$$

Since $Var\left(S\right) = \frac{N\kappa(1-\pi)}{\pi^2}$ and $E\left(x\right) = \frac{S}{N} = \frac{\kappa(1-\pi)}{\pi}$, we have that

$$Var\left(\left(\frac{f_1}{S}\right)^{\frac{\kappa}{\kappa+1}} | f_1\right) \approx f_1^{\frac{2\kappa}{\kappa+1}}\left(-\frac{\kappa}{\kappa+1}\right)^2 S^{2\left(-\frac{\kappa}{\kappa+1}-1\right)} \frac{S}{\pi}.$$

Due to $p_0 = \pi^{\kappa} = \left(\frac{f_1}{S}\right)^{\frac{\kappa}{\kappa+1}}$, we are able to achieve that

$$Var\left(\left(\frac{f_1}{S}\right)^{\frac{\kappa}{\kappa+1}} | f_1\right) \approx f_1^{\frac{2\kappa}{\kappa+1}}\left(-\frac{\kappa}{\kappa+1}\right)^2 S^{-\frac{2\kappa}{\kappa+1}-2+1}\left(\frac{1}{p_0^{\frac{1}{\kappa}}}\right)$$

$$\approx f_1^{\frac{2\kappa}{\kappa+1}}\left(\frac{\kappa}{\kappa+1}\right)^2 S^{-\frac{2\kappa}{\kappa+1}-2+1}\left(\frac{S}{f_1}\right)^{\frac{1}{\kappa+1}}$$

$$\approx S^{-\frac{3\kappa}{\kappa+1}}\left(\frac{\kappa}{\kappa+1}\right)^2 f_1^{\frac{2\kappa-1}{\kappa+1}}. \qquad (11)$$

Hence eq.(9) becomes

$$Var\left(\left(\frac{f_1}{S}\right)^{\frac{\kappa}{\kappa+1}}\right) \approx S^{-\frac{2\kappa}{\kappa+1}}\left(\frac{\kappa}{\kappa+1}\right)^2 f_1^{\frac{\kappa-1}{\kappa+1}}\left(1-\frac{f_1}{\hat{N}}\right)$$

$$+S^{-\frac{3\kappa}{\kappa+1}}\left(\frac{\kappa}{\kappa+1}\right)^2 f_1^{\frac{2\kappa-1}{\kappa+1}}. \qquad (12)$$

Substituting eq.(12) into eq.(8), we have that

$$E_n\{Var(\hat{N}|n)\} = \frac{n^2}{\left(1-\left(\frac{f_1}{S}\right)^{\frac{\kappa}{\kappa+1}}\right)^4}$$

$$\times\left(\frac{\kappa}{\kappa+1}\right)^2 f_1^{\frac{\kappa-1}{\kappa+1}} S^{-\frac{2\kappa}{\kappa+1}}\left\{\left(1-\frac{f_1}{\hat{N}}\right)+\left(\frac{f_1}{S}\right)^{\frac{\kappa}{\kappa+1}}\right\}. \qquad (13)$$

We can estimate eq.(13) by

$$E_n\{Var(\hat{N}|n)\} = \frac{n^2}{\left(1-\left(\frac{f_1}{S}\right)^{\frac{\hat{\kappa}}{\hat{\kappa}+1}}\right)^4}$$

$$\times\left(\frac{\hat{\kappa}}{\hat{\kappa}+1}\right)^2 f_1^{\frac{\hat{\kappa}-1}{\hat{\kappa}+1}} S^{-\frac{2\hat{\kappa}}{\hat{\kappa}+1}}\left\{\left(1-\frac{f_1}{\hat{N}}\right)+\left(\frac{f_1}{S}\right)^{\frac{\hat{\kappa}}{\hat{\kappa}+1}}\right\}. \qquad (14)$$

Substituting eq. (7) and eq. (14) into eq. (5), we finally have that

$$Var(\hat{\text{N}}_{GT}) = \frac{n\hat{p}_0}{(1-\hat{p}_0)^2} +$$

$$\frac{n^2}{(1-\hat{p}_0)^4}\left(\frac{\hat{\kappa}}{\hat{\kappa}+1}\right)^2 f_1^{\frac{\hat{\kappa}-1}{\hat{\kappa}+1}} S^{-\frac{2\hat{\kappa}}{\hat{\kappa}+1}}\left\{\left(1-\frac{f_1}{\hat{N}}\right)+\hat{p}_0\right\}$$

where $\hat{p}_0 = \left(\frac{f_1}{S}\right)^{\frac{\hat{\kappa}}{\hat{\kappa}+1}}$. (15)

It is crucial that estimated variance in eq.(15) is conditioning on estimated $\kappa$.To estimate the value of $\kappa$,this could be obtained by utilizing the ratio plot of neighboring frequency counts and constructing a weighted regression estimator for the regression coefficients in $r_i = \alpha + \beta i$ with a diagonal weight matrix containing the inverse variances of $\hat{r}_i = \frac{(i+1)f_i}{f_i}$.More details are provided in Böhning [10]. The estimate of $\kappa$ could be obtained from $\frac{\hat{\alpha}}{\hat{\beta}}$.

*2.2 Confidence interval*

Base on the estimates and the estimated variance, a confidence interval of the population size $\hat{N}_{GT}$ can be constructed by Normal approximation approach. $100(1-\alpha)\%$ confidence interval for $\hat{N}_{GT}$ is as follows:

$\hat{N}_{GT} \pm Z_{1-\frac{\alpha}{2}} Se\left(\hat{N}_{GT}\right)$ where $Z_{1-\frac{\alpha}{2}}$ is $\left(1-\frac{\alpha}{2}\right)^{th}$ percentile of the standard normal distribution and $Se\left(\hat{N}_{GT}\right) = \sqrt{Var(\hat{N}_{GT})}$ is an estimate of the standard error of $\hat{N}_{GT}$.

## 3. Simulation Study

*3.1 Aim of study*

A Monte Carlo simulation was provided to study the performance of the proposed variance estimation and to examine the performance of confidence intervals obtained from the proposed one and several estimators including the Chao, the Zelterman and the Lanumteang&Böhning estimator.

*3.2 Scope of Study*

Data were generated from negative binomial distribution with parameters $(\kappa,\pi) = (5,0.625)$, $(5,0.42),(7,0.7),(7,0.5)$ indicating small mean and large mean. The population size to be estimated was $N$ = 5,000, 10,000, 50,000 and 100,000. Each condition was repeated 10,000 times by using program R version 3.1.2. The confidence level is 95%.

Performance of the proposed variance estimation was evaluated by comparing an average estimated standard error of $\hat{N}_{GT}$ or $Mean\left(Se\left(\hat{N}_{GT}\right)\right)$ with empirical standard error $Se\left(\hat{N}_{GT}\right)$. The best performance occurs when $Mean\left(Se\left(\hat{N}_{GT}\right)\right)$ is close to $Se\left(\hat{N}_{GT}\right)$.

In addition, the performance of confidence intervals obtained from the proposed one and several estimators were evaluated in terms of coverage probability and average length. The best performance occurs when coverage probability is close to confidence level and provides the shortest average length.

In this study we compared confidence intervals obtained from the proposed one and several estimators as follows:

*1) Chao's estimator*

$$\hat{N}_{Chao} = n + \frac{f_1^2}{2f_2}$$

$$Var(\hat{N}_{Chao}) = \frac{f_1^4}{4f_2^3} + \frac{f_1^3}{f_2^2} + \frac{f_1^2}{2f_2} - \frac{f_1^4}{4nf_2^2} - \frac{f_1^4}{2f_2\left(2nf_2 + f_1^2\right)}$$

, see [6,8].

*2) Zelterman's estimator*

$$\hat{N}_Z = \frac{n}{1-e^{-\hat{\lambda}}}$$

$$Var(\hat{N}_Z) = nG\left(\hat{\lambda}\right)\left[1 + nG\left(\hat{\lambda}\right)\hat{\lambda}^2\left(\frac{1}{f_1}+\frac{1}{f_2}\right)\right]$$

where $G\left(\hat{\lambda}\right) = \frac{e^{-\hat{\lambda}}}{\left(1-e^{-\hat{\lambda}}\right)^2}$ and $\hat{\lambda} = \frac{2f_2}{f_1}$ , see[7,8].

*3) Lanumteang&Böhning'estimator*

$$\hat{N}_{LB} = n + \frac{f_1^2}{2f_2}\left(\frac{3f_1f_3}{2f_2^2}\right)$$

$$Var(\hat{N}_{LB}) = \left(\frac{9}{4}\right)^2 \frac{f_1^5 f_3^2}{f_2^6}\left(\frac{f_1}{f_2}+1\right) + \left(\frac{3}{4}\right)^2 \frac{f_1^6 f_3}{f_2^6}\left(1-\frac{f_3}{n}\right)$$

$$+\left(\frac{3n}{4}\right)\frac{f_1^3 f_3}{nf_2^3 + \frac{3}{4}f_1^3 f_3}$$ , see [8]

### *3.3 Simulation results*

According to Table 1, an average estimated standard error of Generalized Turing estimator $Mean\left(Se\left(\hat{N}_{GT}\right)\right)$ and empirical standard error $Se\left(\hat{N}_{GT}\right)$ are significantly similar in the case of large $\mu$ for all population sizes. However, it can be stated that the proposed variance estimation represents well the true variance of Generalized Turing estimator.

Presented in Table 2 is the coverage probability and average length of 95% confidence interval for population size *N* with small $\mu$ . Lanumteang&Böhning is the best with the highest coverage probability for all population sizes. The proposed CI performs the second best in which the coverage probability increases with increasing *N*. Apparently, CIs of Chao approach do not cover the true *N* at all. The estimated coverage probability of CIs obtained from Zelterman decrease when *N* increases. Although CIs obtained from all estimators are not close to the confidence level, the proposed CI provides the narrowest average length in all situations. As can be seen that CIs of Zelterman and Lanumteang&Böhning produce the widest length of CI.

Table 3 shows the coverage probability and average length of 95% confidence interval for population size *N* with large $\mu$ . The result shows that the coverage probability of CIs obtained from the proposed one and the Lanumteang&Böhning are close to confidence level compared to the others for all population sizes. Although the proposed CI gives the lower coverage probability than the Lanumteang&Böhning, the length of the proposed one is dramatically shorter than CIs of Lanumteang&Böhning. In addition, Zelterman provides the lowest coverage and the widest confidence interval for all sizes of *N*.

Table 1: Estimated standard error of Generalized Turing estimator for $p_i \sim NB(\kappa,\pi)$

| $NB(\kappa,\pi)$ | $N$ | $Se\left(\hat{N}_{GT}\right)$ | $Mean\left(Se\left(\hat{N}_{GT}\right)\right)$ | $Mean\left(Se\left(\hat{N}_{Chao}\right)\right)$ | $Mean\left(Se\left(\hat{N}_Z\right)\right)$ | $Mean\left(Se\left(\hat{N}_{LB}\right)\right)$ |
|---|---|---|---|---|---|---|
| | | | | $\mu = 3$ | | |
| (5,0.625) | 5,000 | 41.96 | 26.38 | 34.65 | 66.61 | 69.85 |
| | 10,000 | 59.23 | 37.62 | 48.93 | 93.93 | 97.82 |
| | 50,000 | 132.62 | 84.86 | 109.31 | 209.69 | 217.36 |
| | 100,000 | 185.27 | 120.18 | 154.53 | 296.41 | 306.98 |
| (7,0.7) | 5,000 | 35.79 | 24.29 | 32.18 | 60.26 | 61.43 |
| | 10,000 | 50.12 | 34.60 | 45.48 | 85.08 | 86.19 |
| | 50,000 | 112.35 | 77.93 | 101.57 | 189.78 | 191.28 |
| | 100,000 | 158.25 | 110.35 | 143.63 | 268.30 | 270.26 |

Table 1: Estimated standard error of Generalized Turing estimator for $p_i \sim NB(\kappa,\pi)$ (Cont.)

| $NB(\kappa,\pi)$ | $N$ | $Se(\hat{N}_{GT})$ | $Mean(Se(\hat{N}_{GT}))$ | $Mean(Se(\hat{N}_{Chao}))$ | $Mean(Se(\hat{N}_Z))$ | $Mean(Se(\hat{N}_{LB}))$ |
|---|---|---|---|---|---|---|
| (5,0.42) | 5,000 | 10.52 | 8.58 | 11.04 | 55.11 | 20.08 |
| | 10,000 | 14.62 | 12.29 | 15.52 | 76.11 | 27.41 |
| | 50,000 | 32.77 | 27.88 | 34.62 | 168.09 | 60.05 |
| | 100,000 | 46.37 | 39.52 | 48.91 | 236.94 | 84.56 |
| (7,0.5) | 5,000 | 7.63 | 6.74 | 8.60 | 43.60 | 14.75 |
| | 10,000 | 10.75 | 9.63 | 12.08 | 59.56 | 20.08 |
| | 50,000 | 24.54 | 21.79 | 26.92 | 129.89 | 43.71 |
| | 100,000 | 34.03 | 30.87 | 38.01 | 182.57 | 61.45 |

Table 2: Coverage Probability and average length of $95\%CI$ of $N$; $p_i \sim NB(\mu=3,\kappa,\pi)$

| $NB(\kappa,p)$ | Estimator | Mean $\hat{N}$ | Coverage Probability | Average Length | Mean $\hat{N}$ | Coverage Probability | Average Length |
|---|---|---|---|---|---|---|---|
| | | | $N=5,000$ | | | $N=10,000$ | |
| (5,0.625) | Chao | 4,921.35 | 0.39 | 135.63 | 9,842.82 | 0.13 | 191.75 |
| | Zelterman | 5,058.78 | 0.89 | 260.38 | 10,115.97 | 0.79 | 367.9 |
| | L&B | 4,991.58 | 0.93 | 271.61 | 9,979.22 | 0.93 | 382.27 |
| | GT | 4,983.52 | 0.74 | 103.26 | 9,981.59 | 0.76 | 147.41 |
| (7,0.7) | Chao | 4,949.11 | 0.62 | 126.02 | 9,898.42 | 0.40 | 178.21 |
| | Zelterman | 5,047.93 | 0.91 | 235.5 | 10,095.19 | 0.83 | 333.02 |
| | L&B | 4,997.08 | 0.94 | 238.75 | 9,991.79 | 0.94 | 336.53 |
| | GT | 4,988.05 | 0.79 | 95.14 | 9,986.51 | 0.80 | 135.6 |
| | | | $N=50,000$ | | | $N=100,000$ | |
| (5,0.625) | Chao | 49,207.51 | 0.00 | 428.48 | 98,408.10 | 0.00 | 605.65 |
| | Zelterman | 50,566.13 | 0.22 | 821.89 | 101,118.59 | 0.03 | 1161.5 |
| | L&B | 49,874.98 | 0.89 | 851.5 | 99,733.61 | 0.85 | 1,202.35 |
| | GT | 49,975.62 | 0.79 | 332.7 | 99,967.68 | 0.79 | 471.06 |
| (7,0.7) | Chao | 49,484.91 | 0.00 | 398.06 | 98,970.86 | 0.00 | 562.94 |
| | Zelterman | 50,459.75 | 0.31 | 743.49 | 100,920.42 | 0.06 | 1,051.51 |
| | L&B | 49,937.42 | 0.93 | 748.99 | 99,873.30 | 0.92 | 1,058.89 |
| | GT | 49,979.48 | 0.82 | 305.45 | 99,977.56 | 0.83 | 432.52 |

Table 3: Coverage Probability and average length of $95\%CI$ of $N$; $p_i \sim NB(\mu=7,\kappa,\pi)$

| $NB(\kappa,p)$ | Estimator | Mean $\hat{N}$ | Coverage Probability | Average Length | Mean $\hat{N}$ | Coverage Probability | Average Length |
|---|---|---|---|---|---|---|---|
| (5,0.42) | Chao | 4,989.93 | 0.80 | 43.06 | 9,979.57 | 0.71 | 60.72 |
| | Zelterman | 5,096.73 | 0.66 | 210.44 | 10,187.49 | 0.22 | 295.20 |
| | L&B | 5,000.37 | 0.93 | 76.24 | 9,998.70 | 0.94 | 106.12 |
| | GT | 4,996.48 | 0.86 | 33.60 | 9,995.88 | 0.88 | 48.16 |

Table 3: Coverage Probability and average length of $95\%CI$ of $N$; $p_i \sim NB(\mu = 7, \kappa, \pi)$ (Cont.)

| $NB(\kappa, p)$ | Estimator | Mean $\hat{N}$ | Coverage Probability | Average Length | Mean $\hat{N}$ | Coverage Probability | Average Length |
|---|---|---|---|---|---|---|---|
| (5,0.42) | Chao | 4,989.93 | 0.80 | 43.06 | 9,979.57 | 0.71 | 60.72 |
| | Zelterman | 5,096.73 | 0.66 | 210.44 | 10,187.49 | 0.22 | 295.20 |
| | L&B | 5,000.37 | 0.93 | 76.24 | 9,998.70 | 0.94 | 106.12 |
| | GT | 4,996.48 | 0.86 | 33.60 | 9,995.88 | 0.88 | 48.16 |
| (7,0.5) | Chao | 4,995.31 | 0.88 | 33.40 | 9,990.75 | 0.85 | 47.30 |
| | Zelterman | 5,058.59 | 0.90 | 162.76 | 10,113.95 | 0.56 | 229.39 |
| | L&B | 5,000.78 | 0.93 | 55.37 | 10,000.81 | 0.94 | 77.61 |
| | GT | 4,998.14 | 0.90 | 26.38 | 9,997.93 | 0.91 | 37.77 |
| (5,0.42) | Chao | 49,895.49 | 0.17 | 135.60 | 99,790.74 | 0.02 | 191.71 |
| | Zelterman | 50,915.04 | 0.00 | 656.88 | 101,823.82 | 0.00 | 928.15 |
| | L&B | 49,984.39 | 0.93 | 234.55 | 99,966.72 | 0.92 | 331.15 |
| | GT | 49,993.76 | 0.90 | 109.26 | 99,992.98 | 0.90 | 154.89 |
| (7,0.5) | Chao | 49,951.21 | 0.55 | 105.32 | 99,903.02 | 0.29 | 148.96 |
| | Zelterman | 50,539.55 | 0.00 | 505.12 | 101,076.10 | 0.00 | 714.04 |
| | L&B | 49,995.04 | 0.94 | 170.21 | 99,989.68 | 0.94 | 240.43 |
| | GT | 49,996.74 | 0.92 | 85.39 | 99,996.92 | 0.92 | 121.00 |

Table 4: The number of heroin users that contacted the treatment centres in 2002; $n = 9{,}302$.

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $f_i$ | 2,176 | 1,600 | 1,278 | 976 | 748 | 570 | 455 | 368 | 281 | 254 | 188 | 138 | 99 | 67 | 44 | 34 | 17 | 3 | 3 | 2 | 1 |

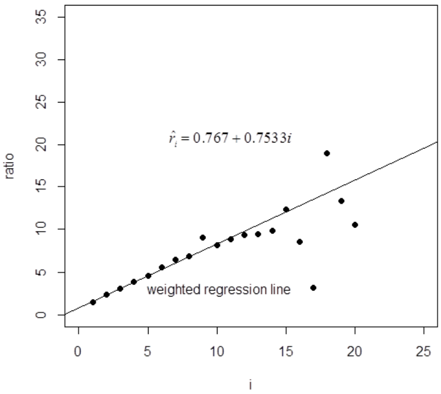


Figure 1: Show the scatter of ratio plot and weighted regression line. (Heroin users in Bangkok)

Table 5: The estimated number of heroin users in Bangkok, Thailand 2002.

| Method | $\hat{N}$ | $Se(\hat{N})$ | $95\%CI$ |
|---|---|---|---|
| Chao | 10,782 | 80.21 | 10,624-10,940 |
| Zelterman | 12,078 | 184.54 | 11,716-12,440 |
| L&B | 11,714 | 250.16 | 11,223-12,205 |
| GT | 12,126 | 73.16 | 11,983-12,270 |

## 4. Real data example

To illustrate the potential of the proposed method, we consider the real data set. The population of interest is drug users addicted to heroin in Bangkok, Thailand in 2002. The list of the surveillance system is from 61 private and public treatment centres in the Bangkok metropolitan area. The information is constructed on the basis of frequencies of the treatment episodes permitted to treat drug addicts and arise from the surveillance system of the Office of the Narcotics Control Board (ONCB) of the Ministry of Public Health (Thailand). More details of the data source are provided in Lanumteang [8].

Presented in Table 4 is the number of heroin users that contacted the treatment centres in 2002. The variables of interest are the number of occasions that a specific drug user contacted the treatment. It is found that the observed number of heroin users n = 9302.

Table 5 shows the estimated total number of the heroin users. All estimators have similar estimated numbers of the heroin users. It clearly reveals that $\hat{N}_{GT}$ gives the highest estimate. In addition, $\hat{N}_{GT}$ provides not only the lowest standard error but also the shortest confidence interval. As a result, $\hat{N}_{GT}$ is the best and becomes the candidate for use. Whereas the Chao estimator yields the smallest estimate and the Lanumteang&Böhning produces the highest standard error.

## 5. Conclusion

With regard to making inference about the unknown size *N* of a closed population, we propose the variance estimation and confidence interval estimation based upon the Generalized Turing estimator. To evaluate the performance of the estimated variance, a simulation study is done comparing with empirical variance of $\hat{N}_{GT}$. The simulation results provide the evidence that the proposed variance estimation represents well the true variance of $\hat{N}_{GT}$ in the case of large population mean $\mu$. Then, we concern construction the confidence interval for population size *N*. A simulation is conducted again to assess performance of CIs obtained from the proposed approach. Simulation results indicate that the coverage probabilities of the proposed confidence interval are close to confidence level in the case of large population mean $\mu$ for all population sizes. The proposed confidence interval tends to perform reasonably well and can be an alternative method for population size estimation in the case of large population.

Estimating confidence interval of *N* is not an easy task for capture-recapture studies in particular derivation of $Var\left(\hat{N}\right)$ which is important for forming the intervals. One could use a bootstrap approach such as percentile bootstrapping. However, there is a drawback of bootstrap method due to long run time for computation. Nonetheless, it might be useful to investigate the performance of CI using the proposed variance estimation and bootstrap percentile method. This would be following up in future work.

## Acknowledgements

We are grateful to the Research Professional Development Project under the Science Achievement Scholarship of Thailand for providing funding for Parawan Pijitrattana.

## References

[1] Chao A. and Huggins R.M. Modern Closed-Population Capture-Recapture Model. Handbook of Capture-Recapture, eds. S.C. Amstrup, T.L. Mcdonald and B.F.J. Manly, Princeton : Princeton University Press. 2005; 58-87.

[2] Tercero F. and Anderson R.Measuring transport injuries in a developing country: an application of the capture-recapture method.Accident Analysis and Prevention.2002;13-20.

[3] van der Heijden P,Cruyff M, and van Houwelingen H. Estimating the Size of a Criminal Population From Police Records Using the Truncated Poisson Regression Model. Statistica Neerlandica, 2003;1-16.

[4] Robles Sc , et al. An application of capture-recapture methods to the estimation of completeness of cancer registration. J Clin Epidemiology.1988; 495-501.

[5] van der Heijden P, Bustami R, Cruyff M, Engbersan G, and Houwelingen H. Point and interval estimation of population size using the truncated Poisson regression model. Statistical Modelling.2003; 3:305-322.

[6] Chao A. Estimating the Population Size for Capture-Recapture Data with Unequal Catchability. Biometrics. 1987; 43, 783-791.

[7] Zelterman D. Robust Estimation in Truncated Discrete Distributions with Application to Capture-Recapture Experiments. Journal of Statistical Planning and Inference**.** 1998**;** 18: 225–237.

[8] Lanumteang K. and Böhning D. An extension of Chao’s estimator of population size based on the first three capture frequency counts. Computational Statistics and Data Analysis,55 .2011;2302-2311.

[9] Böhning D, Baksh M.F, Lerdsuwansri R, and Gallagher. Use of the Ratio Plot in Capture-Recapture Estimation. Journal of Computational and Graphical Statistics,Volume22,Number1 .2013;135-155.

[10]Böhning,D. A Simple Variance Formula for Population Size Estimators by Conditioning. Statistical Methodology,5.2008;410-423.

[11] Y.M.M. Bishop, S.E. Fienberg, P.W. Holland. Discrete Multivariate Analysis : Theory and Practice,MIT Press,Cambridge.1975;493.

# A Robustness Study of Fed-Batch Fermentation for the Plasmid DNA Production Using Definitive-Screening Design

Sarawuth Noppiboon[1], Panit Kitsubun[2, 3*], Lalintip Hocharoen[3] Jindarat Pimsamarn[1], Saengchai Akeprathumchai[4], Yaowaluck Maprang Roshorm[4]

*[1]Department of Chemical Engineering, King Mongkut's University of Technology Thonburi, Thung Khru, Bangkok 10140, Thailand.*

*[2]Biopharmaceutical Research and Development Laboratory, National Center for Genetic Engineering and Biotechnology, Klong Luang, Pathumthani 12120, Thailand, panit.kit@biotec.or.th*

*[3]National Biopharmaceutical Facility, King Mongkut's University of Technology Thonburi, Bang Khun Thian, Bangkok 10150, Thailand, panit.kit@nbf.kmutt.ac.th*

*[4]School of Bioresources and Technology, King Mongkut's University of Technology Thonburi, Bang Khun Thian, Bangkok 10150, Thailand.*

## Abstract

A systematic approach to characterize the fed-batch fermentation process of the plasmid DNA production by *E.coli* DH5α was developed using design of experiments. This study aims to determine parameters which are temperature, pH, dissolved oxygen, cultivation time and feed rate, whether they have important effects on volumetric yield (mg pDNA/L) and supercoiled plasmid content, and define the operational ranges of these parameters. The experimental designs were created using the definitive-screening designs (DSDs) for the parameters at three levels. The DSDs require 16 runs including 3 additional center points. The results were analyzed based on the subset of all possible models with corrected Akaike information criterion to estimate how the parameters relatively affect the responses and build the predictive quadratic models. Finally, the Monte Carlo simulations combined with the predictive models were used to simulate the operational ranges to demonstrate the robustness of the process. Herein the study showed that the important process parameters influenced the volumetric yield and supercoiled content were temperature and cultivation time. The robustness of the process was achieved with acceptable criteria.

*Keywords*: Robustness study, Plasmid DNA production, Definitive-screening designs, corrected Akaike information criterion, Monte Carlo simulations.

*Corresponding Author
E-mail Address: panit.kit@biotec.or.th, panit.kit@nbf.kmutt.ac.th

## 1. Introduction

Implementing a robust fermentation process is a challenge during the early stages of process development due to costly and time-consuming trials. One efficient tool to achieve a proper process robustness is design of experiment (DOE) by using Definitive-screening designs (DSDs) [1, 2]. The main advantage of DSDs is an increased amount of information gained from fewer of trials when compared to conventional DOE.

The goals of the DOE in the robustness studies are identifying which process parameters have important or significant effects on the responses, and thus, define the operational ranges of manufacturing process that delivers products of pre-determined specification [3].

By using definitive-screening designs, the influences of temperature, pH, dissolved oxygen (%DO), cultivation time and feed rate on the volumetric yield and supercoiled plasmid content (%SC) were characterized in this study.

## 2. Methodology

### *2.1 Fermentation Experiments*

The experiments were carried out with *E. coli* strain DH5α bearing the pTH.PRRSV_*GP5* plasmid. The cultivations were carried out in 2 L bioreactors equipped with the BIOSTAT B Plus.

### *2.2 Definitive-Screening Design*

Jones and Nachtsheim [2] proposed a new class of designs called Definitive-screening designs, which had the desirable properties as following;

1. For m factors, the DSDs require 2m+1 runs if m is even and 2m+3 if m is odd.

2. All main effects are uncorrelated with quadratic and two-factor interaction effects.

3. The second-order effect are partially correlated with quadratic and two-way interaction effects.

The correlation plot of Plackett-Burman (PB) design (Fig. 1) was shown that PB designs had the complex aliasing and could not capture any curvature terms comparing to DSDs (Fig.2).

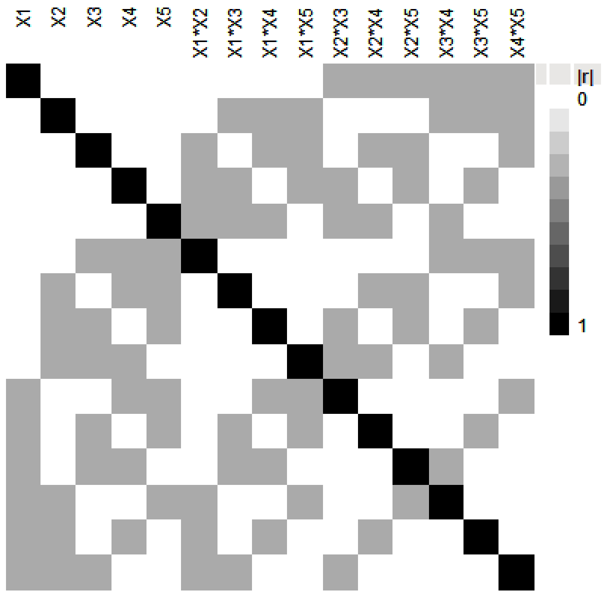


Figure 1: Absolute values of column correlations for 5 factors 12 runs Plackett-Burman design.

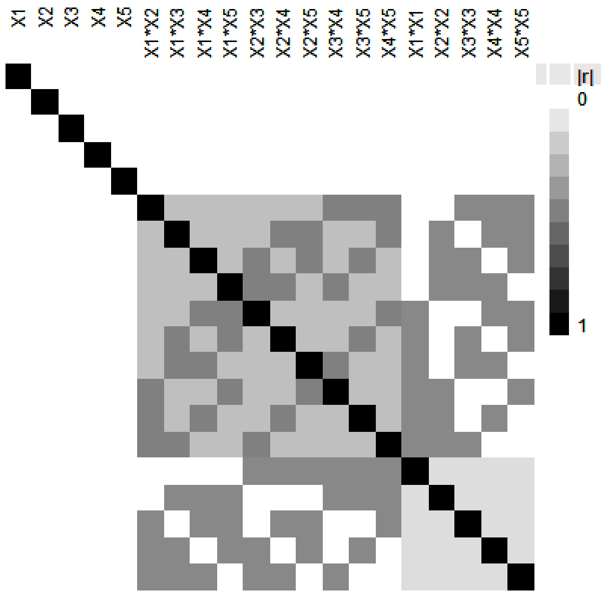


Figure 2: Absolute values of column correlations for 5 factors 13 runs of Definitive-Screening design.

In this study, the DSDs require 16 runs including 3 additional center points to estimate the pure error as shown in Table 1.

### *2.3 Model Building Strategy*

The relationship between the responses and factors can be described by using the quadratic predictive model.

$$y_i=\beta_0+\sum_{i=1}^{m}\beta_j x_{i,j}+\sum_{j=1}^{m-1}\sum_{k=j+1}^{m}\beta_{jk}x_{i,j}x_{i,k}+\sum_{i=1}^{m}\beta_{jj}x_{i,j}^{2}+\epsilon_i\,,$$

*i=1, ..., 2m+1 or 2m+3*

As mentioned earlier, DSDs allow us to estimate the main effects and second-order effects of the quadratic model. The model building strategy was based on all possible model with corrected Akaike information criterion (AICc).

### *2.4 Relative Importance of Variables*

Relative importance of variables [4] were based on the smallest value of AICc. By making all candidate set of *R* models, important effects can be refined by Akaike weights ($w_i$).

Akaike weights ($w_i$) are

$$w_i=\frac{exp(-(AICc_i\text{-}AICc_{min})/2)}{\sum_{r=1}^{R}exp(-(AICc_i\text{-}AICc_{min})/2)}$$

$w_+(j)$ are summed Akaike weights for all candidate set of models containing variable $x_j,j=1,...,R$. The variable with the largest weight, $w_+(j)$, is estimated to be the most important; the variable with the smallest sum is estimated to be the least important variable.

## 3. Results, Data Analysis and Discussion

### *3.1 Predictive Model Building*

In this study, the number of factors was five, and the full quadratic model contained 5 main effects, 5 quadratic effects, 10 two-factor interactions, and an intercept, for a total of 21 terms. Because this DSDs (super-saturated design) could estimate only 11 terms and assume the number of active factor followed by the principle of effect sparsity. Therefore, the models building strategy was used to detect the active factors.

The DSDs data in Table 1 was used to develop a predictive models using subsets of all possible models techniques with corrected Akaike information criterion (AICc) as described by Erler *et al.* [1] to fit the models that best predict the behavior of the process. In figure 3, the AICc value *vs*. the number of terms plot showed that the number of terms with the low value of AICc potentially describes the behavior of the process. Then, the candidate set of potentially best predictive model was chosen with the lowest prediction error of sum square (PRESS). The resulting selected models showed that these models were statistically significant (P value < 0.05) and provided a good description of the process (Fig. 4 and 5). Finally, the predictive models for this fermentation process were as following:

*%SC plasmid content*

$$\begin{aligned}&=77.69-1.06(\%DO)+0.62(Feed\ Rate)\\&-1.60(Cultivation\ Time)-2.04(pH)(pH)\\&+1.44(pH)(Cultivation\ time)\\&+0.86(Temperature)(Cultivation\ time)\end{aligned}$$

*Volumetric yield*

$$\begin{aligned}&=105.34+7.46(pH)+12.04(Temperature)\\&+5.37(Feed\ Rate)-7.70(pH)(Temperature)\\&-16.00(Temperature)(Temperature)\end{aligned}$$

Table 1: Definitive-screening design for pDNA production.

| Run | pH | Temperature | %DO | Feed Rate | Cultivation Time | *Volumetric Yield (mg/L)* | *%SC* |
|---|---|---|---|---|---|---|---|
| 1 | 0 | -1 | -1 | -1 | -1 | 73.12 | 80.35 |
| 2 | 0 | 1 | 1 | 1 | 1 | 98.08 | 76.53 |
| 3 | -1 | 0 | 1 | -1 | -1 | 85.16 | 76.88 |
| 4 | 1 | 0 | -1 | 1 | 1 | 121.60 | 76.73 |
| 5 | -1 | 1 | 0 | 1 | -1 | 117.68 | 78.10 |
| 6 | 1 | -1 | 0 | -1 | 1 | 90.00 | 75.26 |
| 7 | -1 | -1 | 1 | 0 | 1 | 64.56 | 79.18 |
| 8 | 1 | 1 | -1 | 0 | -1 | 96.96 | 76.15 |
| 9 | 1 | 1 | 1 | -1 | 0 | 101.72 | 73.27 |
| 10 | -1 | -1 | -1 | 1 | 0 | 67.32 | 77.40 |
| 11 | 1 | -1 | 1 | 1 | -1 | 91.72 | 77.02 |
| 12 | -1 | 1 | -1 | -1 | 1 | 92.72 | 73.85 |
| 13 | 0 | 0 | 0 | 0 | 0 | 118.76 | 79.45 |
| 14 | 0 | 0 | 0 | 0 | 0 | 103.40 | 76.90 |
| 15 | 0 | 0 | 0 | 0 | 0 | 117.20 | 76.50 |
| 16 | 0 | 0 | 0 | 0 | 0 | 86.20 | 78.13 |

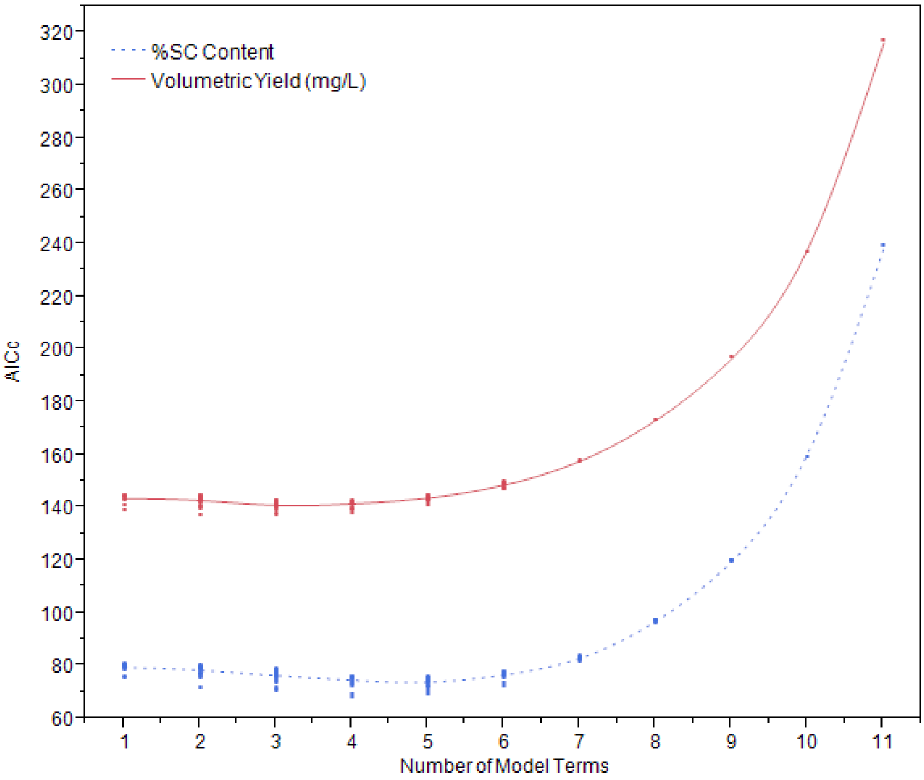


Figure 3: AICc plot for the candidate sets of model for model building.

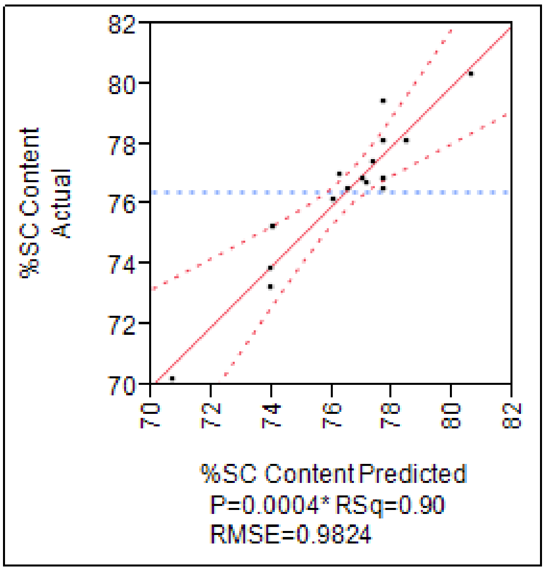


Figure 4: Actual by predict plot of %SC plasmid content.

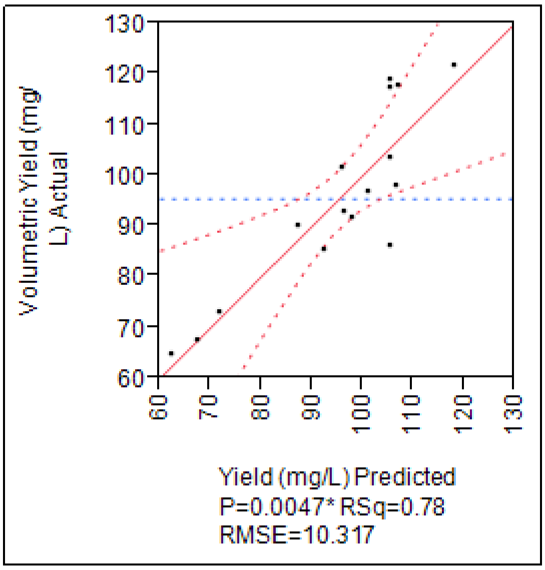


Figure 5: Actual by predict plot of volumetric yield.

*3.2 Monte Carlo Simulations*

To demonstrate the process robustness, the Monte Carlo simulations were then used to simulate 100,000 runs. In figure 6, the assumed distributions of each factor were translated to assess the variability of the volumetric yield and %supercoiled plasmid content.

*3.3 Relative Importance of Variables*

Relative importance of variables analysis, based upon sum of Akaike weights ($w_+(j)$) with the all subset of possible models, was used as a tool that provides an approach to generating a ranked order of each variables.

In figure 7 and 8 showed the estimation of the parameters that relatively affect the volumetric yield and %supercoiled plasmid content.

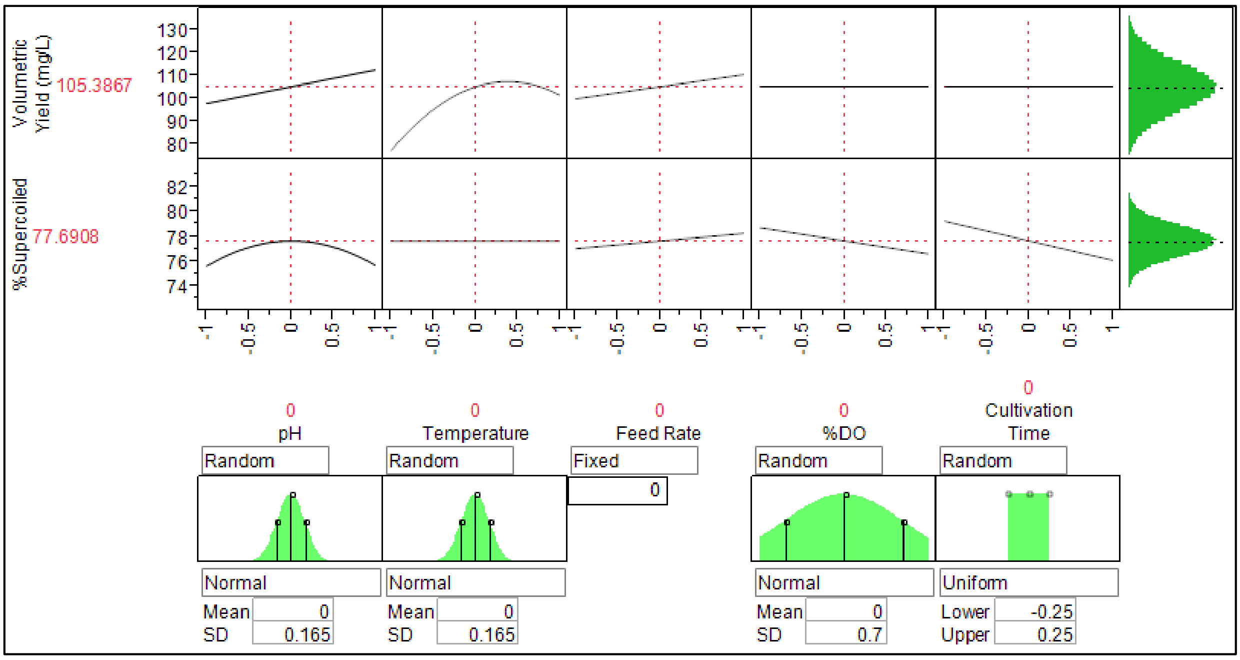


Figure 6: Monte Carlo simulation for a robustness studies.

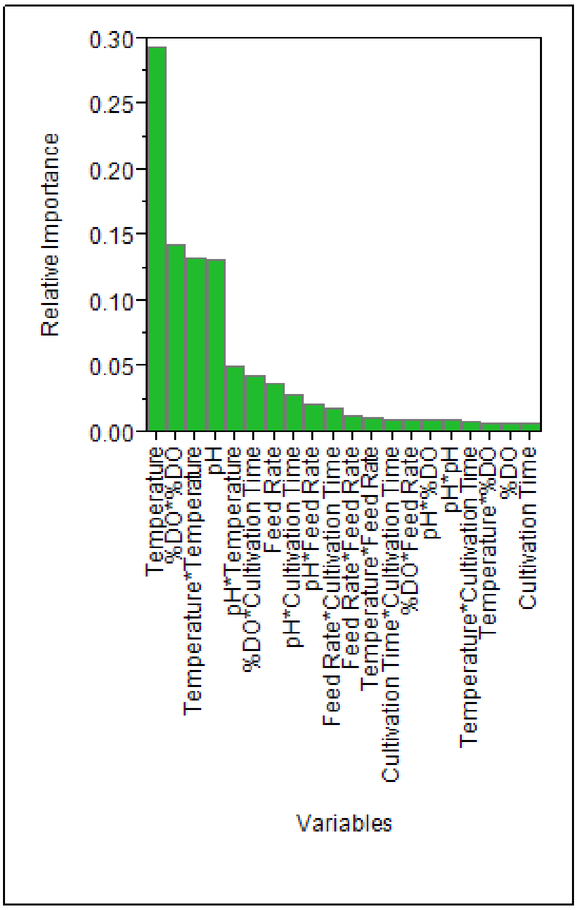


Figure 7: The relative importance of variable for volumetric yield.

## Conclusion

The definitive-screening design was proven as a powerful tool in process robustness of significant fed-batch fermentation process parameters. The predictive models, Monte Carlo simulations and relative importance of variables techniques can be used to increase knowledge gained for the improvement of the process performance during the process development.

## Acknowledgements

We are grateful for a financial support by National Science Technology and Innovation Policy Office. We thank Dr. Yaowaluck Maprang Roshorm for plasmid pTH .We thank School of Bioresources and Technology as well as National Biopharmaceutical Facility, King Mongkut's University of Technology Thonburi for their contributions.

## References

[1] Erler A, de Mas N, Ramsey P, Henderson G. Efficient biological process characterization by definitive-screening designs: the formaldehyde treatment of a therapeutic protein as a case study. Biotechnology letters. 2013;35(3):323-9.

[2] Jones B, Nachtsheim CJ. A Class of Three-Level Designs for Definitive Screening in the Presence of Second-Order Effects. Journal of Quality Technology. 2011;43(1).

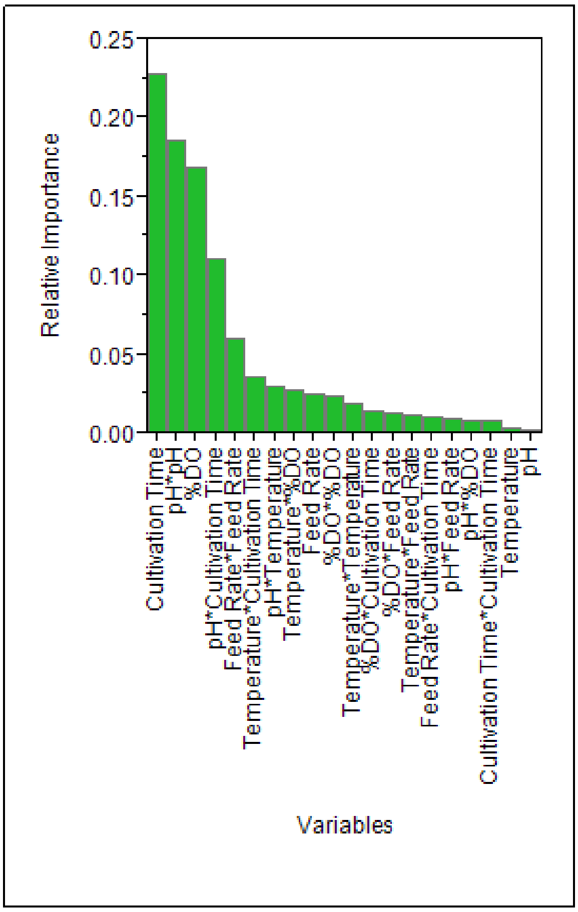


Figure 8: The relative importance of variable for %SC plasmid content.

[3] PDA. PDA Technical Report No. 60 – Process Validation: A Lifecycle Approach 2013.

[4] Burnham KP, Anderson DR. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Second Edition ed. New York: Springer; 2002.

# The Development of Mathematics Learning Achievement on the Topic of Applications of Quadratic Equations for Mathayomsuksa 4 Students by Technique of Mathematical Model

Jaruwan Suwannalai[1], Sukuman Sarikavanij[1]* and Adisak Pongpullponsak[1]
[1] Department of Mathematics, King Mongkut's University of Technology Thonburi, Bangkok, Thailand,
netadin@hotmail.com, sukuman@hotmail.com, adisak.pon@kmutt.ac.th

## Abstract

The objective of this research is to develop mathematics learning achievement on the topic of applications of quadratic equations for Mathayomsuksa 4 students by using the technique of mathematical model. The samples are 72 students in Mathayomsuksa 4 from Satri Si Suriyothai School in semester 2-2014. They are classified with random into two groups. One group is randomly selected as a control group of 36 students taught by conventional method, and another group as an experimental group of 36 students taught by the technique of mathematical model. The research duration is 12 periods of 50 minutes. The results show that the efficiency of lesson created by using the technique of mathematical model on the topic of "application of quadratic equation" is 66.79/60.85. The students have ability to use mathematics in solving problem situations. Moreover, the learning achievement of students taught by the technique of mathematical model is higher than those taught by the conventional method.

*Keywords:* Mathematical model, Mathematics, Learning achievement, Mathematics education

*Corresponding Author
E-mail Address : sukuman@hotmail.com

## 1. Introduction

Mathematics is necessary to sustain human life in working, communication, logically thinking, systematic analysis and problem solving. Mathematics is an abstract concept. If students don't have an imagination, it's hard to understand and study mathematics [1]. This reason makes the students don't attend to mathematics as teacher expect. Teacher should have instructional media such as illustration or demonstration to make the students understand. Moreover, if the students don't acquire knowledge by themselves but learn mathematics by rote, they will not able apply knowledge to solve new problem. Therefore students should practice for thinking and synthesis. The question is whether the use of tools can be used in order to promote students' problem solving ability.

In the past, mathematics learning in Thailand didn't approach to the real-life, teacher taught only by lecture and didn't focus on how to teach students to solve problem. Therefore students couldn't solve the problems by themselves [2]. In present, most teachers in Thailand teach according to the guideline from IPST (Institute for the Promotion of Teaching Science and Technology) follow to The Basic Education Core Curriculum B.E. 2551 (A.D. 2008) that is the education for all. Everyone must have chance to learn mathematics [3].

For high school in Thailand, Mathayomsuksa 4 is the first year. It's important to practice basis in learning of the higher level. Especially, students often have experience difficulties in learning the applications of quadratic equations. Teachers should find proper tools and method to improve teaching and learning of the applications of quadratic equations.

PISA (Programme for International Student Assessment) examination emphasized the real-life problem in the real-world situations. The "world" in this context means situation of nature, social and culture in students' daily life [4]. A central aspect of PISA conception of students as active problem solvers is mathematical model. It is used to help define the mathematical processes in which students solve problem [5]. The PISA reports that the strength in problem solving of Thai's students is interpretation and the weakness is mathematical thinking [6]. Teachers should manage their teaching to develop students' mathematics thinking.

From research of Areti Panaoura, we found that the students can develop the ability to solve the problems by using mathematical model. Students can study by self-learning and face the obstruction of solving problem and the mathematical model is the one of method to increase mathematical ability of students. The mathematical model in figure 1 consists of 4 steps, the first step of the problem solving procedure by the use of the mathematical model students have to consider and decide what elements are essential and what elements are less important to include in the situation model. In the next step, the situation model needs to be mathematised i.e. translated into mathematical form by expressing mathematical equations involving the key quantities and relations. For this, students need to rely on another part of their knowledge base, namely mathematical concepts, formulas, techniques and heuristics. After the mathematical model is constructed and results are obtained by manipulating the model, numerical result needs to be interpreted in relation to the real situation of the model. As a final step, the interpreted and validated result needs to be communicated in a way that is consistent with the goal or the circumstances in which the problem arose [7].

Using mathematical model is another way to improve students' learning skill of problem solving. Consequently, teaching by using mathematical model lead to make the students understand more in the content and help to increase academic achievement of students.

From the problem mentioned above, we are interested in studying the learning achievement of students taught by mathematical model and conventional method. We compare the learning achievement of students taught by two methods and provide lesson plans to develop students' mathematics thinking.

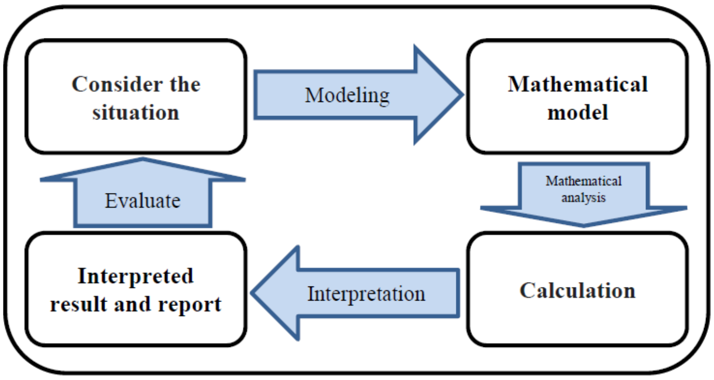


Figure 1: Mathematical model

## 2. Research Methodology

*2.1 Population and samples*

The population in this study is students of Mathayomsuksa 4 from Satri Si Suriyothai School in semester 2-2014.

The samples are two groups of Mathayomsuksa 4 student from Satri Si Suriyothai School in semester 2-2014. We randomly select 72 students from Mathayomsuksa 4 and divide into two groups. The levels of mathematics learning achievement in semester 1-2014 of student are mixed for each group. Therefore these two groups are equivalent in mathematics learning achievement. One group is randomly selected as a control group of 36 students taught by conventional method, and another group as an experimental group of 36 students taught by the technique of mathematical model.

*2.2 Research tools*

*2.2.1 Conventional Lesson Plans*

The conventional lesson plans are created according to mathematics core course textbook for Mathayomsuksa 4 – 6, 2008 [8]. The lesson is divided into five lesson plans consist of review about surface area and volume, relationship between numbers, determining the length and area of triangles and rectangles, volume of rectangular solid, cylindrical and sphere, and the application of another problem for quadratic equation with 12 periods of 50 minutes.

*2.2.2 Mathematical Model Lesson Plans*

The mathematical model lesson plans are created according to the technique of mathematical model. The lesson is divided into five lesson plans consist of review about surface area and volume, relationship between numbers, determining the length and area of triangles and rectangles, volume of rectangular solid, cylindrical and sphere, and the application of another problem for quadratic equation with 12 periods of 50 minutes. In the class, a teacher shows the real-world problems and encourages the students to transform the real-world problems to mathematical model. Students use the mathematical knowledge to solve the problem and interpret the result. These lesson plans are considered by three specialists. The average of IOC is 4.71 out of 5.

*2.2.3 Learning Achievement Test*

The researcher creates the achievement test which covers the content of the topic of "application of quadratic equation" This test is considered by three specialists. The questions of the test are selected 6 questions considering from difficulty index and discriminant index. The reliability of the test is 0.74.

The process of creating the research tools is demonstrated in Figure 2.

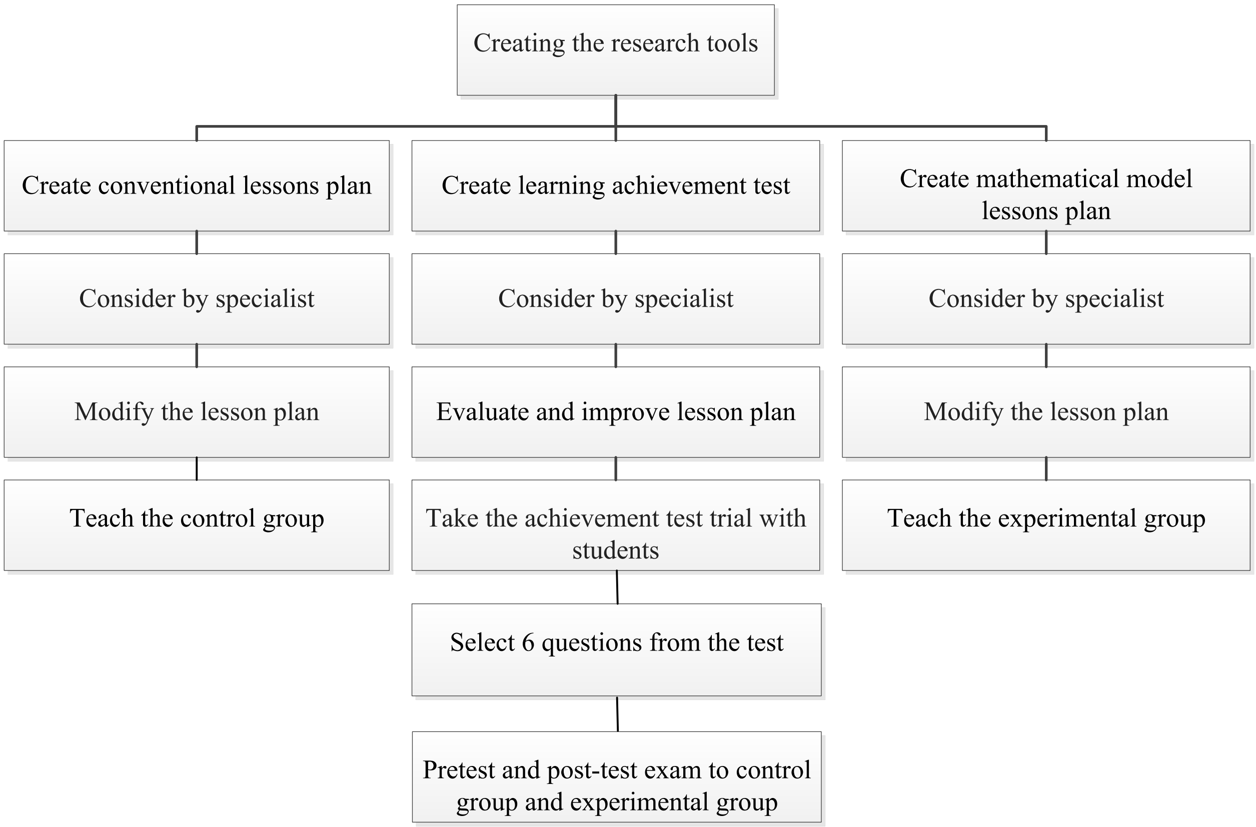

Creating the research tools
Create conventional lessons plan
Create learning achievement test
Create mathematical model lessons plan
Consider by specialist
Consider by specialist
Consider by specialist
Modify the lesson plan
Evaluate and improve lesson plan
Modify the lesson plan
Teach the control group
Take the achievement test trial with students
Teach the experimental group
Select 6 questions from the test
Pretest and post-test exam to control group and experimental group

Figure 2: Process of creating the research tools.

*2.3 Objectives*

*2.3.1* Create the mathematical model lesson plans on the topic of application of quadratic equation.

*2.3.2* Study the learning effectiveness of students on the topic of application of quadratic equation.

*2.3.3* Compare the learning achievement of students taught by mathematical model and conventional method.

## 3. Result

This research was designed in order to study the development of mathematics learning achievement on the topic of applications of quadratic equations for Mathayomsuksa 4 students by technique of mathematical model. Before teaching, the students are tested with pre-test exam to investigate the basic knowledge about applications of quadratic equations. During teaching, teacher will give the assignments to the students and the scores are collected. When teaching is finished, the post-test scores are collected. The results in Table 1 show the analysis of efficiency of lesson created by using the technique of mathematical model. The assignment score mean of 36 students taught by the technique of mathematical model is 86.83 points from the full score of 130. And the post-test score mean is 12.17 points from the full score of 20. These results show that the efficiency of lesson created by using the technique of mathematical model on the topic of “application of quadratic equation” is 66.79/60.85.

Table 1 Analysis of efficiency of lesson created by the technique of mathematical model on the topic of application of quadratic equation.

| Source of scores | Number of student | Full scores | Mean | Efficiency |
|---|---|---|---|---|
| Assignment | 36 | 130 | 86.83 | 66.79 |
| Post test | 36 | 20 | 12.17 | 60.85 |

Now we are interested in the learning effectiveness of students both on the technique of mathematical model and conventional method. Table 2 shows the means of pre-test and post-test scores. The results show that learning effectiveness of students on both methods increase and the learning effectiveness of students taught by the technique of mathematical model is higher than that taught by conventional method about 3.15%.

Table 2 Analysis of learning effectiveness of students on the topic of application of quadratic equation.

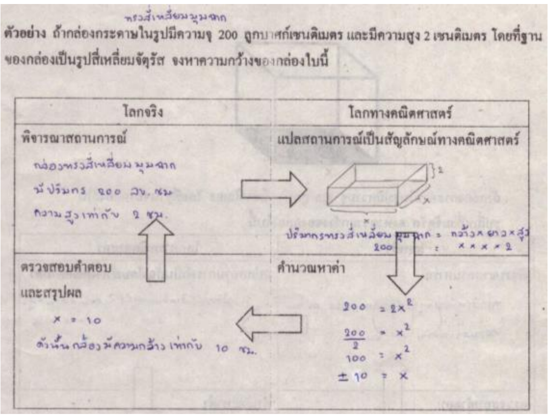
| Teaching method | Number of students | Full scores | Mean of pretest scores | Mean of post-test scores | Different mean | Percentage |
|---|---|---|---|---|---|---|
| Mathematical model | 36 | 20 | 1.39 | 12.17 | 10.78 | 53.90 |
| Conventional method | 36 | 20 | 1.50 | 11.65 | 10.15 | 50.75 |

Moreover, we compare the learning achievement of students taught by the technique of mathematical model and conventional method, the detail is showed in Table 3. From the results we found that the means of post-test scores of students taught by the technique of mathematical model and conventional method are 12.17 and 11.65, with the standard deviation 9.50 and 10.67, respectively. The test of hypothesis $H_1 : \mu_1 > \mu_2$, when $\mu_1$ and $\mu_2$ are the means of post-test score of students taught by the technique of mathematical model and conventional method, respectively is significant with significance level of .05 $(t = 2 \times 2.22,\ p = \frac{0.03}{2})$. This result means that the learning achievement of students taught by the technique of mathematical model is higher than those taught by the conventional method [9].

Table 3 Comparing learning achievement of students taught by the technique of mathematical model and conventional method.

| Teaching way | Mean | SD. | t | Sig.(2-tailed) |
|---|---|---|---|---|
| Mathematical model | 12.17 | 9.50 | 2.22 | 0.03 |
| Conventional method | 11.65 | 10.67 | | |

The examples of problem solving by using the technique of mathematical model are showed in Figure 3. We can see that students who follow the procedure of mathematical model can understand and can solve the problems in class hour.

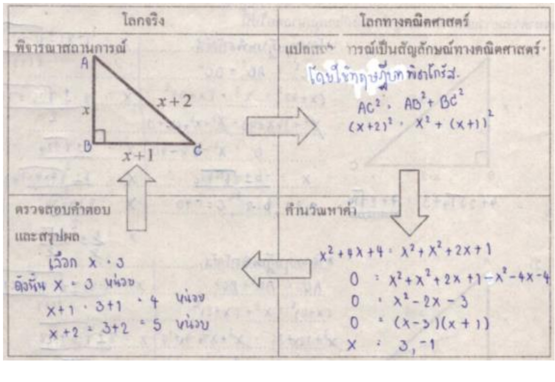


ทรงสี่เหลี่ยมมุมฉาก

ตัวอย่าง ถ้ากล่องกระดาษในรูปมีความจุ 200 ลูกบาศก์เซนติเมตร และมีความสูง 2 เซนติเมตร โดยที่ฐานของกล่องเป็นรูปสี่เหลี่ยมจัตุรัส จงหาความกว้างของกล่องใบนี้

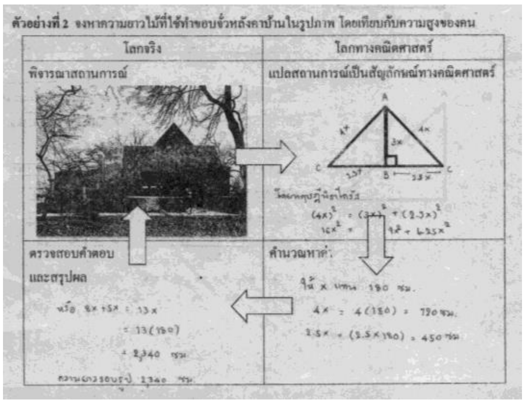
| โลกจริง | โลกทางคณิตศาสตร์ |
|---|---|
| พิจารณาสถานการณ์<br>กล่องทรงสี่เหลี่ยมมุมฉาก<br>มีปริมาตร ๒๐๐ ลบ.ซม.<br>ความสูงเท่ากับ ๒ ซม. | แปลสถานการณ์เป็นสัญลักษณ์ทางคณิตศาสตร์<br>ปริมาตรทรงสี่เหลี่ยมมุมฉาก = กว้าง×ยาว×สูง<br>200 = x × x × 2 |
| ตรวจสอบคำตอบและสรุปผล<br>x = 10<br>ดังนั้นกล่องมีความกว้างเท่ากับ 10 ซม. | คำนวณหาค่า<br>$200 = 2x^2$<br>$\frac{200}{2} = x^2$<br>$100 = x^2$<br>$\pm 10 = x$ |

2
x



| โลกจริง | โลกทางคณิตศาสตร์ |
|---|---|
| พิจารณาสถานการณ์ | แปลสถ การณ์เป็นสัญลักษณ์ทางคณิตศาสตร์<br>โดยใช้ทฤษฎีบทพิทาโกรัส<br>$AC^2 = AB^2 + BC^2$<br>$(x+2)^2 = x^2 + (x+1)^2$ |
| ตรวจสอบคำตอบและสรุปผล<br>เลือก x = 3<br>ดังนั้น x = 3 หน่วย<br>x+1 = 3+1 = 4 หน่วย<br>x+2 = 3+2 = 5 หน่วย | คำนวณหาค่า<br>$x^2+4x+4 = x^2+x^2+2x+1$<br>$0 = x^2+x^2+2x+1-x^2-4x-4$<br>$0 = x^2-2x-3$<br>$0 = (x-3)(x+1)$<br>$x = 3, -1$ |

A
x
x+2
B
x+1
C



ตัวอย่างที่ 2 จงหาความยาวไม้ที่ใช้ทำขอบจั่วหลังคาบ้านในรูปภาพ โดยเทียบกับความสูงของคน

| โลกจริง | โลกทางคณิตศาสตร์ |
|---|---|
| พิจารณาสถานการณ์ | แปลสถานการณ์เป็นสัญลักษณ์ทางคณิตศาสตร์<br>โดยทฤษฎีบทพีทาโกรัส<br>$(4x)^2 = (3x)^2 + (2.5x)^2$<br>$16x^2 = 9x^2 + 6.25x^2$ |
| ตรวจสอบคำตอบและสรุปผล<br>หรือ 8x + 5x = 13x<br>= 13(180)<br>= 2,340 ซม.<br>ความยาวรอบรูป 2,340 ซม. | คำนวณหาค่า<br>ให้ x แทน 180 ซม.<br>4x = 4(180) = 720 ซม.<br>2.5x = (2.5×180) = 450 ซม. |

A
4x
4x
3x
C
B
2.5x
C

Figure 3: Examples of problem solving by using the technique of mathematical model

## 4. Conclusion and Discussion

In this research, we want to study about mathematics learning achievement on the topic of applications of quadratic equations by using the technique of mathematical model. The research tools in this study are conventional lesson plans, mathematical model lesson plans and achievement test. The results show that the efficiency of lesson created by using the technique of mathematical model on the topic of "application of quadratic equation" is 66.79/60.85. Moreover, the learning achievement of students taught by the technique of mathematical model is higher than those taught by the conventional method.

From this research, we can see that the students taught by the technique of mathematical model have logical thinking and they can solve the problems. Students can link the situation in daily life to mathematics. It is possible that mathematical model is appropriate method to teach mathematics in another topic.

## References

[1] Hill. What is Problem Solving? [Internet]. 2001 [Updated 2001 Feb 1; cited 2014 Dec 1]. Available from: http://nzmaths.co.nz/what-problem-solving

[2] Sirinart Sirisuntorn. Why Thai's youth not good at mathematics [Internet]. 2012 [Updated 2012 Aug 27; cited 2014 Dec 5]. Available from: http://daily.bangkokbiznews.com/detail/73678

[3] The Ministry of Education Thailand. The Basic Education Core Curriculum B.E. 2551. Bangkok: Kurusapa Printing Ladphrao; 2008.

[4] IPST. PISA Thailand The Institute for the Promotion of Teaching Science and Technology. Bangkok: The Institute for the Promotion of Teaching Science and Technology publishing; 2009.

[5] OECD (2013). PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy [Internet]. 2013 [Updated 2013; cited 2015 May 2]. Available from: http://dx.doi.org/10.1787/9789264190511-en

[6] IPST. Results of the PISA 2012 mathematics, reading and science. Bangkok: The Institute for the Promotion of Teaching Science and Technology publishing; 2014.

[7] Areti Panaoura. Improving problem solving ability in mathematics by using a mathematical model: A computerized approach. Elsevier. 2012; 28(2012): 2291–2297.

[8] IPST. Book learning basic math Volume 1 Learning Mathematics Mathayomsuksa 4 - 6 Institute of Science and Technology (IPST). Bangkok: The Institute for the Promotion of Teaching Science and Technology publishing; 2012.

[9] Adisak Pongpullponsak. Research methodology. Bangkok: Publishers the company charansnitwong printing limited; 2009.

# The Effect of Organizing Mathematics Learning Activity on the Topic of Applications Using the BSCS 5E Instructional Model on Mathematics Achievement for Prathomsuksa 5 Students at School District 2 Wat Om Yai

Sirinapa Prakoobdee[1*], Angkool Wangwongchai[2] and Adisak Pongpullponsak[3]

*[1]Depaertment of Mathematics, King Mongkut's University of Technology Thonburi, Bangkok, Thailand,*
*muy_sirinapa@hotmail.com*
*[2]Department of Mathematics, King Mongkut's University of Technology Thonburi, Bangkok, Thailand,*
*angkool.wan@kmutt.ac.th*
*[3]Department of Mathematics, King Mongkut's University of Technology Thonburi, Bangkok, Thailand,*
*adisak.pon@kmutt.ac.th*

## Abstract

The objectives of this research were (1) to study the learning achievement between pre-study and post-study on the topic of application for Prathomsuksa 5 students by using the BSCS 5E instructional model; (2) to compare the mathematics learning achievement on the topic of application for Prathomsuksa 5 students who were taught by the BSCS 5E instructional model and the conventional teaching method. The researchers developed a new lesson on the topic of application which is designed to cover the teaching methods of BSCS 5E instructional model. It consists of five steps: step 1 Engagement, step 2 Exploration, step 3 Explanation, step 4 Elaboration, and step 5 Evaluation. Sixty-six students who study in Prathomsuksa 5 at School District 2 Wat Om Yai were randomly classified to an experimental and a control group. Students of each group were mixed ability of advanced, intermediate and elementary levels based on a grade point average. The statistical analysis results revealed that the learning achievement of students using the BSCS 5E instructional model is higher than the conventional teaching method.

*Keywords*: BSCS 5E instructional model, mathematics learning achievement, prathomsuksa 5 students, application

*Corresponding Author
E-mail Address: muy_sirinapa@hotmail.com

## 1. Introduction

Mathematics is a utility tool for human to pursuit their vast knowledge other than mathematic itself. The mathematics is at least improved human general knowledge and ability such as ability to calculate. This ability is using in every society, every nation, every language and religious. With only these abilities, mathematics is invaluable. Moreover, mathematics is also a tool to develop youth to be potential and grow to be valuable adult. It is the nature of the mathematics to develop youth to be reasonable, contemplate, observant, step thinking, discipline and calculated. Furthermore, mathematics is improve the "Mathematical Power" of the learner. This is the ability which will grant the learner to be able to analyze and synthesis, resolve their own problem, induction and deduction their situation or problem, able to predict, able to connect and giving reason as well as having vision and creative thinking [1].

According to the past of mathematics instruction in Thailand, we found that learning and teaching mathematics are not successful enough. A research study from the Third International Mathematics and Science Study (TIMSS) investigate and compare the results with other countries by participating in evaluation of the projects continuously such as the international cooperative study to evaluate an achievement of Mathayomsuksa 2 students from 40-50 countries in 1980-2007, it found that Thai student has science score 55.1, 510, 482 and 471 whereas math score 42.7, 516, 467 and 441 which are both below international average score. Meanwhile the other countries in Asia such as Singapore or South Korea have above average score and also tend to increase continuously (SIMS 1980, SISS 1984, TIMSS 1995, TIMSS 1999, TIMSS 2007). The cooperative study with OECD to evaluate the knowledge of reading, mathematics and science for 15 years old students in the OECD and non-OECD countries in the number of 50-60 countries in 2000, 2003 and 2006 according to PISA (Program for International Student Assessment) project, we found that 15 years old Thai students have knowledge and ability level decreasing thoroughly the project [2]. In addition, the Ordinary National Education Test (O-NET) examination conducted by the National Institute of Education Testing Service (NIETS) for assessing knowledge and idea of Prathomsuksa 6, Mathayomsuksa 3, and Mathayomsuksa 6 students was found that science and mathematics testing scores were below an average for both Prathomsuksa and Mathayomsuksa [3]. In 2001-2005, found that 65 percent of public educational institution are non-standard in terms of student learning, quantity and quality of teacher, and academic management ability of educational institution administrator. The majority of these educational institution are primary school and secondary school [4]. Therefore, it is essential that all sectors must work together to solve the educational quality problem especially, science and mathematics which are the foundation of national development and the social economy have to rely on technology and knowledge. Moreover, the mathematics which is the core subject in qualifying examination to study in higher education is less than 25 percent [5]. Therefore, it is a responsibility of the teacher to find the various teaching methods to help students for increasing their learning. Many of the teaching style are lecture-based learning because it is a traditional method. Teacher explains the lesson and content to students, and there may be a question to check for understanding at the end of the lecture. This is a teacher-centered learning, students become passive learners, and they rarely take responsibility of their own learning. It is difficult to make all students achieve the aim of learning according to established criteria. Because of each student has an individual difference, the students learning are both success and failure depends on the learning management. Teachers are also encouraged to use a variety technique that will make students aware of their defined target [6].

The National Education Act B.E.2542 and Amendments (Second National Education Act B.E. 2545) Section 22 states that education shall be based on the principle that all learners are capable of learning and self-development, and are regarded as being most important. The teaching-learning process shall aim at enabling the learners to develop themselves at their own pace and to the best of their potentiality [7].

The 5E instructional model is the concept of the learning process when there were a problem or intellectual conflict, the person will use the existing intellectual structure to interact with an environment and others around them. Intellectual conflict would be an incentive to contemplate, which is an examine activity and a hypothesis modification with a reason, and lead to the creation of a new intellectual structure [8]. By using the 5E instructional model as a guide of instructional management is consistent with the concept of the Constructivism theory that the learners construct their own knowledge. The teacher acts as a facilitator to motivate, ask questions, providing the necessary resources and create a suitable situation for learners' prior knowledge. The learners are inspired to think, to link knowledge, and a meaningful learning are stored in long term memory. This is suitable for the development of basic thinking skill of Thai children.

The teaching method of the BSCS 5E instructional model consists of five steps as follows:

**Steps 1: Engagement**

This step is a warm up students before starting a lesson with a question, situation or event to encourage enthusiasm which leads to study the issue more clearly.

**Steps 2: Exploration**

This step is an understanding the studied issue. The method might be examine, trial, practice, and inquiry to obtain a sufficient information to be used in the next step.

**Steps 3: Explanation**

This step is an analysis, interpretation and conclusion the collected information which is presented in the form of figures, tables or charts. This stage may support or reject the hypothesis but the results can contribute a knowledge creation and learning.

**Steps 4: Elaboration**

This step is to link a generating knowledge with prior knowledge. The concept and conclusion can explain the various situations to raise a broader knowledge.

**Steps 5: Evaluation**

This step is an evaluation of a learning process that gauge how much students have learned and applied knowledge to other lessons.

According to the aforementioned rationale, if the BSCS 5E instructional model is applied to teaching mathematics, it is an approach that will help to improve and solve problems in learning mathematics more effectively. Therefore, the researcher was interested in studying the learning achievement by activity using the BSCS 5E instructional model, whether there were any difference from the conventional teaching method.

The objectives of this research were (1) to study the learning achievement between pre-study and post-study on the topic of application by using the BSCS 5E instructional model; (2) to compare the mathematics learning achievement on the topic of application for Prathomsuksa 5 students who were taught by the BSCS 5E instructional model and the conventional teaching method

## 2. Research Methodology

This research was to study the achievement effect of mathematics learning activity using the BSCS 5E instructional model compare with a conventional teaching method. The methodology of this research was concluded as shown in Figure 1.

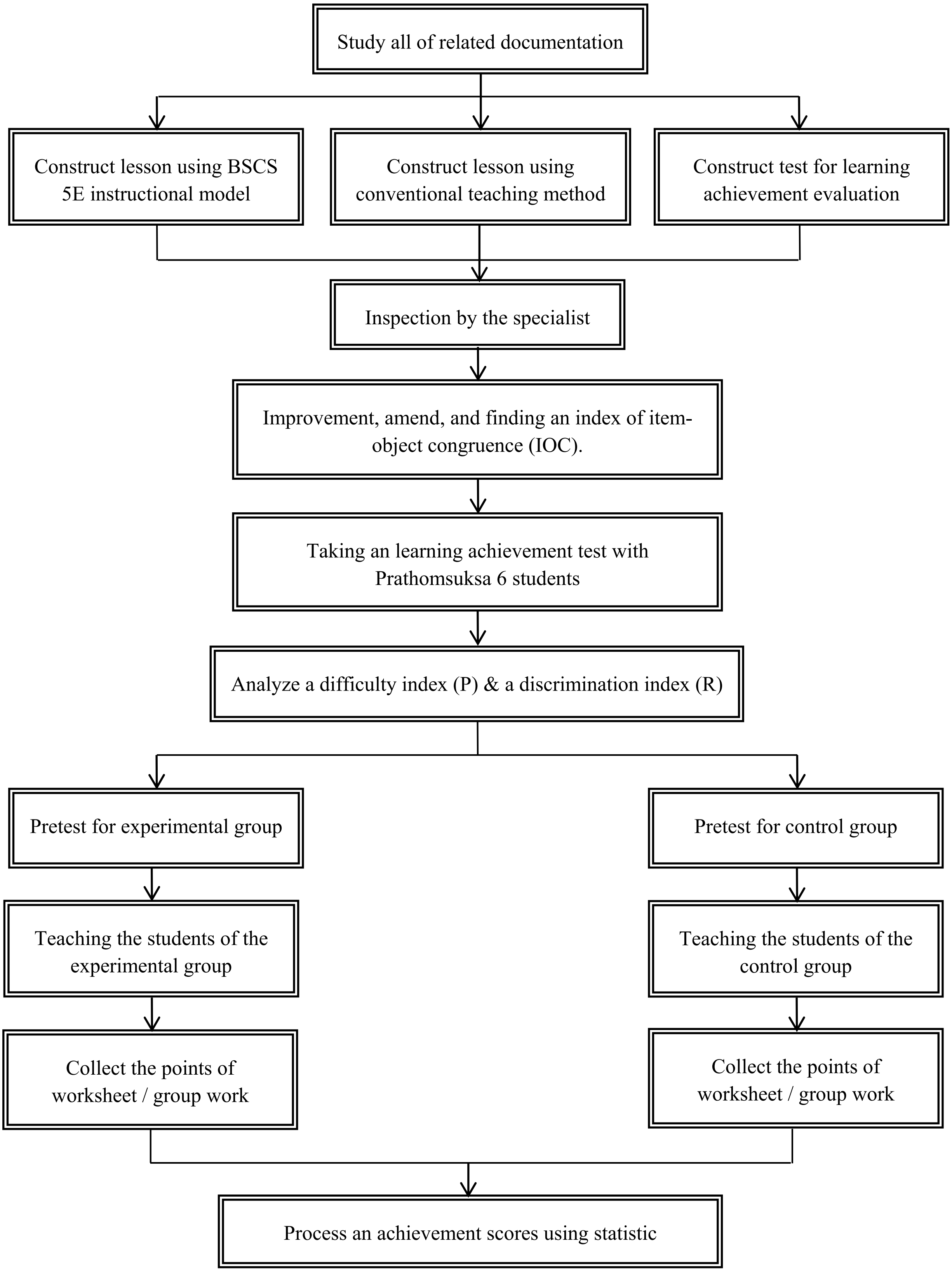

Study all of related documentation
Construct lesson using BSCS 5E instructional model
Construct lesson using conventional teaching method
Construct test for learning achievement evaluation
Inspection by the specialist
Improvement, amend, and finding an index of item-object congruence (IOC).
Taking an learning achievement test with Prathomsuksa 6 students
Analyze a difficulty index (P) & a discrimination index (R)
Pretest for experimental group
Pretest for control group
Teaching the students of the experimental group
Teaching the students of the control group
Collect the points of worksheet / group work
Collect the points of worksheet / group work
Process an achievement scores using statistic

Figure 1: Demonstration of the steps of the research.

*2.1 Samples group*

This is an experimental research. The samples were Prathomsuksa 5 students at School District 2 Wat Om Yai. The 66 students from two classrooms were divided into two equal groups according to cluster sampling, that is, experimental group and control group. The students of each group were equally mixed ability to learn. Each group included students from advanced, intermediate and elementary levels. The experimental group of 33 students was taught by the BSCS 5E instructional model, whereas the control group was taught by the conventional method.

*2.2 Tools used in the research*

*2.2.1 Regular lessons*

Researchers studied a curriculum objectives, a learning objectives and mathematics contents of Prathomsuksa 5 on the application topic which is based on the mathematics learning areas and the core curriculum B.E. 2001 of the Institute for the Promotion of Teaching Science and Technology (IPST). Researchers analyze a learning objective, knowledge, skill, and attribute to construct lessons and learning activities in accordance with a learning objective, evaluation and assessment. A total of 13 lessons contain multiplication, division and fraction, the problems of multiplication, division (the rule of three), interpretation of fractions, reading and writing fractions, writing percentage as fractions and decimals (up to 2 decimal places), percentage of counting number, and percentage problems including discount, profit, loss, sell price. Teaching time is 15 hours.

*2.2.2 5E lessons using the teaching method of BSCS*

Researchers constructed a lessons on the topic of application for an experimental group by using the BSCS 5E instructional model. It consists of title, learning plan, subject, study time, learning standards, indicators, main topics, learning objectives, contents, learning activities, materials and learning equipment, learning assessment, rubric assessment, teaching log, behavior observation form in classroom, behavior observation form in group activities. The lessons was created in accordance with the 5E model of BSCS which consist of 5 steps: step 1 engagement, step 2 exploration, step 3 explanation, step 4 elaboration and step 5 evaluation. There are 13 lessons with the same contents as a regular lessons and a teaching time is 15 hours.

*2.2.3 Creating a learning achievement test*

Researchers constructed a learning achievement test on the topic of application. It was a subjective test which consists of 3 parts of 10 questions, and created by studying the curriculum documents, assessment handbook including how to create achievement tests, data analysis. Creating an analyzing table of learning objective that are consistent with the mathematics content of the course.

## 3. Results

The regular lessons were used to teach students in a control group. Teaching time was 15 hours, pre-learning achievement test for 1 hour, and post-learning achievement test for 1 hour. Teacher collected the points during the class from student's worksheet. According to the average score, the worksheet of percentage problem has the highest average score whereas the worksheet of multiplication and division (the rule of three) has the lowest average score.

On problem solving of students whether worksheet or group work, the appointed problem was set up by teacher via situations or activities to help students to search for their knowledge with principle and reason They can expand their idea and plan how to search for knowledge on their own, without waiting to listen to a lecture from teacher only [9]. The steps of inquiry cycle (5Es) of IPST, Biology [10] was shown in Figure 2.

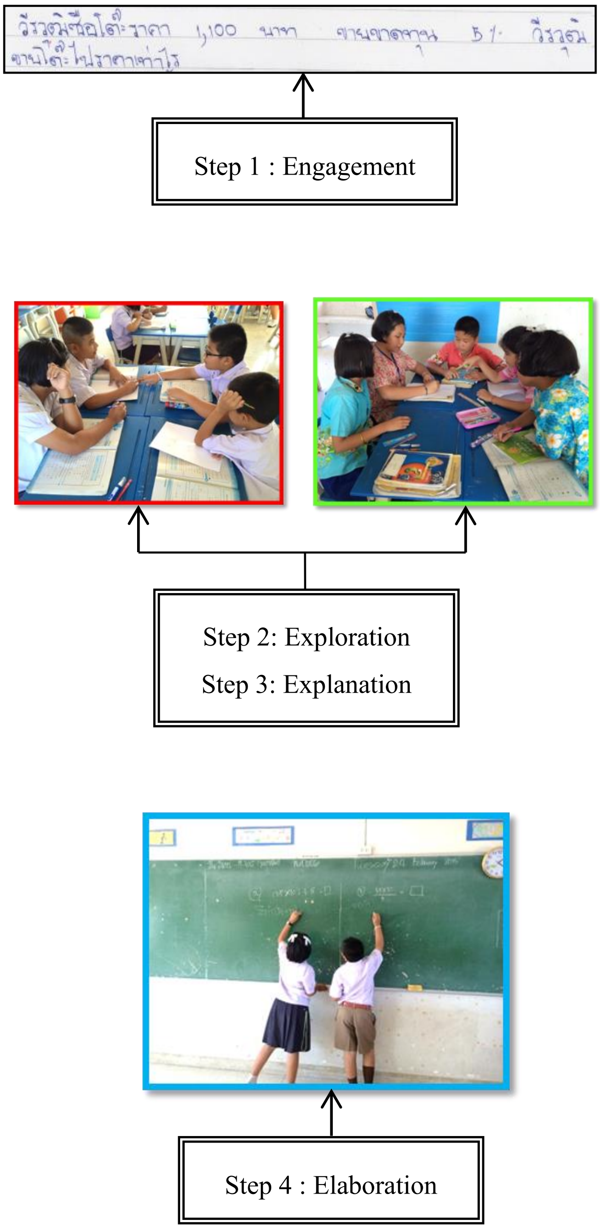
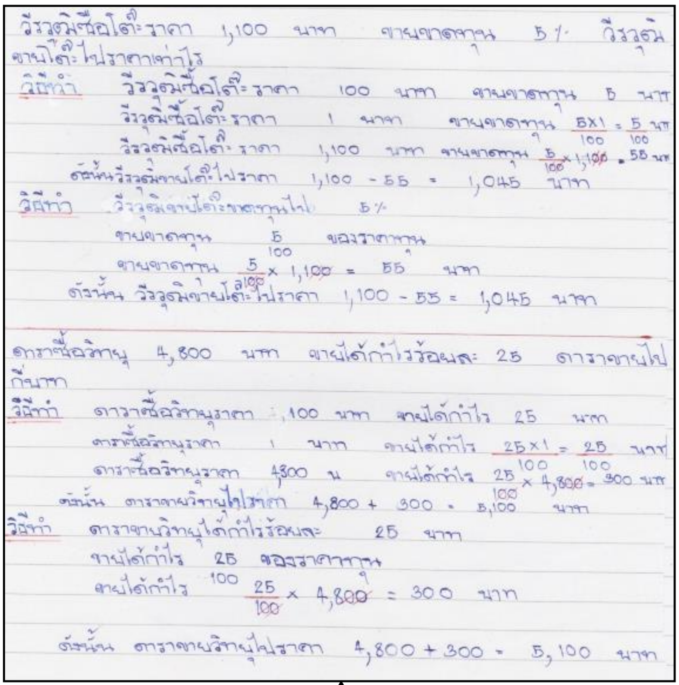


Figure 2: Five steps in the course of Inquiry (BSCS 5E).



วีรวุฒิซื้อโต๊ะราคา 1,100 บาท ขายขาดทุน 5% วีรวุฒิขายโต๊ะไปราคาเท่าไร

วิธีทำ วีรวุฒิซื้อโต๊ะราคา 100 บาท ขายขาดทุน 5 บาท
วีรวุฒิซื้อโต๊ะราคา 1 บาท ขายขาดทุน $\frac{5\times1}{100} = \frac{5}{100}$ บาท
วีรวุฒิซื้อโต๊ะราคา 1,100 บาท ขายขาดทุน $\frac{5}{\cancel{100}}\times 1{,}1\cancel{00} = 55$ บาท
ดังนั้นวีรวุฒิขายโต๊ะไปราคา 1,100 - 55 = 1,045 บาท

วิธีทำ วีรวุฒิขายโต๊ะขาดทุนไป 5%
ขายขาดทุน $\frac{5}{100}$ ของราคาทุน
ขายขาดทุน $\frac{5}{\cancel{100}}\times 1{,}1\cancel{00} = 55$ บาท
ดังนั้น วีรวุฒิขายโต๊ะไปราคา 1,100 - 55 = 1,045 บาท

ดาราซื้อวิทยุ 4,800 บาท ขายได้กำไรร้อยละ 25 ดาราขายไปกี่บาท

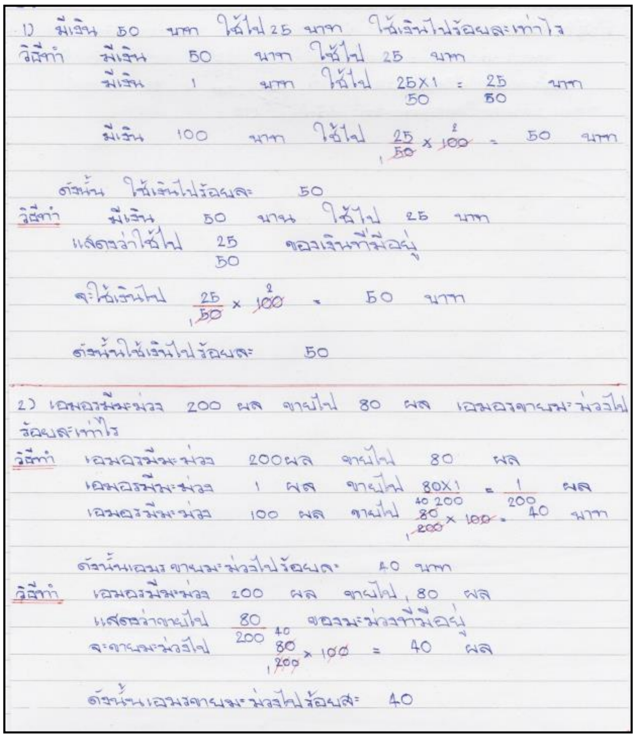
วิธีทำ ดาราซื้อวิทยุราคา 100 บาท ขายได้กำไร 25 บาท
ดาราซื้อวิทยุราคา 1 บาท ขายได้กำไร $\frac{25\times1}{100} = \frac{25}{100}$ บาท
ดาราซื้อวิทยุราคา 4,800 บ ขายได้กำไร $\frac{25}{\cancel{100}}\times 4{,}8\cancel{00} = 300$ บาท
ดังนั้น ดาราขายวิทยุไปราคา 4,800 + 300 = 5,100 บาท

วิธีทำ ดาราขายวิทยุได้กำไรร้อยละ 25 บาท
ขายได้กำไร $\frac{25}{100}$ ของราคาทุน
ขายได้กำไร $\frac{25}{\cancel{100}}\times 4{,}8\cancel{00} = 300$ บาท
ดังนั้น ดาราขายวิทยุไปราคา 4,800 + 300 = 5,100 บาท

Step 5 : Evaluation

Figure 2: (continued).

For the experimental group, it has been found that when students follow the process of the BSCS 5E instructional model, they are able to understand and solve the problems within a given time period. They also use a group brainstorming. They can not only spot problems quickly and agility in solution, but also they have a systematic thinking. The example of student's worksheet was shown in Figure 3.



1) มีเงิน 50 บาท ใช้ไป 25 บาท ใช้เงินไปร้อยละเท่าไร

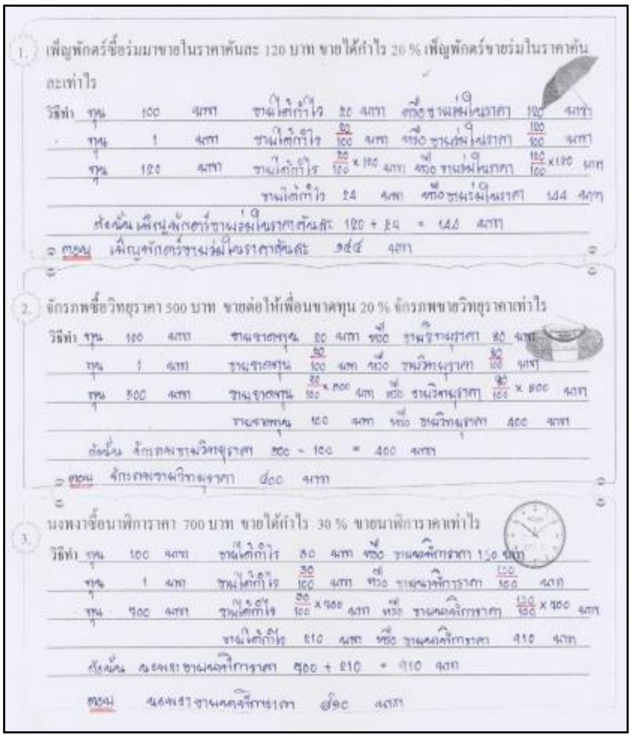
วิธีทำ มีเงิน 50 บาท ใช้ไป 25 บาท
มีเงิน 1 บาท ใช้ไป $\frac{25\times1}{50} = \frac{25}{50}$ บาท
มีเงิน 100 บาท ใช้ไป $\frac{25}{\cancel{50}}\times\cancel{100} = 50$ บาท
ดังนั้น ใช้เงินไปร้อยละ 50

วิธีทำ มีเงิน 50 บาท ใช้ไป 25 บาท
แสดงว่าใช้ไป $\frac{25}{50}$ ของเงินที่มีอยู่
จะใช้เงินไป $\frac{25}{\cancel{50}}\times\cancel{100} = 50$ บาท
ดังนั้นใช้เงินไปร้อยละ 50

2) เฉลิมศรีมีมะม่วง 200 ผล ขายไป 80 ผล เฉลิมศรีขายมะม่วงไปร้อยละเท่าไร

วิธีทำ เฉลิมศรีมีมะม่วง 200 ผล ขายไป 80 ผล
เฉลิมศรีมีมะม่วง 1 ผล ขายไป $\frac{80\times1}{200} = \frac{1}{200}$ ผล
เฉลิมศรีมีมะม่วง 100 ผล ขายไป $\frac{\cancel{80}}{\cancel{200}}\times\cancel{100} = 40$ บาท
ดังนั้นเฉลิมศรีขายมะม่วงไปร้อยละ 40 บาท

วิธีทำ เฉลิมศรีมีมะม่วง 200 ผล ขายไป 80 ผล
แสดงว่าขายไป $\frac{80}{200}$ ของมะม่วงที่มีอยู่
จะขายมะม่วงไป $\frac{\cancel{80}}{\cancel{200}}\times\cancel{100} = 40$ ผล
ดังนั้นเฉลิมศรีขายมะม่วงไปร้อยละ 40



1. เพ็ญพักตร์ซื้อร่มมาขายในราคาคันละ 120 บาท ขายได้กำไร 20 % เพ็ญพักตร์ขายร่มในราคาคันละเท่าไร

วิธีทำ ทุน 100 บาท ขายได้กำไร 20 บาท หรือขายร่มในราคา 120 บาท
ทุน 1 บาท ขายได้กำไร $\frac{20}{100}$ บาท หรือขายร่มในราคา $\frac{120}{100}$ บาท
ทุน 120 บาท ขายได้กำไร $\frac{20}{100}\times120$ บาท หรือขายร่มในราคา $\frac{120}{100}\times120$ บาท
ขายได้กำไร 24 บาท หรือขายร่มในราคา 144 บาท
ดังนั้น เพ็ญพักตร์ขายร่มในราคาคันละ 120 + 24 = 144 บาท
ตอบ เพ็ญพักตร์ขายร่มในราคาคันละ ๑๔๔ บาท

2. จักรภพซื้อวิทยุราคา 500 บาท ขายต่อให้เพื่อนขาดทุน 20 % จักรภพขายวิทยุราคาเท่าไร

วิธีทำ ทุน 100 บาท ขายขาดทุน 20 บาท หรือ ขายวิทยุราคา 80 บาท
ทุน 1 บาท ขายขาดทุน $\frac{20}{100}$ บาท หรือ ขายวิทยุราคา $\frac{80}{100}$ บาท
ทุน 500 บาท ขายขาดทุน $\frac{20}{100}\times500$ บาท หรือ ขายวิทยุราคา $\frac{80}{100}\times500$ บาท
ขายขาดทุน 100 บาท หรือ ขายวิทยุราคา 400 บาท
ดังนั้น จักรภพขายวิทยุราคา 500 - 100 = 400 บาท
ตอบ จักรภพขายวิทยุราคา ๔๐๐ บาท

3. นงพงาซื้อนาฬิการาคา 700 บาท ขายได้กำไร 30 % ขายนาฬิการาคาเท่าไร

วิธีทำ ทุน 100 บาท ขายได้กำไร 30 บาท หรือ ขายนาฬิการาคา 130 บาท
ทุน 1 บาท ขายได้กำไร $\frac{30}{100}$ บาท หรือ ขายนาฬิการาคา $\frac{130}{100}$ บาท
ทุน 700 บาท ขายได้กำไร $\frac{30}{100}\times700$ บาท หรือ ขายนาฬิการาคา $\frac{130}{100}\times700$ บาท
ขายได้กำไร 210 บาท หรือ ขายนาฬิการาคา 910 บาท
ดังนั้น นงพงาขายนาฬิการาคา 700 + 210 = 910 บาท
ตอบ นงพงาขายนาฬิการาคา ๙๑๐ บาท

Figure 3: An example of student's worksheet using the BSCS 5E instructional model.

The researchers designed a lesson on a topic of application based on the teaching method of the BSCS 5E instructional model and conducted the teaching experiment on an experiment group 15 hours. Researchers set the threshold of efficiency of the lesson relationships ( $E_1/E_2$ ) equal to 60/60 [11] where $E_1$ is the efficiency of during learning and $E_2$ is the efficiency of post learning. The results were shown in Table 1.

Table 1: An efficiency analysis of the performance of 5E lessons on topic of application by using the instructional model of BSCS.

| Type of Learning | Student quantity | Total scores | Average scores | Efficiency |
|---|---|---|---|---|
| During learning | 33 | 1,652 | 50.06 | $E_1 = \frac{50.06}{65} \times 100 = 77.02$ |
| Post learning | 33 | 403 | 12.21 | $E_2 = \frac{12.21}{20} \times 100 = 61.06$ |

Table 1 shows 33 students who were taught by the BSCS 5E instructional model, a total scores and average scores of during learning were 1,652 and 50.06 (out of 65 points) respectively. A total scores and average scores of post learning were 403 and 12.21 (out of 20 points) respectively. An Efficiency of during learning ( $E_1$ ) was 77.02, whereas an efficiency of post learning ( $E_2$ ) was 61.06. The lesson efficiency on topic of application using the instructional model of BSCS 5E was at 77.02/61.06 which was higher than the threshold of efficiency criterion of 60/60. Therefore, it is recognized that the lesson of application topic using 5E instructional model of BSCS has more efficiency to use in teaching.

The analysis of the learning achievement between pre-learning and post-learning that students were taught by the conventional teaching method and the instructional method of BSCS 5E was shown in Table 2.

Table 2: The analysis results of the learning achievement between pre-study and post-study on topic of application using the instructional model of BSCS 5E.

| Teaching method | Type of Scores | Student quantity | Average scores | S.D. | t-test | Sig.(1-tailed) |
|---|---|---|---|---|---|---|
| The BSCS 5E instructional model | Posttest | 33 | 44.67 | 15.59 | 16.459** | .000 |
| | Pre-test | | | | | |
| The conventional teaching method | Posttest | 33 | 36.15 | 18.28 | 11.360** | .000 |
| | Pre-test | | | | | |

** the level of statistical significance 0.05

According to the Table 2 we found that 33 students using the 5E instructional model of BSCS has an average score between pre- and post-study equals to 44.67, the t-test equals to 16.459 and the p value was .000 which is less than 0.05. This indicates that a learning efficiency between pre- and post-study was statistically significant difference at 0.05 level. The control group with the conventional teaching method of 33 students has an average score between pre- and post-study equals to 36.15, the t-test 11.360 and the p value was 0.00 which is less than 0.05. This indicates that a learning efficiency between pre- and post-study was statistically significant difference at 0.05 level. In conclusion, the lesson of applied chapter of Prathomsuksa 5 students using the instructional model of BSCS 5E has an average score of 44.67 whereas the conventional teaching method was 36.182, which was the difference of 8 scores, it is maybe related to another factors. However, the lesson on topic of application using the instructional model of BSCS 5E is the new teaching method which entertain and interesting to student, and result in higher average score than the conventional teaching method.

An analysis result for comparison of learning achievement of mathematics on topic “application” for Prathomsuksa 5 students between the instructional model of BSCS 5E and the conventional teaching method, and the t-test analysis was shown in Table 3.

Table 3: The comparison of learning achievement of students who were taught by the instructional model of BSCS 5E and the conventional teaching method.

| Teaching method | Mean( $\overline{X}$ ) | S.D. | t-test | Sig.(1-tailed) |
|---|---|---|---|---|
| The conventional teaching method | 52.67 | 17.38 | 2.029 | .047 |
| The instructional model of BSCS 5E | 61.09 | 16.33 | | |

According to the hypothesis $H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 > \mu_2$

Table 3 shows that the arithmetic mean of learning achievement using the instructional model of BSCS 5E was 61.09, the standard deviation was 16.33. The arithmetic mean of learning achievement using the conventional teaching method was 52.67, the standard deviation was 17.38. The t value equals to 2.029, and p value equals to .047 which was less than 0.05, therefore the hypothesis $H_0$ was rejected, and the hypothesis $H_1$ was accepted. This indicates that the learning achievement of Prathomsuksa 5 students on topic of application using the instructional model of BSCS 5E was higher than the conventional teaching method with a significant level of 0.05.

## 4. Conclusion

The results of this research shows that the lesson on the topic of application for Prathomsuksa 5 students using the 5E instructional model of BSCS has higher learning achievement than the conventional teaching method, and can help students to understand the topic of application better.

## References

[1] Chansomsak S. Mahtematics education. Education Supervisory Unit of the Department of General Education; 1976.

[2] Project TIMSS Thailand IPST. Trends in International Mathematics and Science Study 2007. Office of education standard for science, mathematics and technology, The Institute for the Promotion of Teaching Sciences and Technology publishing; 2008.

[3] National Institute of Educational Testing Service (Public Organization). Thai education crisis [Internet]. 2012 [cited 2014 Feb 20]. Available from: https://www.gotoknow.org/posts/482279

[4] Office for National of Education Standards and Quality Assessment. Problems quality of Thai education [Internet]. 2006 [cited 2014 Nov 10]. Available from: https://mameawmewmew.wikispaces.com

[5] Office of the Basic Education Commission. Learning Integration. Bangkok: Kurusapa Printing Ladphrao; 2010.

[6] Pongpullponsak A. Research methodology. Bangkok: Publishers the company charansnitwong printing Limited; 2009.

[7] Office for National of Education Standards and Quality Assessment (ITD). The National Education Act B.E. 2542, Amendments (No.2) Act B.E. 2545. Bangkok: Prikwarn graphic co. ltd; 2004.

[8] Bureau of Educational Testing, Ministry of Education. A new alternative guideline for evaluation according to basic education curriculum B.E. 2544 subject areas of science. Bangkok: Kurusapa Printing Ladphrao; 2003.

[9] Kiawkaew S. Science teaching at secondary education level. Department of education, Faculty of education, Prince of Songkla University, Pattani campus. Pattani: 1997.

[10] Office of the Basic Education Commission. Inquiry-based learning (5Es) [Internet]. 2007. Available from: http://school.obec.go.th/nitade/data/inquiry%20process.pdf.

[11] Khunprom W. The studying mathematics achievement Title : The statistic of students in high school level who gotten the teaching in the pattern of student's thinking. The 1st International Conference on Educational Reform 2007 (ICER2007). November 9-11. Mahasarakham University: 2007.

# Motivational and Constructivist Teaching Methods in Mathematics on the Topic of Function for Mathayom 4 Students.

Wirakorn Boonchawalit[1], Prof. Adisak Pongpullponsak[2*] and Dr. Teerapol Saleewong[3*]
*Mathematics, King Mongkut's University of Technology Thonburi, Bangkok, Thailand,[1] wirakorn.bo@eisth.org*
*[2]adisak.pon@kmutt.ac.th*
*[3]teerapol.sal@kmutt.ac.th*

### Abstract

Constructivism teaching is an approach to fitting new information together with what they have already known. The purposes of this study were to create constructivism lesson plan on the topic of function. It also compared the learning effectiveness of students who were taught by a constructivist teaching method and a conventional teaching method. Sixty students who studied in Mathayom 4 at Mathayom Watdusitaram School were randomly classified to an experimental group and a control group. An experimental group was taught by a constructivist teaching method and a control group was taught by a conventional teaching method. Statistical comparisons of the learners' performance on the posttest showed an advantage of constructivist teaching method over a conventional teaching method. Thus, the results indicated that constructivism theory based on teaching method encouraged and enhanced students' learning interests, and could also help students to understand the topic better.

*Keywords*: Motivational, Constructivist, Function

*Corresponding Author

## 1. Introduction

Mathematics is a course that teaches students to think logically. It is not just a number, but we can use it to prove the correctness of our thoughts. Moreover, it is a structured model that can explain the reason [1]. Even though we can understand the importance of mathematics; the teaching and learning process of mathematics is not very successful, which can be observed from the results of PISA (Program for International Student Assessment). It is affiliated with the International Student Assessment. It evaluates the quality of the educational system of the country participating by assessing the ability to use the knowledge and skills of students aged 15 in reading. Mathematics and science are evaluated once every 3 years. The most evaluation was carried out in 2012 as the main focus on the mathematics assessment. Thailand's average scores in mathematics were 427 points that is lower than an average of a national member of the OECD (Organization for Economic Co-operation and Development), at 494 points statistically significant. Thailand's average score is ranked at 49 - 52 from a total of 65 countries participated in the assessment [3] and the testing of the Ordinary National Education Test (O-NET). In 2013, Mathayom 3 students had an average score of 25.45 points and Mathayom 6 had an average score of 20.48 points out of 100 points [4].

The test results reflect the performance of teaching in Thailand. One reason of this poor performance is because the students do not understand the concepts in mathematics, and cannot use the knowledge to solve problems. When they cannot solve the problem, they do not like mathematics. The teaching and learning of mathematics for students to understand and love math is difficult because of a high level of abstraction in content of mathematics. There are many different definitions, theorems, and conditions. A curriculum that can encourage students to understand and love to learn mathematics is important [5].

The teaching based on constructivist theory is one of the popular teaching methods. Several studies show that the teaching based on constructivist theory can help improve academic achievement of students [6,7]. The teaching model based on constructivist theory is consisted of the followings [8] :

**Step 1: Create a cognitive conflict.**

A teacher presents a problem A for students to solve individually by using knowledge of existing applications to find solutions. Then, the students are divided into small groups to allow them to share their solutions within their group.

**Step 2: Contemplate activities.**

After the students shared their solutions in their own group and got the responses from other members, they select the solution that is best for their group. Then, the students created a structural problem B which has structure related to a problem A, and solve a problem B. After that, the students present a problem solving method to the rest of the class.

Each student created problem C. Then teacher collected all the problems and randomly distributed to the students. The students solved problem C by using a problem solving method.

**Step 3: Summarize.**

The teacher and students conclude the process together.

The teaching based on constructivist theory encourages students to create new knowledge based on the existing one. Interest and motivation of students are the impetus for students to learning by themselves [9].

The teacher has to consider the nature of the learning of each student, i.e., the teacher will teach in accordance to the motivation of the students. Normally, each student has a motivation to learn. With an appropriated teaching method of the teacher, it will help motivate the students [10]. Thus the objectives of the study are as follows.

1) To create constructivism lesson plan on the topic of function with effective 80/80.

2) To compare a test score between pretest and posttest of students.

3) To compare the learning effectiveness of students who were taught by a constructivist teaching method and a conventional teaching method.

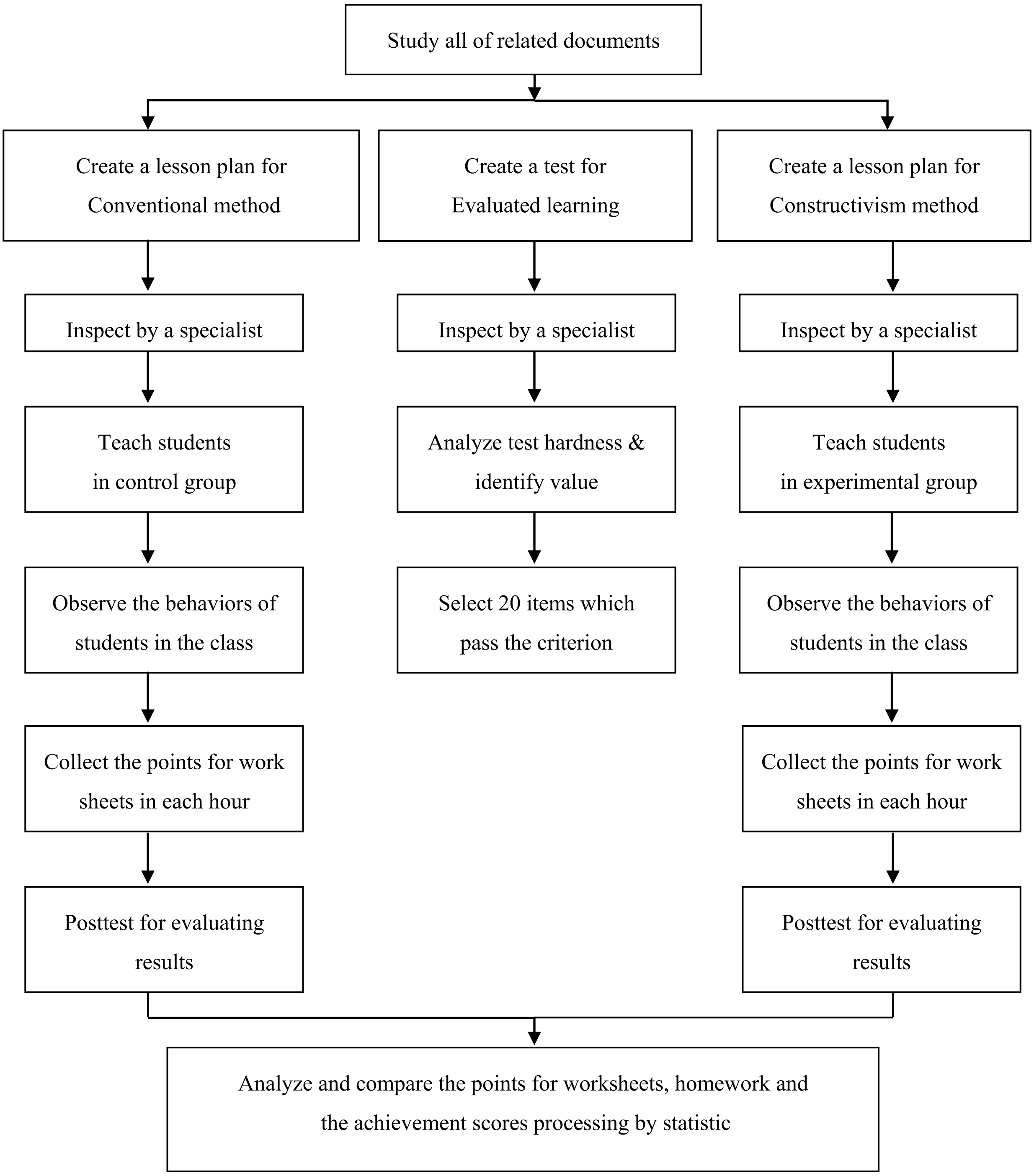

Study all of related documents
Create a lesson plan for Conventional method
Create a test for Evaluated learning
Create a lesson plan for Constructivism method
Inspect by a specialist
Inspect by a specialist
Inspect by a specialist
Teach students in control group
Analyze test hardness & identify value
Teach students in experimental group
Observe the behaviors of students in the class
Select 20 items which pass the criterion
Observe the behaviors of students in the class
Collect the points for work sheets in each hour
Collect the points for work sheets in each hour
Posttest for evaluating results
Posttest for evaluating results
Analyze and compare the points for worksheets, homework and the achievement scores processing by statistic

Figure 1: The outline demonstration of the research

## 2. Research Methodology

*2.1 Samples Group*

The sample group consists of 60 Mathayom 4 students from Mathayom Watdusitaram School. They are randomly divided into two groups. A control group of 30 students who are taught by a conventional teaching method and an experimental group of 30 students who are taught by a constructivist teaching method. Each group is consisted of all three types of students, which are classified based on an achievement of first semester, i.e., high level, middle level and low level.

*2.2 Tools Used in the Research*

*2.2.1 Conventional Lesson Plan*

Conventional lesson plan is used for 30 Mathayom 4 students who are taught by a conventional teaching method in a control group. It has eight subtopics: Definition of function, Domain and Range of function, One to one function and onto function, Increasing and decreasing function, Operation of functions, Composite function, Technical graph writing and Inverse functions.

*2.2.2 Constructivism Lesson Plan*

Constructivism lesson plan is used for 30 Mathayom 4 students who are taught by a constructivist teaching method in an experimental group. It has eight subtopics as same as conventional lesson plan. The learning activities are based on the constructivist theory.

*2.2.3 Achievement tests*

An achievement test is a multiple-choice test on the topic of function according to the learning indicators, which is used for pretest and posttest of the students who are taught by a conventional teaching method and a constructivist teaching method. Then both of the pretest and posttest scores are analyzed to measure learning effectiveness.

We conduct a pretest using an achievement test. Next, we create a conventional lesson plan to apply for a control group of students who are taught by a conventional teaching method, and create a constructivism lesson plan to apply for an experimental group of students who are taught by a constructivist teaching method. Then, we conduct a posttest using an achievement test again. After that, we analyze a pretest and a posttest score for learning effectiveness. Research process is shown in figure 1.

## 3. Results

The evaluation is done by collecting scores of student's work on classroom assignments and posttest scores. Then, the scores were analyzed for efficiency of constructivism lesson plan. The results are in Table 1

Table 1 show up 30 students who were taught by a constructivist teaching method. They had the efficiency of constructivism lesson plan during learning as 81.98. And the efficiency of constructivism lesson plan after learning as 59.67. That shows constructivism lesson plan on the topic of function having the efficiency as 81.98 / 59.67.

After that, we compare scores from pretest and posttest of students who are taught by a constructivist teaching method and a conventional teaching method. The results are shown in Table 2.

Table 1: An analysis of efficiency of constructivism lesson plan.

| Type of scores | Students quantity | Efficiency | Efficiency of constructivism lesson plan |
|---|---|---|---|
| During Learning | 30 | 81.98 | 81.98/59.67 |
| Posttest | 30 | 59.67 | |

Table 2: A comparison of pretest and posttest scores of students who were taught by a constructivist teaching method and students who were taught by a conventional teaching method.

| Teaching method | Scores | Students quantity | Mean | Std. deviation | df | t | Sig (1-tailed) |
|---|---|---|---|---|---|---|---|
| A constructivist teaching method | Posttest | 30 | 59.67 | 7.87 | 29 | 80.79* | .00 |
| | Pretest | 30 | 10.67 | 7.04 | | | |
| A conventional teaching method | Posttest | 30 | 56.17 | 9.26 | 29 | 53.38* | .00 |
| | Pretest | 30 | 10.50 | 7.11 | | | |

* The level of statistical significance is at .05

Table 3: A comparison of the learning effectiveness of students who were taught by a constructivist teaching method and students who were taught by a conventional teaching method.

| Teaching method | Students quantity | Mean | Std. deviation | df | t | Sig (1-tailed) |
|---|---|---|---|---|---|---|
| A constructivist teaching method | 30 | 49.00 | 3.32 | 29 | 3.55* | .00 |
| A conventional teaching method | 30 | 45.67 | 4.69 | | | |

* The level of statistical significance is at .05

Table 2 shows that 30 students who were taught by a constructivist teaching method have an average score of posttest 59.67 points standard deviation 7.87 and the average score of the pretest 10.67 points standard deviation 7.04, t-test 80.79 and p = .00, that is, students who were taught by a constructivist teaching method had posttest scores higher than pretest scores at the .05 significance level. Similarly, 30 students who were taught by a conventional teaching method have an average score of posttest 56.17 points with a standard deviation 9.26 and the average score of pretest 10.50 points standard deviation of 7.11, t-test 53.38 and p =.00, that is, students who were taught by a conventional teaching method posttest scores were higher than pretest scores at the .05 significance level. Likewise, it is concluded that after teaching both methods, students were developed their knowledge on the topic of function increasing before teaching.

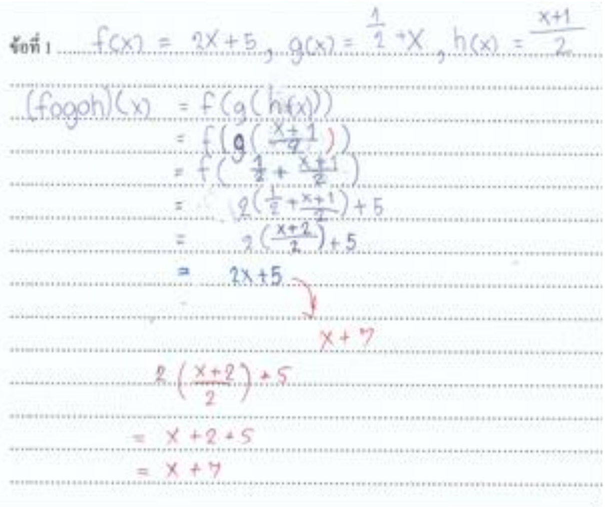
We analyzed pretest and posttest scores to compare to learning effectiveness of students who were taught by a constructivist teaching method and students who were taught by a conventional teaching method. The results are shown in Table 3, an average of learning effectiveness of students who were taught by a constructivist teaching method is 49.00 points with a standard deviation 3.32, an average of learning effectiveness of students who were taught by a conventional teaching method is 45.67 points, with a standard deviation of 4.69, t-test 3.55 and p = .00, that is the learning effectiveness of students who were taught by a constructivist teaching method is greater than students who were taught by a conventional teaching method at the .05 significance level.

## 4. Discussion

From observing students in the experimental group taught by a constructivist teaching method, we found that students are eager to learn and happy with classroom activities. A constructivist teaching method provides students with an idea of creating problems and helping them acquire a new knowledge by sharing their knowledge with the group. That is, students can learn quickly and help explaining their friends in the part of what they do not understand.

Considering the students who are taught by a constructivist teaching method, when the teacher provides a problem A and asks them to find the solution, we found that only some students can solve it properly. Most of students cannot apply their knowledge to solve the problem when they read it.

Since we divided the students into groups of mixed ability, we found that they learn to share and discuss the method used to solve the problem. The students took about 20 minutes for the learning process. After that, the teacher and students summarized the method used to solve the problem together. The example of student's work is shown in Figure 2, suggested that students most often the problem C similar to the problem A. Students can use method to solve the problem as well. But students often miss order of operations.



ข้อที่ 1 $f(x) = 2x+5$, $g(x) = \frac{1}{2}+x$, $h(x) = \frac{x+1}{2}$

$(fogoh)(x) = f(g(h(x)))$

$= f(g(\frac{x+1}{2}))$

$= f(\frac{1}{2}+\frac{x+1}{2})$

$= 2(\frac{1}{2}+\frac{x+1}{2})+5$

$= 2(\frac{x+2}{2})+5$

$= 2x+5$

$x+7$

$2(\frac{x+2}{2})+5$

$= x+2+5$

$= x+7$

Figure 2: An example of student's work.

The lesson taught by a constructivist teaching method on the topic of functions has the effectiveness of 81.98 / 59.67. The effectiveness of a constructivism lesson plan during learning is 81.98 and the effectiveness of a constructivism lesson plan after learning is 59.67, which does not conform to the expectations at 80/80. We expect that the teaching based on constructivist theory will encourage students to learn better than the conventional teaching method. As the students can help each other in the class, they have great classroom assignments scores. On the other hand, the students must use their own knowledge to solve problems that are more complex in the test. As a result, the posttest scores do not meet the expectations. However, the students had been learning this topic before the experiment; the scores might be reasonable accepting for teaching performance.

Analyzing the pretest and posttest scores of students who are taught by a constructivist teaching method and a conventional teaching method found that posttest scores of students who were taught by a constructivist teaching method is greater than the pretest scores at the .05 significance level. The average scores of the posttest and pretest are 59.67 and 10.67 respectively. The posttest of students who were taught by a conventional teaching method is greater than the pretest at the .05 significance level. The average scores of the posttest and pretest are 56.17 and 10.50 respectively.

Finally, the comparison of the learning effectiveness of the students who are taught by a constructivist teaching method and the students who are taught by a conventional teaching method shows that the learning effectiveness of the students who are taught by a constructivist teaching method is 49.00, while the learning effectiveness of the students who are taught by a conventional teaching method is 45.67. As a result, we found that the learning effectiveness of the students who are taught by a constructivist teaching method is greater than the students who are taught by a conventional teaching method at the .05 significance level.

## 5. Conclusion

The results indicate that constructivism theory based on teaching method encouraging and enhancing students' learning interests, and also can help students to understand the topic better.

## Acknowledgement

This research is funded by the Project for the Promotion of Science and Mathematics Talented Teachers (PSMT).

## References

[1] Phiphitthakun Y. Mathematics instruction. Bangkok: Borpith printing; 1980.

[2] Faculty of Education Sukhothai Thammathirat Open University. Lecture on mathematics instruction unit 1-7. 8th ed. Bangkok: Sukhothai Thammathirat Open University publishing; 2003.

[3] PISA Thailand The Institute for the Promotion of Teaching Science and Technology. Results of the PISA 2012 mathematics, reading and science. Bangkok: The Institute for the Promotion of Teaching Science and Technology publishing; 2014.

[4] National Institute of Educational Testing Service (Public Organization). Results of the test O-NET 2014 [Internet]. 2014 [Cited 2014 Dec 4]. Available from: http://www.niets.or.th/index.php/exam_information/view_se/1

[5] Bhanthumnavin D, Makanong A. Causal factors of the behavior of students in secondary school mathematics teacher. Bangkok: National Research Council; 2004.

[6] Ratrongmuang P. The development of learning activities using the instructional model based on constructivist theory emphasizing polya's problem solving processes on linear equation for mathayomsuksa 1 [Dissertation]. Khonkaen: Khonkaen Univ; 2013.

[7] Sopa A. The development of mathematics learning activities based on constructivist instructional emphasizing analytical thinking skills on probability for mattayomsuksa 3 [Dissertation]. Khonkaen: Khonkaen Univ; 2013.

[8] Saduangkarn P. Effects of the mathematical concept of constructivist theory on student achievement in mathematics and the ability to transfer of knowledge of students in secondary school [Dissertation]. Bangkok: Chulalongkorn Univ; 1996.

[9] Khammani T. Art of teaching: knowledge for the effective learning process. 11th ed. Bangkok: Chulalongkorn University Publishing; 2009.

[10] Sagarik R. Strategies and inspiration to work efficiently. Suksasart Paritat Journal. 2001; 16(2): 1-7.

# Forecasting the Level of Rainfall in Chiang Mai Basin Using Grey Systems Theory

Chalermchai Puripat, Sukuman Sarikavanij and Adisak Pongpullponsak*

*Department of Mathematics, King Mongkut's University of Technology Thonburi,*
*Bangkok 10140, Thailand,*
*e-mail: chalermchai.pur@kbu.ac.th, sukuman@hotmail.com, adisak.pon@kmutt.ac.th*

**Abstract**

The Grey systems theory (GST) has been known in China after J. L. Deng's first proposed in early 1980s. The amount of rain is the major factor indicating drought or flood in Chiang Mai (the northern province of Thailand). From the rainfall database of the Thai Meteorological Department, the GM (1, 1) is applied to the monthly average rainfall data in Chiang Mai during 2008-2010. The result shows that grey systems model-GM (1, 1) has high precision in prediction the level of rainfall that leads to drought or flood. The average annual rainfall forecasting by GST together with dams' water management policy will alleviate the drought or flood situations in the future.

*Keywords*: forecasting, Grey systems theory, rainfall, water management

*Corresponding Author
E-mail Address: adisak.pon@kmutt.ac.th

## 1. Introduction

Grey systems theory (GST) was initiated in 1982, for the uncertain systems with small samples and poor information that commonly exist in the natural world. [1]. The GST characteristic is very beneficial to forecast the natural phenomena such as rainfall and the appropriate measures can be taken to prevent damages resulting from seasonal disaster.

This research will be new knowledge in the field of rainfall prediction. The advantage will directly influence upon drought and flood prediction both investigation and development.

GM (1, 1) is an accurate rainfall predictor from GST. This approach can significantly be improved the accuracy of the forecasting results even limited data. This quantitative research uses rainfall data from the Thai Meteorological Department (TMD).The next section initiates to portray the literature review in different prediction algorithms and GST.

To predict the monthly rainfall using GST from incomplete rainfall data during 2008-2011, the test system will forecast their past data in order to understand the efficiency of the system.

## 2. Research Methodology

GM (1, 1) is one of model frequently used in Grey forecasting models. This model is a time series forecasting model, encompassing a group of differential equations adapted for parameter variance, rather than a first-order differential equation. Its difference equations have structures that vary with time rather than being general difference equations [2]. The process of GM (1, 1) is described below. Comparison of all interested data [3, 4] can be represented by:

Let $X = (x_1(0), x_2(0), ..., x_n(0))$ be a sequence of raw data. Denote its accumulation generated sequence by $X_1 = (x_1(1), x_1(2), ..., x_1(n))$. Then $x_0 + ax_1(k) = b$ is referred to as the original form of the GM (1, 1), where the symbol GM (1, 1) stands for first order grey model in one variable.

Let $Z_1 = (z_1(2), z_2(3), ..., z_1(n))$ be the sequence generated from $X_1$ by adjacent neighbour means. That is

$$z_1(k) = \frac{1}{2}(x_1(k) + x_1(k-1)) \qquad (1)$$

$k = 1, 2, ..., n$. Then, $x_0(k) + az_1(k) = b$ is referred to as the basic form of the GM (1, 1) [8].

2.1 The grey relational space (GRS) [5, 6]

A GRS is a binary set denoted by $(X, \Gamma)$

$X$ is a collection composed of sequences $x_i(k)$ to be compared and reference sequence $x_0(k)$.

$\Gamma$ is a grey relational mapping set.

$\gamma \in \Gamma$ (Grey relation grade) is an appointed relational mapping in GRS.

Assume that $\gamma(x_0(k), x_i(k))$ satisfy the following GM (1, 1) conditions and one or more memberships are selected to calculate time of occurrence. Assume that there is an image at point $k$ from the series to real number with map $\gamma$ and $\gamma(x_0(k), x_i(k))$ is an image at all points. The different value of Eq. (2) is classified by Eq. (3) and Eq. (4).

$$\Delta_{ij} = \left| x_i(k) - x_j(k) \right| \qquad (2)$$

$$\Delta \min = \min_{j \in i} \min_{k} \left| x_i(k) - x_j(k) \right| \qquad (3)$$

$$\Delta \max = \max_{j \in i} \max_{k} \left| x_i(k) - x_j(k) \right| \qquad (4)$$

Assume that $x_i(k)$ and $x_j(k)$ are two sequences with the same length, same time distances from that moment to the next, and equal time moment intervals, then the absolute degree of incidence can also be computed.

2.2 Grey Relation Coefficients $\gamma(x_i(k), x_j(k))$

The interested data cluster should be separated as the following and show a sequence existed in the grey relation space each interested group data [7, 8].

$$x_i = (x_i(1), x_i(2), ..., x_i(k)) \qquad (5)$$
$$i = 0,1,2,3,...,m; k = 1,2,3,...,n$$
$$x_0 = (x_0(1), x_0(2), x_0(3), ..., x_0(k))$$
$$x_1 = (x_1(1), x_1(2), x_1(3), ..., x_1(k))$$
$$x_2 = (x_2(1), x_2(2), x_2(3), ..., x_2(k))$$
$$.$$
$$.$$
$$.$$
$$x_m = (x_m(1), x_m(2), x_m(3), ..., x_m(k))$$

The $x_0, ..., x_m$ above implied the classified group buffer operator given in rainfall data for computed example solutions.

2.3 The grey relational method [7]

By using this method, we can compute the cluster data corresponding to GST for each group of our interest.

**Definition 1** Let $X_0 = \left(x_0(1), x_0(2), \ldots, x_0(n)\right)$ be a data sequence of a system's characteristic and $X_i = \left(x_i(1), x_i(2), \ldots, x_i(n)\right)$ , $i = 1,2,\ldots,m$ , relevant factor sequences. For given real numbers $\gamma(x_0(k), x_i(k)), i = 1,2,\ldots,m, k = 1,2,\ldots,n$ ,

$$\gamma(x_i, x_j) = \frac{1}{n}\sum_{k=1}^{n}\gamma(x_i(k), x_j(k)) \qquad (6)$$

Eq. (6) are said to be a grey relational coefficient at the point $k$ and $\gamma(x_0, x_i)$ be a grey relational grade, if satisfies the following fundamental [8].

1) Norm Interval
   $\gamma(x_0(k), x_i(k)) \in (0,1], \forall k$
   $\gamma(x_0(k), x_i(k)) = 1, \ iff \ x_0(k) = x_i(k)$
   $\gamma(x_0(k), x_i(k)) = 0, \quad x_0(k) \in \varphi, x_i(k) \in \varphi$

   where $\varphi$ is an empty set.
2) Duality Symmetric
   $\gamma(x_0(k), x_i(k)) = \gamma(x_i(k), x_0(k))$
   $X = \{x_0, x_i\}$
3) Wholeness
   $\gamma(x_0(k), x_i(k)) \neq \gamma(x_i(k), x_i(k))$
   Almost always iff
   $X = \{x \mid j = 0,1,2,..,n, n > 2\}$
4) Approachability
   $\gamma(x_0(k), x_i(k))$ decreases along with $\Delta_k$ increasing, where

$$\varepsilon(k) = x_0(k) - \hat{x}_0(k), \Delta_k = \frac{|\varepsilon(k)|}{x_0(k)} \qquad (7)$$

If we cannot find the $\gamma$ from Eq. (6) then using by Eq. (8).

$$\gamma(x_i, x_j) = \sum_{k=1}^{n}\beta_k\gamma(x_i(k), x_j(k)) \qquad (8)$$

However, since in real application the effect of each factor on the system is not exactly same, Eq. (8) can be modified as

$\beta_k$ : the weight of the $k$th factor ,

$$\sum_{k=1}^{n}\beta_k = 1$$

**Theorem 1** Assume that $\beta_i$ and $\beta_j$ are nonnegative increasing sequences such that $X_j = X_i + c$ where c is a nonzero constant. Let $D_1$ be an initialing operator and $Y_i = X_i D_1$ and $Y_j = X_j D_1$ . If $\alpha_i$ and $\alpha_j$ are respectively the mean slopes of $X_i$ and $X_j$ , and $\beta_i$ and $\beta_j$ the mean slopes of $Y_i$ and $Y_j$ . Then, the following must be true: $\alpha_i = \alpha_j$ , and when $c < 0$ , $\beta_i < \beta_j$ ; when $c > 0$ , $\beta_i > \beta_j$ [8].

The Eq. (8) can be checked error of initial rainfall data for assessing accurate data. The data was provided for the calculation must be based on conditions and following section introduction.

$$\gamma(x_0(k), x_i(k)) = \frac{\min_i \min_k |x_0(k) - x_i(k)| + \xi \min_i \min_k |x_0(k) - x_i(k)|}{|x_0(k) - x_i(k)| + \xi \min_i \min_k |x_0(k) - x_i(k)|} \qquad (9)$$

Once the equation (5) satisfies all of the previous mentioned axioms, the distinguished coefficient can be established [7].

The calculations in the system of GM (1, 1)

1. There is an initial interested data as follows:
   $X_0((t)|t = 1,2,...,n) = (X_0(1), X_0(2), ..., X_0(n)) \quad (10)$
   From method 1 accumulating generation operator (AGO), we can get
   $X_1 = (X_1(1), X_1(2), ..., X_1(n)) \qquad (11)$

2. The solution, also known as time response function, of the whitenization equation.
   $x_1$ can be established the ability function as
   $$\frac{dx_1}{dt} - ax_1 = b \qquad (12)$$

3. $a$, $b$ are elements of interested data parameters vector that is
   $$\hat{a} = \begin{bmatrix} a \\ b \end{bmatrix} \qquad (13)$$

Building repeated additive matrix B and $\hat{a}$ constant data space $Y_n$ that is topological prediction [6]

$$Y_n = [y_0(1), y_0(2), ..., y_0(n)]^T \qquad (14)$$

and

$$B = \begin{pmatrix} -0.5(x_{(1)}(1) + x_{(1)}(2)) & \cdots & 1 \\ -0.5(x_{(1)}(2) + x_{(1)}(3)) & & 1 \\ \vdots & \ddots & \vdots \\ -0.5(x_{(1)}(n-1) + x_{(1)}(n)) & \cdots & 1 \end{pmatrix} \qquad (15)$$

4. Compute grey systems parameter and continuing all the notations from$\hat{a}$, then

$$\hat{a} = \begin{bmatrix} a \\ b \end{bmatrix} = \left[B^T B\right]^{-1} B^T Y_n \qquad (16)$$

5. The solution, also known as time response function of the whitenization equation:

$$\hat{x}_1(t) = \left( x_0(1) - \frac{b}{a} \right) e^{-at} + \frac{b}{a} \qquad (17)$$

6. Calculate the simulation value of $X_1$ by the following formula:

$$\hat{x}_1(0) = (x_0(1), x_0(2), x_0(3), ..., x_0(n-1)) \ (18)$$

7. Get the simulation value of $X_{(0)}$ by IAGO (Inverse accumulating generation operator)

$$\hat{X}_1(0) = (\hat{x}_0(1), \hat{x}_0(2), \hat{x}_0(3), ..., \hat{x}_0(n-1)) \ (19)$$

$$\hat{x}_0 = \left(\hat{x}_1(k) - \hat{x}_0(k-1)\right)$$

The summarization of all steps for GM (1, 1) is shown in Figure 1

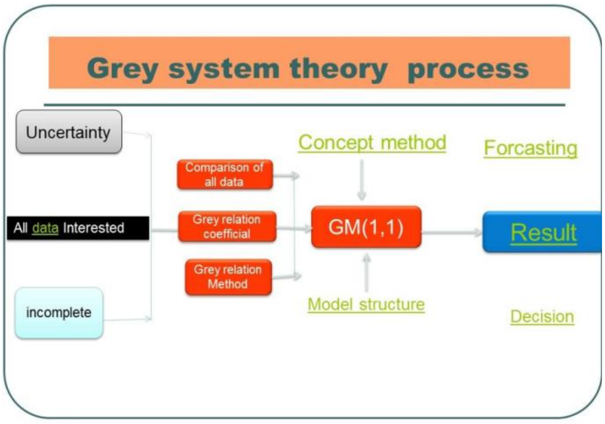


Fig.1 Grey system theory Process

### 3. Results and Discussion

This research is to predict the next year's rainfall. First, Grey systems forecasting shows both methods and calculated following rainfall data in Chiang Mai province during 2008-2010 that designed to explain the manipulation of GM (1, 1) as shown in Table 1.

Table 1 Rainfall in Chiang Mai (mm/month) [9]

| Month | 2008 | 2009 | 2010 |
|---|---|---|---|
| Jan | 16.6 | 0.0 | 21.7 |
| Feb | 13.8 | 0.0 | 0.0 |
| Mar | 9.4 | 16.7 | 4.3 |
| Apr | 57.2 | 97.9 | 3.9 |
| May | 158.7 | 142.0 | 46.4 |
| Jun | 147.1 | 140.2 | 122.7 |
| Jul | 101.6 | 124.0 | 114.5 |
| Aug | 170.9 | 126.8 | 470.6 |
| Sep | 236.4 | 191.7 | 196.2 |
| Oct | 188.1 | 223.4 | 169.6 |
| Nov | 34.1 | 0.0 | 0.0 |
| Dec | 7.1 | 7.5 | 6.1 |

In summer and winter the amount of rainfall are so little that it is drought, but in rainy season there is much more rainfall. The rainfall data were started from 2008 and calculated by GM (1, 1) as followed.

$B_{2008}$= Additive matrix of rainfall data in 2008
$Y_n$ = Initial rainfall data in 2008

$$B_{2008} = \begin{bmatrix} -23.5 & 1 \\ -35.1 & 1 \\ -68.4 & 1 \\ -176.35 & 1 \\ -329.5 & 1 \\ -453.6 & 1 \\ -589.85 & 1 \\ -793.9 & 1 \\ -1005.75 & 1 \\ -1116.85 & 1 \\ -1137.45 & 1 \end{bmatrix}$$

$$Y_n = \{13.8, 9.4, 57.2, 158.7, 147.1, 101.6, 170.9, 236.4, 188.1, 34.1, 7.1\}$$

Calculate $\hat{a}$ as below:

$$\hat{a} = \begin{bmatrix} a \\ b \end{bmatrix} = \left[B^T B\right]^{-1} B^T Y_n$$

$$\hat{a} = \begin{bmatrix} -0.03432 \\ 84.33944 \end{bmatrix}.$$

The solution was known as time response function, of the whitenization equation

$$\frac{dx_1}{dt} - ax_1 = b$$

Use $\hat{a}$ to produce the derivative equation:

$$\frac{dx_1}{dt} - ax_1 = b$$

$$\frac{dx_1}{dt}-(-0.03432)x_1=84.33944$$

Then, time function of the derivative equation is expressed as $\hat{x}_1(t)=\left(x_0(1)-\frac{b}{a}\right)e^{-at}+\frac{b}{a}$

The time response sequence of the GM (1, 1) in equation is given below [6] :

$$\hat{x}_1(t)=\left(x_0(1)-\frac{b}{a}\right)e^{-at}+\frac{b}{a}$$

The predictions of the average rainfall in 2008-2010 show in table 2, 3 and 4 respectively.

From Eq. (17) consist of initial data in 2008 ($x_0(1)$) =16.6

and vector of rainfall data in 2008 $\hat{a}=\begin{bmatrix}-0.0343245\\ 84.33943\end{bmatrix}$

therefore

$$\hat{x}_1(1)=\left(16.6-\frac{84.33943}{-0.0343245}\right)e^{-(-0.034332)}+\frac{84.33943}{-0.0343245}$$
$$=102.983279$$

The average rainfall in 2008 $\left(\hat{x}_1(1)\right)$ is 102.983279 which will divide 11 elements or 11 months as following:

$\hat{x}_1(t)$= {15.25991277, 11.35096642, -0.051173129, 109.6511667, 138.4360297, 174.7772951, 220.6586171, 248.2730705, 128.6031903,66.61528182, 34.50610954}

Table 2 Error checks of rainfall data in 2008

| k | $x_0(k)$ | $\hat{x}_1(k)$ | $x_0(k)-\hat{x}_1(k)$ | $\Delta_k$ |
|---|---|---|---|---|
| 2 | 13.8 | 15.25991277 | 1.459912774 | 0.105790781 |
| 3 | 9.4 | 11.35096642 | 1.950966424 | 0.20754962 |
| 4 | 57.2 | -0.051173129 | -57.25117313 | 1.000894635 |
| 5 | 158.7 | 109.6511667 | -49.04883327 | 0.309066372 |
| 6 | 147.1 | 138.4360297 | -8.663970287 | 0.058898506 |
| 7 | 101.6 | 174.7772951 | 73.17729507 | 0.720248967 |
| 8 | 170.9 | 220.6586171 | 49.75861709 | 0.291156332 |
| 9 | 236.4 | 248.2730705 | 11.87307051 | 0.050224495 |
| 10 | 188.1 | 128.6031903 | -59.49680968 | 0.316304145 |
| 11 | 34.1 | 66.61528182 | 32.51528182 | 0.953527326 |
| 12 | 7.1 | 34.50610954 | 27.40610954 | 3.860015428 |

$$Error\ check\ \varepsilon(k)=x_0(k)-\hat{x}_0(k),\Delta_k=\frac{|\varepsilon(k)|}{x_0(k)}$$

Let us look at the GM (1, 1) the specific model after classified.

$x_0(k)=\beta-\alpha x_1(k-1)$

From table1, it follows that a = −0.03432, *b* = 84.33944. So, we have

$$\alpha=\frac{a}{1+0.5a}=-0.03492$$

$$\beta_1=\frac{b}{1+0.5a}=85.81217$$

The data rainfall of TMD (2008) will be compared between in one year data by cutting in August to September can calculate by using method of GM (1, 1).

The step of predicted rainfall in TMD (2008) data using the original values of the $\beta_1$ to predict the values $\beta_2$ .The result of predicted rainfall in TMD (2008) data after cutting August to September values of $\beta_2$ is less than $\beta_1$ .

$$\hat{a}=\begin{bmatrix}-0.034341235\\ 74.44274126\end{bmatrix}$$

$$\hat{x}_1(1)=\left(16.6-\frac{74.4427426}{-0.03434125}\right)e^{-(-0.03434125)}+\frac{74.4427426}{-0.03434125}$$
$$=92.91569$$

$$\beta_2=\frac{74.44274126}{1+0.5(-0.034341235)}=73.18609$$

The first calculation of $\beta_1$ and $\beta_2$ of rainfall can compare by calculated from TMD (2009) as the following;

$B_{2009}$= Additive matrix of rainfall data in 2009

$Y_n$= Initial rainfall data in 2009

$$B_{2009}=\begin{bmatrix}0 & 1\\ -8.35 & 1\\ -65.65 & 1\\ -185.6 & 1\\ -326.7 & 1\\ -458.8 & 1\\ -584.2 & 1\\ -743.45 & 1\\ -951 & 1\\ -1062.7 & 1\\ -1066.45 & 1\end{bmatrix}$$

$$Y_n=\{0,16.7,97.9,142,140.2,124,126.7,191.7,223.4,0,7.5\}$$

Calculate $\hat{a}$ as below:

$$\hat{a}=\begin{bmatrix}a\\ b\end{bmatrix}=\left[B^TB\right]^{-1}B^TY_n$$

$$\hat{a}=\begin{bmatrix}-0.0245321\\ 85.1299006\end{bmatrix}$$

The solution was known as time response function, of the whitenization equation

$$\frac{dx_1}{dt}-ax_1=b$$

Use $\hat{a}$ to produce the derivative equation:

$$\frac{dx_1}{dt}-(-0.0245321)x_1=85.1299006$$

Then, time function of the derivative equation is expressed as

$$\hat{x}_1(t) = \left( x_0(1) - \frac{b}{a} \right) e^{-at} + \frac{b}{a}$$

The time response sequence of the GM (1, 1)[6] model in equation is given below:

From Eq. (17) consist of initial data in 2009 ($x_0(1)$) = 0

and vector of rainfall data in 2009 $\hat{a} = \begin{bmatrix} -0.0245321 \\ 85.129906 \end{bmatrix}$

therefore

$$\hat{x}_1(1) = \left( 0 - \frac{85.129906}{-0.0245321} \right) e^{-(-0.0245321)} + \frac{85.129906}{-0.0245321}$$

$$= 100.38947$$

The average rainfall in 2009 $\hat{x}_1(1)$ is 100.38947 which will divide 11 elements or 11 months as following:

$\hat{x}_1(t)$ ={ 0, 23.8644685, 115.8002528, 122.3761774, 129.3255276, 136.6695092, 143.2919047, 79.0991024, 223.8541566, 0, 12.27678412}

Table 3 Error checks of rainfall data in 2009

| k | $x_0(k)$ | $\hat{x}_1(k)$ | $x_0(k) - \hat{x}_1(k)$ | $\Delta_k$ |
|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 |
| 2 | 16.7 | 23.8644685 | 7.164468499 | 0.42901009 |
| 3 | 97.9 | 115.8002528 | 17.90025275 | 0.182842214 |
| 4 | 142 | 122.3761774 | -19.62382263 | 0.138195934 |
| 5 | 140.2 | 129.3255276 | -10.87447237 | 0.077563997 |
| 6 | 124 | 136.6695092 | 12.66950919 | 0.102173461 |
| 7 | 126.8 | 143.2919047 | 16.49190468 | 0.13006234 |
| 8 | 191.7 | 179.0991024 | -12.60089764 | 0.065732382 |
| 9 | 223.4 | 223.8541566 | 0.454156574 | 0.00203293 |
| 10 | 0 | 0 | 0 | 0 |
| 11 | 7.5 | 12.27678412 | 4.776784121 | 0.636904549 |

$$Error\ check \varepsilon(k) = x_0(k) - \hat{x}_0(k), \Delta_k = \frac{|\varepsilon(k)|}{x_0(k)}$$

$\Delta_k$ directions of rainfall that has increased or decreased. Data may be adjusted correctly.

Let us look at the GM (1, 1) the specific model

$$x_0(k) = \beta - \alpha x_1(k-1)$$

From Table 3, it follows that
a = –0.03432, b= 85.12990056. So, we have

$$\alpha = \frac{-0.0245321}{1\text{-}0.5(-0.0245321)} = 0.03492$$

$$\beta_{2009} = \frac{85.12990056}{1\text{-}0.5(-0.0245321)} = 85.81217$$

The testing of the predicted results are much rainfall effect. The initial values of the first value to forecast rainfall that are expected to be a précised rainfall data. But there are some errors in correcting rainfall data. Some of rainfall cannot be measured. The predicted rainfall is less than 2008 levels, but also there is with high rainfall.

B<sub>2010</sub> = Additive matrix of rainfall data in 2010
Y<sub>n</sub> = Initial rainfall data in 2010

$$B_{2010} = \begin{bmatrix} -21.7 & 1 \\ -23.85 & 1 \\ -27.95 & 1 \\ -53.1 & 1 \\ -137.65 & 1 \\ -256.25 & 1 \\ -548.8 & 1 \\ -882.2 & 1 \\ -1065.1 & 1 \\ -1149.9 & 1 \\ -1152.95 & 1 \end{bmatrix}$$

$$Y_n = \{0,4.3,3.9,46.4,122.7,114.5,470.6,196.2,169.6,0,6.1\}$$

Calculate $\hat{a}$ as below:

$$\hat{a} = \begin{bmatrix} a \\ b \end{bmatrix} = \left[ B^T B \right] B^T Y_n$$

$$\hat{a} = \begin{bmatrix} -0.050119889 \\ 78.8808869 \end{bmatrix}$$

The solution was known as time response function, of the whitenization equation $\frac{dx_1}{dt} - ax_1 = b$

Use $\hat{a}$ to produce the derivative equation: $\frac{dx_1}{dt} - (-0.050119889)x_1 = -78.88088699$

Then, time function of the derivative equation is expressed as

$$\hat{x}_1(t) = \left( x_0(1) - \frac{b}{a} \right) e^{-at} + \frac{b}{a}$$

The time response sequence of the GM (1, 1) model in equation is given below: [6]

From Eq. (17) consist of initial data in 2010 ($x_0(1)$) =0

and vector of rainfall data in 2010 $\hat{a} = \begin{bmatrix} \text{-}0.0501198 \\ 78.8808869 \end{bmatrix}$

therefore

$$\hat{x}(1) = \left( 21.7 - \frac{78.8808869}{(\text{-}0.0501198)} \right) e^{-(-0.0501198)} + \frac{78.8808869}{(\text{-}0.0501198)}$$

$$= 103.70639$$

The average rainfall in 2010 is 103.70639 which will divide 11 elements or 11 months as following:

$\hat{x}_1(t)$={0, 34.10792295, 4.691023279,10.94412889, 167.2229233,203.5421942,247.7496745,301.5586102,102.2381735, 0, 24.63237336}

The Table4 showed the predicted rainfall in period 10, 11 and 12 month.

Table 4 Error checks of rainfall data in 2010

| k | $x_0(k)$ | $\hat{x}_{01}(k)$ | $x_0 - \hat{x}_1(k)$ | $\Delta_k$ |
|---|---|---|---|---|
| 1 | 21.7 | 21.61195867 | -0.088041333 | 0.004057204 |
| 2 | 0 | 189.9277 | -189.9277 | 0 |
| 3 | 4.3 | 34.10792295 | 29.80792295 | 6.932075106 |
| 4 | 3.9 | 4.691023279 | 0.791023279 | 0.202826482 |
| 5 | 46.4 | 10.94412889 | -35.45587111 | 0.764135153 |
| 6 | 122.7 | 167.2229233 | 44.52292334 | 0.362860011 |
| 7 | 114.5 | 203.5421942 | 89.04219421 | 0.777661085 |
| 8 | 470.6 | 247.7496745 | -222.8503255 | 0.473545103 |
| 9 | 196.2 | 301.5586102 | 105.3586102 | 0.536995974 |
| 10 | 0 | 1059.92 | -1059.92 | 0 |
| 11 | 6.1 | 24.63237336 | 18.53237336 | 3.038093994 |

$$Error\ check\varepsilon(k) = x_0(k) - \hat{x}_0(k), \Delta_k = \frac{|\varepsilon(k)|}{x_0(k)}$$

The last step of calculation checked the specific model by Eq. (18).

$$x_{(0)}(k) = \beta - \alpha x_{(1)}(k-1) \qquad (18)$$

The parameters of rainfall data in 2010 consist of a = -0.050119889, b= 84.33944 which calculated as follow;

$$\alpha = \frac{-0.050119889}{1+0.5(-0.050119889)} = -0.03492$$

$$\beta = \frac{78.88088699}{1+0.5(-0.050119889)} = 85.81217$$

The 2008 to 2009 rainfall data were not completed due to there are very low or unable to measure (0.0 mm/month). In order to predict the rainfall level in 2010, the loss data are simulated and adjusted to GM (1, 1).

Researcher has some correction to adjust a new simulate values that can predict the rainfall in 2010.We have to conclude that the demonstration of all data for forecasted values calculation and then compare each year.

Rainfall data in 2010 can't be calculated because of some months without rain or not measurable. We may need a new data model to calculate for the full year. Therefore, we should be defined limitation in information for prediction.

The value of the information that was updated to allow for the replacement of data with no change and according to initial data.

Table 5 Loss data rainfall simulation.

| Month | 2008 | 2009 | 2010 |
|---|---|---|---|
| Jan | 16.6 | 0.0 | 21.7 |
| Feb | 13.8 | 0.0 | 13.4 |
| Mar | 9.4 | 16.7 | 4.3 |
| Apr | 57.2 | 97.9 | 3.9 |
| May | 158.7 | 142.0 | 46.4 |
| Jun | 147.1 | 140.2 | 122.7 |
| Jul | 101.6 | 124.0 | 114.5 |
| Aug | 170.9 | 126.8 | 470.6 |
| Sep | 236.4 | 191.7 | 196.2 |
| Oct | 188.1 | 223.4 | 169.6 |
| Nov | 34.1 | 0 | 87.85 |
| Dec | 7.1 | 7.5 | 6.1 |

The experimental results

$$\hat{a} = \begin{bmatrix} 0.031482 \\ 80.31396 \end{bmatrix}$$

Prediction adjust in 2010

$$\hat{x}(1) = \left(21.7 - \frac{80.31396}{(0.031482)}\right)e^{-(0.031482)} + \frac{80.31396}{(0.031482)}$$

=100.09035

From Table5, it follows that

a= 0.031482, b= 80.31396. So, we have

$$\alpha = \frac{0.031482}{1+0.5(0.031482)} = 0.031986$$

$$\beta = \frac{80.31396}{1+0.5(0.031482)} = 81.59842$$

The adjusted parameters are recalculated using Grey Systems Model again, the result show that rainfall in 2010 is level very torrent rain.

## 4. Conclusion

The prediction of 2011 will be torrent rain in Chiang Mai that indicates the water level in the northern region because the Chiang Mai province consist of Ping river, Maengad, Maekuang and Maekuang udom thara Dam. The grey system described the prediction and systematic analysis of rainfall and distributes data and differential equation. The parameter of $\beta$ and $\hat{a}$ can be predict the rainfall data which is possible to analyze the effect factor of process in 2011.

Table 6 The five levels of rainfall.

| | Levels of rainfall | rainfall(mm) |
|---|---|---|
| 1 | Little rain could not measure rainfall | <0.1 mm |
| 2 | Little rain | 0.1-10 mm |
| 3 | Moderate rain | 10.1-35 mm |
| 4 | Torrent rain | 35.1-90mm |
| 5 | Very torrent rain | > 90 mm |

Source: The Thai Meteorology Department.

The results show rainfall level as follow:

(1) The predicted values almost matched with the measured values under the rainfall from 35.1tomore than 90 mm (Torrent rain and Very torrent rain).

(2) The method of grey systems forecasting is difficult with high precision in forecasting and may be achieved to be simulated by a Grey Incidence and Evaluations. Might be to add more information to add to predict the outcome correctly. Beginning Data rainfall in 2010 can be used for different rainfall types of level rain of The Thai Meteorological Department. Forecasting the rainfall and it is also helpful for determining other parameters of very torrent rainfall in the season next year.

## Acknowledgements

The authors wish to thanks The Thai Meteorological Department and Department of Disaster Prevention and Mitigation for the valuable co-operation.

## References

[1] Julong D., Introduction to grey system theory.The Journal of Grey System 1(1), 1989, p.1–7.

[2]Li .W, Xie. H. Geometrical Variable Weights Buffer GM(1,1) Model and Its Application in Forecasting of China’s Energy Consumption. China.2014.p. 1-6.

[3] Julong D.The least parameters controller of decentralized control system. Acta Aurtomatica Sinica. China. 1, 1982.p. 49-54.

[4] Julong D. Grey System theories and measurement futurics, Future and Development, China.1983.3, 20-3.

[5] Julong D. Grey fuzzy forecast and control for grain. Journal of Huazhong University of Science and Technoloy. China.1983.2, p. 1-8 .

[6] Julong D. Relational space of Grey Systems. Fuzzy Mathematics, (Special Issue of Grey Systems) China.1985.2, p. 1-10.

[7] Hong G. Identification coefficient of relational grade of Grey Systems. FUZZY Mathematics. China. 1985. 2, p. 55-58.

[8] Si-Feng L., Lin Y., Grey Systems Theory and Applications. Chinese: Springer; 2010.p. 51-130.

[9] The Thai Meteorological Department. 2014. Monthly Annual Rainfall : Selected Location by Region at Meteorology station, Chiamai Mai Province: 2008 – 2010. Available from http://www.tmd.go.th/

# Authors Index

# Sponsors

1. Thailand Research Fund (TRF)
2. Solution Center Co., Ltd. (Minitab)
3. Sift Analytics Group Pte. Ltd. (SIFT)
4. Intouch (AIS)
5. Thaksin University
6. Mr. Tinakorn (Energica Co. Ltd.)
7. Infitel Co., Ltd. (Thailand)
8. KMUTT Saving and credit cooperative limited
9. Office of the Higher Education Commission

ICAS 2015

15-17 July 2015 Pattaya, Thailand