INTERNATIONAL CONFERENCE ON APPLIED STATISTICS 2018

"Improving Data Value & Data Quality using Data Science"

ICAS 2018

Conference Proceedings

24-26 October 2018

Centra by Centara, Government Complex Hotel & Convention Centre Chaeng Watthana Bangkok

Conference Proceeding

6th International Conference on Applied Statistics 2018

ICAS 2018

24-26 October 2018
Centra by Centara, Government Complex Hotel
& Convention Centre Chaeng Watthana Bangkok

All authors and co-authors are individually and collectively responsible for the contents of theirs papers published by ICAS 2018

Honorary Academic Speakers

Keynote Speakers

Tippawan Liabsuetrakul, Ph.D.
Professor, Epidemiology Unit, Faculty of Medicine, Prince of Songkla University, Thailand

Kamarulzaman Ibrahim, Ph.D.
Professor, School of Mathematical Sciences, Universiti Kebangsaan Malaysia

Professor Dr.Dankmar Bohning
Mathematical Sciences, University of Southampton, United Kingdom

Chair Forums

Virasakdi Chongsuvivatwong, Ph.D.
Professor, Epidemiology Unit, Faculty of Medicine, Prince of Songkla University, Thailand

Kanitta Bundhamchareon, Ph.D.
Senior Researcher, International Health Policy Program, Ministry of Public Health, Thailand

Toh Hock Chai
Senior Director of Statistical Services Department, Bank Negara Malaysia

Somsajee Siksamat, Ph.D.
Senior Director of Statistics and Data Management Department, Bank of Thailand

Alfredo Huete, Ph.D.
Professor, University of Technology Sydney, Australia

Jirawan Jitthavech, Ph.D.
Professor, President of Thai Statistical Association

Ajin Jirachiefpattana, Ph.D.
Deputy Director-General, National Statistical Office of Thailand

Halimah Binti Awang, Ph.D.
Associate Professor, Social Security Research Centre, University of Malaya, Malaysia

Don McNeil, Ph.D.
Professor, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Thailand

Invited Speakers

Wichai Aekplakorn, Ph.D.
Professor, Facuty of Medicine at Ramathibodi Hospital, Mahidol University

Khachon Mongkonchoo
National Health Security Office Region 13 Bangkok

Seksan Kiatsupaibul, Ph.D.
Associate Professor, Faculty of Commerce and Accountancy Statistics, Chulalongkorn University

Apiradee Lim, Ph.D.
Associate Professor, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus

Scientific Committees

Halimah Awang, Ph.D.
Associate Prefessor, Social Security Research Centre,
University of Malaya, Malaysia

Binita Kumari Paudel, Ph.D.
Associate Professor, College of Medical and Allied Science,
Purbanchal University, Gothguan, Nepal

Arjun Mani Guragain, Ph.D.
Founding Chair and Director,
Civic Independence Development (CID), Nepal

Preecha Lamchang
Associate Professor, Department of Statistics,
Faculty of Science, Chiang Mai University, Thailand

Manachai Rodchuen, Ph.D.
Assistant Professor, Department of Statistics,
Faculty of Science, Chiang Mai University, Thailand

Manad Khamkong, Ph.D.
Assistant Professor, Department of Statistics, Faculty of Sciences,
Chiang Mai University, Thailand

Patrinee Traisathit, Ph.D.
Assistant Professor, Department of Statistics,
Faculty of Science, Chiang Mai University, Thailand

Suree Chooprateep, Ph.D.
Assistant Professor, Department of Statistics, Faculty of Science,
Chiang Mai University, Thailand

Boonorm Chomtee, Ph.D.
Associate Professor, Department of Statistics,
Faculty of Science, Kasetsart University, Thailand

Winai Bodhisuwan, Ph.D.
Assistant Professor, Department of Statistics,
Faculty of Science, Kasetsart University, Thailand

Wandee Wanishsakpong, Ph.D.
Department of Statistics, Faculty of Science,
Kasetsart University, Thailand

Supunnee Ungpansattawong, Ph.D.
Associate Professor, Department of Statistics,
Faculty of Science, Khon Kaen University, Thailand

Wuttichai Srisodaphol, Ph.D.
Asistant Professor, Department of Statistics,
Faculty of Science, Khon Kaen University, Thailand

Autcha Araveeporn, Ph.D.
Assistant Professor, Department of Statistics,
King Mongkut's Institute of Technology Ladkrabang, Thailand

Sa-aat Niwitpong, Ph.D.
Associate Professor, Department of Applied Statistics, Faculty of Applied Science,
King Mongkut's University of Technology North Bangkok, Thailand

Yupaporn Areepong, Ph.D.
Associate Professor, Department of Applied Statistics,
Faculty of Applied Science,
King Mongkut's University of Technology North Bangkok, Thailand

Orathai Polsen, Ph.D.
Assistant Professor, Department of Applied Statistics, Faculty of Applied Science,
King Mongkut's University of Technology North Bangkok, Thailand

Siraprapa Manomat, Ph.D.
Assistant Professor, Department of Applied Statistics, Faculty of Applied Science,
King Mongkut's University of Technology North Bangkok, Thailand

Wirawan Chinviriyasit, Ph.D.
Associate Professor, Department of Mathematics, Faculty of Science,
King Mongkut's University of Technology Thonburi, Thailand

Krisana Lanumteang, Ph.D.
Assistant Professor, Department of Statistics, Faculty of Science,
Maejo University, Thailand

Bungon Kumphon, Ph.D.
Associate Professor, Department of Mathematics, Faculty of Science,
Mahasarakham University, Thailand

Nipaporn Chutiman, Ph.D.
Assistant Professor, Department of Mathematics,
Faculty of Science, Mahasarakham University, Thailand

Sujitta Suraphee, Ph.D.
Assistant Professor, Department of Mathematics,
Faculty of Science, Mahasarakham University, Thailand

Montip Tiensuwan, Ph.D.
Associate Professor, Department of Mathematics, Faculty of Science,
Mahidol University, Thailand

Jirawan Jitthavech, Ph.D.
Professor, Graduate School of Applied Statistics,
National Institute of Development Administration, Thailand

Samruam Chongcharoen, Ph.D.
Professor, Graduate School of Applied Statistics,
National Institute of Development Administration, Thailand

Preecha Vichitthamaros, Ph.D.
Assistant Professor, Graduate School of Applied Statistics,
National Institute of Development Administration, Thailand

Watchareeporn Chaimongkol, Ph.D.
Graduate School of Applied Statistics,
National Institute of Development Administration, Thailand

Naratip Jansakul, Ph.D.
Assistant Professor, Department of Mathematics, Faculty of Science,
Prince of Songkla University, Thailand

Don Roy McNeil, Ph.D.
Professor, Department of Mathematics and Computer Science,
Faculty of Science and Technology, Prince of Songkla University, Thailand

Apiradee Saelim, Ph.D.
Associate Professor, Department of Mathematics and Computer Science,
Faculty of Science and Technology,
Prince of Songkla University, Thailand

Nifatamah Makaje, Ph.D.
Assistant Professor, Department of Mathematics and Computer Science,
Faculty of Science and Technology,
Prince of Songkla University, Thailand

Nittaya McNeil, Ph.D.
Assistant Professor, Department of Mathematics and Computer Science,
Faculty of Science and Technology,
Prince of Songkla University, Thailand

Phattrawan Tongkumchum, Ph.D.
Assistant Professor, Department of Mathematics and Computer Science,
Faculty of Science and Technology,
Prince of Songkla University, Thailand

Areena Hazanee, Ph.D.
Department of Mathematics and Computer Science,
Faculty of Science and Technology,
Prince of Songkla University, Thailand

Arinda Ma-a-lee, Ph.D.
Department of Mathematics and Computer Science,
Faculty of Science and Technology,
Prince of Songkla University, Thailand

Attachai Ueranantasun, Ph.D.
Department of Mathematics and Computer Science,
Faculty of Science and Technology,
Prince of Songkla University, Thailand

Mayuening Eso, Ph.D.
Department of Mathematics and Computer Science,
Faculty of Science and Technology,
Prince of Songkla University, Thailand

Nurin Dureh, Ph.D.
Department of Mathematics and Computer Science,
Faculty of Science and Technology,
Prince of Songkla University, Thailand

Rattikan Saelim, Ph.D.
Department of Mathematics and Computer Science,
Faculty of Science and Technology,
Prince of Songkla University, Thailand

Salang Musikasuwan, Ph.D.
Department of Mathematics and Computer Science,
Faculty of Science and Technology,
Prince of Songkla University, Thailand

Santhana Chaimontree, Ph.D.
Department of Mathematics and Computer Science,
Faculty of Science and Technology,
Prince of Songkla University, Thailand

Kamolchanok Panishkan, Ph.D.
Assistant Professor, Department of Statistics, Faculty of Science,
Silpakorn University, Thailand

Kusaya Plungpongpun, Ph.D.
Assistant Professor, Department of Statistics, Faculty of Science,
Silpakorn University, Thailand

Naowarut Meejun, Ph.D.
Department of Statistics, Faculty of Science,
Silpakorn University, Thailand

Wassana Suwanvijit, Ph.D.
Assistant Professor, Department of Economics,
Faculty of Economics and Business administration,
Thaksin University, Thailand

Kamon Budsaba, Ph.D.
Associate Professor, Department of Mathematics and Statistics,
Faculty of Science and Technology, Thammasat University, Thailand

Wanyok Atisattapong, Ph.D.
Assistant Professor, Department of Mathematics and Statistics,
Faculty of Science and Technology, Thammasat University, Thailand

Sulawan Yotthanoo, Ph.D.
Department of Statistics, School of Science,
University of Phayao, Thailand

Organizing Committees

Thai Statistical Association and Thailand Statistical Collaborative Research Network	Jirawan Jitthavech, Ph.D. Professor
National Statistical Office of Thailand	Director-General National Statistical Office of Thailand Deputy Director-General National Statistical Office of Thailand
Prince of Songkla University Committee	Dean of Faculty of Science and Technology Vice Dean for Administrative Faculty of Science and Technology Head of Department of Mathematics and Computer Science Head of Faculty Secretary Amran Hayeewaenawae Apiradee Saelim Arinda Ma-a-lee Attachai Ueranantasun Chanpen Phokaew Charan Khwanfai Don Roy McNeil Mareekee Madeng Mariyah Bumamah Mayuening Eso Montri Watthanapradith Nihafisee Binwaedoloh Nittaya McNeil Nurin Dureh Pariyah Mhadmean Phattrawan Tongkumchum Potjamas Chuangchang Rattikan Saelim Reesuwan Waehama Salang Musikasuwan Santhana Chaimontree Sarawuth Chesoh Suhaimee Buya Sujunya Boonpradit Sunaree Suwanro

Contents

Comparison of Distance Measures for Clustering Population Pyramids: A Simulation Study

Akarin Phaibulpanich[1*] and Nat Kulvanich[1]

[1]Department of Statistics, Faculty of Commerce and Accountancy, Chulalongkorn University, Wangmai, Phahumwan, Bangkok, Thailand
*Corresponding email: akarin@cbs.chula.ac.th
Email: nat@cbs.chula.ac.th

ABSTRACT

Clustering Population Pyramids, the age histograms of certain populations, is often of interest in many researches. In this paper, we conduct a simulation study to compare ten histogram distance measures for the purpose of clustering population pyramids, where K-medoid algorithm is used as the clustering tool. The results show that there is no single "best" distance measure that outperforms others across all data types. In general, the higher variation, the harder clustering becomes. For the unimodal pyramids with high variances, cross-bin types of measures such as Wasserstein's or Earth mover distance might be a better choice, while for the bimodal pyramids, bin to bin type of measures, such as Euclidean, Hellinger, and Jaccard generally outperform others.

Keywords: Wasserstein metric; Hellinger distance; Population Pyramid; Cluster analysis; K-medoid

1 INTRODUCTION

Population pyramids refer to histograms, usually shown vertically to illustrate the frequency of various age groups of a population, where each bar's size can be presented in terms of percentages or counts (Kahn, 2007). An example of such pyramids is shown in Figure 1, displaying age distribution of Thai population in 2016 for both males and females.

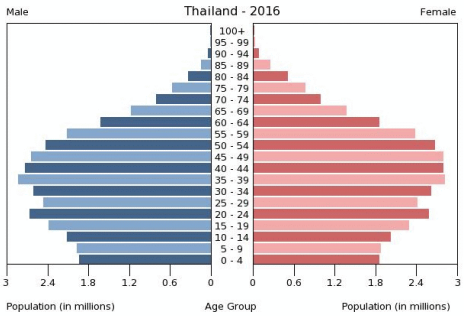


Figure 1: Age distribution of Thai population in 2016

(Source: https://www.indexmundi.com/thailand/age_structure.html)

Such a population pyramid or histogram can tell us many aspects of a population. For example, high birthrate population would have a wide-based pyramid, while aging society would have a wider top.

This research paper is motivated upon investigating the Thai population's age data, publicly available from the Official Statistics Registration System, under Bureau of Registration Administration, Department of Provincial Administration, Ministry of Interior. The dataset contains frequency distributions for different age groups for each of the provinces in Thailand, collected from 1993 to 2017.

Initially, we were interested in studying age distributions of Thai provinces through performing cluster analysis on their population pyramids. However, one important question needed to be asked before implementing such an analysis is: what would be an appropriate distance measure or metric to quantify the differences between two population pyramids? For a clustering to be valid, similar pyramids need to be clustered within the same group, while different-looking pyramids should be in the different ones.

The main objective of this paper is therefore to investigate several histogram distance measures for the purpose of clustering population pyramids, through a simulation study. K-medoid algorithm, a more robust version of K-means, is used as the clustering tool. In section 1, we first review histogram distance measures that have been purposed, and we also provide some background information on different types of population pyramids, which will be used as the basis for our data generation.

In section 2, framework for generating our histogram data is discussed, together with the evaluation method. Clustering performances of ten distance measures on each of the simulated datasets are compared in section 3. Lastly, conclusion will be presented in section 4.

For our specific clustering problem, pyramids or histograms to be compared are of the same number of intervals, d, and the intervals are of the same size, usually equal to 5 or 10 years period. Following (Cha, 2002)'s notations, let P and Q be two histograms of same interval structure and let P_i and Q_i be the corresponding relative frequencies for each interval i. Note that $\sum_{i=1}^{d} P_i = 1$ and $\sum_{i=1}^{d} Q_i = 1$.

Since a distance measure plays a central role in any clustering or classification problems, several distance measures have been proposed over the years. (Cha, 2002) has made a comprehensive summary of nominal type histogram distance measures and group them into eight types: L_p Minkowski family, L_1 family, Intersection family, Inner Product family, Fidelity family, χ^2 family, Shannon's entropy family, and Combinations. Each type has its own distinct characteristics which can be summarized as follows.

1. **L_p Minkowski family:** This group of measures can be seen as an extension of the familiar Euclidean distance, which relies on the square of the difference between two data points (p = 2). The power parameter p can be generalized to other positive numbers. Examples of such measures are Euclidean L_2 (eq.1), Minkowski L_p, and Chebyshev L_∞.
2. **L_1 family:** This type of measures is similar to the L_p Minkowski family, but focuses on the absolute difference instead. Examples of such measures are Sorensen, Gower (eq.2), Soergel, and Canberra.
3. **Intersection family:** The idea behind this distance family is the histogram similarity measure based on their intersection. This family actually has a strong connection to L_1 family types. Examples of such measures are Intersection (eq.3), Wave Hedges, and Czekanowski.
4. **Inner Product family:** Measures in this family are based on the inner or scalar products $P \cdot Q$. Examples of such measures are Inner Product, Harmonic mean, Cosine, and Jaccard (eq.4).
5. **Fidelity family:** All Fidelity-based distances are based on the square-root of the probabilities or relative frequency values. This family is also called Squared-chord family. Examples of such measures are Fidelity, Bhattacharyya, Hellinger (eq.5), and Matusita.
6. **χ^2 family:** As the name suggests, measures in this family are related to the χ^2 statistic. This family is also called Squared L_2 family. Examples of such measures are Squared χ^2, Divergence (eq.6), and Clark.

7. **Shannon's entropy family:** Measures in this family is based on the Shannon's entropy, or the degree of uncertainty from information theory. Examples of such measures are Kullback-Leiber (eq.7), Jeffreys, K divergence, and Jensen-Shannon.
8. **Combinations:** Measures in this family are built on combination of two or more distance measures. For example, Taneja measure (eq.8) is a combination of arithmetic and geometric means. Other examples include Kumar-Johnson (eq.9) and Avg(L_1, L_∞).

The above eight types of distance measures are of nominal types, meaning that they focus on comparing interval to interval or "bin to bin" information only. Zhao, 2011 has discussed another type of histogram distance measure which utilizes information from neighboring intervals, called "cross-bin" distances. Such measures are far more complicated and time-consuming to calculate than the nominal types. Examples of cross-bin measures are earth mover's distance which also goes by the names L_2 Mallow's distance, or Wasserstein and Kantonovich's distance (eq.10).

There are altogether more than 40 types of histogram distance measures available. In this paper, we choose to compare ten histogram distance measures selected from the above list by choosing at least one measure from each type. The comparisons are made on the basis of clustering performance which will be discussed in Section 3.

Table 1: Distance measures used in our analysis

Family of measure	Distance measures used in our analysis
L_p Minkowski	Euclidean: $$d = \sqrt{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert^2} \quad (1)$$
L_1	Gower: $$d = \frac{1}{d}\sum_{i=1}^{d} \lvert P_i - Q_i \rvert \quad (2)$$
Intersection	Intersection: $$d = 1 - \sum_{i=1}^{d} \min(P_i, Q_i) \quad (3)$$
Inner Product	Jaccard: $$d = 1 - \frac{\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i} \quad (4)$$
Fidelity	Hellinger: $$d = 2\sqrt{1 - \sum_{i=1}^{d} \sqrt{P_i Q_i}} \quad (5)$$
χ^2	Divergence: $$d = 2\sum_{i=1}^{d} \frac{(P_i - Q_i)^2}{(P_i + Q_i)^2} \quad (6)$$
Shannon's entropy	Kullback-Leibler: $$d = \sum_{i=1}^{d} P_i ln \frac{P_i}{Q_i} \quad (7)$$
Combinations	Taneja: $$d = \sum_{i=1}^{d} \left(\frac{P_i + Q_i}{2}\right) ln\left(\frac{P_i + Q_i}{2\sqrt{P_i Q_i}}\right) \quad (8)$$ Kumar-Johnson: $$d = \sum_{i=1}^{d} \left(\frac{(P_i^2 - Q_i^2)^2}{2(P_i Q_i)^{3/2}}\right) \quad (9)$$
Wasserstein	$$d = \min_{F=\{F_{ij}\}} \frac{\sum_{i,j} F_{ij} D_{ij}}{F_{ij}} \quad (10)$$ such that: $$\sum_j F_{ij} \le P_i, \sum_i F_{ij} \le Q_i,$$ $$\sum_{i,j} F_{ij} = \min(\sum_i P_i, \sum_j Q_j,),$$ and $F_{ij} \ge 0$

For equations 4, 6, 7, 8, and 9, frequencies of 0 would lead to invalid calculations. The distances are treated as 0 in such cases.

For the background on different types of population pyramids, typically, population pyramids are classified into three main types: expansive, stationary, and constrictive (Boucher, 2016).

As shown in Figure 2, an expansive pyramid shows a highly skewed to the right population where the majority of people are young. Birthrate and death rate are both high and life expectancy is low, usually characterizing a developing society.

A stationary pyramid shows a bimodal population with an increasing adult population. The birthrate and death rate are declining, with longer life expectancy, usually characterizing a more developed populations with better life qualities when compared to an expansive one.

Lastly, a constrictive pyramid shows an aging population with very low birth rate and death rate, and longer life expectancy, usually characterizing a developed society. The pyramid resemblances a mixture of two distributions.

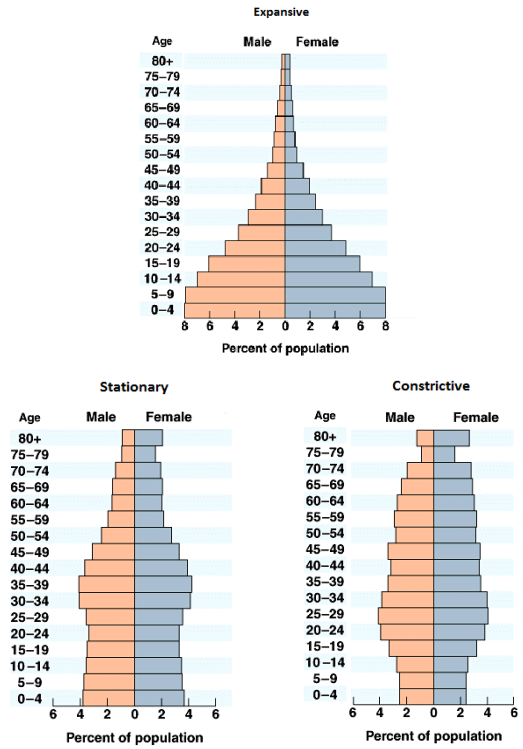


Figure 2: Examples of three main types of population pyramids: expansive, stationary, and contractive.
(Source: https://populationeducation.org)

2 METHODS

2.1 Experimental design

We generated seven different histogram datasets to study the effect of certain parameters on clustering performances. Each dataset contains four groups (or clusters) of histograms, where 100 histograms are generated from each group. That means, there are a total of 400 histograms in each dataset.

Table 2 shows the underlying distributions which we use to generate histogram data. For the so-called "Four type" dataset, we mimic the general types of population pyramids as discussed in the previous section, where group 1, generated from an exponential distribution with rate 1/30, represents a highly right-skewed or highly expansive population. Group 2, generated from a gamma distribution with shape = 1.29 and rate = 0.061, represents a moderately expansive population. Group 3, generated from a mixture between two normal distributions with means of 5 and 40 and stand deviations both equal to 20, represents a relatively stationary pyramid. Lastly, group 3, generated from a mixture between two normal distributions with means

of 15 and 55 and stand deviations both equal to 20, represents a constrictive pyramid.

With the "Normal SD = 15", "Normal SD = 20", and "Normal SD = 25" datasets, we aim to compare the clustering performances when the pyramids are only unimodal and also have different degrees of variations. The four groups are only differentiable by the means at 5, 10, 15, and 20. The means are designed this way to make the clustering task harder than the Four type" dataset.

With the "Mixed Normal SD = 15", "Mixed Normal SD = 20", and "Mixed Normal SD = 25" datasets, we aim to compare the clustering performances when the pyramids are bimodal or are a mixture of two normal distributions, and also have different degrees of variations. The four groups are only differentiable by the means as shown in Table 2.

To create a histogram P in each group, we implement the following:

1. Generate 1,000 observations from the specified distribution.
2. Truncate the values at above 100 and below 0.
3. Create a relative frequency histogram based on the remaining observations using 20 intervals, each of size 5.

P_i, the relative frequency for an interval i of the histogram P, for i = 1, …, 20, are used as the input data for our K-medoid clustering analysis.

Table 2: Underlying distributions for generating histogram data

Dataset	Histogram data generation
"Four types"	Group 1: Exponential, rate = 1/30 Group 2: Gamma, shape = 1.29 and rate = 0.061 Group 3: 0.5N(μ = 5, σ = 20) +0.5 N(μ = 40, σ = 20) Group 4: 0.5N(μ = 15, σ = 20)+0.5 N(μ = 55, σ = 20)
"Normal SD = 15"	Group 1: N(μ = 5, σ = 15) Group 2: N(μ = 10, σ = 15) Group 3: N(μ = 15, σ = 15) Group 4: N(μ = 20, σ = 15)
"Normal SD = 20"	Group 1: N(μ = 5, σ = 20) Group 2: N(μ = 10, σ = 20) Group 3: N(μ = 15, σ = 20) Group 4: N(μ = 20, σ = 20)
"Normal SD = 25"	Group 1: N(μ = 5, σ = 25) Group 2: N(μ = 10, σ = 25) Group 3: N(μ = 15, σ = 25) Group 4: N(μ = 20, σ = 25)
"Mixed Normal SD = 15"	Group 1: 0.5N(μ = 10, σ = 15)+0.5 N(μ = 40, σ = 15) Group 2: 0.5N(μ = 15, σ = 15)+0.5 N(μ = 45, σ = 15) Group 3: 0.5N(μ = 15, σ = 15)+0.5 N(μ = 50, σ = 15) Group 4: 0.5N(μ = 20, σ = 15)+0.5 N(μ = 50, σ = 15)
"Mixed Normal SD = 20"	Group 1: 0.5N(μ = 10, σ = 20)+0.5 N(μ = 40, σ = 20) Group 2: 0.5N(μ = 15, σ = 20)+0.5 N(μ = 45, σ = 20) Group 3: 0.5N(μ = 15, σ = 20)+0.5 N(μ = 50, σ = 20) Group 4: 0.5N(μ = 20, σ = 20)+0.5 N(μ = 50, σ = 20)
"Mixed Normal SD = 25"	Group 1: 0.5N(μ = 10, σ = 25)+0.5 N(μ = 40, σ = 25) Group 2: 0.5N(μ = 15, σ = 25)+0.5 N(μ = 45, σ = 25) Group 3: 0.5N(μ = 15, σ = 25)+0.5 N(μ = 50, σ = 25) Group 4: 0.5N(μ = 20, σ = 25)+0.5 N(μ = 50, σ = 25)

2.2 Evaluation

Since within each dataset, we know the "correct" group to which the histograms generated belonged, after performing K-medoid clustering on each dataset, using each of the ten histogram distance measures, fixing K = 4, we create a confusion matrix from the clustering results and the "true" group information. We use the "correct" classification rate to compare the clustering performances of the ten histogram distance measures. This correct classification rate is defined as the ratio between sum of the diagonal values in the confusion matrix to the sample size (or 400) as explained in Table 3. We then repeat these steps 40 times for each dataset.

Table 3: Example of confusion matrix. Correct Prediction Rate is computed as (A+B+C+D)/400.

		True group			
		1	2	3	4
Cluster Results	1	A			
	2		B		
	3			C	
	4				D

3 RESULTS

The results, shown in Figures 3 to 9, display the mean prediction rate, with a 95% confidence interval, for each of the ten distance measures on each of the datasets, from which we discuss the effects of types of data and sizes of standard deviation on the histogram distance measures' clustering performances.

Figure 3 shows the mean prediction rates on the "Four types" dataset for each of the ten histogram distance measures. In general, almost all distance measures give nearly perfect clustering as the four groups are quite easily distinguishable. The worst two measures are Kullback-Leiber and Intersection.

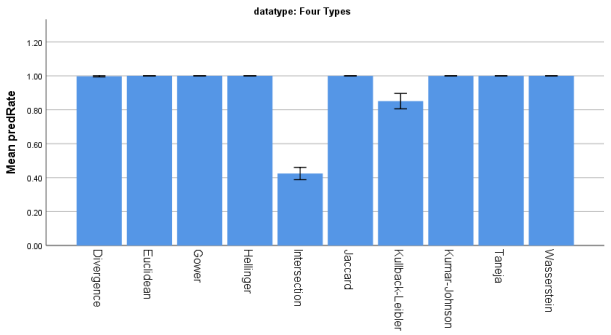


Figure 3: The mean prediction rates for "Four types" dataset

Figures 4 to 6 show the mean prediction rates on the "Normal" datasets. In general, Euclidean, Gower, Hellinger, Jaccard, Kumar-Johnson, and Wasserstein give better clustering performances than the rest. Moreover, as the standard deviations increase from 15, to 20, to 25, all measures' performances decrease noticeably. At the highest standard deviation of 25, Wasserstein measure emerges as the winner.

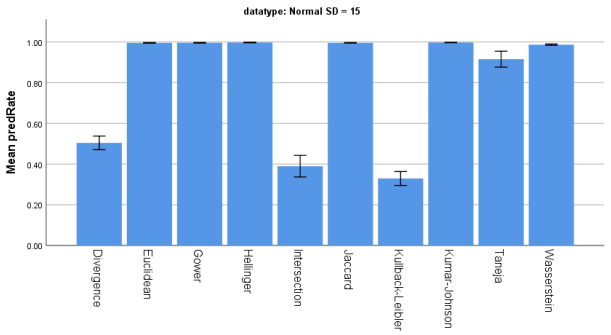


Figure 4: The mean prediction rates for "Normal SD = 15" dataset

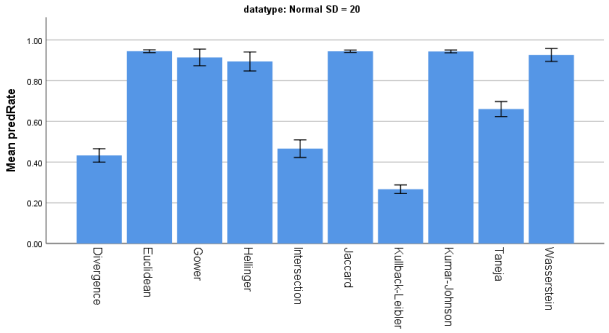


Figure 5: The mean prediction rates for "Normal SD = 20" dataset

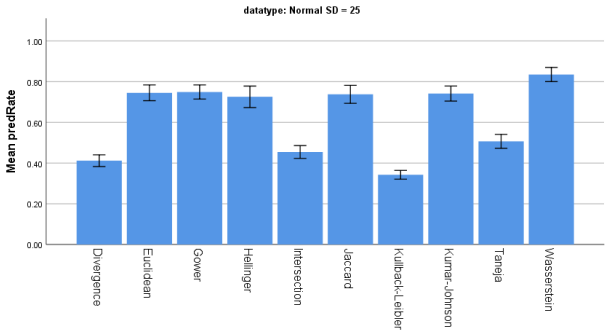


Figure 6: The mean prediction rates for "Normal SD = 25" dataset

Figures 7 to 9 show the mean prediction rates on the "Mixed Normal" datasets. In general, Euclidean, Gower, Hellinger, Jaccard, and Kumar-Johnson give better clustering performances overall. Moreover, as the standard deviations increase from 15, to 20, to 25, all measures' performances drop noticeably. At the highest standard deviation of 25, all measures perform poorly at below 60% accuracy, with Hellinger distance appearing to perform better than others. Divergence, intersection, and Kullback-Leibier tend to perform poorly across all mixed data type.

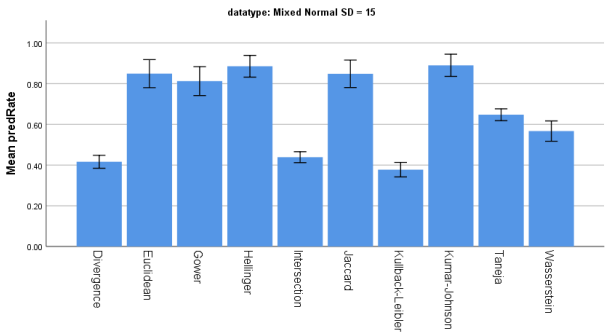


Figure 7: The mean prediction rates for "Mixed Normal SD = 15" dataset

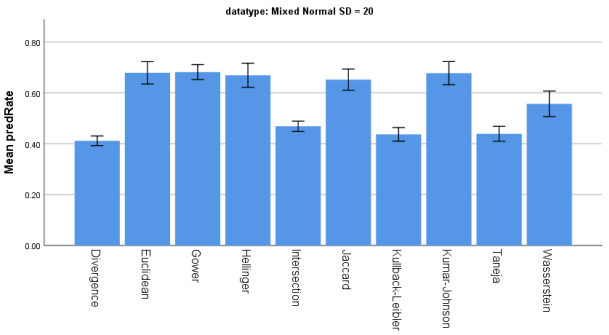


Figure 8: The mean prediction rates for "Mixed Normal SD = 20" dataset

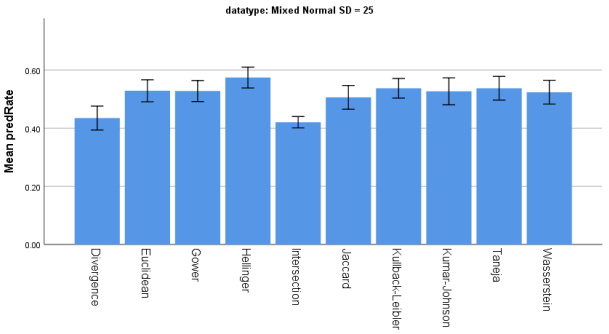


Figure 9: The mean prediction rates for "Mixed Normal SD = 25" dataset

4 CONCLUSIONS

In this paper, we compare ten histogram distance measures for the purpose of clustering population pyramids, through a simulation study, where K-medoid algorithm is used as the clustering tool. The results show that there is no single "best" distance measure that outperforms other across all data types. In general, the higher the variation, the harder the clustering becomes. For the unimodal pyramids with high variances, cross-bin types of measures might be a better choice. On the other hand, for the bimodal pyramids, bin to bin type of measures, such as Euclidean, Hellinger, and Jaccard generally perform better than the rest. In conclusion, when one wants to conduct cluster analysis on histogram data, the shapes and the variations of such histograms should be investigated first so that an appropriate distance measures can be selected for the analysis, accordingly.

ACKNOWLEDGEMENTS

The authors would like to thank Department of Statistics, Faculty of Commerce and Accountancy, Chulalongkorn university for the invaluable supports throughout this work. The analysis was performed in R and SPSS software.

REFERENCES

Boucher, L. (2016). What are the different types of population pyramids?". www.populationeducation.org. Retrieved 29 March 2017.

Cha, S., & Srihari, S.N. (2002). On measuring the distance between histograms. *Pattern Recognition, 35*(6), 1355-1370.

Everitt, B.S. (2012). *Cluster analysis*. Chichester: Wiley.

Kahn, J.R. (2007). Demographic Techniques: Population Pyramids and Age/Sex Structure. *The Blackwell Encyclopedia of Sociology*. doi:10.1002/9781405165518.wbeosd023

Pfanzagl, J. (1982). Distance Functions for Probability Measures. *Contributions to a General Asymptotic Statistical Theory Lecture Notes in Statistics*, 90-98. doi:10.1007/978-1-4612-5769-1_7

Cha, S. (2008) Taxonomy of Nominal Type Histogram Distance Measures. *American Conference on Applied Mathematics (MATH '08)*, Harvard, Massachusetts, USA.

Zhao, X., & Chen, Y. (2011). Evaluation on color spaces and distance measures for color histogram-based image retrieval. *International Conference on Graphic and Image Processing (ICGIP 2011)*. doi:10.1117/12.914398

A Comparison of Competing Risk Analysis to Study the Risk Factors of Loss to Follow-up in HIV-Infected Children on Antiretroviral Therapy

Suttipong Kawilapat[1*] and Patrinee Traisathit[2]

[1]Graduate Program in Applied Statistics, Department of Statistics, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand; Prevention and Treatment of HIV Infection and Virus-associated Cancers in South East Asia (PHPT), Chiang Mai, Thailand
*Corresponding Email: suttipong.kawilapat@gmail.com
[2]Department of Statistics, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand; Center of Excellence in Bioresources for Agriculture, Industry and Medicine, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand; Prevention and Treatment of HIV Infection and Virus-associated Cancers in South East Asia (PHPT), Chiang Mai, Thailand
Email: patrinee.t@cmu.ac.th

ABSTRACT

In the presence of competing events in a time-to-event analysis, a competing risk analysis has been suggested as an alternative method due to treating competing events as censored in standard survival analysis methods such as Kaplan-Meier (KM) and Cox regression, which could overestimate the true failure probability and lead to biased results. However, there has been some disagreement on the suitability of using competing risk analysis if competing events are present. To examine the effect of accounting for competing events in a time-to-event analysis, we compared the cumulative incidence of loss to follow-up (LTFU) between KM and the cumulative incidence function (CIF) to account for competing events (i.e. referral to another hospital and death) and compared the strength of association of potential risk factors between univariable Cox regression and univariable competing risk regression. For the analyses, we included characteristics at the initiation of antiretroviral therapy (ART) of 873 HIV-infected children (aged <18 years) who enrolled in the Prevention and Treatment of HIV Infection and Virus-associated Cancers in a South East Asia cohort study in Thailand and received ART between January 1999 and December 2014. At the median of the follow-up (8.6 years, interquartile range: 4.5-10.6), the estimated cumulative incidence of LTFU using KM was 20.3%, which was higher than the 16.7% estimated using CIF. Indeed, KM overestimated the cumulative incidence of LTFU at any point in the follow-up period. The factors which were significantly associated in the Cox regression lost about 8-30% of their association on the risk of LTFU when accounting for referral to another hospital and death as competing events in the analysis. The findings of this study point toward CIF and competing risk regression as the appropriate statistical methods to deal with competing events in a time-to-event analysis.

Keywords: competing risk analysis; survival analysis; HIV; loss to follow-up

1 INTRODUCTION

A time-to-event analysis is commonly applied in medical research, especially in cohort studies, to follow up on the occurrence of the event of interest in each participant. The observations might have been terminated prior to the occurrence of the event of interest with the termination during follow-up being due to several reasons such as loss to follow-up (LTFU), withdrawal from the study, or death. The occurrence of other outcomes, named competing events, may preclude the occurrence of the event of interest. In standard survival analysis methods such as Kaplan-Meier (KM) and Cox regression, events other than the one of interest are ignored and are subsequently defined as being censored. Since the failure rate in a standard survival analysis is defined as the number of occurrences of the event of interest divided by the total exposure time and precludes the occurrences of competing events as censoring, this approach could overestimate the true failure rate because of the denominator being too small, which it can lead to biased results (Holme et al., 2013). In the presence of competing events, the cumulative incidence function (CIF) (Coviello & Boggess, 2004; Gooley et al., 1999; Moraes et al., 2012) and competing risks regression (Bakoyannis & Touloumi, 2012; Fine & Gray, 1999; Grover et al., 2014; Holme et al., 2013; Noordzij et al., 2013; Wolkewitz et al., 2014) have been specifically designed to handle data with competing events and have been suggested as an alternative approach to estimate and identify any associated factors on the risk of outcomes.

In a study of the risk factors of LTFU in HIV-infected children on antiretroviral therapy (ART), we were interested in using CIF and competing risk regression in the analyses because referral to another hospital and death are competing events for LTFU. However, there has been some disagreement on the suitability of using a competing risk analysis if the competing events are presented. In a previous study comparing the results of Cox regression and Fine and Gray competing risk regression, the authors stated that the Cox regression was better than Fine and Gray competing risk regression because different types of events of interest could result to different patterns of competing events, which makes it difficult in practice to take competing events into account in the analysis (Ranstam & Robertsson, 2017). Therefore, to examine the effect of accounting for competing events in a time-to-event analysis, our aim was to compare the cumulative incidence and strength of association on the risk of LTFU between a standard survival analysis and a competing risk analysis using data on HIV-infected children who received ART in a cohort study in Thailand.

2 METHODS

2.1 Study Population

HIV-infected children (aged <18 years) who enrolled in the Prevention and Treatment of HIV Infection and Virus-associated Cancers in a South East Asia (PHPT) prospective multicenter cohort study (ClinicalTrials.gov: NCT00433030) in Thailand and received ART between 1 January 1999 and 31 December 2014 were included in the analysis. After the ART initiation, children were followed up at 2 weeks, then 1, 3, and 6 months, and every 6 months thereafter. Children who did not attend on regular visit were traced up with telephone calls and home visits.

The caregivers of the participating children provided written informed consent before entry into the cohort study. All data were collected anonymously using patient identification numbers. The PHPT cohort study protocol was approved by the ethics committees at the Thai Ministry of Public Health and at the local hospitals.

2.2 Outcomes

LTFU, which was the primary outcome, was defined as the failure to attend the HIV clinic for more than 9 months despite repeated attempts to reach the child or caregiver via telephone calls or home visits. For the children who LTFU and could be traced after, in case of moving to live in other places and referring to get the treatment in other hospitals were defined as referral to another hospital which accounting as one of competing events in the competing risk analysis. Death was considered as the children in the PHPT cohort who died from any cause during follow-up.

2.3 Variables

Variables assessed for association with the risk of LTFU at ART initiation were sex, age, country of birth, type of hospital at birth,

relationship with the caregiver, height-for-age and weight-for-age using the standard growth values of Thai children (Working Group on Using Weight and Height References in Evaluating the Growth Status of Thai Children, 2000), ART experience, the Centers for Disease Control and Prevention HIV Classification, HIV-RNA load, CD4 percentage, and anemia status using the World Health Organization (WHO) criteria (World Health Organization, 2011). Missing values were imputed using the nearest available data to ART initiation (within 1 year before or, if missing, within 15 days after).

2.4 Statistical Analyses

Referral to another hospital and death from any cause were considered censored in the survival analysis and as competing events of LTFU in the competing risk analysis. In the survival analysis, the cumulative incidence of LTFU was estimated using the KM method whereas the cumulative incidence was estimated using CIF accounting for the competing events in the competing risk analysis (Coviello & Boggess, 2004).

The changes in strength of association of the potential risk factors for LTFU were considered using the differences in hazard ratio (HR) in a univariable Cox regression and the subhazard ratio (SHR) in a univariable Fine and Gray competing risk regression (Fine & Gray, 1999).

We also compared the LTFU rate, competing events rate, and overall risk ratios for each potential variable based on the calculations of the rates as follows (Wolkewitz et al., 2014):

- LTFU rate = (number of children with LTFU) / (number of children-years at risk)
- Competing events rate = (number of children with competing events) / (number of children-years at risk)
- Overall risk = (LTFU rate) / (LTFU rate + competing events rate)

All analyses were performed using Stata 12.0 (StataCorp LP).

3 RESULTS

3.1 Cumulative Incidence of LTFU

Of 873 children who included in the analyses, 196 (22%) were LTFU, 195 (22%) had been referred to another hospital, and 73 (8%) had died. The median follow-up duration was 8.6 years (interquartile range: 4.5-10.6). At the median follow-up duration, the estimated cumulative incidence of LTFU using KM was 20.3%, which was higher than the 16.7% estimated using CIF. Hence, KM overestimated the cumulative incidence of LTFU at any point in the follow-up period (Table 1 and Figure 1).

Table 1: Cumulative incidence of LTFU after 10 years of ART initiation

Year	KM		CIF	
	Cumulative incidence (95% CI)		Cumulative incidence (95% CI)	
1	0.030	(0.020-0.044)	0.029	(0.019-0.413)
2	0.046	(0.034-0.063)	0.044	(0.031-0.059)
3	0.057	(0.043-0.075)	0.053	(0.039-0.691)
4	0.068	(0.052-0.087)	0.062	(0.047-0.080)
5	0.080	(0.063-0.102)	0.073	(0.057-0.091)
6	0.098	(0.079-0.121)	0.087	(0.070-0.107)
7	0.129	(0.106-0.155)	0.112	(0.092-0.134)
8	0.174	(0.148-0.205)	0.146	(0.123-0.171)
8.6[a]	0.203	(0.174-0.235)	0.167	(0.143-0.193)
9	0.222	(0.191-0.256)	0.181	(0.155-0.208)
10	0.280	(0.245-0.318)	0.222	(0.193-0.252)

Abbreviations: KM = Kaplan-Meier; CIF = Cumulative incidence function; CI = Confidence intervals
[a] Median of follow-up

3.2 Strength of Association on the Risk of LTFU

The strength of association of potential risk factors significantly associated with the risk of LTFU in the Cox regression being lower in the competing risk regression but higher when not associated (except for CD4 percentage). The factors significantly associated in the Cox regression (age at ART initiation ≥7 years, country of birth other than Thailand, born in a district hospital, living with relatives, and initiation of ART at or after enrollment) lost 8.0-29.6% of their association on the risk of LTFU after accounting for referral to other hospitals and death as competing events in the analysis (Table 2).

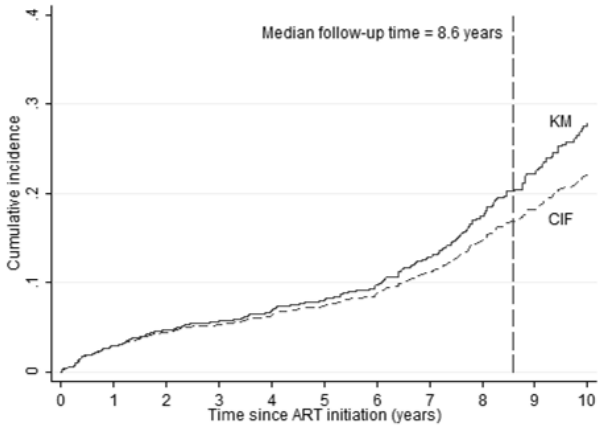


Figure 1: Cumulative incidence of LTFU after 10 years of ART initiation

3.3 LTFU Rate, Competing Events Rate, and Overall Risk Ratios

The LTFU rate ratio was high (rate ratio ≥1.50) among children aged ≥ 7 years at ART initiation, born in a country other than Thailand, lived with relatives, and initiated ART at or after enrollment in the cohort study, while the competing events rate ratio was high (rate ratio ≥1.50) for the same factors other than age ≥7 years at ART initiation. When referral and death were accounted in the calculation, the overall risk ratio was lower than the LTFU rate ratio, notably for children born in a country other than Thailand and who lived with relatives (difference >0.50) (Table 3).

4 DISCUSSIONS

At any point in time during the follow-up period, the estimates of cumulative incidence using KM were overestimated compared to CIF. The cumulative incidence of these methods was close at the beginning of follow-up and then the distance between them increased slightly after 6 years. Since a number of competing events occurred in the first 6 years of follow-up (162 (60%) of 268 children who had competing events in all follow-up periods), it might have affected the estimates of KM to make them similar to those of CIF. This is consistent with Moraes et al. who suggested using CIF instead of KM to avoid the overestimation (Moraes et al., 2012). Gooley et al. suggested that the comparison of cumulative incidences is not enough to clarify the effect of competing events (Gooley et al., 1999), rather it should be evaluated on the covariate's effect on the risk of the event of interest and the competing events.

Age ≥7 years at ART initiation, country of birth other than Thailand, born in a district hospital, living with relatives, and initiation of ART at or after enrollment in the cohort study were associated with the risk of LTFU in the univariable Cox regression. Meanwhile, the strength of association of these factors was decreased when the competing events were accounted in the Fine and Gray competing risk regression, especially for children born in a country other than Thailand, lived with relatives, and initiated ART at or after enrollment in the cohort study. The comparison in our study is in agreement with the one in Holme et al. who found that the strength of association in the competing risk regression could both decrease and increase compared to the Cox regression (Holme et al., 2013).

When we considered the association between the LTFU rate, competing events rate, and overall risk ratios; the factors with a high competing events rate: born in a country other than Thailand, living with relatives, and initiating ART at or after enrollment in the cohort

Table 2: Comparison of association of potential risk factors of LTFU between univariable Cox regression and univariable competing risk regression

Characteristics at ART initiation		Cox regression [a]			Competing risk regression [a, b]			Changes in strength of association [c]
	n/N	HR	(95% CI)	*p*	SHR	(95% CI)	*p*	
Female	111/471	1.14	(0.87-1.52)	0.334	1.17	(0.84-1.48)	0.439	2.6%
Age ≥7 years	114/405	1.88	(1.41-2.51)	<0.001	1.73	(1.30-2.30)	<0.001	-8.0%
Country of birth other than Thailand	5/16	3.04	(1.25-7.43)	0.015	2.14	(0.93-4.92)	0.074	-29.6%
Born in a district hospital	64/252	1.44	(1.01-2.08)	0.046	1.30	(0.90-1.86)	0.159	-9.7%
Living with relatives (versus orphanage)	159/665	2.39	(1.17-4.86)	0.016	2.08	(1.03-4.21)	0.042	-13.0%
Height-for-age Z-scores ≥-2 SD [d]	75/298	1.03	(0.74-1.43)	0.861	1.09	(0.79-1.51)	0.576	5.8%
Weight-for-age Z-scores ≥-2 SD [d]	126/503	1.05	(0.66-1.69)	0.836	1.30	(0.81-2.07)	0.278	23.8%
ART initiation at or after enrollment to cohort	141/613	1.56	(1.14-2.15)	0.006	1.26	(0.92-1.71)	0.144	-19.2%
CD4 ≤10%	85/364	1.11	(0.80-1.55)	0.529	1.05	(0.76-1.46)	0.751	-5.4%
HIV RNA load ≤100,000 copies/mL	51/199	1.08	(0.75-1.55)	0.685	1.13	(0.79-1.62)	0.500	4.6%
CDC HIV classification (class N or A)	93/350	1.17	(0.88-1.57)	0.274	1.23	(0.92-1.63)	0.162	5.1%
Mild or no anemia (WHO criteria) [e]	76/303	1.16	(0.83-1.61)	0.390	1.22	(0.88-1.70)	0.240	5.2%

Abbreviations: ART = Antiretroviral treatment; n = Number of children who were LTFU in the category; N = Number of children in the category; HR = Hazard ratio; SHR = Subhazard ratio; CI = Confidence intervals; SD = Standard deviation; CDC = Centers for Disease Control and Prevention; WHO = World Health Organization

[a] Missing baseline values were imputed using the nearest available data within 1 year before or within 15 days after ART initiation.

[b] Competing events: referral to another hospital, death

[c] The differences between the HR in Cox regression and the SHR in competing risk regression

[d] According to Thai weight and height reference values for children

[e] Anemia classification is based on hemoglobin level, sex, and age following the WHO criteria.

Table 3: LTFU rate, competing events rate, and overall risk ratios for potential risk factors of LTFU

Characteristics at ART initiation	LTFU rate			Competing events rate [a]			Overall risk ratio
	Exposed	Unexposed	Rate ratio	Exposed	Unexposed	Rate ratio	
Female	0.031	0.027	1.148	0.041	0.038	1.079	1.037
Age ≥7 years	0.038	0.022	1.727	0.040	0.039	1.026	1.351
Country of birth other than Thailand	0.076	0.030	2.533	0.061	0.039	1.564	1.276
Born in a district hospital	0.036	0.026	1.385	0.046	0.037	1.243	1.064
Living with relatives (versus orphanage)	0.032	0.014	2.286	0.041	0.020	2.050	1.065
Height-for-age Z-scores ≥-2 SD [b]	0.036	0.034	1.059	0.042	0.050	0.840	1.140
Weight-for-age Z-scores ≥-2 SD [b]	0.035	0.034	1.029	0.041	0.073	0.562	1.449
ART initiation at or after enrollment to cohort	0.033	0.022	1.500	0.048	0.027	1.778	0.907
CD4 ≤10%	0.033	0.030	1.100	0.049	0.043	1.140	0.979
HIV RNA load ≤100,000 copies/mL	0.035	0.033	1.061	0.044	0.050	0.880	1.114
CDC HIV classification (class N or A)	0.035	0.030	1.167	0.037	0.045	0.822	1.215
Mild or no anemia (WHO criteria) [c]	0.035	0.031	1.129	0.042	0.049	0.857	1.173

Abbreviations: ART = Antiretroviral treatment; LTFU = Loss to follow-up; SD = Standard deviation; CDC = Centers for Disease Control and Prevention; WHO = World Health Organization

[a] Competing events: referral to another hospital, death

[b] According to Thai weight and height reference values for children

[c] Anemia classification is based on hemoglobin level, sex, and age following the WHO criteria.

study were notably decreased in the overall risk ratio. This is consistent with the results of the comparison of strength of association that showed decreases of up to 29.6%. However, even though the competing event rates of some variables such as age and location of hospital at birth were not very high, they were still affected by a slight decrease in the strength of association. The association, level, and direction of differences varied depending on the covariate effects on both the event of interest and competing events. Dignam et al. illustrated the modeling approaches with and without accounting for competing events and found that the effects of covariate could be different depending on their relationship to events. Even if the covariate effects were shared between event of interest and competing events, both approaches might be informative. Thus, the choice of approaches depends on questions of interest (Dignam et al., 2012).

The aim of this study was only to examine the effect of accounting for competing events in a time-to-event analysis; we analyzed this using only the characteristics of HIV-infected children at ART initiation. To identify the risk factors of LTFU with an optimal model, we need to include more information such as time-dependent variables that will be recorded with the follow-up data during the study or use a multivariable analysis to adjust the effect of the covariate.

5 CONCLUSIONS

The findings of this study point toward CIF and competing risk regression as the appropriate statistical methods to deal with competing events in a time-to-event analysis. However, the choice of analysis approaches depends on the covariate effects to events and the questions of interest.

ACKNOWLEDGEMENTS

We would like to thank all the children and their caregivers who participated in the PHPT cohort study. We are also grateful to all investigators in each hospital site and PHPT staff for their involvement in the PHPT cohort study and Sukon Prasitwattanaseree who provided statistical advice.

The PHPT prospective multicenter cohort study was funded by the Global Fund to AIDS, Tuberculosis and Malaria, Thailand (PR-A-N-008); Oxfam Great Britain, Thailand (THAA51); Ministry of Public Health, Thailand; and Institut de Recherche pour le Développement (IRD), France.

REFERENCES

Bakoyannis, G., & Touloumi, G. (2012). Practical methods for competing risks data: a review. *Statistical Methods in Medical Research*, *21*(3), 257–272.

Coviello, V., & Boggess, M. (2004). Cumulative incidence estimation in the presence of competing risks. *Stata Journal*, *4*(2), 103–112.

Dignam, J. J., Zhang, Q., & Kocherginsky, M. (2012). The use and interpretation of competing risks regression models. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *18*(8), 2301–2308.

Fine, J. P., & Gray, R. J. (1999). A Proportional hazards model for the subdistribution of a competing Risk. *Journal of the American Statistical Association*, *94*(446), 496–509.

Gooley, T. A., Leisenring, W., Crowley, J., & Storer, B. E. (1999). Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Statistics in Medicine*, *18*(6), 695–706.

Grover, Gurprit, Swain, Prafulla, & Vajala, Ravi. (2014). A competing risk approach with censoring to estimate the probability of death of HIV/AIDS patients on antiretroviral therapy in the presence of covariates. *Statistics Research Letters*, *3*(1), 7–16.

Holme, I., Fellström, B. C., Jardine, A. G., Hartmann, A., & Holdaas, H. (2013). Model comparisons of competing risk and recurrent events for graft failure in renal transplant recipients. *Clinical Journal of the American Society of Nephrology : CJASN*, *8*(2), 241–247.

Moraes, R. B., Friedman, G., Lisboa, T., Viana, M. V., Hirakata, V., & Czepielewski, M. A. (2012). Comparison of cumulative incidence analysis and Kaplan-Meier for analysis of shock reversal in patients with septic shock. *Journal of Critical Care*, *27*(3), 317.e7-11.

Noordzij, M., Leffondré, K., van Stralen, K. J., Zoccali, C., Dekker, F. W., & Jager, K. J. (2013). When do we need competing risks methods for survival analysis in nephrology? *Nephrology, Dialysis, Transplantation: Official Publication of the European Dialysis and Transplant Association - European Renal Association*, *28*(11), 2670–2677.

Wolkewitz, M., Cooper, B. S., Bonten, M. J. M., Barnett, A. G., & Schumacher, M. (2014). Interpreting and comparing risks in the presence of competing events. *BMJ : British Medical Journal*, *349*. https://doi.org/10.1136/bmj.g5060

Working Group on Using Weight and Height References in Evaluating the Growth Status of Thai Children. (2000). *Manual on using weight and height references in evaluating the growth status of Thai children*. Bangkok: Department of Health, Ministry of Public Health.

World Health Organization. (2011). *Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity* (Vitamin and Mineral Nutrition Information System No. WHO/NMH/NHD/MNM/11.1). Geneva: World Health Organization. Retrieved from www.who.int/vmnis/indicators/haemoglobin.pdf

Factors Associated with Medical Pneumonia Cost using Linear Regression and Neural Network

Akemat Wongpairin[1*] and Phattrawan Tongkumchum[2]

[1]Department of Medical Information
Satun Hospital Satun Province, Thailand 91000
*Corresponding Email: akemat.w@kkumail.com
[2]Department of Mathematics and Computer Science Faculty of Science and Technology,
Prince of Songkla University, Pattani Province Thailand 94000
Email: phattrawan@gmail.com

ABSTRACT

Pneumonia is a global significant issue which remains one of the major causes of mortality. This study aims to investigate medical cost of pneumonia among 2,082 patients in Satun province. The outcome variable is medical cost of pneumonia and determinants are demographic and medical factors. Multiple linear regression (MLR) and neural network (nnet) were used. The higher medical cost was associated older age. The cost was higher in year 2013. Having cancer, chronic renal failure and sepsis as co-morbidity was also associated with higher cost. Using antibiotic and ventilator treatment was associated with higher cost. Hospital acquired pneumonia had higher medical cost than average. According to methods used, the nnet had higher performance prediction than MLR with the original data, but lower performance prediction than MLR with training and testing data sets.

Keywords: pneumonia; medical cost; neural network

1. INTRODUCTION

Pneumonia is a global significant issue which remains one of the major causes of mortality among patient in South East Asia. In 2010, it was estimated that there were 2,675 per million and 9.6% of mortality in Thailand. The elderlies and adults with co-morbidity conditions are also at increased risk of pneumonia. These include people who get diabetes mellitus, ischemic heart disease, chronic heart failure, chronic renal failure, cancer, human immunodeficiency virus, sepsis and those who have stayed for long term at hospital care and ventilator treatment. The classification of pneumonia may be classified by based on the location of prior exposure and pathogen which can be categorized as community acquired pneumonia and hospital acquired pneumonia. Pneumonia studies have compared the cost of treatment between pneumonia and a lot of research that conducted in linear regression method and explored the cost of CAP, HAP and co-morbidity of pneumonia by using hospital information system data.

An unprecedented health organization policy worldwide effort is now to boost the adoption of electronic medical records and stimulate innovations in health information. Health information technology has the potential to improve the health innovation and the performance of health care quality and cost savings. Despite the evidences of these benefits on big data, the innovation of health information is less than business studies or other studies. Specifically, establishing the machine learning study and big data of health emphasized on the importance of identifying and testing the innovative and implement models. There are many studies still use in the statistical standard in the big data project.

Machine learning applied to electronic medical records can analyze actionable insights from improving data classification, to predict the cost of treatment. Machine learning models that leverage the variety and richness of health data are still relatively rare and offer an exciting avenue for further research. While Narendra and Parthasarathyhad (1990) described a memory-based network that provided the estimates of continuous variables with identification and control of dynamic systems using neural networks, there were reported about linear models, which included multi-layer neural networks as well as linear dynamics. It could be viewed as generalized neural networks. In this study, we have presented an overview of comparison between linear regression and neural network. These two methods have been applied in cost prediction. We describe the performance challenges of using machine learning in research and comparison practice. Lastly, we offer our perspective on future application areas for machine learning that will significantly impact health data and big data.

2. METHODS

Data source: A cross-sectional analytical study was conducted by using administrative databases obtained from Satun hospital, Thailand, during 2010 and 2014. These data were included demographic characteristics, medical history and cost of treatment in 2,082 cases.

Patient selection: Adult pneumonia patients who have at least one claim with the primary diagnosis code of pneumonia (J12-J18) and age more than 4 years old. Determinant comprises age, gender, diabetes mellitus (DM), ischemic heart disease (IHD), chronic heart failure (CHF), chronic renal failure (CRF), stroke, asthma, human immunodeficiency virus (HIV), cancer, sepsis, ventilator, antibiotic and the type of pneumonia (community acquired pneumonia (CAP), hospital acquired pneumonia (HAP) and unspecified pneumonia (unspec)).

Statistical methods: We used linear regression to model medical cost and its determinants and compared with neural network.

3. RESULTS

This study found that adult pneumonia were 2,082 patients with average age 58.4 years old (standard deviation= 23.2) and male 53.7 percent. Community acquired pneumonia 25.7 percent, hospital acquired pneumonia 2.9 percent and unspecified acquired pneumonia 71.4 percent, treatment with antibiotic 98.2 percent, ventilator treatment 34.2 percent and co-morbidity including diabetes mellitus 16.2 percent, ischemic heart disease 9.8 percent, chronic heart failure 10.8 percent, chronic renal failure 7.7 percent, asthma 5.4 percent, cancer 5.1 percent, human immune virus 6.0 percent, Stroke 0.4 percent and sepsis 7.3 percent.

Figure 1 showed factors associated with medical cost of pneumonia were age group, year, co-morbidity including cancer, chronic renal failure, sepsis, as well as antibiotic, ventilator treatment, and pneumonia type. The cost increased with age. During the five-year period, medical cost had increased for year 2013. The cost was higher than average for patients having cancer, chronic renal failure, and sepsis as comorbidity. Using antibiotic and ventilator treat were also associated with higher cost. Hospital pneumonia type is also associated with higher cost.

Figure 2 showed plot of fitted and observed values from linear regression and neural network using original data. The R^2 for linear regression was 47.7 percent while for the neural network was 71.4 percent.

Figure 3 showed plot of fitted and observed values using testing and training data. The data were splitted into 50:50 ratios for training and testing data sets. The models were fitted to the training 1,401 records using linear regression and neural network. Then the models were fitted to the testing data for 1,401 records. The performance of the model using testing data set was shown in Figure 3. The linear regression showed the difference with training model at 48.9 percent and testing at 47.3 percent while the neural network model was decreased the performance from training model 83.2 percent to testing model 21.4 percent.

4. DISCUSSIONS

The first finding in this study was the factors associated of medical cost on the occurrence of adult pneumonia. These found that hospital acquired pneumonia had maximal influence to medical cost in MLR model. Hospital acquired pneumonia still was the main problem in hospital which increased medical cost among ventilator treatment. Elderlies were the risk factors for increasing the medical care.

This study showed chronic renal failure which was significant with medical cost among pneumonia patient. Taiwan national Insurance studied the report chronic renal failure which could be increased risk 2.17-fold higher for inpatient pneumonia compared and these had not only an increased risk of pneumonia but also an increased severity of pneumonia.

Cancer disease among pneumonia patient was the major medical cost issue with long term care and antibiotic therapy because survival study about the associated parts between pneumonia and lung cancer found that lung cancer patients were died with pneumonia and all the lung cancer patients with a diagnosed infection were subjected to the antibiotic therapy.

Many reports suggested that inappropriate initial antibiotic therapy of microbiologically confirmed that hospital acquired pneumonia was associated with mortality. Attributed to antibiotic-resistant bacteria, this can increase medical cost which will be the patient experience with adverse events from health care service in hospital. Therefore, we should reduce the adverse events in health care system and their consequences which are needed the health workers on the best common competency and the equipment suitability. HAP continues to be a commonly encountered challenge among pneumonia patient and carries significant burdens of morbidity, antibiotic utilization and medical cost. Health development on prevention strategies directed towards the pathophysiological mechanisms of HAP can be shown the success of their health care.

We have presented an overview of comparison of neural network methods which has been applied in medical cost prediction. It found that it depended on data number and data complex. Traditional statistical analysis methods have lower performance than machine learning on big data. However, these still are (dependable) than machine learning when we analyze on the small data. American Medical Association journal published about the inevitable application of big data with health care. It found that the transition of data from small data to big data was important in the research. Innovations in analytic techniques on the computer sciences, especially in machine learning, had been a major catalyst for dealing with these complex data sets. These analytic techniques are in contrast with traditional statistical methods which derived from the social and physical sciences. These are not useful for analysis of unstructured data, such as text-based documents because these do not fit into relational tables.

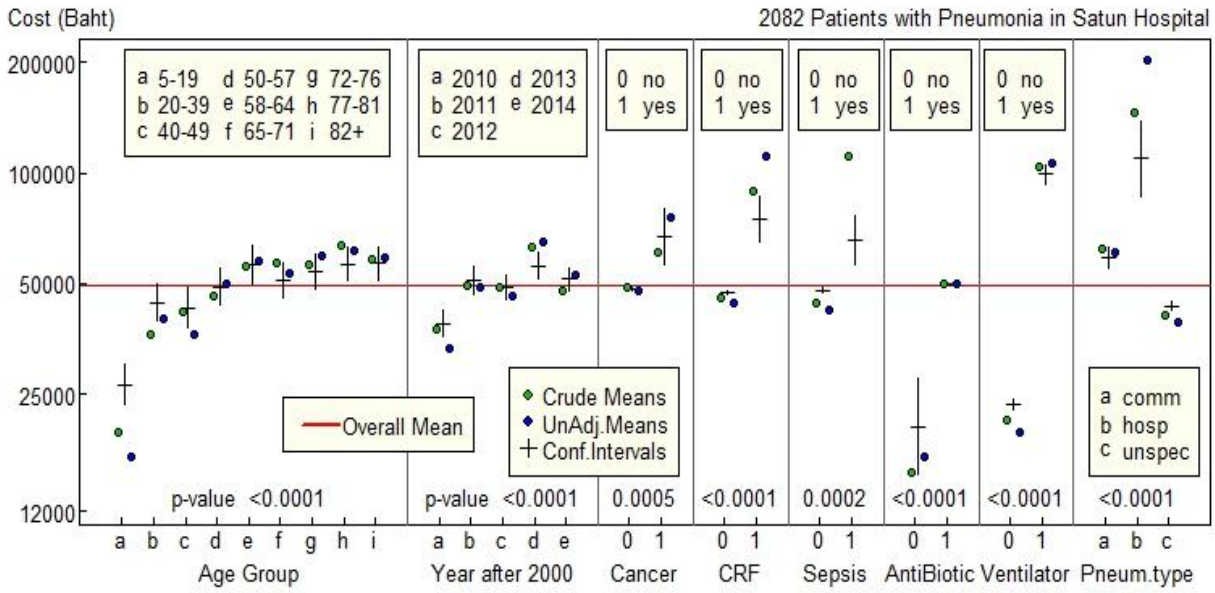


Figure 1 Linear regression model using original data

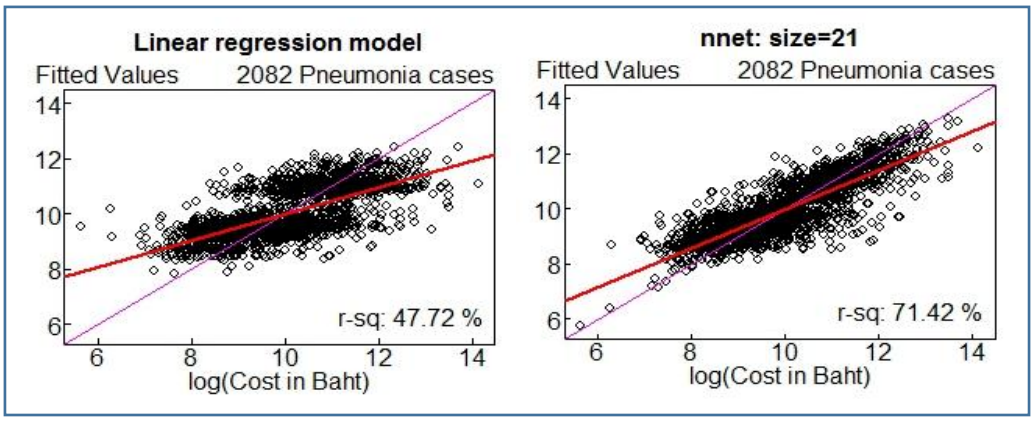


Figure 2 Plot of observed and fitted values using original data

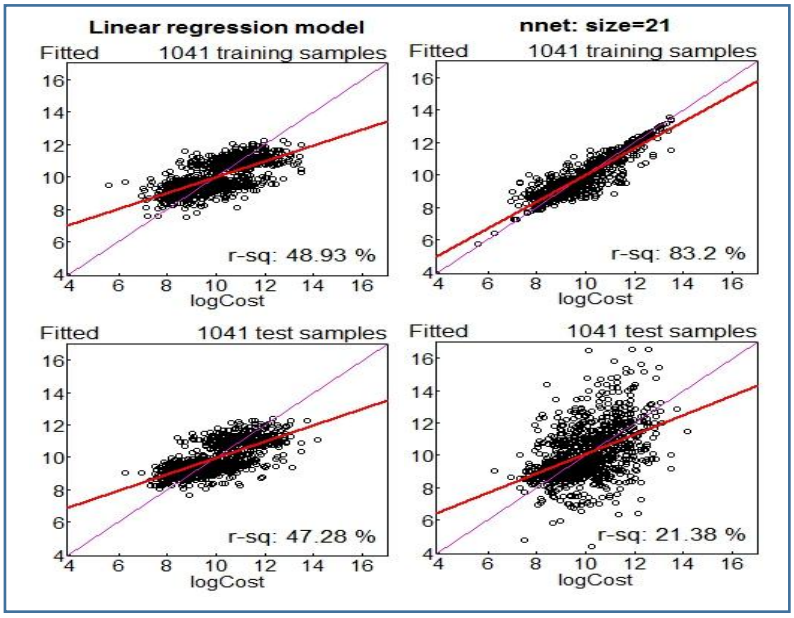


Figure 3 Plot observed and fitted values using split data

ACKNOWLEDGEMENTS

The authors would like to gratefully thank director of Satun hospital, my advisor and Mr. Wiroj Yommuang, the director of the department of Health Services Development in Satun hospital for their valuable suggestions and support. This study is the data supported by the department of Medical Informatics of Satun hospital.

REFERENCES

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. 2014. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, *33*(7), 1123-1131.

Boonsawat, W., Boonma, P., Tangdajahiran, T., Paupermpoonsiri, S., Wongpratoom, W., & Romphryk, A. (1990).Community-acquired pneumonia in adults at Srinagarind Hospital. *Journal of the Medical Association of Thailand*, *73*(6), 345-352.

Buntin, M. B., Burke, M. F., Hoaglin, M. C., & Blumenthal, D. 2011. The benefits of health information technology: a review of the recent literature shows predominantly positive results. *Health affairs*, *30*(3), 464-471.

Callahan, A., & Shah, N. H. 2018.Machine Learning in Health care. In Key Advances in Clinical Informatics pp. 279-291.

Cho, S. H., Ketefian, S., Barkauskas, V. H., & Smith, D. G. (2003). The effects of nurse staffing on adverse events, morbidity, mortality, and medical costs. *Nursing research*, *52*(2), 71-79.

Chou, C. Y., Wang, S. M., Liang, C. C., Chang, C. T., Liu, J. H., Wang, I. K., & Wang, R. Y. (2014). Risk of pneumonia among patients with chronic kidney disease in outpatient and inpatient settings: a nationwide population-based study. *Medicine*, *93*(27).

Fry, A. M., Lu, X., Chittaganpitch, M., Peret, T., Fischer, J., Dowell, S. F., & Olsen, S. J. (2007). Human bocavirus: a novel parvovirus epidemiologically associated with pneumonia requiring hospitalization in Thailand. *The Journal of infectious diseases*, *195*(7), 1038-1045.

Jain, S., Self, W. H., Wunderink, R. G., Fakhran, S., Balk, R., Bramley, A. M., & Chappell, J. D. (2015). Community-acquired pneumonia requiring hospitalization among US adults. *New England Journal of Medicine*, *373*(5), 415-427.

Lacher, R. C., Coats, P. K., Sharma, S. C., & Fant, L. F. 1995.A neural network for classifying the financial health of a firm. *European Journal of Operational Research*, *85*(1), 53-65.

Murdoch, T. B., & Detsky, A. S. 2013.The inevitable application of big data to health care. Jama, *309*(13), 1351-1352.

Narendra, K. S., & Parthasarathy, K. (1990).Identification and control of dynamical systems using neural networks. *IEEE Transactions on neural networks*, *1*(1), 4-27.

Olsen, S. J., Thamthitiwat, S., Chantra, S., Chittaganpitch, M., Fry, A. M., Simmerman, J. M., & Talkington, D. 2010. Incidence of respiratory pathogens in persons hospitalized with pneumonia in two provinces in Thailand. *Epidemiology & Infection*, 138(12), 1811-1822.

Pelletier, A. J., Mansbach, J. M., & Camargo, C. A. (2006).Direct medical costs of bronchiolitis hospitalizations in the United States. *Pediatrics*, *118*(6), 2418-2423.

Raghupathi, W., & Raghupathi, V. 2014. Big data analytics in healthcare: promise and potential. *Health information science and systems*, *2*(1), 3.

Recknagel, F., French, M., Harkonen, P., & Yabunaka, K. I. 1997. Artificial neural network approach for modelling and prediction of algal blooms. *Ecological Modelling*, 96(1-3), 11-28.

Song, J. H., Thamlikitkul, V., & Hsueh, P. R. (2011).Clinical and economic burden of community-acquired pneumonia amongst adults in the Asia-Pacific region. *International journal of antimicrobial agents*, *38*(2), 108-117.

Specht, D. F. (1991). A general regression neural network. *IEEE transactions on neural networks*, *2*(6), 568-576.

Zieba, M., Baranowska, A., Krawczyk, M., Noweta, K., Grzelewska-Rzymowska, I., & Kwiatkowska, S. (2003). Pneumonia as a cause of death in patients with lung cancer. *Radiology and Oncology*, *37*(3).

An Assessment of Knowledge on Biostatistics among the Postgraduate Students of a Medical College in Northeast India

RajkumariSanatombi Devi*

Department of Community Medicine
Sikkim Manipal Institute of Medical Sciences,
5th Mile, Tadong, Gangtok, Sikkim, India
*Corresponding Email: rajkumari.sd@gmail.com

ABSTRACT

The objective of this study was to assess level of knowledge in understanding the basic concepts of descriptive and inferential Biostatistics among the Postgraduate (PG) students of Sikkim Manipal Institute of Medical Sciences, Gangtok, Sikkim, India. The study was a cross sectional survey study. All the Postgraduate (MD/MS/DNB) students enrolled in the first, second and third years during the academic session April 2013 to April 2017 were the study population. A questionnaire developed by the researcher was used for collecting the data. The questionnaire consisted of 47 items, including 38 knowledge based questions on Biostatistics. The reliability of the tool was done by Cronbach's alpha test. Statistical analysis consisted of frequency distribution for correct responses of each items converted into percentage were used to assess knowledge on Biostatistics. Out of 73 students enrolled, 66 participated in the study. Data was analyzed for 64 students as 2 students were excluded in the study. The response rate was 88%. The number of male and females were the same in the study. The average age of the students was 30 old. Majority of the students were from Clinical department (64.1%) and 43.8% of the students were from 1st year of their three year academic course. More than half of the students had previous training on Biostatistics/ research and had attended conference. The percentage of overall mean correct answer was 45.72 with a standard deviation of 11.81% (range 14.81 to 70.37%). More than 50% of the students answered the questions correctly in 9 questions out of 27 questions. The average number of correct answer responded by the students was 12. It was observed that more than half of the respondents were not able to answer the questions correctly that was designed to measure their knowledge on Biostatistics. Their lack of knowledge on Biostatistics can be improved by encouraging them to attend short term training on Biostatistics and motivating them to be involved in research activities.

Keywords: Biostatistics; Knowledge; Questionnaire; Northeast India

1 INTRODUCTION

In the Edinburg Declaration of World Conference on Medial Education held in the year 1998, it was made clear that the mandate of medical education should be to produce professionals who are capable of the undertaking needs of the community (The Edinburg Declaration, 1998).

Higher levels of statistical methods are being used in contemporary medical literature, but basic concepts, frequently occurring tests, and interpretation of results are not well understood by resident physicians. If physicians cannot detect appropriate statistical analyses and accurately understand their results, the risk of incorrect interpretation may lead to erroneous applications of clinical research (Windish et al., 2007).

It should be to be noted that anyone who is involved in medical research should always keep in mind that science is a search for the truth and that, in searching for the truth, there is no room for bias or inaccuracy in statistical analysis and interpretation of data (Peat & Michael, 2005). Any errors in statistical analysis will mean that the conclusions of the study may be incorrect (Altman, G. D. 1991). As per the Medical Council of India (MCI) requirements, postgraduate students have to carry out a dissertation project as a part of their Doctor of Medicine/ Master of Surgery (MD/MS) curriculum. Above that MCI also made it mandatory to not only attend one international /national conference, but also give an oral/poster presentation and send the article for publication (Giri et al., 2014).

Although Biostatistics is taught into several teaching-learning activities at undergraduate (MBBS) level, there is no recommendation for structured lessons in many Medical Institutes in India. At undergraduate level, this subject is taught as a part of the Community Medicine and the convention is to allocate 15/20 didactic lectures (spread across first year of the professional course) and 10/20 practical sessions in third semester of the second year to this subject. The purpose of this scientific study is to provide baseline knowledge about Biostatistics of medical postgraduate students of Sikkim Manipal Institutes of Medical Sciences, Sikkim Manipal University, India.

2 METHODS

The study was conducted among the 1st, 2nd, and 3rd year PG (MD/MS/DNB) students of Sikkim Manipal Institute of Medical sciences. Students enrolled during 2013 to 2017 were included in the study. A self administered questionnaire consisting of 47 items including 38 questions measuring knowledge on Biostatistics developed by the researcher was used for data collection. Most of the questions were based on the topics of undergraduate level that are followed in most of the medical colleges in India. The content validity was done with the help of subject experts as well as from literature searched from internet similar to the objective of the present study. The researcher also checked the pattern of previous questions asked in All India PG entrance test from the Preventive and Social Medicine where Biostatistics is a part of this subject. After obtaining the requisite permission from the institution Ethics committee, the researcher distributed the questionnaire to the PG students while they were attending the research methodology classes conducted by the department of Community Medicine, SMIMS. Those absentees were also tried to contact at their respective departments and questionnaires were distributed to them. The fill up form was collected directly from them by the researcher. Thirty minutes was given to fill up the form. Verbal consent was taken in collecting the data from the head of department and students after providing the purpose of the study. Anonymity of the respondents was not fully maintained as decision to write their name was given to the respondents. The reliability of the tool was done by calculating Cronbach's alpha test. The value of test was 0.597 indicating moderate reliability. The questionnaire consisted of two parts. The first part contained the demographic information like age and sex and current year of their academic course, previous exposure to Biostatistics/research training, publication of paper in journal. Information on their attendance and paper presentation in the conference was also included in the questionnaire. The second part of the questionnaire consisted of questions about knowledge on Biostatistics. There were 38 multiple choice questions (MCQs) having one correct answer and three false choices. Students marking the correct answer were given a score of 1 and the ones giving wrong answer were given a score of 0. Non-response items were counted as incorrect responses. Questions contained in the tool were type of variable, scale of measurement, source of data, concept, formula and computation of measure of central tendency. Two questions each on measure of

dispersion, identification and interpretation of correlation analysis were also included in the questionnaire. Question on inferential Biostatistics viz. statement of null hypothesis, interpretation of type I-error and p value, application of Chi-square , and t-test, and interpretation of Chi-square test based on calculated test value compared with table value were included in the questionnaire. The total items on descriptive and inferential Biostatistics were 19 each. In the present analysis results, 11 questions on Sampling were not included. Therefore, the total number of items in this study was 27.The assessment of knowledge on Biostatistics was performed by determining the frequencies of correct responses to each items of the questionnaire. The interpretation of the results was done in the form of percentage and proportion. Data were entered in Microsoft Excel 2007 and was analysed using SPSS version 16.

3 RESULTS

Out of total 73 students enrolled, 66 students had participated in this study. However, data was analyzed for 64 students as 2 students were excluded in the study because of the incompleteness in their responses. The response rate was 88%. As shown in Table 1, the number of male and females are the same in the study. The average age of the PG students was found to be 30 years (range 25 to 41) with a standard deviation of 4 year. The maximum number of respondents were in the age grouped of 29 to 32 (35.9%) followed by 25 to 28 old (34.4%). Majority of the students were from Clinical department (64.1%) followed by Para-clinical students (25.0%). Pre-clinical students constituted 10.9% of the study population. Table 1 also revealed that 43.8% of the students were from 1st year of their three year academic course followed by 3rd year students 31.8%. More than half of them (59.4%) had taken training course in Biostatistics/research and 40.6% had not taken training course in Biostatistics/research. Out of 64 medical PG students, 70.3% had attended and at least one paper (18.8%) was presented in the conference. Only one student had published more than 3 papers (1.6%) in the journal. Majority of them had not published any paper in the journal (87.5%).

The percentage of overall mean correct answer for 27 items was 45.72 with standard deviation of 11.81% (range 14.81 to 70.37%).More than 50% of the students answered the questions correctly in 9 questions out of 27 questions. The average number of correct answer respondents by the students was 12 in the study. Table 2 depicted the distribution of the number of students with correct answer of descriptive Biostatistics. There were 19 questions in this category. More than 50% students gave the answer correctly in 7 questions (ranging 53.1 to 81.1%) out of total 19 questions in descriptive Biostatistics. The maximum percentage on correct answer was on measure of dispersion (81.2%). Out of 64 students, 52 (81.1%) students were able to identify that standard deviation is a measure of variation. The minimum scored on correct answer was on the application of scatter diagram (12.5%) that was followed by the application of sub-divided bar diagram by illustrating in a data (17.2%).

Table 1: Characteristics of the study population

Characteristics	Frequency	Percent
Gender		
Male	32	50
Female	32	50
Age (years)		
25 - 28	22	34.4
29 - 32	23	35.9
33 - 36	15	23.4
37 – 41	4	6.2
Specialization		
Clinical	41	64.1
Para-clinical	16	25
Pre-clinical	7	10.9
Academic year		
1	28	43.8
2	16	25
3	20	31.8
Training on Biostatistics/ research attended		
Yes	38	59.4
No	26	40.6
Conference attended		
Yes	45	70.3
No	19	29.3
Paper presentation in conference		
1	12	18.8
2	2	14.1
More than 3	4	6.2
Nil	4	60.9
Paper publication in journal		
Nil	56	87.5
1	7	10.9
More than 3	1	1.6

Table 2 also revealed that out of 64 students, 13 (30.3%) students can identified the method used in correlation analysis and 22 (34.4%) understand that the range of correlation coefficient can't exceed more than 1. Forty two percent of the students understand the correct answer in example of continuous quantitative variable while 37 (57.8%) responded correctly in the ordinal scale of measurement. Forty two percent of the students answered correctly in the question related to source of data and 28 (43.8%) students understand the concept of measure of central tendency.

Table 2: Number of students with correct answer on descriptive biostatistics

Questions No.	Descriptive Statistics	Frequency	Percent
6	Graphical representation of data	8	12.5
9	Graphical representation of data	11	17.2
13	Application of median, mode	12	18.8
17	Identification of method of correlation analysis	13	20.3
11	Application of median, mode	21	32.8
19	Interpretation of correlation coefficient	22	34.4
7	Graphical representation of data	26	40.6
5	Source of data	27	42.2
1	Type of variable	27	42.2

Questions No.	Descriptive Statistics	Frequency	Percent
10	Concept of measure of central tendency	28	43.8
4	Source of data	28	43.8
20	Formula of mean	29	45.3
14	Application of median, mode	34	53.1
2	Scale of measurement	37	57.8
18	Example of measuring variability	42	65.6
3	Type of variable	47	73.4
21	computation of median	48	75
8	Graphical representation of data	50	78.1
12	Concept of measure of dispersion	52	81.2

Table 3 showed the correct response to each of the individual questions on inferential Biostatistics among the 64 students. Out of 64 respondents, the maximum percentage scored (75%) of correct answer was on statement of null hypothesis followed by interpretation of the test result when P-value less than and equal to 0.05 (64.7%). The minimum percentage scored on correct answer was in interpretation of the test value of Chi-square test compared with the table of the test (14, 21.9%) followed by the interpretation of P = 0.05 (19, 29.7%). As shown in Table 3, less than 50% of the PG students answered the question correctly in areas related to the interpretation of P less than 0.001, Type –I error and application of Chi-square test and t- test. Out of 8 questions related to inferential statistics more than 50 percent of the students can answered correctly in two questions viz. statement of null hypothesis and interpretation of test result when P-value is equal to 0.05.

Table 3: Number of students with correct answer on inferential Biostatistics

Questions no.	Inferential Statistics	Frequency	Percent
38	Interpretation of Chi-square test	14	21.9
37	Interpretation of P-value	19	29.7
36	Interpretation of P-value	23	35.9
15	Application of Chi-square test	24	37.5
16	Application of t-test	29	45.3
34	Interpretation of Type - I error	30	46.9
35	Interpretation of P-value	41	64.1
33	Statement of null hypothesis	48	75.0

4 DISCUSSIONS

In the present study, the percentage of overall mean correct answer was 47.5% with standard deviation of 11.54% (range 13.16 to 73.68%). Similar results were obtained in a study that the overall mean (SD) biostatistics knowledge score was 47.3% (18.50%; range 0-90) (Bookstaver et al., 2012).

In a cross- sectional study in the United States, it was found that the percentage of overall mean correct answers was 51.3% (SD: 17.3%, range, 22.2-88.9). The majority of the participants (66%) thought that Biostatistics knowledge is important for evidence-based practice (Alonaizan, 2013). In the present study, the response rate was 88% and the reliability of the tool was 0.597. A study conducted in Vadodara, Gujarat, India revealed that the level of biostatistics knowledge was moderate. The response rate of the study was 99.44% and the reliability of the tool was found to be 0.65 (Wadhwa et al., 2015).

In the present study, the number of male and females were found to be 32 each and the average age of the students was 30 year. In a survey study among pharmacy residents' knowledge of Biostatistics and the research study design conducted by Bookstaver et al. (2012) revealed that overall, respondents were predominantly female (74%) and younger than 30 years (81%). A study conducted by Windish et al. (2007) also revealed that sex was associated with a difference in scores. A study conducted in the United States revealed that participants who were younger than 30 years of age and those who had statistical education during residency were more likely to have higher mean score (Alonaizan, 2013). In the present study, more than half of them (59.4%) had taken training course in biostatistics/research and 58% students had correctly answered in the identification of continuous variable, 57.8 % were able to understand ordinal data, 37.5% and 45.3% of the students can identified the application of Chi-square and t-test respectively.

A study on medicine residents' understanding of the biostatistics and results in the medical literature revealed that nearly one-third of trainees indicated that they never received Biostatistics teaching at any point in their career. Residents with prior biostatistics training scored better than their counterparts. In their study 43.7% of students give the correct answer in the identification of contentious variable, 41.5% in ordinal data , 25.6% in Chi-square test, 58.1% in t-test, 58.8% and 50.2% can interpreted the meaning of p value and standard deviation respectively. Twelve percent of the respondents can interpret the 95% CI and statistical significance (Windish et al., 2007). In another study conducted at Pondicherry, south India revealed that students knowledge were not satisfactory in the area of null hypothesis (39.5%), measurement of scale (16.3%), but had a satisfactory level in the question of type of variable (81.4%), graphical representation of data (55.8%) (Majumdar et al., 2015).

5 LIMITATIONS OF THE STUDY

The information presented in this paper represents only a small proportion of the medical students. It is not sufficient enough to generalize the finding to other medical college. The study did not perform any statistical significant conclusion in all the selected variables. Less than half of the respondents had correct knowledge on biostatistics in this study may be due the reasons that most of the students had passed out their undergraduate course from different medical colleges. There is no uniform syllabus and allotment of time for classes for the subject biostatistics in all medical institutes in India. Some students did their undergraduate in abroad medical institute therefore the researcher had no ideas whether there was biostatistics in their medical undergraduate course.

6 CONCLUSIONS

The finding of the present study were similar with the results of previous studies The percentage of overall mean correct answer was found to be less than half of the respondents. More than half of the students answered the questions correctly in 9 questions. Majority of the students had correct knowledge on graphs, measure of central tendency, and measure of dispersion and scale of measurement. But the in-depth knowledge of the same topic was unable to understand by the students like scatter diagram and subdivided bar diagram. A large number of students were not able to interpret the p value, Type –I error, application of Chi-square test, and t-test. But most of the students understood the meaning of the statement of null hypothesis and statistically significance. It was also observed that majority of medical PG students had difficulty in understanding the basic concepts of inferential biostatistics as compared with descriptive biostatistics. Emphasis should be given while teaching that most of the important terminology used in inferential biostatistics should be clearly explained with understanding to the students. Orientation programme on research methodology and biostatistics should be conducted in the early period of their three year professional course that could help them to become an independent and a successful researcher in their profession.

7 RECOMMENDATIONS

Our medical education unit should take a step towards the education of research methodology and biostatistics. This subject should be taught in graduation within a specific syllabus and time for both medical undergraduate and postgraduate students so that they would not face problem in submitting their dissertation during their PG course. Therefore, there is a need for incorporating biostatistics as a subject in the undergraduate and postgraduate curriculum of medical education in India.

ACKNOWLEDGEMENTS

The author gratefully acknowledges the PG students for their participation in this study and thanks the staff of HRD department for providing the list of all medical PG students of SMIMS. The author acknowledges the help and support of Dr. Asutosh Kumar Dixit for reviewing the manuscript of the study.

REFERENCES

Alonaizan, F. (2013). Assessment of the knowledge of biostatistics and the attitude toward evidence based practice among endodontic residents in the United States: A Cross-sectional Survey (Master's thesis). Retrieved from http://digitallibrary.usc.edu/cdm/ref/collection/p15799coll3/id/308957.

Altman, D. G. (1990). *Practical Statistics for Medical Research*. CRC Press.

Bookstaver, P. B., Miller, A. D., Felder, T. M., Tice, D. L., Norris, L. B., & Sutton, S. S. (2012). Assessing pharmacy residents' knowledge of biostatistics and research study design. *Annals of Pharmacotherapy*, 46(7-8), 991- 999.

The Edinburg Declaration. (1998*). In Proceeding of the World Conference on Medical Education: 1998*, Edinburg World Medical Association.

Giri, P. A., Bangal, V. B., & Phalke, D. B. (2014). Knowledge, attitude and practices towards medical research amongst the postgraduate students of Pravara Institute of Medical Sciences University of Central India. *Journal of family medicine and primary care*, 3(1), 22-24.

Mazumdar, A., Kumar, S. K., & Roy, G. (2015). Knowledge, attitude and perception of medical students regarding community oriented research. *National Journal of Community Medicine*, 6(2), 97-102.

Peat, J., & Barton, B. (2008). *Medical Statistics: A Guide to Data Analysis and Critical Appraisal*. John Wiley & Sons.

Wadhwa, M., Kalyan, P., & Kalantharakath,T. (2015). Knowledge and attitude of medical and dental postgraduate students toward practice of biostatistics. *Journal of Postgraduate Medicine, Education and Research*, 49(1), 1-4.

Windish, M. D., Huot, S. J., & Green, M. L. (2007). Medicine residents' understanding of the biostatistics and results in the medical literature. *JAMA*, 298(9), 1010-1022.

Factors Associated with Agreement on Discriminatory Statements toward People Living with HIV in Participants Presenting for HIV Testing in Chiang Mai, Thailand

Chanapat Pateekhum[1,2*], Anouar Nechba[2], Wanlee Kongnim[2], Nirattiya Jaisieng[2], Woottichai Khamduang[3], Wasna Sirirungsi[3] and Patrinee Traisathit[2,4]

[1]Master's Degree Program in Applied Statistics, Department of Statistics, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand
*Corresponding Email: chanapat_p@cmu.ac.th
[2]Prevention and treatment of HIV infection and virus-associated cancers in south East Asia (PHPT), Chiang Mai, Thailand
Email: a.nechba@gmail.com
Email: wkongnim@gmail.com
Email: nirattiya.jaisieng@phpt.org
[3]Faculty of Associated Medical Sciences, Chiang Mai University, Chiang Mai, Thailand
Email: Woottichai.khamduang@phpt.org
Email: wasna.s@cmu.ac.th
[4]Center of Excellence in Bioresources for Agriculture, Industry and Medicine, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand
Email: patrinee.t@cmu.ac.th

ABSTRACT

HIV-related stigma and discrimination is a known obstacle to accessing medical care and participating in the community. The purpose of this study was to identify factors associated with agreement on discriminatory statements toward people living with HIV (PLHIV). The participants were consenting adults who had never tested positive for HIV. They were tested anonymously and free of charge for HIV, hepatitis B, hepatitis C, and syphilis between October 2015 and April 2016. Data were collected using a questionnaire completed by participants on Tablet PC to obtain information about socio-demographic characteristics, HIV knowledge (HIV-KQ-18 score), risk behaviors regarding HIV, anxiety score and depression score (Patient Health Questionnaire-4). Participants rated 8 questions using Subscale: "Discrimination" with a 4-point Likert scale used by Genberg and coll. The sum of the scores of all 8 questions provided an overall score, which was dichotomized according to the third quartile. We performed a binary logistic regression analysis to identify factors associated with agreement on discriminatory statements. We used multivariate imputation using a chained equation for missing data. From 494 participants, 8% were hill tribe members, and 71% had never previously tested for HIV. The median age of 24.0 years (interquartile range: 21.5-32.8). The results indicate that agreement on discriminatory statements toward PLHIV was associated with not being a hill tribe member [adjusted odds ratio (aOR) 6.11; 95% confidence interval (CI): 1.14-32.86], received money or other benefits in exchange for sex [aOR 5.17; 95% CI: 1.21-22.06], HIV knowledge above the median score [aOR 1.79; 95% CI: 1.08-2.97], and injection drug use [aOR 4.64; 95% CI: 1.33-16.26]. Since participants were knowledgeable about HIV seemed to have a negative attitude toward PLHIV, educating on HIV might be insufficient to reduce the negative attitude. Therefore, our study suggests educating on HIV simultaneously with finding interventions to improve the society's attitudes.

Keywords: people living with HIV; discriminatory statements; Thailand

1 INTRODUCTION

In Asia and the Pacific region, Thailand is one of six countries with the largest number of people living with HIV (PLHIV) (UNAIDS, 2014a), it has been estimated that 450,000 people of more than 15 years of age were living with HIV in 2016 and the number of new PLHIV decreased from 13,000 in 2010 to 6,400 in 2016 (UNAIDS, 2017b). However, the estimated number of Thai PLHIV from hospitals was only part of the total number since the remaining PLHIV are unaware of their HIV status because they have not sought the HIV treatment, it has been found that the majority of Thai PLHIV only present for HIV treatment when they are in an advanced stage of HIV infection (The Asian Epidemic Model, 2008). The reasons mentioned for this include fear of disclosure of their HIV status, and denial of compliance with medical advice relate to HIV-related stigma and discrimination (UNAIDS, 2017c), which can affect the efficacy of HIV treatment and HIV transmission (Genberg et al., 2008).

Discrimination as a result of stigma refers to the exclusion and restriction of access of a particular group to family, community activities (UNAIDS, 2005), and medical care, such as PLHIV receiving unfair treatment because of their HIV status. HIV-related discrimination is "often based on beliefs related to stigmatizing attitudes, their behaviors and sex" (UNAIDS, 2014b). In addition, regarding the chance of HIV infection, other people might also discrimination against those engaged in risky sexual behavior in the same way as PLHIV (WHO, 2016). In a study in Hong Kong, almost all of the people had a negative attitude toward PLHIV and agreed that PLHIV deserved punishment (Lau & Tsui, 2005). The reported percentages of PLHIV subjected to discrimination based on HIV status in Thailand are 32% losing their job, 15% unable to rent accommodation, 26% social marginalization, 12% family exclusion, and 20% denied medical care (UNAIDS, 2011). UNAIDS (UNAIDS, 2016) considered that the key population consists of commercial sex workers (CSW), men who have sex with men (MSM), transgender people, and injection drug users (IDU). They were vulnerable to HIV infection, had restricted medical care access, and were also stigmatized by society (UNAIDS, 2017a).

Previous studies found discriminatory attitudes toward PLHIV in people with a poor level of education, low wealth index, and ethnicity (Dahlui et al., 2015; Wong, 2013). Other studies also reported that medical providers with poor level of basic and in-depth knowledge on HIV transmission and prevention, unreasonable fear of HIV infection, and increasing age had discriminatory attitudes toward PLHIV (Feyissa et al., 2012; Harapan et al., 2013). Gender could be associated with a discriminatory attitude, as was found in a study in Brazil which reported that females had a stronger discriminatory attitude (Garcia & Koyama, 2008). However, a study in Thailand found that strong agreement on discriminatory statements was reported among people lacking in knowledge about antiretrovirals (ARVs) (Genberg et al., 2009). The authors mentioned an interesting trend when exploring factors associated with agreement on discriminatory statements.

To our knowledge, a study of factors associated with agreement on discriminatory statements concerning PLHIV in Thailand are lacking. Consequently, our study was aimed at identifying factors associated with agreement on discriminatory statements toward PLHIV in participants presenting in a test for HIV and other sexually transmitted diseases in Chiang Mai, Thailand.

2 METHODS

Consenting adults (age ≥ 18 years) residing in Thailand and able to communicate with the counselor were included in this study. They were tested anonymously and free of charge for HIV, hepatitis B, hepatitis C, and syphilis between October 2015 and April 2016 as part of the Napneung project (ClinicalTrial.gov: NCT02752152) in Chiang Mai, Thailand. Participants who were previously HIV-positive aware, responded with the same answer to all of the questions, did not respond to any questions, or foreigners were excluded from the analysis.

The outcome of interest was the measure of the agreement on discriminatory statements toward PLHIV resulting from the answers to 8 questions with a 4-point Likert scale of the "Discrimination" subscale (Genberg et al., 2008). The overall score was calculated as the sum of the scores to each question graded 1 to 4 according to the strength of the participant's agreement on the discriminatory statements. The overall score was dichotomized to the third quartile.

Each participant was invited to respond to a questionnaire on a Tablet PC that included basic socio-demographic characteristics, HIV knowledge (HIV-KQ-18 score) (Carey & Schroder, 2002), his/her attitude related to discriminatory statements toward PLHIV, risk behaviors regarding HIV, place of current living, place of birth, and signs of depression before received counseling consisting of an anxiety score ≥ 3 and a depression score ≥ 3 (calculated from answer to Patient Health Questionnaire-4) (Kroenke et al., 2009).

Independent variables which included in the analysis for association with agreement on discriminatory statements were consisted of age, education, HIV knowledge score (before receiving counseling), income (per month), occupation, hill tribe member, had a regular partner, living alone, injection drug use, substance drugs use, received money or other benefits in exchange for sex, gave money or other benefits in exchange for sex, always used a condom in previous 3 months, had more than 1 sexual partner in the previous 3 months, amount of time to ever get previously HIV tested, sexual orientation, had anxiety, had depression, place of current living, and place of birth.

The questionnaires and the participant consent documents were reviewed and approved by the Ethics Committee of the Faculty of Associated Medical Sciences, Chiang Mai University. Participants provided informed consent before entry in the Napneung project.

The characteristics of the study population were presented as medians and interquartile ranges (IQRs) for continuous variables and as frequencies and percentages for categorical variables. Fisher's exact test was used to compare the differences of proportion for categorical variables. If R x C table contingency table can be classified into 2 x 2 table (T), the 2-sided p-value of the test would be the sum of probability of all possible tables which can be computed using Equation (1) as follows (Freeman & Halton, 1951):

$$p = \sum_{T=1}^{A} Pr(T) \quad (1)$$

Where Pr(T) is the probability of each classified table and A is the number of all possible tables.

$$Pr(T) = \frac{\prod_{i=1}^{r} n_{i\cdot}! \prod_{j=1}^{c} n_{\cdot j}!}{n! \prod_{i=1}^{r} \prod_{j=1}^{c} n_{ij}!}$$

Where i = number of row; i = 1, 2, … , r
j = number of column; j = 1, 2, … , c
n_{ij} = observed number in row i and column j
$n_{i\cdot}$ = total observed number of row i
$n_{\cdot j}$ = total observed number of column j
n = total observed number

The proportion of each category will be different when p less than the significant level.

Factors associated with agreement on discriminatory statements toward PLHIV were identified using a binary logistic regression analysis. The logit transformation of logistic regression model (Kleinbaum & Klein, 2010) is given by

$$\log\left[\frac{P}{1-P}\right] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \quad (2)$$

Where P = probability statement determined agreement on discriminatory statements, as either 1 if "agreeing on" or 0 if "not agreeing on"
$\frac{P}{1-P}$ = the ratio of the probability that agreeing on discriminatory statements will occur over the probability that not agreeing on
β_0 = intercept coefficient
β_k = logistic regression coefficients of the k^{th} independent variable
X_k = the k^{th} independent variable
k = 1,2,…,p

All variables with a p-value < 0.25 in the univariable analysis were included in the multivariable analysis. Afterward, a backward elimination was performed to identify the significant factors.

Missing data should be imputed when the percentage of missing data is 10% to 30% (Marshall et al., 2010; Scheffer, 2002). Hence, missing values were imputed using multivariate imputation by chained equation (MICE) (Raghunathan et al., 2001; van Buuren, 2007) based on the predicted missing data using linear regression for continuous variables and binary or ordinal logistic regression for categorical variables. First, we classified the assumption of the missing data as missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). Second, testing the assumption found that the missing data was classified as MAR. The probability of the missing data depended on the observed data (Mislevy, 1991). Consequently, MICE was performed to address the missing data and we produced the imputed datasets to 1 dataset by randomizing the imputed dataset.

All reported p-values were 2-sided and p-value < 0.05 were considered as statistically significant. The analyses were conducted using Stata version 14 (StataCorp LP, Texas, USA).

3 RESULTS

From October 2015 to April 2016, 510 participants presented for testing. Of the 510 participants, 494 (96.9%) were included in the analysis (Figure 1).

Of the 494 participants, their median age was 24.0 years (IQR: 21.5-32.8), 52% were male, 71% had attained a higher education level, 50% were students, 38% had no income, 8% were hill tribe members, 64% lived in an urban environment, 71% had never

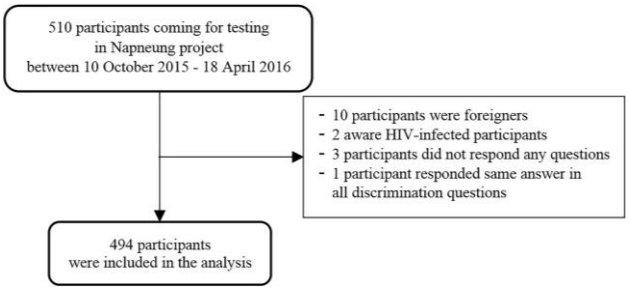


Figure 1: Population disposition

previously tested for HIV, 18% were MSM, 20% had suffered from anxiety in the previous 2 weeks, and 16% had been depressed in the previous 2 weeks. In terms of risk behavior regarding HIV infected, 2% used injection drug, 23% used substance drugs, 35% had had sexual activity without a condom in the previous 3 months, and 21% had had more than 1 sexual partner in the previous 3 months. Their median HIV knowledge score was 12 (IQR: 8-15) (Table 1).

The percentage of participants agreeing on the discriminatory statements was 17% and the percentage of missing data varied from 0.4% to 17.2%. Proportion of people with HIV knowledge score below the median and proportion of IDU were significantly different between the agreeing on and not agreeing on discriminatory statements groups (Table 2).

The univariable and multivariable analysis show that agreement on discriminatory statements toward PLHIV was associated with not being a hill tribe member [adjusted odds ratio (aOR) 6.11; 95% confidence interval (CI): 1.14-32.86, p=0.04], received money or other benefits in exchange for sex [aOR 5.17; 95% CI: 1.21-22.06, p=0.03], HIV knowledge above the median

score [aOR 1.79; 95% CI: 1.08-2.97, p=0.02], and injection drug use [aOR 4.64; 95% CI: 1.33-16.26, p=0.02]. However, age, living alone, depression, place of current living, and place of birth were not associated with agreement on discriminatory statements (Table 2).

Table 1: The characteristics of the study population

Characteristics	Median (IQR) or number (%)	Characteristics	Median (IQR) or number (%)
Sex at birth (male)	257 (52)	Hill tribe member (n=487)	37 (8)
Gender		Had a regular partner (n=492)	264 (54)
Male	245 (50)	Sexual orientation (n=486)	
Female	232 (47)	Straight	364 (75)
Male to female transgender	11 (2)	Gay	77 (16)
Female to male transgender	4 (1)	Lesbian	10 (2)
Not sure	2 (0)	Bisexual	22 (4)
Age (years)	24.0 (21.5 – 32.8)	Not sure	13 (3)
Age		Living alone (n=490)	127 (26)
≤ 22	149 (30)	Injection drug use (n=491)	12 (2)
> 22-24	98 (20)	Substance drugs use (n=490)	111 (23)
> 24-33	127 (26)	Received money or other benefits in exchange for sex (n=486)	10 (2)
> 33	120 (24)	Gave money or other benefits in exchange for sex (n=486)	22 (5)
Education (n=492)		Always used a condom in the previous 3 months (n=483)	
Primary	36 (7)	No	171 (35)
Secondary	110 (22)	Yes	125 (26)
Higher	346 (71)	No intercourse	187 (39)
HIV knowledge score out of 18 (before receiving counseling)	12 (8 - 15)	Had more than 1 sexual partner in the previous 3 months	106 (21)
Low HIV knowledge (scores below median)	220 (44)	Amount of time to ever get previously HIV tested (n=487)	
Income (per month) (n=492)		Never	348 (71)
No income	187 (38)	Once or more times	139 (29)
1-5,000 Bahts	53 (11)	Men who have sex with men	87 (18)
5,001-10,000 Bahts	79 (16)	Had anxiety (n=412)	82 (20)
10,001-15,000 Bahts	72 (15)	Had depression (n=409)	64 (16)
> 15000 Bahts	101 (20)	Place of current living (n=486)	
Occupation (n=491)		Suburb	174 (36)
Student	248 (50)	Urban	312 (64)
Employed	145 (30)	Place of birth (n=460)	
Self-employed	68 (14)	Suburb	249 (54)
Unemployed or housewife	26 (5)	Urban	211 (46)
Retired	4 (1)		

IQR, interquartile range.

Table 2: Factors associated with the agreement on discriminatory statements toward PLHIV

Characteristics	Agreement/ Total (%)	p-value[b]	Univariable analysis		Multivariable analysis	
			OR (95% CI)	p-value	aOR (95% CI)	p-value
Age		0.19		**0.19**		
≤ 22	24/149 (16.1)		1.00			
> 22-24	17/98 (17.3)		1.09 (0.55-2.16)			
> 24-33	29/127 (22.8)		1.54 (0.84-2.81)			
> 33	15/120 (12.5)		0.74 (0.37-1.49)			
Education		1.00		0.87		
Primary	6/37 (16.2)		1.00			
Above primary	79/457 (17.3)		1.08 (0.44-2.67)			
HIV knowledge score (before receiving counseling)		**0.02**		**0.02**		**0.02**
Below the median	28/220 (12.7)		1.00		1.00	
Above the median	57/274 (20.8)		1.80 (1.10-2.95)		1.79 (1.08-2.97)	
Income (per month)		0.61		0.60		
No income	29/187 (15.5)		1.00			
≤ 10,000 Bahts	22/132 (16.7)		1.09 (0.59-2.00)			
> 10,000 Bahts	34/175 (19.4)		1.31 (0.76-2.27)			
Occupation		0.28		0.30		
Student	45/248 (18.1)		1.00			
Employed	38/215 (17.7)		0.97 (0.60-1.56)			
Unemployed or housewife	2/31 (6.4)		0.31 (0.07-1.35)			
Hill tribe member		0.07		**0.06**		**0.04**
No	83/457 (18.2)		1.00		1.00	
Yes	2/37 (5.4)		0.26 (0.06-1.09)		0.16 (0.03-0.88)	
Had a regular partner		0.55		0.51		
No	42/228 (18.4)		1.00			
Yes	43/266 (16.2)		0.85 (0.53-1.36)			
Living alone		0.13		**0.10**		
No	57/366 (15.6)		1.00			
Yes	28/128 (21.9)		1.52 (0.92-2.52)			
Injection drug use		**0.04**		**0.03**		**0.02**
No	80/482 (16.6)		1.00		1.00	
Yes	5/12 (41.7)		3.59 (1.11-11.59)		4.64 (1.33-16.26)	
Substance drugs use		0.48		0.44		
No	63/382 (16.5)		1.00			
Yes	22/112 (19.6)		1.24 (0.72-2.12)			

Characteristics	Agreement/ Total (%)	p-value[b]	Univariable analysis		Multivariable analysis	
			OR (95% CI)	p-value	aOR (95% CI)	p-value
Received money or other benefits in exchange for sex		0.08		**0.07**		**0.03**
No	81/484 (16.7)		1.00		1.00	
Yes	4/10 (40.0)		3.32 (0.92-12.02)		5.17 (1.21-22.06)	
Gave money or other benefits in exchange for sex		0.57		0.56		
No	80/471 (17.0)		1.00			
Yes	5/23 (21.7)		1.36 (0.49-3.76)			
Always used a condom in previous 3 months		0.81		0.80		
No	28/178 (15.7)		1.00			
Yes	23/126 (18.2)		1.20 (0.65-2.19)			
No intercourse	34/190 (17.9)		1.17 (0.67-2.02)			
Had more than 1 sexual partner in the previous 3 months		0.39		0.35		
No	70/388 (18.0)		1.00			
Yes	15/106 (14.2)		0.75 (0.41-1.37)			
Amount of time to ever get previously HIV tested		0.51		0.47		
Never	63/350 (18.0)		1.00			
Once or more times	22/144 (15.3)		0.82 (0.48-1.40)			
Sexual orientation		0.44		0.46		
Men who have never had sex with men	31/170 (18.2)		1.00			
Men who have sex with men	18/87 (20.7)		1.17 (0.61-2.24)			
Female	36/237 (15.2)		0.80 (0.47-1.36)			
Had anxiety		1.00		0.96		
No	69/400 (17.2)		1.00			
Yes	16/94 (17.0)		0.98 (0.54-1.78)			
Had depression		0.31		**0.23**		
No	76/421 (18.1)		1.00			
Yes	9/73 (12.3)		0.64 (0.30-1.34)			
Place of current living		0.11		**0.08**		
Suburb	24/180 (13.3)		1.00			
Urban	61/314 (19.4)		1.57 (0.94-2.62)			
Place of birth		0.15		**0.14**		
Suburb	41/274 (15.0)		1.00			
Urban	44/220 (20.0)		1.42(0.89-2.27)			

OR, odds ratio.
[b] Fisher's exact test.

4 DISCUSSIONS

Our study reported the results of identification factors for agreement on discriminatory statements toward PLHIV in Thailand. We found that participants were knowledgeable about HIV, they were more to agree with discriminatory statements toward PLHIV. This is consistent with a study in Kenya, from which it can be deduced that although HIV knowledge had increased, HIV-caregivers and HIV-positive children and adolescents recognized that the community still had a negative attitude toward them and a

misunderstanding about HIV (McHenry et al., 2017). A study in Zimbabwe and South Africa also reported that people who had knowledge about ARVs were more to agree with discriminatory statements (Genberg et al., 2009). However, the finding on HIV knowledge in our study is inconsistent with a study in Indonesia reported that high HIV knowledge was related to low level of discriminatory attitude toward PLHIV (Harapan et al., 2013). In addition, a study in Thailand reported that there was strong agreement on discriminatory statements among people lacking in knowledge about ARVs (Genberg et al., 2009).

Previous studies have reported that hill tribe members lacked HIV knowledge (Apidechkul, 2011; Kunstadter, 2013). Their understanding of HIV was different due to their beliefs, HIV knowledge, and ability to reach medical services in Thailand (Kunstadter, 2013). HIV-related stigma, which could lead to discriminatory attitudes, might be resulted from the fear of being infected with HIV and a misunderstanding about HIV transmission (UNAIDS, 2005). This finding is inconsistent with our study which found that non-hill tribe members were more to agree with discriminatory statements toward PLHIV since most of them (71%) were living in an urban environment. In addition, the finding in our study is consistent with a study in Vietnam which found that people living in an urban environment had a more negative attitude toward PLHIV than those living rurally (Tuan et al., 2008).

A study in Russia reported that most of female CSW sympathized with PLHIV (King et al., 2013). This is inconsistent with our study which found that CSW and IDU were more to agree with discriminatory statements toward PLHIV. However, we did not find any studies on IDU. Some CSW had faced the discriminatory attitudes from medical providers (NSWP, 2017). These prejudice attitudes toward CSW and IDU might be results from the effects of stereotypical attitudes on HIV which might lead to their negative attitudes on PLHIV. Evaluation the effects of prejudice attitudes of other people toward CSW and IDU on their attitudes toward PLHIV is interesting in further study.

The questionnaire of attitude related to discriminatory statements toward PLHIV using in our study might not be reflected only participants' attitude, but it might also be mixed with their perception. Some participants might answer the question according to their perception from observation or trend in their societies which is inconsistent with their own attitude. More agreement on discriminatory statements might be also results from perception. The negative perception might lead to discriminant or negative attitude toward PLHIV even if people were knowledgeable about HIV. We supposed that only educating on HIV might insufficient to reduce the negative attitude toward PLHIV. Therefore, our study suggests educating HIV knowledge simultaneously with find interventions to improve the society's attitudes related to the stigma and discrimination.

ACKNOWLEDGEMENTS

We would like to thank the participants in the Napneung project. We are also grateful to Dr. Gonzague Jourdain, who are principal investigator, provide insight advice, and comments for improvement of manuscript, Natthanidnan Sricharoen, and Kanchana Than-in-at, who performed data management, and Nicolas Salvadori, who provided statistical advice, Dr. Luc Decker, who are information technologists. We are also thank the physicians, virologists, a coordinator/project manager, nurses and medical technologists, clinical research assistants, administrative/logistic team, and advisory board: Sumet Ongwandee, Thitipong Sangyoung, Nakorn Premsri, Maleerat Vandriesten, Suchada Chaivoot, Sombat Thanprasertsuk, Christine Rouzioux, Eric Fleutelot, Pongthorn Chanlearn, Mukta Sharma, Ratchadet Reaukhamfu, Brahm Press, Monthira Metha, Boonium Wongjaikham, and Chutima Jaruwat from the Napneung project. We are also thank Asst. Prof. Dr. Sukon Prasitwattanaseree and Asst. Prof. Dr. Bandhita Plubin from Chiang Mai University. The Napneung project is funded by Expertise France (Initiative 5%).Finally, Chanapat Pateekhum received the funding from Graduate School, Chiang Mai University, Chiang Mai, Thailand.

REFERENCES

Apidechkul, T. (2011). P2-358 HIV/AIDS status in a hill-tribe population, Thailand. *Journal of Epidemiology & Community Health*, *65*, A322–A322.

Carey, M. P., & Schroder, K. E. E. (2002). Development and psychometric evaluation of the brief HIV knowledge questionnaire. *AIDS Education and Prevention : Official Publication of the International Society for AIDS Education*, *14*(2), 172–182. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2423729/

Dahlui, M., Azahar, N., Bulgiba, A., Zaki, R., Oche, O. M., Adekunjo, F. O., & Chinna, K. (2015). HIV/AIDS related stigma and discrimination against PLWHA in Nigerian population. *PLOS ONE*, *10*(12), e0143749. https://doi.org/10.1371/journal.pone.0143749

Feyissa, G. T., Abebe, L., Girma, E., & Woldie, M. (2012). Stigma and discrimination against people living with HIV by healthcare providers, southwest Ethiopia. *BMC Public Health*, *12*, 522.

Freeman, G. H., & Halton, J. H. (1951). Note on an exact treatment of contingency, goodness of fit and other problems of significance on JSTOR. *Biometrika*, *38*, 141–149. Retrieved from https://www.jstor.org/stable/2332323

Garcia, S., & Koyama, M. A. H. (2008). Stigma, discrimination and HIV/AIDS in the Brazilian context, 1998 and 2005. *Revista de Saúde Pública*, *42*, 72–83.

Genberg, B. L., Hlavka, Z., Konda, K. A., Maman, S., Chariyalertsak, S., Chingono, A., … Celentano, D. D. (2009). A comparison of HIV/AIDS-related stigma in four countries: Negative attitudes and perceived acts of discrimination towards people living with HIV/AIDS. *Social Science & Medicine (1982)*, *68*(12), 2279–2287.

Genberg, B. L., Kawichai, S., Chingono, A., Sendah, M., Chariyalertsak, S., Konda, K. A., & Celentano, D. D. (2008). Assessing HIV/AIDS stigma and discrimination in developing countries. *AIDS and Behavior*, *12*(5), 772–780.

Harapan, H., Feramuhawan, S., Kurniawan, H., Anwar, S., Andalas, M., & Hossain, M. B. (2013). HIV-related stigma and discrimination: A study of health care workers in Banda Aceh, Indonesia. *Medical Journal of Indonesia*, *22*(1), 22–29.

King, E. J., Maman, S., Bowling, J. M., Moracco, K. E., & Dudina, V. (2013). The influence of stigma and discrimination on female sex workers' access to HIV services in St. Petersburg, Russia. *AIDS and Behavior*, *17*(8). https://doi.org/10.1007/s10461-013-0447-7

Kleinbaum, D. G., & Klein, M. (2010). *Logistic regression: A self-learning text* (3rd ed.). New York: Springer-Verlag. Retrieved from //www.springer.com/gp/book/9781441917416

Kroenke, K., Spitzer, R. L., Williams, J. B. W., & Löwe, B. (2009). An ultra-brief screening scale for anxiety and depression: The PHQ-4. *Psychosomatics*, *50*(6), 613–621.

Kunstadter, P. (2013). Ethnicity, socioeconomic characteristics and knowledge, beliefs and attitudes about HIV among Yunnanese Chinese, Hmong, Lahu and Northern Thai in a north-western Thailand border district. *Culture, Health & Sexuality*, *15*, S383-400.

Lau, J. T. F., & Tsui, H. Y. (2005). Discriminatory attitudes towards people living with HIV/AIDS and associated factors: A population based study in the Chinese general population. *Sexually Transmitted Infections*, *81*(2), 113–119.

Marshall, A., Altman, D. G., Royston, P., & Holder, R. L. (2010). Comparison of techniques for handling missing covariate data within prognostic modelling studies: A simulation study. *BMC Medical Research Methodology*, *10*, 7.

McHenry, M. S., Nyandiko, W. M., Scanlon, M. L., Fischer, L. J., McAteer, C. I., Aluoch, J., … Vreeman, R. C. (2017). HIV stigma: Perspectives from Kenyan child caregivers and adolescents living with HIV. *Journal of the International Association of Providers of AIDS Care*, *16*(3), 215–225.

Mislevy, R. J. (1991). Review of *review of statistical analysis with missing data*, by D. B. Rubin & R. J. A. Little. *Journal of Educational Statistics*, *16*(2), 150–155.

NSWP. (2017, November 29). Briefing paper: The meaningful involvement of sex workers in the development of health services aimed at them [Text]. Retrieved August 25, 2018,

from http://www.nswp.org/resource/briefing-paper-the-meaningful-involvement-sex-workers-the-development-health-services-aimed

Raghunathan, T. E., Lepkowski, J. M., Hoewyk, J. V., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*(1), 85–95.

Scheffer, J. (2002). Dealing with missing data. Retrieved from https://mro.massey.ac.nz/handle/10179/4355

Analysis and Advocacy Project and Thai Working Group on HIV/AIDS Projections. (2008). *The Asian Epidemic Model (AEM) projections for HIV/AIDS in Thailand: 2005-2025 | UNESCO HIV and health education clearinghouse*. Retrieved from https://hivhealthclearinghouse.unesco.org/library/documents/asian-epidemic-model-aem-projections-hivaids-thailand-2005-2025

Tuan, N. A., Ha, N. T. T., Diep, V. T. B., Thang, P. H., Long, N. T., Huong, P. T. T., … Hien, N. T. (2008). Household survey in two provinces in Viet Nam estimates HIV prevalence in an urban and a rural population. *AIDS Research and Human Retroviruses*, *24*(8), 1017–1026.

UNAIDS. (2005). *HIV-related stigma, discrimination and human rights violations*. Retrieved from http://www.unaids.org/en/resources/documents/2005/20051005_jc999-humrightsviol_en.pdf

UNAIDS. (2011). *People living with HIV stigma index: Asia Pacific regional analysis 2011*. Retrieved from http://www.unaids.org/en/resources/documents/2011/20110829_PLHIVStigmaIndex

UNAIDS. (2014a). *Gap report*. Retrieved from http://www.unaids.org/en/resources/campaigns/2014/2014gapreport/gapreport

UNAIDS. (2014b). *Reduction of HIV-related stigma and discrimination*. Retrieved from http://www.unaids.org/en/resources/documents/2014/ReductionofHIV-relatedstigmaanddiscrimination

UNAIDS. (2016). Key population groups, including gay men and other men who have sex with men, sex workers, transgender people and people who inject drugs. Retrieved June 5, 2018, from http://www.unaids.org/en/topic/key-populations

UNAIDS. (2017a). *Confronting discrimination*. Retrieved from http://www.unaids.org/en/resources/documents/2017/confronting-discrimination

UNAIDS. (2017b). UNAIDS data 2017. Retrieved August 7, 2018, from http://www.unaids.org/en/resources/documents/2017/2017_data_book

UNAIDS. (2017c). UNAIDS warns that HIV-related stigma and discrimination is preventing people from accessing HIV services. Retrieved August 9, 2018, from http://www.unaids.org/en/resources/presscentre/pressreleaseandstatementarchive/2017/october/20171002_confronting-discrimination

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*(3), 219–242.

WHO. (2016). Striving for zero discrimination in health care. Retrieved June 5, 2018, from http://www.who.int/mediacentre/commentaries/zero-discrimination-day/en/

Wong, L. P. (2013). Prevalence and factors associated with HIV/AIDS-related stigma and discriminatory attitudes: A cross-sectional nationwide study. *Preventive Medicine*, *57*, S60-63.

Data mining for knowledge extraction from violent incidents in Thailand's deep South

Bunjira Makond[1,2]*

[1]Faculty of Commerce and Management, Prince of Songkla University, Trang campus, Trang 92000, Thailand
[2]Centre of Excellence in Mathematics, Commission on Higher Education, Bangkok, 10400 Thailand
*Corresponding Email: bunjira.m@psu.ac.th

ABSTRACT

Nowadays, due to advances in computer technology huge amounts of data can be efficient collected and stored on a reasonable budget and many domains have relevant data stored in databases. It is very difficult and cumbersome to manually process such enormous amounts of data without the using computers. This study aims to extract the hidden patterns, relationships and knowledge from the Deep South Coordination Centre database collected between 2004 and the beginning of January 2016 using data mining techniques. Using the WEKA tool, this study has performed experiments on 31 dummy variables of a data set containing 21,424 violent cases. Two data mining techniques, clustering and association rule, were used on the data. Cluster analysis assigned the data into two clusters (i.e. "non-physical injury" and "physical injury"); meanwhile, ten rules were generated using the association rule technique. The useful information gained will provide decision makers with relevant information for devising suitable prevention and intervention efforts to address risk factors for violence.

Keywords: data mining; violence; clustering; association rule

1 INTRODUCTION

Violence has been recognized as a global problem and public concern because of its significant impact on the health and lives of all humans. Yearly, over a million people are killed and many additional people suffer non-fatal injuries because of self-inflicted, interpersonal, or collective violence. Generally, violence is among the leading causes of death globally for people aged between 15-44 years (Dahlberg & Krug, 2006). In the deep south of Thailand violence also results in a public health problem which has direct and indirect effects on victims. The statistics show that, between January 2004 and March 2013, around 13,000 violent events occurred, leading to 15,574 victims, which includes 5,614 people who died and 9,960 people who were injured. Simultaneously, people involved can be affected by mental conditions and need psychological counseling because they are unable to cope with the situation that they are confronted with. Moreover, violence leads to economic consequences, for example, decreases job creation and increased poverty (Makond, 2018).

Nevertheless, in regard to the public health consequences on people, violence is predictable and preventable. Successes in preventing violence can be achieved through reliable surveillance and monitoring systems, generally referred to as the systematic collection, analysis, interpretation, and dissemination of data on violence, and its determinants (Schuurman et al., 2015). The manual implementation of traditional statistical approaches to the data has prompted the development of effective prevention programs and policies over the past several decades. Subsequently, the data about victims and perpetrators has been inspected to detect social, environmental, and risk factors for violence. Therefore, relevant information is available to devise suitable prevention and intervention efforts to address the risk factors (Comstock et al., 2005; Espinosa et al., 2008)

Nowadays, due to the advance computer technology, it is possible to efficiently collect and store huge amounts of data with a reasonable budget and many domains already have relevant data stored in databases. It is very difficult and cumbersome to manually process such an enormous amount of data without the use of computers (Patil et al., 2015). Moreover, the use of traditional statistical approaches is not sufficient to discover the knowledge hidden in the data (Karrar et al., 2016). However, data mining is able to deal with this problem and make it possible to uncover useful, relevant information which can be provided to decision makers to assist in devising suitable prevention and intervention efforts to address the risk factors for violence that occurs in Thialand's deep south. This study aims to extract these hidden patterns, relationships, and knowledge from the Deep South Coordination Centre (DSCC) database which was collected between 2004 and the beginning of January 2016 using data mining techniques.

The paper is constructed as follows. Section 2 reviews the literature on data mining, clustering, and the association rule. Section 3 briefly proposes the methodology used in the study. Experiment results and conclusions are presented in sections 4 and 5, respectively.

2 LITERATURE REVIEW

Data mining techniques can be categorized into supervised, or predictive, learning techniques and unsupervised, or descriptive, learning techniques. Supervised learning is based on associating an example from a data source with a correct classification that has already been determined. Thus, there is a special attribute, namely class, existent in all cases. It identifies whether a case is in a certain class, which will be the focus of learning. Supervised learning is divided into classification techniques and regression. The common description of unsupervised learning is learning without pre-classified examples. It is divided into clustering analysis and association rule (Martin et al., 2014).

2.1 Clustering analysis

Clustering analysis is an unsupervised learning technique that operates on data objects without referring to a known class label. The technique is useful for uncovering trends and patterns in data when there are no pre-defined classes (Singh et al., 2015). The goal of the technique is to classify similar (or related) objects into a group and different (or unrelated) objects into other groups. The greater the similarity (or homogeneity) within a group, and the greater the difference between groups, the "better" or more distinct the clustering (Gupta & Kaur, 2017). The algorithms that are generally used to achieve a cluster analysis can be classified as hierarchical algorithms or non-hierarchical algorithms (i.e. k-means algorithm). For hierarchical algorithms, each object is initially in its own cluster and then clusters are successively joined to create a clustering structure. In a non-hierarchical algorithm, the number of clusters must be theoretically known. The algorithm is done by minimizing a measure of dissimilarity within each cluster and maximizing the dissimilarity between different clusters (Cornish, 2007).

2.2 K-Means algorithm

K-means is a typical, and commonly used, algorithm for clustering operations because of many advantages, namely its simplicity, its efficiency, and being less time consumption (Sharma et al., 2012). The goal of this technique is to allocate n observations into k-clusters, with each cluster's center denoted by the mean value of the objects in the cluster. Consequently, each object is allocated to its nearest cluster center. Thus the general outcome is to reduce the total squared distance of all objects from their cluster centers. The selection of the mean as the cluster center is able to reduce the total squared distance from every cluster point to its center. Therefore, it is practical

to characterize and isolate every object in the data set into different groups by reducing the mathematical distance between the substantial groups or clusters (Singh et al., 2015; Nieves & Cruz, 2011). Although many descriptions of mathematical distance are presented, Euclidean distance is used in WEKA (Aksenova, 2004).

Recently, the k-means algorithm has been applied in various research domains. Gowri et al. (2017) applied k-means procedure to a database of students, obtained from four government schools in Vellore district, Tamil Nadu, in order to generate a grouping of students based on their personal characteristics in addition to their academic record. The results showed that the performance of students was clustered as good and poor. Bach et al. (2018) studied the impact of the economic costs of violence in 119 countries worldwide. The economic cost of violence and on the economic development levels in the countries was measured by Cultural, Administrative, Geographic, and Economic differences and 10 additional variables. In order to group the observed countries according to their economic costs of violence and economic development levels, the k-means clustering procedure was applied. According to the conducted nonhierarchical cluster analysis, in which a k-means approach was used, it was determined that there were six groups of countries. The study of Wakoli et al. (2014) applied the k-means clustering algorithm to medical claims records related to data obtained from 15 insurance companies in Kenya. The results showed the successful application of the k-means clustering algorithm to medical claims records. The k-means method was also applied to classify the popularity of tourist destinations based on the number of instagram accounts from explorejogja. which consists of 121 tourist destinations. The results presented 3 groups of tourist destinations, each of which consists of cluster 1, which included 9 tourist destinations, cluster 2, consisted of 60 tourist destinations, and cluster 3, which contained as many as 52 tourist destinations (Iswandhani & Muhajir, 2018).

2.3 Association Rule

Association rule is one of the most important data mining techniques which was initiated from market basket analysis problems (Nieves & Cruz, 2011). This technique is used to discover the associations among a set of items. The process of applying the association rule is divided into steps as follows (Al-Maolegi & Arkok, 2014). In the first step, each set of items is known as an itemset. If the total that each item occurs is greater than the minimum support which has been pre-specified, this itemset is named a frequent itemset. In the second step, many rules can be created from one itemset. For example, the rule, $X \rightarrow Y$ where X and Y are items validated based on a confidence threshold. The value is defined as the probability of transactions containing X which contain also Y. The support and confidence thresholds can be specified in advance by users. The support and confidence can be expressed as follows:

$$\text{support } (X \rightarrow Y) = P(X \cup Y) \quad (1)$$

$$\text{confidence } (X \rightarrow Y) = P(Y \mid X) = \frac{P(X \cup Y)}{P(X)} \quad (2)$$

2.4 Apriori algorithm

Apriori algorithm is simple and easy to use to find all of the frequent itemsets in a database. The algorithm iteratively finds frequent itemsets where k-itemsets are used to generate k+1-itemsets. Each k-itemset must be greater than, or equal to, the minimum support threshold specified by the user. Otherwise, they are called candidate itemsets which were pruned. Firstly, the algorithm scans the database to find frequency of 1-itemsets that contains only one item by counting each item in database. The frequency of 1-itemsets is used to find the itemsets in 2-itemsets which is in turn used to find 3-itemsets and so on until no more k-itemsets are found. Finding frequent itemsets is time consuming when dealing with a huge number of candidate sets containing frequent itemsets, low minimum support, or large itemsets. In order to efficiently find frequent itemsets and reduce time consumption, the algorithm employs a criterion which states that, if any k-itemset is frequent, all of its subsets must be frequent, and conversely, if an itemset is not frequent, any large subset within it must also be non-frequent. The property, therefore, enables researchers to reduce the search space in the database (Buczak & Gifford, 2010; Al-Maolegi & Arkok, 2014).

Currently, association rule technique is not confined to market basket analysis. Margono et al. (2013) used association rule technique to discover trends in human rights violations which occurred in Indonesia by mining data in a human rights violations database. The outcomes of this work provided people involved with information regarding the kinds of violations that occurred within Indonesian society. Nieves and Cruz (2011) discovered inherent information related to different terrorist organizations based on the different types of terrorist acts they committed by using association rule. The results should be useful, not only for counter terrorism security analysts, but also in determining the prioritization and geographical allocation of military and law enforcement resources. This technique has also been used to identify disease co-occurrences based on ICD-9-CM codes in a statewide hospital discharge data set. In the study of Rani, Vohra and Gulia (2014) the association rule was applied to a simulated primary database of passport and visa information, to discover the travel patterns in different segments. The results revealed total 10 rules supporting each other according to their dominant category.

3 METHODS

3.1 Data and data preprocessing

A large amount of detailed data about violent incidents between 2004 and 2016 was obtained from the Deep South Coordination Centre (DSCC) database, Prince of Songkla University, Pattani, Thailand. All types of violent events were equally considered in the study. Meanwhile, the input data had to be preprocessed with the purpose of decreasing the consequence of noise, missing values, and discrepancies before executing the data mining techniques. The data mining process consisted of data cleaning, data integration, data transformation, data reduction, discretization, and concept hierarchies to enable data preprocessing.

Data cleaning was employed to handle missing values for variables, namely, "day","time ", "zone", and "province". The incidents with variables which lacked the information were deleted from the data set. Data transformation was implemented to transformation data into another form appropriate for mining; for example, the variable "day" was derived from the date of the event. All the processes of data transformation were executed using SQL and Microsoft Excel. Ultimately, the total number of events in the study was 21,424.

3.2 Variables

The ten related variables “time”, “day”, “quarter”, “zone”, “district”, “province”, “arson”, “gun”, “bomb”, and “outcome” included in this study were similar to Makond (2018). In addition, all variables were converted into dummy variables, as presented in Table 1.

Table 1: Variables, dummy variables with descriptions

Variables	Descriptions	Dummy variable	Number of events
time	time of the day	"t1 "represents time period from 00:01 a.m .to 03.00 a.m.	1,437
		"t2 "represents time period from 03:01 a.m .to 06:00 a.m.	1,912
		"t3 "represents time period from 06:01 a.m .to 09:00 a.m.	3,824
		"t4 "represents time period from 09:01 a.m .to 12:00 a.m.	2,538
		"t5 "represents time period from 12.01 p.m .to 15.00 p.m.	2,116
		"t6 "represents time period from 15:01 p.m .to 18.00 p.m.	2,564
		"t7 "represents time period from 18:01 p.m .to 21.00 p.m.	4,841
		"t8 "represents time period from 21:01 p.m .to 24.00 a.m.	2,192

Variables	Descriptions	Dummy variable	Number of events
Day	day of the week	"d1" represents Sunday	2,773
		"d2" represents Monday	3,316
		"d3" represents Tuesday	3,130
		"d4" represents Wednesday	3,358
		"d5" represents Thursday	3,323
		"d6" represents Friday	3,030
		"d7" represents Saturday	2,494
quarter	quarter of the year	"q1" represents January to March	5,081
		"q2 "represents April to June	5,824
		"q3 "represents July to September	5,599
		"q4 "represents October to December	4,920
zone	place of the incident	"zone1 "represents road/highway	11,792
		"zone2 "represents residential area/personal area or shop	4,603
		"zone3 "represents Other/unspecified	5,029
province	province in this study	"PR_N "represents Narathiwat	7,614
		"PR_P "represents Pattani	7,040
		"PR_S" represents Songkhla	1,092
		"PR_Y "represents Yala	5,678
district	district that is adjacent to neighbour province/country	district (0,1)	(7,174, 14,250)
arson	means used in the incident was arson	arson (0,1)	(18,886, 2,538)
Gun	weapon used in the incident was one or more guns	gun (0,1)	(11,082, 10,342)
bomb	weapon used in the incident was one or more bombs	bomb (0,1)	(17,437, 3,987)
outcome	the outcome of violence is physical injury	outcome (0,1)	(10,684, 10,740)

3.2 Tool and algorithm

SQL and Microsoft Excel were the tools used for data preprocessing. For analyzing the data, two WEKA algorythms were employed, namely the SimpleKMeans algorithm for clustering analysis and the Apriori algorithm for association rules. Some implementations of k-means only allow numerical values for variables. In such circumstances the data set may need to be transformed into a standard spreadsheet format. Similarity, categorical attributes would need to be transformed to binary data. Moreover, data measured on considerably different scales were needed to normalize the values. However, WEKA provides a SimpleKMeans algorithm which automatically handles a mixture of categorical and numerical variables and normalizes numerical variables when computing distance. The SimpleKMeans algorithm computes distances between instances and clusters using Euclidean distance. The Apriori algorithm for learning association rules in Weka works only with discrete data and will identify statistical dependencies between groups of attributes using confidence and support measurements. Apriori can compute all of the rules that have a specified minimum support and are greater than a specified confidence level (Aksenova, 2004; Sharma et al., 2012).

4 EXPERIMENT RESULTS

Using the WEKA tool the experiments were performed on a data set containing 31 dummy variables and 21,424 violent cases. Two data mining techniques, clustering and association rule, were used on the data.

4.1 Results from clustering analysis

In the current analysis of 21,424 violent cases, using cluster analysis, two clusters were initially assigned and the data was classified into the two clusters (i.e. non-physical injury and physical injury) with 10,927 (51%) and 10,497 (49%) falling within each group, respectively. The results shown in figure 1 can be explained as follows. The first cluster categorized the violent event as the occurrence of non-physical injury (denoted with "no") where the event did not involve the use of arson, guns, or bombs, and took place in a district that is adjacent to a neighbouring province/country. The second cluster categorized the violent event as the occurrence of physical injury (denoted with "yes") where the event involved the use of one, or more guns, but not arsons or bombs, occurred in a district that is adjacent to a neighbouring province/country and took place on a road/highway.

```
Cluster#
Attribute Full Data  0  1
  (21424.0) (10927.0) (10497.0)
=======================================
arson  no no no
gun  no no yes
bomb  no no no
district  yes yes yes
t1   no no no
t2   no no no
t3   no no no
t4   no no no
t5   no no no
t6   no no no
t7   no no no
t8   no no no
PR_P  no no no
PR_Y  no no no
PR_N  no no no
PR_S  no no no
d1   no no no
d2   no no no
d3   no no no
d4   no no no
d5   no no no
d6   no no no
d7   no no no
q1   no no no
q2   no no no
q3   no no no
q4   no no no
zone1  yes no yes
zone2  no no no
zone3  no no no
```

```
outcome  yes no yes


Time taken to build model (full training data) :
0.17 seconds
=== Model and evaluation on training set ===
Clustered Instances
0 10927 ( 51%)
1 10497 ( 49%)
```

Figure1: The results from clustering analysis

4.2 Results from the association rule

Apriori algorithm processes data by mining the rules to extract patterns that are similar along their associations in relation to several set of records. A minimum support of 20% was applied to the data set. The rules generated were ranked by confidence metric which was specified at 0.8. The output of this analysis is shown in figure 2. The findings are as follows:

For rule 1, of the violent incidents in which guns were used as weapons, but bombs were not use as weapons, and the incidents did not happen during the period from 21:01 p.m. to 24:00 a.m., 85% (confidence) involved physical injury.

For rule 2, of the violent incidents in which guns were used as weapons, but bombs were not use as weapons, and arson was not used as a mean, and the incidents did not happen during the period either from 00:01 a.m. to 03:00 a.m. or from 03:01 a.m. to 06:00 a.m., 85% (confidence) involved physical injury.

For rule 3, of the violent incidents in which guns were used as weapons, but bombs were not use as weapons, and the incidents did not happen during the period either from 00:01 a.m. to 03:00 a.m. or from 03:01 a.m. to 06:00 a.m., 84% (confidence) involved physical injury.

For rule 4, of the violent incidents in which guns were used as weapons, but bombs were not use as weapons, and arson were not used as a mean, and the incidents did not happen during the period from 03:01 a.m. to 06:00 a.m., 84% (confidence) involved physical injury.

For rule 5, of the violent incidents in which guns were used as weapons, but bombs were not use as weapons, and arson were not used as a mean, and the incidents did not happen during the period from 00:01 a.m. to 03:00 a.m., 84% (confidence) involved physical injury.

For rule 6, of the violent incidents in which guns were used as weapons, but arson was not use as a mean, and the incidents did not happen during the period from 21:01 p.m. to 24:00 a.m., 84% (confidence) involved physical injury.

For rule 7, of the violent incidents in which guns were used as weapons, but bombs were not use as weapons, and the incidents did not happen during the period from 00:01 a.m. to 03:00 a.m., 84% (confidence) involved physical injury.

For rule 8, of the violent incidents in which guns were used as weapons, but bombs were not use as weapons, and arson was not used as a mean, 84% (confidence) involved physical injury.

For rule 9, of the violent incidents in which guns were used as weapons, and the incidents did not happen during the period from 21:01 p.m. to 24:00 a.m., 84% (confidence) involved physical injury.

For rule 10, of the violent incidents in which guns were used as weapons, but bombs were not use as weapons, and the incidents did not happen during the period from 03:01 a.m. to 06:00 a.m., 84% (confidence) involved physical injury.

```
Apriori
=======
Minimum support: 0.35 (7498 instances)
Minimum metric <confidence>: 0.8
Number of cycles performed: 13

Generated sets of large itemsets:

Size of set of large itemsets L(1): 50

Size of set of large itemsets L(2): 322

Size of set of large itemsets L(3): 405

Size of set of large itemsets L(4): 153

Size of set of large itemsets L(5): 20

Best rules found:

 1. gun=yes bomb=no t8=no 8859 ==> outcome=yes
7505 conf:(0.85)
 2. arson=no gun=yes bomb=no t1=no t2=no 8936
==> outcome=yes 7563 conf:(0.85)
 3. gun=yes bomb=no t1=no t2=no 9033 ==>
outcome=yes 7632 conf:(0.84)
 4. arson=no gun=yes bomb=no t2=no 9485 ==>
outcome=yes 8012 conf:(0.84)
 5. arson=no gun=yes bomb=no t1=no 9365 ==>
outcome=yes 7907 conf:(0.84)
 6. arson=no gun=yes t8=no 9034 ==> outcome=yes
7625 conf:(0.84)
 7. gun=yes bomb=no t1=no 9466 ==> outcome=yes
7980 conf:(0.84)
 8. arson=no gun=yes bomb=no 9914 ==>
outcome=yes 8356 conf:(0.84)
 9. gun=yes t8=no 9141 ==> outcome=yes 7702
conf:(0.84)
10. gun=yes bomb=no t2=no 9598 ==> outcome=yes
8086 conf:(0.84)
```

Figure2: The results from association rule

5 CONCLUSIONS

This study's purpose was to apply two data mining techniques to extract the hidden patterns, relationships, and knowledge from the DSCC database which was collected between 2004 and the beginning of January 2016. The data consisted of 21,424 violent cases and 31 dummy variables.

Cluster analysis assigned the data into two clusters, a physical injury cluster and a non-physical injury cluster, based on the use of arson, guns, or bombs; time period of the day; day of the week; quarter of the year; place of the incident; district; province and outcome. Meanwhile, the Apriori algorithm generated 10 rules. Overall results obtained in this study, whether using either cluster analysis or the association rule, revealed out that violent events involving the use of a gun results in physical injury. This result is consistent with the results of Chirtkiatsakul et al. (2014) and Makond (2018).

This study concluded that clustering analysis and association rule are essential data mining techniques for exploring useful, relevant information which can be provided to decision makers to assist in devising suitable prevention and intervention efforts to address the risk factors for violence that occurs in Thialand's deep south. However, this is an exploratory study rather than explanation study; therefore, more knowledge discovery is needed using the other data mining techniques.

ACKNOWLEDGEMENTS

The author would like to thank and gratitude to Assistant Prof.Metta Kuning, the former Director of the DSCC, Prince of Songkla University, Pattani campus. This study is funded by the Centre of Excellence in Mathematics, Commission on Higher Education, Thailand.

REFERENCES

Aksenova, S. S. (2004). Machine Learning with WEKA WEKA Explorer Tutorial for WEKA Version 3.4.3. School of Engineering and Computer Science. Department of Computer Science California State University, Sacramento California.

Al-Maolegi, M., & Arkok, B. (2014). An improved apriori algorithm for association rules. International *Journal on Natural Language Computing*, 3(1), 21-29.

Bach, M. P., Dumicic, K., Jakovic, B., Nikolic, H., & Berislav, Z. (2018). Exploring impact of economic cost of violence on internationalization: Cluster analysis approach. International *Journal of Engineering Business Management*, 10, 1-15.

Buczak, A. L., Gifford, C. M. (2010). Fuzzy association rule mining for community crime pattern discovery. Proceedings of the Acm Sigkdd Workshop on Intelligence and Security Informatics. 10.1145/1938606.1938608.

Chirtkiatsakul, B., Kuning, M., McNeil, N., & Eso, M. (2014). Risk factors for mortality among victims of provincial unrest in

Southern Thailand. *Kasetsart Journal of Social Sciences*, 35, 84-91.

Comstock, R. D., Mallonee, S., & Jordan, F. (2005). A comparison of two surveillance systems for deaths related to violent injury. *Injury Prevention*, 11, 58-63.

Cornish, R. (2007). Statistics: 3.1 Cluster Analysis. Mathematics learning support centre.

Espinosa, R., Gutiérrez, M. I., Mena-Muñoz, J. H., & Córdoba, P. (2008). Domestic violence surveillance system: a model. *Salud Publica Mex*, 50(Suppl 1), 1-11.

Dahlberg, L. L. & Krug, E. G. (2006). Violence a global public health problem. *Ciência & Saúde Coletiva*, 11(2), 277-292.

Gowri, G.S., Thulasiram, R., & Baburao, M. A. (2017). Educational data mining application for estimating students performance in weka environment. *Preceedings of IOP Conf. Series: Materials Science and Engineering* 263, 1-9. doi:10.1088/1757-899X/263/3/032002

Gupta, A., & Kaur, V. (2017). Implementation of proposed clustering algorithm (ECBA) on criminal dataset. *International Journal of Engineering Science and Computing*, 7(5). 11234-11235.

Iswandhani, N., & Muhajir, M. (2018). K-means cluster analysis of tourist destination in special region of Yogyakarta using spatial approach and social network analysis. *Proceeding of International Conference on Mathematics: Pure, Applied and Computation. IOP Conf. Series: Journal of Physics: Conf*, 1-8.doi:10.1088/1742-6596/974/1/012033

Karrar, A. E., Abdalrahman, M. A., & Ali, M. M. (2016). Applying K-means clustering algorithm to discover knowledge from insurance dataset using WEKA Tool. *The International Journal Of Engineering And Science*, 5(10), 35-39.

Makond, B. (2018). A comparison of statistical techniques in predicting violent outcomes in Thailand's deep South. *Kasem Bundit Journal, 19 (special edition)*, 284-300.

Margono, H., Yi, X., & Raikundalia, G. K. (2013). Using association rules mining to analyze human rights violations in Indonesia. *International Journal of Computer Science and Electronics Engineering*, 1(1), 65-70.

Martín, L., Baena, L., Garach, L., López, G., & de Oña, J. (2014). Using data mining techniques to road safety improvement in Spanish roads. *Procedia - Social and Behavioral Sciences*, 160, 607-614.

Nieves, S., Cruz, A. (2011). Finding patterns of terrorist groups in Iraq: a knowledge discovery analysis. *Proceedings of Ninth LACCEI Latin American and Caribbean Conference, Engineering for a Smart Planet, Innovation, Information Technology and Computational Tools for Sustainable Development*, 1-10.

Patil, R., Deshmukh, S., & Rajeswari, K. (2015). Analysis of simple K-means with multiple dimensions using WEKA. *International Journal of Computer Applications*, 110(1), 14-17.

Rani, P., Vohra, R., & Gulia, A. (2014). Association rule mining in discovering travel pattern in passport data analysis. *International Journal of Computer Science and Information Technologies*, 5(4), 5015-5019.

Schuurman, N., Cinnamon, J., Walker, B. B., Fawcett, V., Nicol, A., Hameed, S. M., & Matzopoulos, R. (2015). Intentional injury and violence in Cape Town, South Africa: an epidemiological analysis of trauma admissions data. *Global Health Action*, 8(27016), 1-9.

Sharma, N., Bajpai, A., Litoriya, R. (2012). Comparison the various clustering algorithms of weka tools. *International Journal of Emerging Technology and Advanced Engineering*, 2(5), 73-80.

Sharma, R., Alam, M. A., & Rani, A. (2012). K-means clustering in spatial data mining using Weka Interface. *Proceeding of International Conference on Advances in Communication and Computing Technologies*, 26-30.

Singh, S., Shrivastava, N., & Tiwari, R. K. (2015). K-means cluster model for climate prediction of Vindhya region. *International Journal of Emerging Technologies in Computational and Applied Sciences*, 12(1), 70-72.

Wakoli, L. W., Orto, A., & Mageto, S. (2014). Application of the K-means clustering algorithm in medical claims fraud/abuse detection. *International Journal of Application or Innovation in Engineering & Management*, 3(7), 142-151.

Structural Model of Opportunity Management (OM) towards Corporate Governance (CG) and Enterprise Risk Management (ERM)

Patipan Sae-Lim

Graduate School of Management and Innovation (GMI) King Mongkut's University of Technology Thonburi (KMUTT), BKK
Corresponding Email: patipanlim7@gmail.com

ABSTRACT

Historically, the study of Corporate Governance (CG) and Enterprise Risk Management (ERM) has focused on the business assurance aspect. The research question in the study was about if it could be possible to conceptualize CG and ERM in a proactive manner? There are two objectives in this study: 1) to study the relationship between CG and ERM and 2) to determine to what extent CG and ERM could be possible to improve long term growth by enhancing opportunity management (OM) for both financial and non-financial aspects. Approximately 700 Thai-listed companies were considered. A mixed-method approach through structural equation modelling (SEM) and interviews was employed. With 175 organizations and eight interviewees, it could be confirmed that CG and ERM have a significant relationship. The convergence between the quantitative and qualitative analyses displayed that systemic CG and ERM could enhance management for better decision making in new business arenas as seized opportunities. To be precise, CG, ERM and OP were themselves correlated. However, different experts interpreted and quantified OP in distinctive ways. From an empirical finding, CG and ERM were insignificantly associated with some financial indicators.

Keywords: Opportunity Management; Corporate Governance; Enterprise Risk Management; Structural Equation Modelling (SEM.)

1 RATIONALE OF STUDY

Corporate Governance (CG) is a multifaceted term depending on one's view of the world (Stuart, 2006: 382). Some theories focus CG on the view of laws, rules and factors that control operations at a company. Others defined CG as an organizational complex system given many indicators.

Systematic measurement of CG was initially undertaken by Ross et al. (2005) (Figure 1). They defined CG from compositions between internal and external governance. Other articles tried to explore similar related indicators under its framework. Stuart (2006) divided internal governance into five basic categories: 1) the Board of Directors (and their role, structure and incentives), 2) Managerial Incentives, 3) Capital Structure, 4) Bylaw and Charter and 5) Internal Control System, while the external governance was divided into two groups: law and markets.

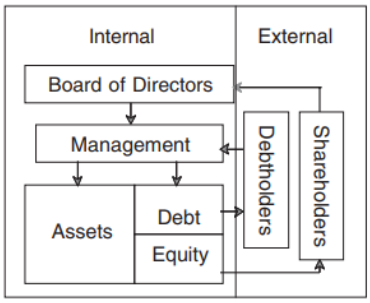


Figure 1: CG Dimensions (Ross et al., 2005)

CG was widely known from the bankruptcy of Enron, an American energy, commodities, and services company. It became a well-known example of corporate fraud and corruption due to artificial financial reports and accounting fraud. The concepts of CG were then derived after the enactment of the Sarbanes–Oxley Act of 2002 that guides and controls accounting practice as well as corporate activities including corporate governance directly measured from the leaders. Later on, the concept of Risk Management (RM) was popularized to prevent business losses as well as develop regulatory alignment (Fraser et al. 2010). Thus, CG was the driving factor for the birth of ERM, and this is one of the paths in this research.

Historically, the adoption of risk was about achieving business goals as well as preventing loss, while modern business theories try to propose the concept of "Enterprise Risk and Opportunity Management (EROM)" that refers to the approach adopted by corporates to manage risks and seize the opportunities related to the accomplishment of business objectives (Benjamin, 2017). Such a pioneer concept could be beneficial to enhance the level of participation from leaders, but the problem is about obtaining the empirical data to test the concept of EROM. Furthermore, as ERM is derived somehow from CG, why do researchers not study both CG and ERM effects on corporate opportunity management (OM) the level of participation from leaders, but the problem is about obtaining the empirical data to test the concept of EROM.

Furthermore, as ERM is derived somehow from CG, why do researchers not study both CG and ERM effects on corporate opportunity management (OM).

To summarize, this study has two objectives: 1) confirm previous studies about the relationship between CG and ERM and 2) to empirically investigate/explore the relationship among CG, ERM and OM via a mixed method. Practically, the findings will be related to both CG and ERM in corporations and theoretically, and such path results could be used to create a new concept of EROM.

2 THEORIES CONSTRUCTION AND CONCEPTUAL FRAMEWORK

2.1 Defined Opportunity Management

Risk is multifaceted; however, most theories define risk as a negative event. Fraser et al. (2010) defined risk as "the possibility of future performance shortfalls with respect to reach explicitly stated objectives thru organizations"; while, opportunity is "the possibility of future performance improvement with respect to reach explicitly stated objectives thru organizations".

To measure OM, it depends on the organizational views. Based on interviewing people in executive management positions in Thai-listed companies, it became clear that they view OM in two dimensions. First of all, they interpret OM as the ability to incline shareholder values throughout the growth of revenue and earnings, the growth of capital and reducing loss. Secondly, based on expert views, OM is about the ability to promote better decision making opportunities.

2.2 Process of Enterprise Risk and Opportunity Management (EROM)

Historically, the adoption of Enterprise Risk Management (ERM) aimed to prevent loss via an early warning when negative events disrupt the corporate goals. Alternatively, today, ERM can be employed as a strategic tool. COSO (2004) encompassed ERM as 1) an aligning of risk appetite and strategy, 2) enhancing risk response decisions, 3) reducing operational surprises and losses, 4) identifying and managing multiple and cross enterprise risks, 5) seizing opportunity and 6) improving deployment of capital.

As mentioned above, COSO initially defined the more proactive benefits of ERM, but there was a lack of empirical analysis of proactive ERM. Benjamin (2017) defined EROM as a process of "seeking an optimal balance between minimizing the potential for loss (risk) while maximizing the potential for gain (opportunity) with the respect to organizational mission". Achievement of this optimization implies the

agility to make a new decision to the maximum tolerable levels for risk, minimum desirable levels for opportunity and trade-offs between them.

For conceptualization, the paradigm of RM had shifted from a quantitative method in traditional RM to the new definition of ERM. Traditional RM has focused on the risk management process: identification of risk, assessment of risk, response to risk and monitoring of risk. With these steps, it lacks a role in the internal environment, and that is why ERM -a new paradigm of RM- now tries to consolidate the readiness of the internal environment as part of the process of implementing RM.

Figure 2: Risk Management (RM) Paradigm's shift

For conceptualization, ERM divides as follows:

- Internal Environment
- Risk Identification
- Risk Assessment
- Risk Response
- Risk Monitoring

Figure 3: ERM Conceptualization

2.3 Corporate Governance Indicators

Many articles try to quantify the indicators of CG and they depend on the theories and contexts. However, as this study focused on Thai listed companies, it considered the context of CG in the Thai industrial environment.

This article triangulates research findings with secondary data from the Thai Institute of Directors (IOD) about CG indicators (CRG report, 2017). The CRG report presented the CG score with five bands: pass, satisfactory, good, very good and excellence. The indicators used to categorize the five bands came from: Rights of Shareholders, Equitable Treatment of Shareholders, Role of Stakeholders, Disclosure and Transparency and Board Responsibilities. Moreover, each mentioned indicators has several questions for the management of Thai-listed companies. The descriptive statistics for the CGR 2017 are shown in table 1.

This research used five bands for CG indicators for each respondent company in the process of the structural equation analysis. CG was the exogenous variable in the model presented in figure 4.

2.4 Relationship between Corporate Governance and Enterprise Risk Management

Prior to the development of ERM, it is significant to completely understand the relationship between CG and ERM (Marchetti, 2012). The relationship between them has historically been mentioned since the tragedy of Enron (Robert, 2003). Due to a lack of a governance system as well as accountability for Enron, a RM system has since then been required by the Security Exchange Commission (SEC), which directly regulates listed-companies.

Beside the regulatory base, CG is an indispensable component of ERM. It supplies the top-down monitoring and management of risk in organizations. Therefore, in terms of the correlation, they are both closely related. Both focus on strategy and support the strategy direction of the organization. Some empirical studies disclosed that good governance by the Board of Directors (BOD) should be developed and implement comprehensively in a RM policy, used when determining the risk appetite and establishing the overall corporate culture that supports RM.

To confirm the relationship between CG and ERM, the Organization for Economic Co-operation and Development (OECD) (2014) reviewed the corporate governance frameworks and practices in the 27 jurisdictions that participated in the OECD Corporate Governance Committee. This article found a strong relationship between the CG and ERM systems. Moreover, the OECD revealed that the cost of ERM failure is still underestimated both internally and externally, including the cost in terms of management time needed to rectify the situation. Therefore, good CG thru the role of the BOD should include the maturity level of the ERM.

Table 1: Descriptive Statistics of CGR 2017

Survey Category	Average	Median	Maximum	Minimum
Right of Shareholders	93	95	100	43
Equitable Treatment of Shareholders	92	96	100	59
Role of Stakeholder	78	82	98	19
Disclosure & Transparency	84	86	100	35
Board Responsibility	71	71	95	37
Overall Scores	80	81	97	48

Source: Thai Institute of Directors (IOD) of CG indicators (CRG report, 2017)

2.5 Proposed Path of Conceptual Framework and Hypothesis Testing

As described in a previous literature review, the author proposed a path conceptual framework and three hypotheses as in figure 4.

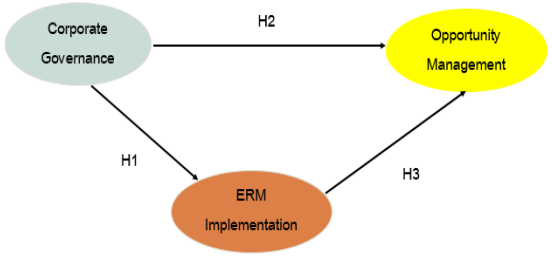


Figure 4: Conceptual Framework

3 METHODS

3.1 Quantitative Methodology

3.1.1 Research Design

In Thailand, CG and ERM are not at a high maturity level. However, corporations or listed companies have been required to embed CG and ERM, due to the SEC as a compulsory system of intentionally disclosed company information to shareholders. The unit of analysis thus accounted for approximately 749 Thai-listed corporations in both SET and MAI (referred to in May 2018). The portions of each sector are shown in figure 5.

A survey was employed as a generalization process (Babbie, 2007) to gather respondent preference on ERM and EROM. Validity and reliability testing were both verify philosophically. As the population is quite small, all the population was selected to rectify the problem of a low respondent rate.

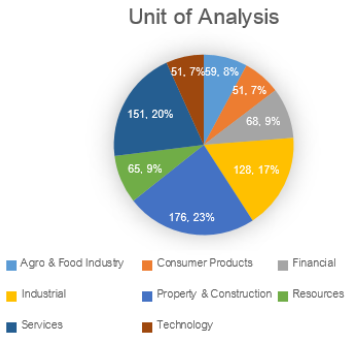


Figure 5: Unit of Analysis

3.1.2 Data Management and Measurement Items

Both primary and secondary data were adopted for triangulation with qualitative data. The survey instrument was used as the primary data while the secondary data was gathered from reliable reports: financial statements, CG and so on. As the majority of the analysis tool was multivariate analysis thru Structural Equation Modeling (SEM), the data violation of the assumptions was verified via normality, multicollinearity, homoscedasticity and sample size

For the measurement items, there were three latent variables: CG, ERM and OP, and their observed variables are presented in Table 2.

Table 2: Details of Variables

Latent Variable	Observed Variable
CG	1. Rights of Shareholders, 2. Equitable Treatment of Shareholders, 3. Role of Stakeholders, 4. Disclosure and Transparency, 5. Board Responsibilities
ERM	1. Internal Environment 2. Risk Identification 3. Risk Assessment 4. Risk Mitigation 5. Risk Monitoring
OP	1. Non-Financial OP (Deducing a better decision making in new business arena, proactive strategy) 2. Financial OP (Return of Equity)

3.1.3 Statistical Model

Both descriptive and inferential statistics were used to analyze the data. For inferential statistics, SEM is suitable as this research looks at the relationship between the observed and the latent variables (Foster et al., 2006: 103). The latent variables were CG, ERM and OM. To determine the path of the conceptual framework, the SEM model used in this research was as follows:

$$CG = \beta_0 + \beta_1 ERM \quad (1)$$
$$CG = \gamma_0 + \gamma_1 OM \quad (2)$$
$$ERM = \partial_0 + \partial_1 OM \quad (3)$$

where β_0 = intercept between CG and ERM
β_1 = regression of coefficient between CG and ERM
γ_0 = intercept between CG and OM
γ_1 = regression of coefficient between CG and OM
$\partial(0)$ = intercept between ERM and OM
∂_1 = regression of coefficient between ERM and OM

3.2Qualitative Methodology

In-depth interviews were employed as a qualitative data collection process. Thru path analysis was used for the qualitative part of this study to determine which factors were related to each other. The in-person interview was selected as the method of data gathering to build a stronger rapport and trust with the participants. This method has potential benefits as the researcher considered both verbal and non-verbal communication (Aurini, Heath and Howells, 2016). The qualitative result will triangulate and support the research implications later.

4 RESULTS

4.1 Descriptive Statistics

There were 175 organizations that responded to the research questionnaire from 749 Thai listed-companies (23.4%). The respondent rate was not very high as some Thai-listed companies had not implement ERM formally and its maturity level seemed quite low. Fortunately, nearly 90 percent of the respondents are currently working in the ERM field (RM Committee and ERM Department), which suggest high accuracy in the research findings (figure 6). Furthermore, 80% of the respondents have adopted COSO as an ERM standard and principle. Most of the sectors returned the questionnaires at a rate of approximately 10%, except the financial sector that had a higher respondent rate (24%), due to the higher maturity level compared to the other sectors. The greater the sample size the better the inferential statistics. (Kumar, 2005) (Foster et al., 2006 :105)

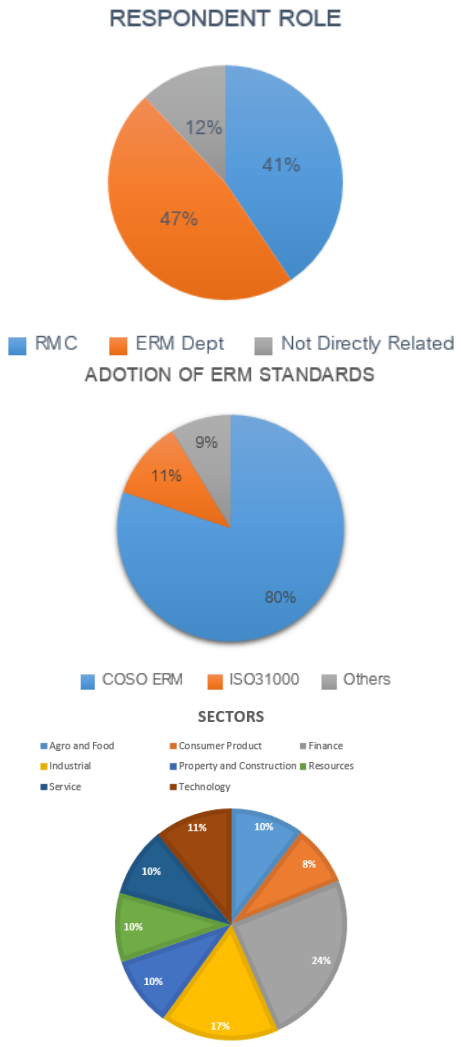


Figure 6: Descriptive Statistics

4.2 Data Violation of Assumption Testing

Starting with the reliability and validity construction, all primary data from the questionnaires had high reliability in the range 0.78 to 0.87. The construct validity accounted for above 0.80. The greater the sampling size the better, in terms of inferential statistics. Kumar (2005) stated that a large sample size should be employed, but it depends on the number of observed variables. The sample size of 175 in this study was suitable $(12\times(13/2)=78)$.

For the multivariate normality, it was difficult to test; fortunately, the author then tested the univariate normality thru a normality plot. The result displayed that the observed value for each variable against the expected value was located on a straight line; hence, it suggested a normal distribution (Pallant, 2005). In terms of the relationship among the independent variables themselves -multicollinearity- the variance inflation factors (VIF) ranged from 1.8 to 7.1, which were less than 10. Therefore, the multicollinearity was marginal. Lastly, the empirical data did not violate homoscedasticity (figure 7).

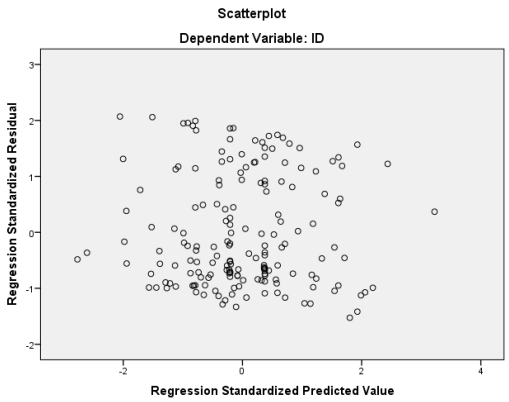
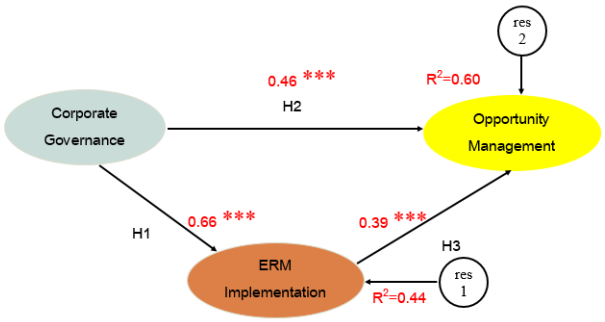


Figure 7: Homoscedasticity Testing

4.3 Hypothesis Testing Result

SEM is composed of two types of model: measurement and structural. Both models had high values for the regression weigh, as shown in the appendix. The hypothesis fixing is shown in figure 8.



Chi square =32.787, df =18 p =0.02,
CMIN/df =1.822, GFI =0.956, CFI =0.986 and RMSEA =0.07
*p<.05, **p<.01, ***p<.001

Figure 8: Hypothesis Testing Result

Table 3: Hypothesis Result

Research Hypothesis	Standardized Regression Weights	P-Value	Interpretation Compared to Sig 0.05
Corporate Governance→ ERM Implementation	.66	<.001	Support Hypothesis
Corporate Governance → Opportunity Management	.46	<.001	Support Hypothesis
ERM Implementation → Opportunity Management	.39	<.001	Support Hypothesis

4.4 Qualitative Findings

Apart from quantitative analysis, qualitative analysis was conducted with interviews from eight managers across the industries with nine in-depth questions. Based on the findings, they all agreed that CG is a driving factor to implement ERM successfully. To be precise, a high maturity in CG could be brought about by high maturity in ERM. Interviewees interpreted "Opportunity Management (OP)" in different ways but converged on the theme of OP: to their point of view, it is about seizing the opportunity not only for the investment aspect, but it also concerns new business arenas. Some interviewees from banking agreed that successfully implementing ERM could lead to a high rate of ROE. Finally, unfortunately, other interviewees from other industries suggested that a good ERM system significantly increases good management decisions. Therefore, the majority of interviewees interpreted the direct relationship between ERM and OP in terms of non-financial op. To CG, most of the interviewees concluded that CG is not directly linked to good OP, but CG can lead to better business assurance. Finally, they agreed that both CG and ERM are both partial improvements for good OP. OP, as well as seizing opportunities, in business needs multidisciplinary teams with a high level of cooperation.

5 CONCLUSIONS

Based on multivariate analysis, the collected empirical data could be fitted to a model with acceptable statistical indices that gives high explanatory power. The first objective in this study was to confirm the relationship between CG and ERM. In Thai-listed companies across industries, there was a convergence finding for both quantitative and qualitative analyses that accounted for the strong relationship between CG and ERM. For the empirical data, the factor loading (standardized regression weight) between CG and ERM was significantly high (0.66 with p-value <0.001). All eight interviewees agreed that CG is a driving factor for high ERM.

The second objective related to the causality among CG, ERM and OP while ERM was a mediate variable. Based on the second & third hypotheses, for the quantitative data, both ERM and CG were found to be significantly associated with OP. Therefore, both ERM and CG are strategic tools to identify and capitalize on opportunity in business. Even CG and ERM both have positive relationships with OP, from the quantitative analysis, they are significantly improved only for non-financial OP, while the empirical data displayed an insignificant correlation with the financial OP (p-value=0.145>0.05) (appendix2). Fortunately, this finding is similar to the qualitative analysis that ERM and CG are associated with OP in terms of non-financial OP.

6 DISCUSSIONS AND POLICY RECOMMENDATIONS

Thai-listed companies aim to enhance managerial and shareholder performance as well as overcoming their competitors, while other sources of funds could be derived from creating new business arenas and inclining ROE to develop new shareholder OP. There are many strategic tools to improve OP, yet, often, business assurance aspects, like CG and ERM, are ignored. Previous studies quantified the tangible benefits of CG and ERM in terms of preventative tools, while this article tried to challenge the previous finding that CG and ERM could significantly improve the OP by promoting good decisions in new business arenas via a proactive strategy.

However, based on a mixed-method, both CG and ERM were perceived to have a low correlation in terms of stimulating organizations for long term growth due to an insignificant correlation to ROE. Importantly, the low correlations among CG, ERM and financial OP were derived from the low maturity level in CG and ERM in Thai-listed companies. Some organizations only conducted CG and ERM to comply with standards and regulations without understanding the other prospective benefits of them. Thai-listed companies then conducted CG and ERM only to align with regulators so they had low quality, low staff cooperation and low support from leaders. Therefore, it could be possible that if listed-companies undertook ERM and CG as an end-to-end process, they could generate long term growth by inclining ROE.

Moreover, this research identified the same finding as in prior studies about a strong correlation between CG and ERM. This means that board responsibility, disclosure and transparency are significantly associated with the implementation of ERM. Therefore, the role of the

governance system from the management to the BOD can determine the successful implementation of ERM.

For policy recommendations, the author will propose organizational strategies from the research findings as follows:

- One tangible benefit of ERM and CG is to manage opportunities by balancing risk versus opportunity. Hence, management should communicate and display such tangible benefits to related staff and the BOD to increase the level of cooperation as well as participation.
- Nowadays, most Thai-listed companies undertake CG and ERM in a piece-meal way: without coverage throughout organization or an end-to-end process while being conducted as individual projects, and this is why the empirical data showed a low level of correlation among CG, ERM and financial indicators. Accordingly, to stimulate growth and develop the maturity of CG and ERM, organizations should conduct CG and ERM following a top-down approach, as an end-to-end process and in a regular system.
- A strong correlation between CG and ERM displays that organizations should integrate these two systems to utilize resources, reduce silos and make a strong governance system.

7 LIMITATIONS AND FUTURE RESEARCHS

This research empirically studied only two concepts relating to improving OM: CG and ERM, while other factors can also lead to better managing opportunities. Future research should find other theories related to improving opportunity management. This research also found that risk management can somehow seize opportunity; therefore, how to determine an appropriate risk appetite for firms where they need to balance taking risks and seizing opportunity? Therefore, future research should focus on these two areas.

REFERENCES

Babbie, E. (2007). *The practice of social research (Vol. 11)*. Belmont, CA:Wadsworth. publishing company Committee of Sponsoring Organizations of the Treadway Commission. (2004). *Enterprise risk management-integrated framework: executive summary & framework*. American Institute of Certified Public Accountants (AICPA).

Aurini, J. D., Heath, M., & Howells, S. (2016). *The how to of qualitative research: Strategies for executing high quality projects*. SAGE Publications, Inc.

Barkus, E., Yavorsky, C., & Foster, J. (2006). Understanding and using advanced statistics. *Faculty of Health & Behavioural Sciences-Papers*, 393.

Fraser, J. R., Fraser, J., & Simkins, B. (2010). *Enterprise risk management: Today's leading research and best practices for tomorrow's executives* (Vol. 3). John Wiley & Sons.

Kumar, R. (2005). *Research Methodology: Step by step guide for beginners*. Thousand Oak: SAGE Publications, Inc.

Marchetti, A. M. (2012). *Enterprise Risk Management Best Practices*. New Jersey: JohnWiley& Sons.

Pallant, J. (2005). *SPSS Survival Manual (2nd ed.)*. London: McGraw Hill.

Rosen, R. (2003). Risk Management and Organizational Governance: The Case of Enron. *Conn. L. Rev.*, *35*, 1157-1180.

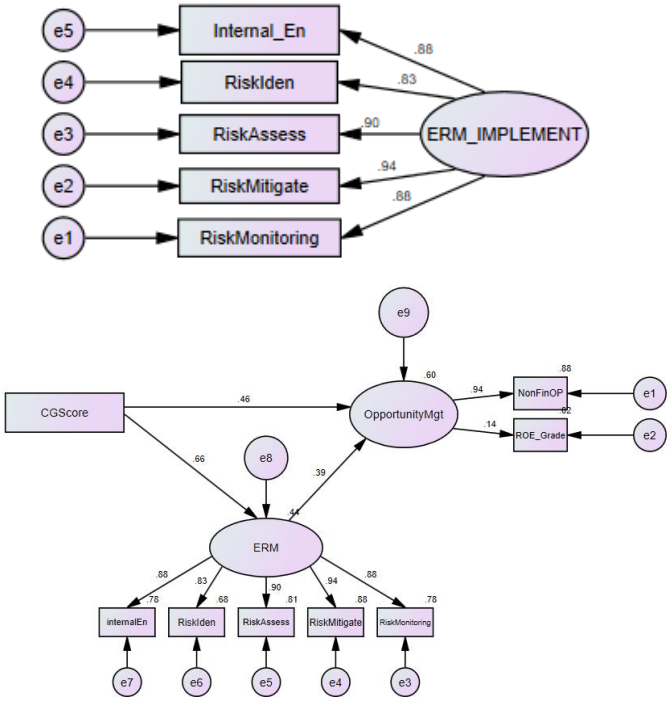
Ross, S.A., Westerfield, R.W., Jaffe, J., (2005). *Corporate Finance, 7th edition*. New York: McGraw Hill Irwin.

Gillan, S. L. (2006). Recent developments in corporate governance: An overview. *Journal of Corporate Finance*. 12(1), 117-130.

The Institute of Directors Association. (2017). *Corporate Governance Report of Thai* Listed Companies (CRG). Bangkok: Thai Institute of Directors Association.

OECD (2014), *Risk Management and Corporate Governance*, Corporate Governance, OECD Publishing. http://dx.doi.org/10.1787/9789264208636-en

APPENDIX



Modified Estimator for Right-Censored Data in Multiple Linear Regression Model

Sinjai Wisetdee* and Uthumporn Domthong

Department of Statistics, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand
*Corresponding Email: sinjai9wisetdee@gmail.com
Email: uthudo@kku.ac.th

ABSTRACT

The objective of this research is to propose a parameter estimation method for Type I right-censored data in multiple linear regression model. The proposed method, namely the Weighted Buckley and James (WBJ) method was developed from Buckley and James (BJ) method by using the Kaplan-Meier Weighted method for weighting. This appears to obtain an effective parameter estimation method. The WBJ method is compared the efficiency with two methods which are the BJ method and Smith (SM) method by using the simulated data set by the Monte Carlo simulation. The performance criterion of each method is an average mean square error (AMSE). The result indicates that the WBJ method has AMSE smaller than the compared methods in most situations.

Keywords: Multiple Linear Regression Model; Right-Censored Data; Parameter Estimation

1 INTRODUCTION

Linear regression analysis is a model that expresses the relationship between independent variables and response variable. Some data set in linear regression analysis is related to the lifetime study which often results in censored data. Censored data is one of characteristic of incomplete data. That is, we cannot determine the duration of a failure or event of genuine interest within the study period. This will affect the analysis and conclusion. In addition, if there is a large amount of censored data, the accuracy of the estimation can be reduced. In general, there are two types of right-censored data including Type I right-censored data which cause from fixed censoring time in advance, and Type II right-censored data which cause from fixed number of uncensored failure in advance. Censored data is usually found in medical and public health research that is often interested in survival time when given a drug to experimental units. The insurance industry has studied the life insurance policy which relate to survival time as well.

Researchers have studied the method of parameter estimation when response variable was censored in several ways. For example, Buckley and James (1979) presented a parameter estimation method for right-censored data in simple linear regression model. The Survival function proposed by Kaplan and Meier (1985) is used to calculate the weighted values for the least squares formula of parameter estimation. Smith (1986) presented the parameter estimation method for right-censored data in simple linear regression model, adapted from Buckley and James method using a constant value added to the parameter estimation. Ersin and Dursun (2017) compared multiple linear regression to linear regression when the response variable was censored between weighted least squares using Kaplan-Meier weights for weighting, and the synthetic data transformations using the Kaplan-Meier estimator to calculate the weighted values for the response variables. The researcher concludes that the parameter estimation method for Type I right-censored data in the multiple linear regression model by Buckley and James method, and Smith method give the unbiased estimators.

In this paper, we propose the parameter estimation method for Type I right-censored data in multiple linear regression model. Namely, the Weighted Buckley and James (WBJ) method was developed from Buckley and James (BJ) method using the Kaplan-Meier Weighted method for weighting. This appears to obtain an effective parameter estimation method. The WBJ method is compared the efficiency with two methods which are the BJ method and Smith (SM) method in various situations in term of sample size, censoring level of data.

We describe the estimation of linear model with right censored data and the parameter estimation method in section 2, and the Weighted Buckley and James Method and simulation study in section 3. We present results in section 4, and we summarize this finding in section 5.

2 ESTIMATION OF LINEAR MODEL WITH RIGHT CENSORED DATA

We consider the linear regression model

$$T_i = \beta_0 + X_{ij}\beta_j + \varepsilon_i \quad ;i = 1, 2, ..., n,\ j = 1, 2, ..., k \tag{1}$$

where X_{ij} are the independent variables of realized covariate which are fully observed, T_i are the response variables, β_0, β_j are the parameters to be estimated, and ε_i are independent and identically distributed with mean zero and constant variance.

In matrix form, the model (1) is given by

$$\begin{bmatrix} T_1 \\ T_2 \\ . \\ . \\ . \\ T_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & ... & X_{k1} \\ 1 & X_{11} & X_{21} & ... & X_{k1} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & X_{1n} & X_{2n} & ... & X_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix} \quad \text{or} \quad \mathbf{T} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \tag{2}$$

In practice, T_i may be incompletely observed and the right-censored by a variables at the location of the censored data (T_R). We consider the location of the censored data (R) from $R = P_{100-c}$ of the variables in data set by T_i where P_{100-c} is the location of the percentile at $100-c$, c is a censoring level of data, and the information at this position is T_R. In this case, instead of observing (T_i, X_i), we observe the data sets $\{(T_i, X_i, \delta_i), i = 1, 2, ..., n\}$ with

Define Y_i as the following:

$$Y_i = \begin{cases} T_i \ ; i < R \\ T_R \ ; i \geq R \end{cases} \quad ;i = 1, 2, ..., n \tag{3}$$

Let δ_i be the index of the censored data.

$$\delta_i = \begin{cases} 1 \ ; Y_i = T_i \\ 0 \ ; Y_i = T_R \end{cases} \quad ;i = 1, 2, ..., n \tag{4}$$

where i is the location of data and Y_i are the response variables of censored data.

In the procedure of estimation in censored data, it can be said that there are also two important assumptions:

I. T_i and X_i are independent.

II. T_i are the observed lifetimes, and X_i are quantitative variables.

We cannot be applied the ordinary least squares method for estimating the model (2) because variable Y_i includes censored observations influence the biased estimators. To overcome this problem, there are two popular methods such as Buckley and James (BJ) method

and Smith (SM) method. In this context, several important studies can be shown as follows:

2.1 Buckley and James method

Buckley and James (1979) presented the parameter estimation method for right-censored data in simple linear regression model. The survival function proposed by Kaplan-Meier (1985) is used to calculate the weighted values for the least squares formula of parameter estimation.

The parameter estimation procedure by the Buckley and James (BJ) method is performed as follows:

Step 1 The initial parameters are estimated by using the least squares method:

$$\hat{\beta}^u = \frac{\sum_{i=1}^{u} Y_i^u (X_i^u - \bar{X}^u)}{\sum_{i=1}^{u} (X_i^u - \bar{X}^u)^2} \tag{5}$$

$$\hat{\alpha}^u = \bar{Y}_i^u - \hat{\beta}^u \bar{X}^u \tag{6}$$

where u is a number of uncensored data, X_i^u are the independent variables of uncensored, Y_i^u are the response variables of uncensored for $i = 1, 2, ..., u$, $\bar{X}^u$ is a mean of independent variable of uncensored, and $\bar{Y}^u$ is a mean of response variable of uncensored.

Step 2 The uncensored data give the partial error, and define $\hat{\alpha}^u = 0$ as follows:

$$e_i = Y_i - \hat{\beta}^u X_i \tag{7}$$

where Y_i are the response variables of censored data, and X_i are the independent variables $; i = 1, 2, ..., n$.

Step 3 The partial error sort from the lowest number to the largest number.

Step 4 The survival function is derived from

$$\hat{S}_i = \prod_{l=1}^{i} \left(\frac{n-l}{n-l+1} \right)^{\delta(l)}, \; i = 1, 2, ..., n \tag{8}$$

where l are the rank of error values, n is a sample size $; n = r + u$, and r is a number of censored data. δ_i is the index of the censored data $; i = 1, 2, ..., n$ where $\delta_i = 1$ is for the uncensored data, and $\delta_i = 0$ is for the censored data.

Step 5 The survival functions are applied to the cumulative distribution function.

$$\hat{F}_i = 1 - \hat{S}_i \tag{9}$$

Step 6 The cumulative distribution functions are applied to the weighted values.

$$\begin{aligned} W_1 &= \hat{F}_1 \\ W_2 &= \hat{F}_2 - \hat{F}_1 \\ &\cdots \\ W_n &= \hat{F}_n - \hat{F}_{n-1} \end{aligned} \tag{10}$$

Step 7 We take the weighted value from step 6 to obtain Y_i^*

$$Y_i^* = \hat{\beta}^u X_i + \frac{\sum_{i=1}^{n} W_i (Y_i - \hat{\beta}^u X_i)}{1 - \hat{F}_i} \tag{11}$$

, and the parameter estimate by BJ method is as following:

$$\hat{\beta}_{BJ} = \frac{\sum_{i=1}^{u} Y_i (X_i - \bar{X}) + \sum_{i=u+1}^{n} Y_i^* (X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \tag{12}$$

where $\bar{X}$ is a mean of independent variable.

Step 8 The estimator is replaced in step 2, and iterate from step 2-8 until $\left| \hat{\beta}_{BJ}^{(m)} - \hat{\beta}_{BJ}^{(m+1)} \right| < 0.001$, then we stop the process where m is the number of iteration in the estimation of $\hat{\beta}_{BJ}$.

Step 9 We computed the fixed parameter estimator of the regression model $(\hat{\alpha}_{BJ}^*)$ from

$$\hat{\alpha}_{BJ}^* = \frac{1}{n} \{ \sum_{i=1}^{u} Y_i + \sum_{i=u+1}^{n} Y_i^* \} - (\hat{\beta}_{BJ} \bar{X}) \; . \tag{13}$$

Step 10 The regression model with the estimators from step 8 and 9 is

$$\hat{Y}_i = \hat{\alpha}_{BJ}^* + \hat{\beta}_{BJ} X_i \tag{14}$$

where $\hat{Y}_i$ are the predicted response variables as derived from Buckley and James method $i \; ; i = 1, 2, ..., n$.

2.2 Smith method

Smith (1986) presented the parameter estimation method for right-censored data in simple linear regression model. This method was adapted from Buckley and James (1979) using the constant value added to the parameter estimation.

The parameter estimation procedure by the Smith (SM) method is the same as for Buckley and James in steps 1-6. Then, step 7 is performed as follows:

Step 7 We take the weighted value from step 6 to obtain Y_i^*

$$Y_i^* = \hat{\beta}^u X_i + \frac{\sum_{i=1}^{n} W_i (Y_i - \hat{\beta}^u X_i)}{1 - \hat{F}_i} \tag{15}$$

then, the predictive value was calculated from the response variable of the censored data.

$$Y_i^{**} = Y_i \delta_i + Y_i^* (1 - \delta_i) \tag{16}$$

Thus, we have $$\rho_i = 1 + [Y_i - \hat{\beta}^u X_i + Y_i^{**}] \tag{17}$$

and the parameter estimate by SM method is as following:

$$\hat{\beta}_{SM} = \frac{\sum_{i=1}^{u} (X_i - \bar{X}) [((\hat{S}_i \rho_i) - 1)]}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \tag{18}$$

Then, perform steps 8-10 as the procedure of BJ method in section 2.1.

3 METHODS

3.1 Weighted Buckley and James Method

The Weighted Buckley and James (WBJ) method was developed from Buckley and James (BJ) method using the Kaplan-Meier Weighted method for weighting.

The parameter estimation procedure by the WBJ method is performed as follows:

Step 1 We estimate initial parameters using the least squares method:

$$\hat{\boldsymbol{\beta}}^u = (\mathbf{X}^{u'} \mathbf{X}^u)^{-1} (\mathbf{X}^{u'} \mathbf{Y}^u) \tag{19}$$

where $\hat{\boldsymbol{\beta}}^u$ is a (4×1) parameter vector to be estimated of uncensored data, $\mathbf{X}^u$ is a $(u \times 4)$ independent variable matrix of uncensored, $\mathbf{Y}^u$ is a $(u \times 1)$ response variable vector of uncensored, r is a number of censored data, u is a number of uncensored data and n is a sample size where $n = r + u$

Step 2 The uncensored data give the partial error, and define $\hat{\alpha}^u = 0$ as follows:

$$e_i = Y_i - \hat{\beta}_1^u X_{i1} - \hat{\beta}_2^u X_{i2} - \hat{\beta}_3^u X_{i3} \tag{20}$$

where X_{ij} are the independent variables $i = 1, 2, ..., n, \; j = 1, 2, 3$.

Step 3 The partial error sort from the lowest number to the largest number.

Step 4 We find W_i by using Kaplan-Meier Weights.

$$W_1 = \frac{\delta_{(i)}}{n-l+1} \prod_{l=1}^{i-1} \left(\frac{n-l}{n-l+1} \right)^{\delta_{(l)}}, \; i = 2, 3, ..., n \tag{21}$$

$$W_1 = \frac{\delta_1}{n}$$

Step 5 We take the weighted value from step 4 to obtain Y_i^*

$$Y_i^* = \hat{\beta}_1^u X_{i1} + \hat{\beta}_2^u X_{i2} + \hat{\beta}_3^u X_{i3} + W_i \quad (22)$$

, and the parameter estimate by WBJ method is as following:

$$\hat{\beta}_{WBJ_j} = \frac{\sum_{i=1}^{u} Y_i (X_{ij} - \bar{X}_j) + \sum_{i=u+1}^{n} Y_i^* (X_{ij} - \bar{X}_j)}{\sum_{i=1}^{n} (X_{ij} - \bar{X}_j)^2} \quad (23)$$

where $\bar{X}_j$ are the mean of independent variables j ; $j = 1, 2, 3$.

Step 6 The estimator is replaced in step 2, and iterate from step 2-6 until $\left| \hat{\beta}_{WBJ_j}^{(m)} - \hat{\beta}_{WBJ_j}^{(m+1)} \right| < 0.001$ for $j = 1, 2, 3$, then we stop the process where m is the number of iterate in the estimation $\hat{\beta}_{WBJ}$.

Step 7 We computed the fixed parameter estimator of the regression $(\hat{\alpha}_{BJ}^*)$ from

$$\hat{\alpha}_{WBJ}^* = \frac{1}{n} \{ \sum_{i=1}^{u} Y_i + \sum_{i=u+1}^{n} Y_i^* \} - (\hat{\beta}_{WBJ_1} \bar{X}_1 + \hat{\beta}_{WBJ_2} \bar{X}_2 + \hat{\beta}_{WBJ_3} \bar{X}_3) \quad (24)$$

Step 8 The regression model with the estimators from step 6 and 7 is

$$\hat{Y}_i = \hat{\alpha}_{WBJ}^* + \hat{\beta}_{WBJ_1} X_{i1} + \hat{\beta}_{WBJ_2} X_{i2} + \hat{\beta}_{WBJ_3} X_{i3} \quad (25)$$

where $\hat{Y}_i$ are the predicted response variables as derived from Weighted Buckley and James method i ; $i = 1, 2, ..., n$.

To gain some understanding of how well the mentioned methods work, we obtained the predicted response variables obtained by the BJ method, SM method and WBJ method under the four different censoring levels. Moreover, in order to assess the performance of the regression parameters, we use mean squared error (MSE) and average mean square error (AMSE) which can be calculated as,

$$\text{MSE}_t = \frac{\sum_{i=1}^{n} (T_{ti} - \hat{Y}_{ti})^2}{n} \quad ; i = 1, 2, ..., n \quad ; t = 1, 2, ..., L \quad (26)$$

$$\text{AMSE} = \frac{\sum_{i=1}^{L} MSE_t}{L} \quad ; \text{respectively} \quad (27)$$

where T_{ti} are the response variables before censored, $\hat{Y}_{ti}$ are the predicted values of the response variables, L is a number of iteration, MSE_t are the mean squared errors of the predicted response variables, and $AMSE$ is a average mean square error of the predicted response variables.

3.2 Simulation Study

We carried out a simulation study to compare the performance of three methods which are BJ method, SM method and WBJ method. There are 1000 simulation runs $(L = 1000)$ in four different sample sizes $n = 20, 50, 100, 250$ and censoring levels $c = 5\%, 10\%, 25\%, 40\%$. The simulation process is as follows:

Step 1 We generate $X_{i1}, X_{i2}, X_{i3}, \varepsilon_i, i = 1, 2, ..., n$ from $X_{i1} \sim N(50,5)$, $X_{i2} \sim \exp(1)$, $X_{i3} \sim \exp(2)$ and $\varepsilon_i \sim N(0, 1)$ respectively.

Step 2 The response variables T_i are computed according to a censored multiple linear regression model

$$T_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \quad ; i = 1, 2, ..., n \quad (28)$$

where $\alpha = 6.63$, $\beta_1 = -0.09$, $\beta_2 = 0.15$, $\beta_3 = 0.06$.

The correlation coefficients between this three independent variables are test by hypothesis test of correlation. We fail to reject the null hypothesis that is the population correlation coefficient equals 0. That is independent variables do not correlate among all three variables.

Step 3 Determine the variables at the location of the censored data (T_R) . We consider the location of the censored data from $R = P_{100-c}$ of the variable data set by T_i where c is the censoring level of data at the predetermined, and the information at this position is T_R .

Step 4 Define Y_i as the following:

$$Y_i = \begin{cases} T_i \; ; i < R \\ T_R \; ; i \geq R \end{cases} \quad ; i = 1, 2, ..., n \quad (29)$$

Let δ_i be the index of the censored data.

$$\delta_i = \begin{cases} 1 \; ; Y_i = T_i \\ 0 \; ; Y_i = T_R \end{cases} \quad ; i = 1, 2, ..., n \quad (30)$$

where i is the location of data, and Y_i are the response variables of censored data.

Step 5 We replace estimator $Y_i, X_{i1}, X_{i2}, X_{i3}$ to parameter estimate from multiple linear regression models by BJ method, SM method and WBJ method, and apply to the regression model to predict the value of the response variable.

Step 6 The mean square errors (MSE) are computed for predicted response variables in all three methods.

Step 7 We do steps 1-6 until the iteration L .

Step 8 The average mean square errors (AMSE) are computed for the predicted response variables, then the AMSE was obtained to compared the efficiency among with three methods.

4 RESULTS

We consider the results to compare the performance of the proposed method, the Weighted Buckley and James (WBJ) method with two other methods which are Buckley and James (BJ) method, and Smith (SM) method based on multiple linear regression model. Results are summarized in Table 1 that is obtained from 1000 simulated data sets for each of sample size.

Table 1: AMSE of the predicted response variables for comparing the BJ method against the SM method and WBJ method for each of sample size

SAMPLE SIZE (n)	CENSORING LEVEL (c)	METHOD		
		BJ	SM	WBJ
20	5	1.745164	2.602128	**1.738794**
	10	1.822769	2.592627	**1.814571**
	25	2.177049	2.563205	**2.163901**
	40	2.740161	**2.469241**	2.726279
50	5	1.519240	6.468427	**1.517165**
	10	1.575361	6.204307	**1.573417**
	25	1.828830	3.536587	**1.820504**
	40	2.362435	2.886356	**2.354043**
100	5	1.483347	7.766724	**1.481595**
	10	1.512707	6.959025	**1.508147**
	25	1.769666	3.851242	**1.764586**
	40	2.261319	3.010281	**2.252690**
250	5	1.464766	13.826271	**1.462846**
	10	1.487494	7.065841	**1.485306**
	25	1.715469	5.966119	**1.710157**
	40	2.205418	3.663860	**2.197949**

Table 1 shows AMSE for comparing the BJ method against the SM method and WBJ method. The results show that for all sample size, the WBJ method gives higher AMSE when the censoring levels of data increases. The WBJ method has smaller AMSE than the compared methods in most situations, except the SM method has smallest AMSE for sample size equal to 20 and high censoring level of data $c = 40\%$.

5 CONCLUSIONS

In this study, we proposed a parameter estimation method for Type I right-censored data in multiple linear regression model, namely the Weighted Buckley and James (WBJ) method. The WBJ method is proposed for predicted response variables. The performance criterion of each method is an average mean square error (AMSE). The result indicates that for all sample size, the WBJ method gives higher AMSE when the censoring levels of data increases. Moreover, the WBJ method has AMSE smaller than the compared methods in most situations. So, the WBJ method is efficient parameter estimation method for Type I right-censored data in multiple linear regression model.

REFERENCES

Buckley, J., & James, I. (1979). Linear regression with censored data. *Biometrika*, *66*(3), 429-436.

Ersin, Y., & Dursun, A. (2017). A Comparison of two methods for estimating censored linear regression models. *International Journal of Statistics in Medical and Biological Research*, 1, 1-8.

Schneider, H., & Weissfeld, L. (1986). Estimation in linear models with censored data. *Biometrika*, *73*(3), 741-745.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, *53*(282), 457-481.

Smith, P. J. (1986). Estimation in linear regression with censored response. In *Pacific Statistical Congress, Amsterdam, Holland* (pp. 261-265).

Weisberg, S. (2005). *Applied Linear Regression* (Vol. 528). John Wiley & Sons.

Testing the Accuracy of Paddy Productivity Data to Support Indonesian Food Tenacity (Case Study in Subang Regency, West Java Province)

Yulianto Antonius[1*], Suryanto Aloysius[2] and Risni Julaeni[3]

[1]Institute of Statistics, Jakarta, Indonesia
*Corresponding Email: yuliantoyorki@yahoo.com
[2] Institute of Statistics, Jakarta, Indonesia
Email: suryanto@stis.ac.id
[3] Institute of Statistics, Jakarta, Indonesia
Email: risnij@stis.ac.id

ABSTRACT

Indonesia is a country with the fourth largest population in the world where the majority of the population consumes rice as its staple food. Indonesian people need approximately 82.0 million tons of paddy in the year of 2019. Stability in providing rice is one of the most important things to create a stable social-economic and political condition. Paddy production is estimated by BPS-Statistics Indonesia cooperated with the Indonesian Ministry of Agriculture (MoA). MoA estimates paddy field areas using eye estimate, and BPS together with MoA estimate paddy productivity using a crop cutting technique that called *ubinan*. By multiplying those two information, we get paddy production estimation. *Ubinan* is conducted by using a square equipment consisted of twelve related sticks and a scale. The weight of *ubinan*'s equipment is about twelve kg. In addition to its heavy equipment, field officer should also follow the *ubinan*'s procedure included how to get *ubinan*'s plots randomly. By following the right procedure then the productivity estimation will have minimum error or in other words we can get data estimation that could represent the real condition. Bearing this in mind, we conducted research in Subang-regency to see whether the data got from field officers could represent the real condition. By using paired sample t-test, researchers concluded that there are no differences between *ubinan* results did by field officers a year ago and by officers under researcher's supervision. It means that field officers had done *ubinan* using the right procedures. Implicitly, paddy productivity estimation could represent the real condition. By using independent sample t-test, researchers concluded that there are no differences between *ubinan* results did by field officer from BPS and from MoA. It means that all field officers either from BPS or from MoA had done *ubinan* using the same procedures.

Keywords: crop cutting; productivity; t-test.

1 INTRODUCTION

1.1 Background

Indonesia is a country that its people are depended on rice as their main food. So if we talk about Indonesian food tenacity then we talk about rice production respectively which is it should be enough for all Indonesian people. Based on Indonesian development planning in food and agriculture for 2015-2019, Indonesian people would need about 82.0 million tons of paddy in the year of 2019, 24.1 million tons of corn, and 1.92 million tons of soybean (RPJMN 2015-2019 page 8). Although Indonesian government has had food diversification policy, rice is still as main food for Indonesian people. About 95 percent Indonesian people are still consuming rice as their main food. For Indonesian people, rice is not only as the main food commodity but it also as income source for about 21 millions households (Suryana, 2002 in Dewa, 2007). Even, eating rice becomes food habit for Indonesian people. They think that they do not eat yet if they do not eat rice yet. Based on Indonesian social economic survey in 2013, rice consumption per capita for Indonesian people is about 85.51 kg a year. It is higher than rice standard consumption per capita from FAO (Food and Agriculture Organization) which is about 60 to 65 kg a year.

Food position in development era has very strategy position, we could not delay to provide it, and we could not substitute it with other material as well. Providing enough food for people is an important part in national development to prevent socio-economic and political stability. So, having accurate data related to provide rice production data is very important.

Data are the basic information for government to make policy. Using good and accurate data, government would have good policy as well. Furthermore, having good and accurate food production data, government would have good policy in food sector and it would have strong tenacity food at the end.

Paddy production could be improved through extensification such as makes cultivation area wider, rehabilitates irrigation facilities. It also could be done by increasing its productivity. Productivity itself is one variable to estimate paddy production besides harvesting areas. Till now, Paddy production is estimated by BPS-Statistics Indonesia cooperated with the Indonesian Ministry of Agriculture (MoA). MoA estimates paddy field areas using eye estimate, and BPS together with MoA estimate paddy productivity using a crop cutting technique, which is called *ubinan*. By multiplying those two information, we get paddy production estimation. Ubinan tool has measurement in square form 2.5 m x 2.5 m by connecting 12 pipe sticks one to the other. If field officers do ubinan following right procedure, then paddy productivity estimation would not have leaning over or under estimation. However, the heavy tool is one possibility that could make data inaccurate (based on report of comparative study done by Japan International Cooperation Agency (JICA) and MoA), ubinan plot samples that located far away form the road which make field officers should do walking to reach that location is also barrier to get accurate data. Based on those conditions, it's looked that researchers need to do research to test the accuracy of paddy productivity data.

1.2 Problem identification

Based on explanation in background above, there is tendency that some field officers conducting ubinan plot do not follow ubinan procedure because of the heaviness of ubinan tool. Some field officers have tendency to increase production in ubinan as well. So researchers have questions as follows:

1) Do ubinan results conducted by field officers and ubinan results controlled by research team have the same result in average?
2) Do ubinan results conducted by field officers from BPS and MoA have the same result in average?

1.3 Research goals and benefits

Research goals:

1) To identify the mean difference between ubinan results conducted by field officers and ubinan results controlled by research team.
2) To identify the mean difference between ubinan results conducted by two groups of field officers those are between field officers from BPS and MoA.

Research benefits:

1) It could give information and fact for development of production statistics and agriculture in the future, so

Indonesian agriculture sector would have higher competitive power and food self-sufficient that could be achieved.

2) It could give knowledge and information for BPS Indonesia and Ministry of Agriculture so that it could be useful for the people in general.

2 THEORY

2.1 Data productivity collecting

Data productivity collecting of food crops have been done using sample through "ubinan" survey by household approach. Food crops are wet land paddy, dry land paddy, corn, soybean, ground nuts, green nuts, cassava, and sweet potatoes. In this research, we only use paddy crop. Data productivity collecting method uses direct measurement to the sample plots of ubinan and interview to the selected farmers asking for paddy planting related characteristics such as the usage of fertilizer, seeds, pesticide, and others.

Ubinan survey is conducted every year. Its implementation is divided into three periods or sub rounds; those are sub round I (January-April), sub round II (May-June), and sub round III (September-December). Ubinan samples are chosen from the agriculture households that have paddy plantation in a certain sub round. So before samples are chosen, household listing is conducted in every chosen census blocks in the last month before sub round. Census block is part of village as working sample area consisted of 80 to 120 households that has clear natural boundary such as river, train road, and etc. The sample unit of ubinan survey is the agriculture household that has paddy harvesting.

Paddy production data is estimated by multiplying harvested area with production per hectare (productivity). Harvested area data is estimated from agriculture survey reports conducted by field officers from agriculture ministry and productivity data is estimated using ubinan survey conducted by field officers from BPS and agriculture ministry.

2.2 How to do Ubinan

There are several things that should be understood by field officers in doing ubinan, those are about ubinan tool, ubinan procedure, and farmer interviewing related to the paddy production in the paddy area sample where ubinan plot is conducted.

2.2.1 Ubinan tool

The whole ubinan tool is consisted of ubinan tool its self, scale with tripod to hold the scale in the field, and an ubinan bag. Ubinan tool is a square form tool consisted from 12 sticks of measurement pipes and 4 iron pin that used for holding or fasten every angel curve pipe. 12 measurement pipes are consisted of 4 top pipe sticks, 4 middle pipe sticks, and 4 base (starting point) pipe sticks. It also has 4 angle curve pipes. Every pipe could be connected to each other so it would make square shape that has 2.5 meter by 2.5 meter in area.

2.2.1.1 Scale and tripod

Scaling tool is used to weight paddy seed got from ubinan. Weighting should be done in high carefulness to get high accuracy in productivity data (scale provides measurement level till gram in weigh). Tripod is used to hold the scale in the field, so field officers could do measurement in high accuracy.

2.2.1.2 Ubinan bag

This bag is used as a place to put ubinan tool, scale, and tripod. The purpose is to make field officers easier to bring the whole ubinan tool in the field. This bag is made from material that could not be emerged by water.

2.2.2 Ubinan procedure

Ubinan plots are conducted by field officers from BPS and ministry of agriculture in district level. Steps in doing ubinan survey are as follow :

- Deciding starting point in paddy field plot
- Deciding starting point of ubinan plot
- Doing ubinan plot (paddy crops cutting)
- Thrashing and cleaning of paddy seeds
- Weighting the clean paddy seeds
- Interviewing farmer

The first step in doing ubinan is field officer decides starting point in paddy field plot. Starting point is in south-west corner of the plot. The second step, field officer decides starting point of ubinan plot. How to do it, firstly field officer does stepping forward on the side of paddy field plot start from west to the east then from south to the north. Secondly field officer decides ubinan starting point randomly by using random table. After getting starting point of ubinan plot, field officer starts to do ubinan plot procedurely using ubinan tool. Furthermore, paddy clumps in ubinan plot is harvested/cutted by farmer as he/she does as usual (harvesting is not done by field officer). After paddy be harvested, field officer do counting how many number of paddy clumps in ubinan plot.

Thrashing and cleaning of paddy seeds got from ubinan plot is conducted by farmer soon after paddy being harvested as he/she does as usual (thrashing and cleaning is not done by field officer). The last but not least in doing ubinan is to do weighting (scaling) of paddy production that have been cleaned by the farmer. Weighting should be done in high carefulness, because having accurate measurement we could get high accuracy estimation in productivity data of paddy. Weighting should be done by field officer.

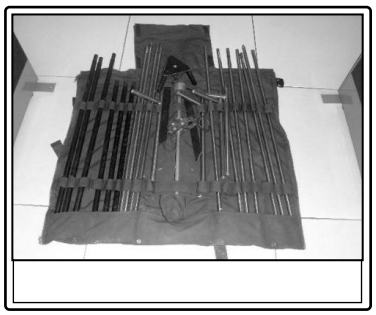


Figure 1: a set of ubinan tool

After finishing all ubinan working till weighting, field officer does interview to farmer (the owner of paddy field) to get the supporting information in cultivating paddy such as the variety of paddy, the usage of fertilizer, pesticide, and so on.

3 METHODS

3.1 Method analysis

Research location was in Subang regency, West Java. Subang is one of paddy production center in West Java. Sample of ubinan plots in this research are subsample that were selected from last year sample of Subang's ubinan plots from BPS Indonesia. From three sub rounds provided in a year, researchers only used two sub rounds those are sub round III (September-December) 2013 and sub round I (January-April) 2014 because of limitation of time and research expenses. The reason why using subsample from last year sample of Subang's ubinan plots because researchers want to compare the ubinan results got from field officer and ubinan results controlled by researchers in the same plots. Paired samples t-test was used to answer the research question. Plot sample used in this research were selected randomly from Subang's ubinan plot sample one year ago (only sub round III and I sample) using systematic linear random sampling. 13 plots sample were chosen.

Paired sample t-test is used in this research to see the significance of paired sample mean between ubinan results got from field officer and ubinan results controlled by researchers in the same plots. Hypothesis in paired sample t-test is as follows:

$$H_0 : \mu_d = 0 \quad vs \quad H_1 : \mu_d \neq 0 \quad \text{(for two ways testing)}$$
$$H_1 : \mu_d < 0 \quad \text{(for one way testing)}$$
$$H_1 : \mu_d > 0 \quad \text{(for one way testing)}$$

In this research, we only use two ways testing. Statistical testing is t-test with formula as follows:

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} \sim t_{(n-1)} \qquad (1)$$

$\bar{d} = \frac{1}{n}\sum d_i \quad ; \quad d_i = (x_{1i} - x_{2i})$

$S_d^2 = \frac{1}{n-1}(\sum d_i^2 - n\bar{d}^2)$

n = number of paired samples

x_{1i} = the i observation of the first group (ubinan result got from field officer)

x_{2i} = the i observation of the second group (ubinan result controlled by researcher)

Rejection area for two ways testing is as follows:

if $t > t_{(\alpha/2,n-1)}$ or $t < -t_{(\alpha/2,n-1)}$

Rejection area for one way testing is as follows:

if $t > t_{(\alpha,n-1)}$ and $t < -t_{(\alpha,n-1)}$

this testing has assumption normality in different data (d_i), violation to this assumption then we can not use this statistical testing as the method analysis. As its replacement, we could use a nonparametric statistical method that theoretically it doesn't need any assumptions. That statistical method is Wilcoxon signed rank sum test.

Wilcoxon signed rank sum test can be useful to see the comparation of the paired sample data. The data could be in ranking or in quantitative but not normal. It uses differences between each paired data (d_i). Let d_i be the difference score for any matched pair, represetating the difference between the pair's scores under two treatments X and Y. That is, $d_i = X_i - Y_i$. If $d_i = 0$, such pairs are dropped from the analysis and the sample size is reduced accordingly. Rank all of the d_i's without regard to sign; give the rank of 1 to the smallest Id_iI, the rank of 2 to the next smallest, etc. when two or more d_i have the same magnitude, the rank assigned is the average of the ranks which would have been assigned. After every d_i has its rank, then to each rank affix the sign of the difference. That is, indicate which ranks arose from negative d_i's and which ranks arose from positive d_i's. Thus, we sum those ranks having plus sign (T^+) and sum those ranks having minus sign (T^-). when H_0 is true, we would expect the two sums to be about equal. But if the sum of the positive ranks is very much different from the sum of the negative ranks, we would infer that treatment X differs from treatment Y, and thus we would reject H_0. For small samples ($n \leq 25$), to decide either accepting or rejecting H_0 we use statistical test T. T is equal sum of positives ranks or negatives ranks that has smaller value. If $T \leq$ value from Wilcoxon table in a certained α thus we would reject H_0. For large samples, when n is larger than 25 then Wilcoxon table can no longer be used. However, it can be shown that in such cases the sum of ranks is approximately normally distributed. In this research, because our samples are 13 thus we use Wilcoxon table to decide rejecting H_0.

Independent samples t-test in this research is used to see the significance of independent sample mean between ubinan results done by field officers from BPS and MoA. The reason why researchers want to compare ubinan results from those two groups, we want to see if there is no tendecious to increase ubinan results and they follow the right ubinan procedures that have been made. this testing requires that the variables used are approximately normally distributed and assumes the variances of the two groups used are equal in the population (homoscedasticity of variance). violation to these assumption either one or both of the group data, then we can not use this statistical testing as the method analysis. As its replacement, we could use a nonparametric statistical method that theoretically it doesn't need any assumptions. That statistical method is Mann-Whitney U test. To test normality data, researchers use the Shapiro-Wilks testing; and to test homoscedasticity of variance, researchers use the Levene test of equality of variances.

The null hypothesis for the independent t-test is that the population means from the two unrelated groups are equal, as follows:

$H_0 : \mu_1 - \mu_2 = 0$ *vs* $H_1 : \mu_1 - \mu_2 \neq 0$ (for two ways testing)

$H_1 : \mu_1 - \mu_2 < 0$ (for one way testing)

$H_1 : \mu_1 - \mu_2 > 0$ (for one way testing)

when variances population (σ_1^2 and σ_2^2) are not known assuming they are both the same and sample is small, the test statistics are as follows:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{Sp\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (2)$$

Where:

$$Sp = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

And df = $n_1 + n_2 - 2$

Sp is pooled standard deviation

Rejection area for two ways testing is as follows:

if $t > t_{(\alpha/2,n1+n2-2)}$ or $t < -t_{(\alpha/2,n1+n2-2)}$

Rejection area for one way testing is as follows:

if $t > t_{(\alpha,n1+n2-2)}$ and $t < -t_{(\alpha,n1+n2-2)}$

Mann-Whitney U test could be used in this research when there is an assumption violation in using independent samples t-test. Using this method, it decides that n_1 is the number of observation from the smaller group in those two independent groups; and n_2 is the number of observation from the larger group accordingly.

Say we have two samples from two populations, A and B. then the Hypothesis:

H_0 : A and B have the same distribution.

H_1 : A and B have not the same distribution.

Steps of testing:

- firstly, all observations from those two groups are merged
- then observations are ranked from the smallest score till the biggest (if there is an observation with the biggest negative score, it has the first ranking). If there are the same scores, then they have the same ranking by making them averaged.
- Calculate the U value.

For small sample ($n_2 < 9$), The value of U is calculated by counting how many score from the smaller group precedes the score from larger group. For example: we have two independent groups, A and B. A has three observations 9, 11, 15 and B has four observations 6, 8, 10, 13; so $n_1 = 3$ and $n_2 = 4$. By these kind of data, the value of U is 3. One score from A preceded score 10 from B, and two scores from A preceded score 13 from B. By using n_1, n_2, and U values we can get p-value from Mann-Whitney table for small sample. With significant α, we can conclude to reject H_0 if p-value $< \alpha$.

For medium sample ($9 \leq n_2 \leq 20$), The value of U is calculated using the formula as follows:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \quad or \quad (3)$$

$$U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \quad (4)$$

R_1 is the total ranking from smaller group (n_1)

R_2 is the total ranking from larger group (n_2)

The smaller value of U is used as statistics testing. There is a relationship between those two U's as follows:

$$U = n_1 n_2 - U^I$$

if the value of U calculation is less than the value of U from table than we decide to reject H_0 in a certain significant level (α).

For large sample ($n_2 > 20$), the statistics analysis could be approximately estimated by standard normal distribution Z using the formula as follows:

$$Z = \frac{U - \mu_U}{\sigma_U}$$

where: $\mu_u = \frac{1}{2}(n_1 . n_2)$ and

$\sigma_u^2 = 1/12 \{n_1 . n_2 (n_1 + n_2 + 1)\}$

In this research, because our samples are 13 thus we use Mann-Whitney table to decide rejecting H_0.

Shapiro-Wilk test is used to test normality data. The reason why this test is used in this research, because it is suitable for small simple ($n < 30$) besides it has more simple formula compare to other normality testing. Its formula is as follows:

$$W = \frac{b^2}{SS} \quad (5)$$

Where:

$$b = \sum_{i=1}^{m} a_i (x_{n+1-i} - x_i)$$

$$SS = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

n is the total sample
m = n/2 if n is even
m = (n-1)/2 if n is odd

a_i is the weighting score from Shapiro-Wilk table1

Hypothesis in Shapiro-Wilk test is as follows:
Ho: data are normally distributed
H_1: data are not normally distributed
From the W value, we can get p-value from Shapiro-Wilk table2. if p-value is less than α, we decide to reject H_0.

Levene test is used to test homoscedasticity data. The null hypothesis for this test is that the population variances from the two or more unrelated groups are equal, as follows:
$H_0 : \sigma_1 = \sigma_2 = \ldots\ldots = \sigma_k$
$H_1 : \sigma_i \neq \sigma_j$ for at least one pair (i,j).
The formula of Levene test is

$$F = \frac{(n-k)\sum_{i=1}^{k} n_i (\bar{Z}_{i.} - \bar{Z}_{..})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_{i.})^2} \quad (6)$$

Where:

n is the number of total observations
k is the number of groups
$Z_{ij} = |Y_{ij} - \bar{Y}_{i.}|$
$\bar{Y}_{i.}$ is the average of groups i
$\bar{Z}_{i.}$ is the average of groups $Z_{i.}$
$\bar{Z}_{..}$ is the overall average of Z_{ij}

Rejection area for this testing is as follows:
if $F > F_{(\alpha;k-1,n-k)}$

This research covers paddy fields which are ready to be harvested in Subang regency, West Java province. Sample ubinan plots in this research are subsample ubinan plots a year ago from Subang regency. Collecting data uses direct research doing harvesting to the paddy plots. Field research was conducted by team supported by field officers using formal procedure under supervising research team leader.

Data were processed using SPSS and Excel in Institute of Statistics. So decision in rejecting hypothesis was based on statistical testing for the final result.

4 RESULTS

13 agriculture households from ubinan samples chosen as samples in this research would be as ubinan control, one agriculture household will give one ubinan plot. And 13 ubinan regular which are conducted a year before as samples are seven plots from MoA's field officers and six plots are from BPS field officers. The data are as follow (in Kgs):

Table 1: ubinan control and ubinan regular results (in Kg)

Sample Number	UBINAN CONTROL	UBINAN REGULAR
1	4.69	5.50 (KCD)
2	5.15	6.00 (KCD)
3	3.41	5.35 (KSK)
4	2.30	2.50 (KCD)
5	5.15	3.52 (KSK)
6	2.19	3.87 (KSK)
7	3.97	4.12 (KSK)
8	4.17	3.23 (KSK)
9	3.15	3.60 (KSK)
10	4.80	4.60 (KCD)
11	4.07	4.21 (KCD)
12	4.83	4.41 (KCD)
13	4.84	3.90 (KCD)

4.1 Normality testing

The distribution of data difference between ubinan regular results conducted by field officers and ubinan control results supervised by researchers is followed normal distribution with 5% significant level (see **Table 2** below). From Shapiro-Wilk testing we got p-value 0.984, it means that paired samples t-test can be used to test the mean difference between those two groups of ubinan results

Table 2: Normality testing of data difference between ubinan regular and ubinan control

	Shapiro-Wilk		
	Statistic W	n	Sig.
Diff	0.981	13	0.984

Moreover, the distributions of those two group ubinan results data conducted by field officers from BPS and conducted by field officers from MoA followed normal distribution with 5% significant level (see **Table 3** below). From Shapiro-Wilk testing we got p-value 0.739 and 0.181, both of them mean that independent samples t-test can be used to test the mean difference between those two groups of ubinan results.

Table 3: Normality testing of ubinan results conducted by KCD (officers from MoA) and KSK (officers from BPS)

	Shapiro-Wilk		
	Statistic W	n	Sig.
KCD	0.962	7	0.839
KSK	0.858	6	0.181

4.2 paired samples t test

This testing is used to test whether ubinan results conducted by field officers are different in average with ubinan control results supervised by researchers. From paired samples t-test, we got t-value 0.683 (see **Table 4** below), and we got $t_{(0.025,12)}$ from t-table 2.179. It can be concluded that there is no different between those two ubinan results group conducted by field officers and supervised by researchers with 5% significant level.

Table 4: Paired Samples Test

		Pair 1
		Regular – control
Paired Differences	Mean	0.18846
	Std. Deviation	0.99502
	Std. Error Mean	0.27597
t		0.683
df		12

The result from Levene's test showed that there is not significantly different in variances of those two groups between ubinan conducted by field officers from BPS and MoA with 5% significant level. From **Table 5** below we can see that the value of F calculation is 0.554.

Table 5: Test of homoscedasticity of variance

		Ubinan result
		Equal variances assumed
Levene's Test for Equality of Variances	F	0.554
	Sig.	0.472

The result from t test between ubinan conducted by field officers from BPS and MoA showed that there is no significantly different in the average of ubinan results from those two groups with 5% significant level. From **Table 6** below we can see that the value of t calculation is 0.914.

Table 6: Independent Samples Test

		Ubinan result
		Equal variances assumed
t-test for Equality of Means	t	0.914
	df	11
	Mean Difference	0.49738
	Std. Error Difference	0.54388

5 CONCLUSIONS AND SUGGESTIONS

5.1 Conclusions

The difference of paired sample mean between ubinan results got from field officer and ubinan results controlled by researchers is not statistically significant with 5% significant level. It could be indication that field officers have already followed procedure in doing ubinan; in other word, data of paddy productivity in Subang regency represent real field condition.

Furthermore, ubinan results conducted by field officers from BPS and MoA in average are not statistically different with 5% significant level.

5.2 Suggestions

working spirit of field officers and the way they do paddy ubinan could be hold as usual. However, upgrading of their working quality could be done to refresh their knowledge and to remain them about updating procedure and methodology by giving them trainings such as ubinan training especially for new field officers.

Supervising and controlling to them (field officers) in doing field workings could be done continuously to reduce improperly mistakes.

Government needs to have integrated field activities schedules among surveys, so field officers could be more concentrate when they do some working (survey).

ACKNOWLEDGMENTS

This research was financially suported by Institute of Statistics, Jakarta. We would like to thank to the Head of Institute of Statistics who has already supported this research. We would also like to thank to the Head of Subang BPS office who has already let his staffs and field officers to help us in the field. and finally we would thank to our colleagues who have already provided their expertise to support this research and some of my students from Institute of Statitics who have already helped this research.

REFERENCES

Keller, G., Warrack, B. (2000). Statistics for Management and Economics, fifth edition. *Duxbury-CA 93950 USA*.

Maksum, C., Yulianto, A., ect. (2008). Production Statistics. *Institute of Statistics, Jakarta*.

Niino, K., Muroi, T., ect. (2001). The Agriculture Statistical Technology Improvement and Training Project (ASTIT) in Indonesia. *Final Report of Research Cooperation Among BPS Statistics Indonesia, Ministry of Agriculture Indonesia, and JICA (Japan International Cooperation Agency) Japan*.

Rusono, N., Sunari, A,, ect. (2014). *Penyususnan RPJMN 2015-2019 bidang pangan dan pertanian, Direktorat pangan dan pertanian Kementerian Perencanaan Pembangunan Nasional/BAPPENAS*.

Saridewi, R., Nani, A. (2010), (Suryana 2002 in Dewa 2007), Hubungan antara penyuluhan dan adopsi teknologi oleh petani terhadap peningkatan produksi padi di Kabupaten Tasikmalaya.

Siegel, S., Castellan, J., Jr. (1988). Nonparametric Statistics for the Behavioral Sciences, second edition. *McGraw-Hill Book Co*.

Siegel Sidney (1992). *Statistik Nonparametrik untuk Ilmu-Ilmu Sosial. PT Gramedia, Jakarta*.

Association of Body Mass Index and Blood Chemistries

Vadhana Jayathavaj[1*] and Pranee Boonya[2]

[1]College of Oriental Medicine, Rangsit University, Pathumthani, Thailand
[*]Corresponding Email: vadhana.j@rsu.ac.th
[2]Office of Health Welfare, Rangsit University, Pathumthani, Thailand
Email: pranee.b@rsu.ac.th

ABSTRACT

Body mass index (BMI) and blood chemistries are the indicators of both noncommunicable diseases (NCD) and metabolic syndrome (MS). BMI is the most convenience to access and will be the independent variable to monitor the dependent blood chemistries: fasting blood sugar (FBS), triglycerides (TG), Total Cholesterol (TC), high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol (LDL). A private university has a healthy campus vision and gathered the annual individual personnel medical check-up reports into the personnel health database. The prevalence of BMI $\geq$ 30 kg/m^2 was 13% in 2015. The association of BMI and blood chemistries is the information to create a university campaign against obesity and noncommunicable diseases. The 1,020 medical check-up records of university personnel in 2015 (females, n=568, 56%) with age profiles: minimum, maximum, and average are 24, 78, and 47.60 years, respectively. The descriptive statistics, Shapiro-Wilk normality test, log transformation, and Pearson correlation are applied to the medical check-up data. There was no association between BMI and Total Cholesterol (TC) (r=0.031 p=0.316), a negative association with high-density lipoprotein cholesterol (HDL) (r=-0.371, p<0.001), and a positive association with triglycerides (TG) and the log transformation of TG (r=0.326, p<0.001, and r=0.369, p<0.001), and fasting blood sugar (FBS) and the log transformation of FBS (r=0.274, p<0.001, and r=0.318, p<0.001). BMI showed very weak association with LDL (r=0.080, p=0.011). Creating health awareness campaign for university community members, the development of the multivariate model on age, sex, BMI, and daily behaviors (eating, smoking, alcohol, exercise, and etc.) to predict biochemistries of a person and verify with the medical check-up reports would be designed.

Keywords: body mass index; blood chemistries; correlation; normality test

1 INTRODUCTION

World Health organization (WHO) has been publicized that noncommunicable diseases (NCDs) kill 40 million people globally each year, equivalent to 70% of all deaths. (WHO, 2017a). The total number of NCDs deaths in Thailand were 393,000 of 68.6 Million population (5.72:1,000), WHO survey shows that Thailand is among the top ten performers for NCDs prevention and control (WHO, 2017b).

In Thailand health surveys by medical check-up 2004, the survey classified metabolic syndrome when a patient has diagnosed at least 3 of the following 5 conditions: fasting glucose $\geqslant$ 100 mg/dL, blood pressure $\geqslant$130/85 mm Hg, triglycerides $\geqslant$150 mg/dL, high density lipoprotein cholesterol <40 mg/dL in men or <50 mg/dL in women, waist circumference $\geqslant$ 90 cm (40 in) in men or $\geqslant$80 cm (35 in) in women; if Asian or Body Mass Index (BMI) >30 kg/m^2 (Ekpalakorn, 2016; Alberti et al., 2009).

BMI is anthropometric measurement tools for obesity, to facilitate the prevention of metabolic syndrome. (Shiwaku et al., 2004). Appropriate BMI for Asian population is shown in Table 1 (Ekpalakorn, 2016; WHO expert consultation, 2004).

Table 1: Obesity classification for Asian

Classification	BMI (kg/m2)
Underweight	<18.50
Normal range	18.50 - 22.99
Over weight	$\geq$ 23
Pre-obese	23.00 - 24.99
Obese Level I	25.00 - 29.99
Obese Level II	$\geq$30.00

Although overweight people have a tendency to develop metabolic disease, an overweight condition is only a part of metabolic disease, so predicting metabolic syndrome by anthropometric measurements only might be problematic. The most practical for screening metabolic syndrome is the use of plasma triglyceride and HDL concentrations together with BMI. (McLaughlin et al., 2003).

Body mass index (BMI) and blood chemistries are the indicators of both noncommunicable diseases (NCD) and metabolic syndrome (MS). BMI is the most convenience to access and will be the independent variable to monitor the dependent blood chemistries: fasting blood sugar (FBS), triglycerides (TG), Total Cholesterol (TC), high-density lipoprotein cholesterol (HDL), and low-density lipoprotein cholesterol (LDL). Jayathavaj & Boonya (2018) studied the annual medical medical check-up reports from the personnel health database of a private university and projected that the steady state prevalence of MS in this private university was at 7.5%. This private university needs to build awareness on MS among their university personnel in order to support their healthy campus vision. The preliminary statistical report on the association of BMI and blood chemistries is needed to setup the campaign against NCDs and MS effectively.

2 METHODS

2.1 Medical Check-up Reports

In 2015, 1,508 university personnel had been attended annual physical examination. The responsible unit gathered annual personnel medical check-up reports into the digital personnel health database, the record contains employee code, name, age, sex, weight, height, the laboratory blood tests and urine tests, and etc. The personal identities (employee code, name, and affiliation) were removed from the data in order to provide confidentiality and privacy of personal data before transfer to the research team. The medical check-up records are edited and only 1,020 of have completed the fields for this study: age, sex, weight, height, and blood tests (FBS, TG, TC, HDL, and LDL). BMI is derived from weight in kilogram and height in metre.

2.2 Data Processing

The 1,020 personnel medical check-up records are processed as shown in Figure 1. The reports are as follow.

The frequency classified by sex and age, and descriptive statistics to describe the subjects' biodata.

The descriptive statistics including with skewness, kurtosis, and Shapiro-Wilk normality test of BMI, FBS, TG, TC, HDL, and LDL are processed.

The correlations: Pearson, Kendall's tau, and Spearman's rho, and scatter plot to illustrate the association between BMI and blood chemistries.

All statistical data analyses are performed using SPSS version 21.0 (IBM Corp, 2012).

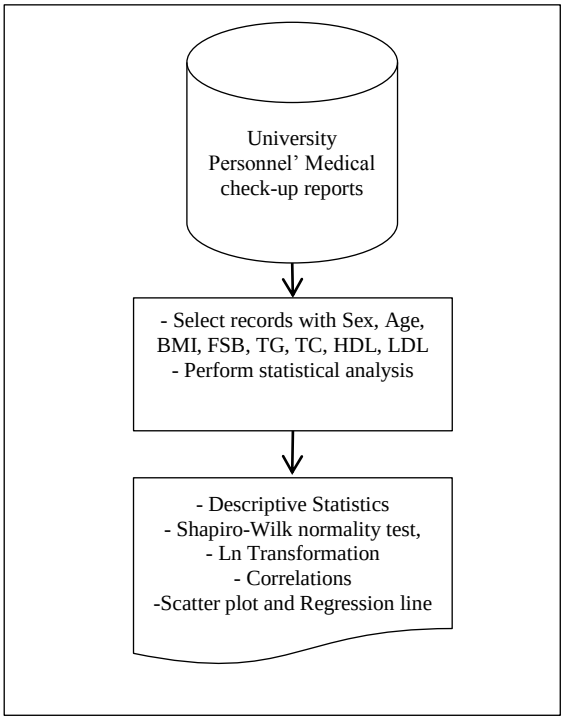


Figure 1: Data Processing

2.3 Statistical Analysis

2.3.1 Normality Assumption and Correlations

To measure the degree of the relationship between linearly related variables, Pearson *r* correlation is the most widely used correlation statistic in all disciplines. But when a parametric test of the correlation coefficient is being used, assumptions of bivariate normality and homogeneity of variances must be met. If data violate the normality assumptions, a test of the significance of Pearson *r* correlation may inflate Type I error rates and reduce power. (Bishara & Hittner, 2012). If the data are not normally distributed, a non-parametric test which often called distribution free tests can be used instead, a non-parametric correlation coefficient would have been more appropriate. Common transformations include taking the log or square root of the dependent variable (Influential Points, n.d.), and also nonlinear transformations can improve the power of Pearson's *r* (Dunlap et al., 1995; Rasmussen, 1989).

In 2011, the multivariable linear analysis was performed to evaluate the relation between BMI and TC, LDL, HDL, and TG as continuous variables. For TG, a log-transformed outcome was created because this variable did not have a normal distribution. Pearson coefficients evaluated the correlation between continuous BMI with each of the blood chemistries. (Shamai et al., 2011).

Kendall rank correlation and Spearman rank correlation are the non-parametric tests that measure the strength of dependence or the degree of association between two variables. The Pearson is most appropriate for measurements taken from an interval scale, while the Spearman is more appropriate for measurements taken from ordinal scales. Spearman's rho and Kendall's tau can both be used for non-parametric data, as they are both measures of rank correlation. Spearman's rho was used in the study of blood lipids correlate to Body Mass Index (Lautsch, et al., 2018).

The Shapiro-Wilk normality test is more appropriate for small sample sizes (< 50 samples), but can also handle sample sizes as large as 2000. For this reason, the Shapiro-Wilk test is used as numerical means of assessing normality. (Lund Research Ltd, 2018; Shapiro & Wilk, 1965).

3 RESULTS

From the university personnel medical check-up data in 2015, the number of subjects are 1,020 total (452 males, 568 females), age average 47.6 years in the range from 24 to 78 years as shown in Table 2. The mean ± standard deviation of BMI for male, female, and total are 25.37±3.77, 24.31±4.67, and 24.78±4.32 kg/m^2, respectively as shown in Table 3.

The over weighs: Pre-obese, Obese Level I, and Obese Level II are at 19%, 30%, and 13%, respectively. The details of obesity classification by sex are shown Table 4.

The descriptive statistics skewness, kurtosis, and Shapiro-Wilk normality test of BMI and blood chemistries including BMI, FBS, TG, ln(TG), TC, HDL, and LDL are shown in Table 5.

The Pearson, Kendall's tau, and Spearman's rho correlations between BMI and blood chemistries are shown in Table 6. The scatter plot and regression line for each bivariate between BMI and blood chemistries are shown in Figure 2.

Table 2: Descriptive statistics of age

Description	Male	Female	Total
Number of records	452	568	1,020
Age (years)			
- Minimum	25	24	24
- Maximum	74	78	78
- Mean	47.61	47.59	47.60
- Std. Deviation	8.95	8.20	8.54

Table 3: Descriptive statistics of BMI

Description	Male	Female	Total
Number of records	452	568	1,020
BMI (kg/m^2)			
- Minimum	14.06	17.05	14.06
- Maximum	39.04	42.20	42.20
- Mean	25.37	24.31	24.78
- Std. Deviation	3.77	4.67	4.32

Table 4: Descriptive statistics of BMI level

BMI (kg/m^2)	Male	Female	Total
	Number of persons		
<18.50	11	24	35
18.50 - 22.99	111	247	358
23.00 – 24.99	102	91	193
25.00 - 29.99	177	128	305
≥30.00	51	78	129
Total	452	568	1,020
	Percentage		
<18.50	2%	4%	3%
18.50 - 22.99	25%	43%	35%
23.00- 24.99	23%	16%	19%
25.00 - 29.99	39%	23%	30%
≥30.00	11%	14%	13%
Total	100%	100%	100%

Table 5: Descriptive statistics of BMI and blood chemistries

	BMI kg/m^2	FBS mg/dl	TG mg/dl	ln(TG) mg/dl
N	1,020	1,020	1,020	1,020
Mean	24.78	91.70	129.65	4.75
Std. Error	0.14	0.67	2.07	0.01
Std. Deviation	4.33	21.26	66.01	0.47
Minimum	14.06	67.00	32.00	3.47
Maximum	42.20	281.00	399.00	5.99
Skewness	0.80	4.57	1.48	0.18
Std. Error	0.08	0.08	0.08	0.08
Kurtosis	0.66	28.87	2.61	-0.25
Std. Error	0.15	0.15	0.15	0.15
Tests of Normality				
Shapiro-Wilk	0.065	0.220	0.126	0.037
p-value	.000	.000	.000	.000

Table 5: Descriptive statistics of BMI and blood chemistries (Cont.)

	TC mg/dl	HDL mg/dl	LDL mg/dl
N	1,020	1,020	1,020
Mean	224.93	60.58	138.43
Std. Error	1.13	0.50	1.05
Std. Deviation	36.19	15.83	33.66
Minimum	108.00	25.00	33.00
Maximum	380.00	130.00	284.00
Skewness	0.36	0.67	0.29
Std. Error	0.08	0.08	0.08
Kurtosis	0.46	0.46	0.54
Std. Error	0.15	0.15	0.15
Shapiro-Wilk	0.037	0.079	0.032
p-value	.000	.000	.000

4 DISCUSSION

The personnel obesity in 2015 of this private university are Pre-obese 23.00- 24.99 kg/m^2, Obese Level I 25.00 - 29.99 kg/m^2, and Obese Level II ≥30.00 kg/m^2 at 19%, 30%, and 13%, respectively. The average BMI for male and female were 25.37 and 24.31 kg/m^2, respectively, while the surveyed in 2014, the average BMI of Thai population age more than 15 years for male and female was 23.6 and 24.6 kg/m^2, respectively. (Ekpalakorn, 2016).

The distribution of every variable is positively skew, the skewness of FBS and TG are >+1 highly skewed, and the skewness of BMI, TC, HDL, and LDL are in the range of +0.5 to +1 moderately skewed.

The kurtoses of FBS equaled to 28.87 which > +3 is leptokurtic, and the kurtosis of the rest variables (BMI 0.66, TG 2.61, TC 0.46, HDL 0.46, and LDL 0.54) had < 3 are platykurtic. (Brown, 2016).

Shapiro-Wilk shows p-value <0.01 for every variable, the data significantly deviate from a normal distribution. (Lund Research Ltd, 2018).

Correlations of BMI with each blood chemistry are significant at the 0.01 level (2-tailed), except BMI and TC is significant at the 0.05 level (2-tailed). The Spearman's rho of BMI with TC and LDL are at 0.070 and 0.126 show a weak uphill (positive) linear relationship; BMI with FBS and TG are at 0.339 and 0.415 show a moderate uphill (positive) relationship; and only BMI with HDL has -0.397 shows a moderate downhill (negative) relationship. (Lund Research Ltd., 2018).

5 CONCLUSIONS

A private university has a healthy campus vision and gathered the annual individual personnel medical check-up reports into the digital personnel health database. The prevalence of BMI ≥ 30 kg/sq.m. was 13% in 2015. The 1,020 medical check-up records of university personnel in 2015 (females, n=568, 56%) with age profiles: minimum, maximum, and average are 24, 78, and 47.60 years, respectively. The descriptive statistics, Shapiro-Wilk normality test, log transformation, and Pearson, Kendall's tau, and Spearman's rho correlations are applied with the medical check-up data. The Pearson is most appropriate for measurements taken from an interval scale, The association with BMI, there are no association with Cholesterol (TC) (r=0.031 p=0.316), negative association with high-density lipoprotein cholesterol (HDL) (r=-0.371, p<0.001), positive association with triglycerides (TG) and the log transformation of TG (r=0.326, p<0.001, and r=0.369, p<0.001), and fasting blood sugar (FBS) and the log transformation of FBS (r=0.274, p<0.001, and r=0.318, p<0.001). BMI shows very weak association with LDL (r=0.080, p=0.011). Creating health awareness campaign for university personnel, the development of the multivariate on age, sex, BMI, and daily behaviors (eating, smoking, alcohol, exercise, and etc.) to predict biochemistries of a person from the medical check-up reports would be designed.

Table 6: Correlations of BMI and blood chemistries

	FBS	TG	ln(FBS)
Pearson	0.274**	0.326**	0.318**
Sig. (2-tailed)	.000	.000	0.000
Kendall's tau	0.234**	0.279**	
Sig. (2-tailed)	0.000	0.000	
Spearman's rho	0.339**	0.415**	
Sig. (2-tailed)	0.000	0.000	

Table 6: Correlations of BMI and blood chemistries (Cont.)

	TC	HDL	LDL
Pearson	0.031	-0.371**	0.080*
Sig. (2-tailed)	0.316	0.000	0.011
Kendall's tau	0.048*	-0.272**	.086**
Sig. (2-tailed)	0.022	0.000	0.000
Spearman's rho	0.070*	-0.397**	.0126**
Sig. (2-tailed)	0.025	0.000	0.000

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed)

ACKNOWLEDGEMENTS

The authors would like to thank Supachai Kunarattanapruek, M.D. for their valuable initiative ideas. This study was funded by Office Health Welfare, Rangsit University.

ETHICAL APPROVAL

The employee code, name, and affiliation were removed from the annual medical record in order to provide confidentiality and privacy of personal data. The use of medical check-up data for the project "Medical Examination Reports Data Analysis" was approved by the head of Office of Health Welfare, Rangsit University. The study protocol was approved by Ethical committee of Research Institute of Rangsit University, Thailand – project number "RSPE 03/2560". This study was performed in accordance with guidelines prescribed by the

Declaration of Helsinki as developed by the World Medical Association.

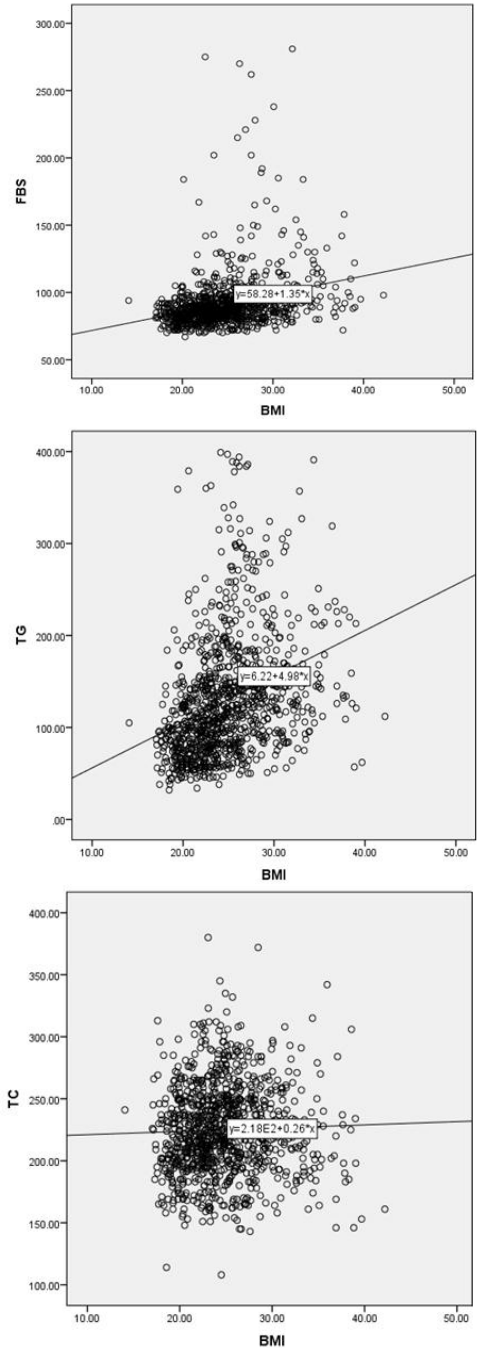


Figure 2: Scatter plot of BMI and blood Chemistries

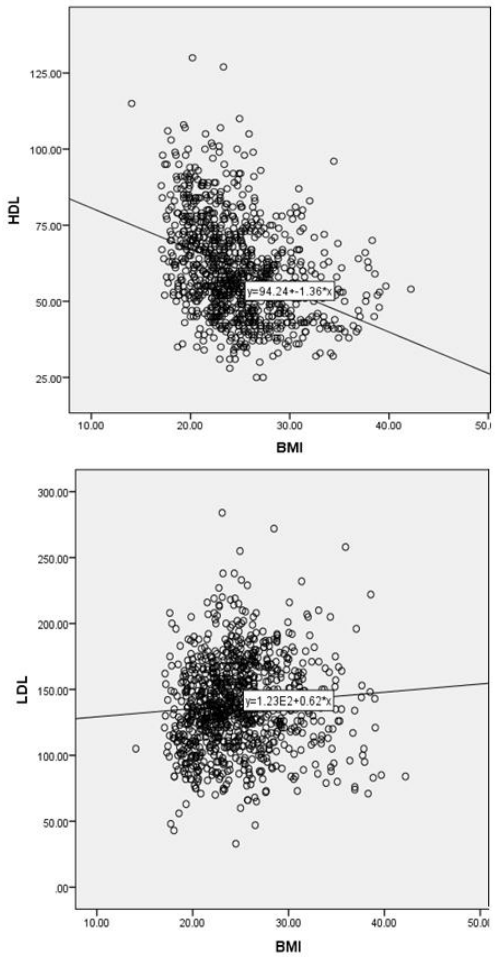


Figure 2: Scatter plot of BMI and blood Chemistries (Cont.)

REFERENCES

Alberti, K.G.M.M., Eckel, R.H., Grundy, S.M., Zimmet, P.Z., Cleeman, J.I., Donato, K.A., & Smith S.C. (2009). Harmonizing the metabolic syndrome a joint interim statement of the international diabetes federation task force on epidemiology and prevention; national heart, lung, and blood institute; american heart association; world heart federation; international atherosclerosis society; and international association for the study of obesity. *Circulation*, *120*, 1640-1645.

Al-Bachir, M., & Bakir, M.A. (2017). Predictive value of body mass index to metabolic syndrome risk factors in Syrian adolescents. *Journal of Medical Case Reports*, *11*(1), 170.

Bishara, A.J., & Hittner, J.B. (2012). Testing the significance of a correlation with non-normal data: comparison of pearson, spearman, transformation, and resampling approaches. *Psychological Methods*, *17*, 399-417.

Dunlap, W.P., Burke, M.J., & Greer, T. (1995). The effect of skew on the magnitude of product-moment correlations. *Journal of General Psychology*, *122*(4), 365-377.

Ekpalakorn, V., (Ed.). (2016). Report on the 5th Thailand health surveys by medical check-up B.E. 2557 (p. 170) [in Thai]. Bangkok: Health Systems Research Institute.

IBM Corp. (2012). IBM SPSS Statistics for Windows, Version 21.0.

Influential Points**.** (n.d.). Pearson's correlation coefficient: Use & misuse (scatterplot, bivariate normality, homogeneity of variances, linearity, causality, association versus agreement). Retrieved from http://influentialpoints.com/Training/Pearsons_correlation_coefficient_use_and_misuse.htm

Jayathavaj, V., & Boonya, P. (2018). Projecting the Prevalence and Distribution of Metabolic Syndrome in a Private University.

Proceedings of RSU International Research Conference 2018 (pp. 20-27). Pathumthani, Thailand: Rangsit University.

Lautsch, D., Gitt, A.K., Ferrieres, J., Horack, M., Brudi1, P., & Ambegaonkar, B. (2018). Do blood lipids correlate to Body Mass Index? Findings from 52.916 statin treated patients. Results of the Dyslipidemia International Study. National Lipid Association. Retrieved from https://www.lipid.org/util/eposters /PDF/128.pdf

Lund Research Ltd. (2018). Testing for normality using SPSS statistics. retrieved from: https://statistics.laerd.com/spss-tutorials/testing-for-normality-using-spss-statistics.php

McLaughlin, T., Abbasi, F., Cheal, K., Chu, J., Lamendola, C., & Reaven, G. (2003). Use of metabolic markers to identify overweight individuals who are insulin resistant. *Annals of Internal Medecine, 139*(10), 802–809.

Rasmussen, J. (1989). Data transformation, Type I error rate and power. *British Journal of Mathematical and Statistical Psychology*, *42*(2), 203-213.

Razali, N.M., & Wah, Y.B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Anlytics*, *2*(1), 21-33.

Shamai, L., Lurix, E., Shen, M., Novaro, G.M., Szomstein, S., Rosenthal,R., Asher, C.R. (2011). Association of body mass index and lipid profiles: evaluation of a broad spectrum of body mass index patients including the morbidly obese. *Obesity Surgery*, *21*(1):42–47.

Shapiro, S.S., & Wilk, M.B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4): 591-611.

Shiwaku, K., Anuurad, E., Enkhmaa, B., Kitajima, K., & Yamane, Y. (2004). Appropriate BMI for Asian populations. *The Lancet*, *363*(9414), 1077.

World Health Organization. (2000). Obesity: preventing and managing the global epidemic (No. 894). World Health Organization.

WHO, E.C. (2004). Appropriate body-mass index for ASIAN populations and its implications for policy and intervention strategies. *The Lancet* , *363*(9403), 157.

World Health Organization. (2017a). Noncommunicable diseases fact sheet. Retrieved from http://www.who.int/mediacentre/factsheets /fs355/en/

World Health Organization. (2017b). Noncommunicable diseases progress monitor 2017.

Brown, S. (2016). Measures of Shape: Skewness and Kurtosis. Retrieved from https://brownmath.com/stat/shape.htm

Evaluation of Loss Disabled Workers on Compensation of Occupational in 2016

Krieng Kitbumrungrat[1*]

[1]Department of Mathematics, Faculty of Science, Dhonburi Rajabat University, Bangkok 10600, Thailand
*Corresponding Email: Kriengstat@yahoo.com

ABSTRACT

The purpose of this research is the evaluation of loss disabled workers on compensation of occupational in 2016 aimed to examine: 1) study lifestyle characteristics of the workers who had occupational loss finger, hand and arm after receiving compensation; 2) study variables that affect the compensation base on the worker's organ, variables that affect the income of workers before the accident, variables that affect the income of workers after rehabilitation or receiving compensation and variables that affect the income of workers and injured work related workers. The sample was consisted of 6 groups in area Rayong, Puthum thani, Samut pakran and Samut sakon involving 300 workers. Methodology of research is the evaluation of loss disabled workers on compensation of occupational by using Multiple Linear Regression Model for occupational loss.

We found that the compensation paid to the worker was in accordance to the type of organ lost and the worker's age. The income model of the workers before the accident was associated with the worker's gender, education level, age, and work experience. The income model of the workers after rehabilitation was associated with the worker's gender, education level, age, and work experience before the accident. The income model difference between the wages of regular and degraded workers was associated with the education level and occupation of the former employee. In this research, we will develop and improve the workmen's compensation according to type of worker organ loss and promote equality.

Keywords: Evaluation of socio-economic impact, Occupational Loss, Analysis of variance (ANOVA), Multiple regression model

1 INTRODUCTION

The workers in an establishment where there are greater risks of a work related illness and an increased level of dangerous work can include but are not limited to extensive use of machinery, chemicals and an improper production working environment. However, awareness of potential accidents at work or prevent disease from the work environment were limited. Workers also lacked the knowledge and availability of personal protective equipment (PPE) and safeguards when using machinery or knives. Statistics from the compensation fund data found: the causes of workers injury related to the severity of cuts, bumps or chemical splashes into the eye (Piyanut Ratkul (2005)).

The Workmen's Compensation Act B.E. 2537 (1994) covers the monies which are paid as indemnity, medical expenses, rehabilitation expenses and funeral expenses. The Labors and Social Welfare Law have a rate for medical expenses when the workers are hurt or develop a work related illness, the employer is required to fund the medical expenses, and this has been effective in Thailand since 13 May 2008, by increasing the amount of medical expenses for the workers that are injured from 35,000 to 45,000 bath (Anantaya Neamklay (2008)). If this is not enough, it is raised from 50,000 to 65,000 baths, and the highest medical expenses from 200,000 baht. The maximum is 300,000 baht in case of an inpatient, expenses for the room, food, nursing services and general service charges the employer's responsibility is limited to no more than 1,300 baht per day. For the workers who are required to undergo rehabilitation after a serious injury or illness, the employer will pay the cost. The rehabilitation of medical and vocational training, the actual cost of not more than 20,000 baht and the cost of surgery for rehabilitation not exceeding 20,000 baht as prescribed under this announcement.

If a worker suffers a serious accident for example loss of fingers, hands or arms, this will affect the body, mind and society. The most obvious physical impact is not being able to use fingers, hands and arms like normal people. It affects their daily activities and the injured person needs to rely on others. Mental effects include: feeling frustrated with both themselves and others. The worker who loses fingers, hands or arms will lose self-confidence. The worker who experiences losing his finger, hand or arm will try to adapt to the changes that have occurred (Srisak Sonthonchai. (2005)).

Adaptation and acceptability by employees that lost a finger or arm losses the function or the control of their body, this promotes a lack of opportunity to live in society as they are faced with the negative attitudes of the society towards the disabled. Rejected by society and sometimes being dismissed will make them feel discouraged, worthless or they have a lack of ability (Lertrit Arayasajjapong (2007)).

2 OBJECTIVES

2.1 To describe life style, physical, mental, socioeconomic, and social status of employees with a work related disability after receiving compensation.

2.2 To find the models and variables affecting the compensation for the type of worker's loss, income from employee's wages before an accident and revenue of employee wages after rehabilitation or compensation. The income of the difference between a normal fully fit employee and one that has sustained a work related injury.

3 METHODS

This research is a survey to study the behavior of disabled employees on occupational compensation. To study the root mean cause of mechanical hazards and the lack of personal protective equipment use. Also, to study the economic and social environment that affects these employees.

3.1. Population and Samples

Sample group of employees who suffer work related physical disabilities, such as lose of fingers, hands or arms, in Rayong, Pathum Thani, Samut Prakan and Samut Sakhon provinces. Around 300 people were affected because of dangerous working conditions, according to data from 2016, the affected workers lost fingers or hands.

3.2 Research methodology

The assessments of the economic and social impact on employees with disabilities who lost of their fingers or arms and received compensation for loss of occupation during the year 2016 are as follows:

1. Descriptive statistics of economic impact and the social status of workers who lost their fingers or arm to the compensation for loss of occupation for the year 2016. The statistics used: Percent, Mean and Standard Deviation.

2. To create a multiple linear regression model of compensation according to the type of employee's loss. This used the income of employee before an accident and the revenue of employee after rehabilitation. The income was the difference between the pay rate of a normal fully fit worker and a worker who had loss of physical fitness due to a work related incident.

The model in the study of this research can be written as follows.

$$Y_t = f(L, G, t, Y_{t-1}) \quad (1)$$

By L = Level of Education
G = Gender
t = Experience
Y_{t-1} = Wage rate in past, when t-1
Y_t = Wage rate, when t

The model and finite methodology used to estimate lifetime wages in years t of normal people (Y_t^n) and people with disabilities (Y_t^c) share the solution to the expected problems. The details are as follows:
The model is Linear Regression Model based on N * T.
[Number of Worker x time which annual pay can be collected (Panel Data)]

$$Yt = f (L,G, t, Yt-1)$$

$$Yit=\beta_0+\beta_1 X_{it}+\varepsilon_{it} \quad (2)$$

Assume that ε_{it} it has the following E (ε_{it}) = 0
Then, the data is estimated Coefficient, a type of Fixed Effect Model, or may be estimated in other ways depending on the appropriateness of the results. The solution Autocorrelation and finally equation 2. Equation (3) is the equation for the wage of a regular worker. Equation (4) is the equation for estimating the wage of the worker with a disability, fingers, hands or arms from work. However, both solutions must take into account whether the new estimate is reasonable.

$$Y_t^n = \beta_0^n + \beta_1^n Y_{t-1}^n + \beta_2^n G + \beta_3^n t + \beta_4^n L \quad (3)$$

$$Y_t^C = \beta_0^C + \beta_1^C Y_{t-1}^C + \beta_2^C G + \beta_3^C t + \beta_4^C L \quad (4)$$

When Least Squares Estimates of Normal Employees in Equation (3) Remove with Least Squares Estimates of Employees who loss fingers, hands, or arms from working in Equation (4). The loss of a disabled employee's fingers, hands or arms from work for earn money each year.

4 RESULTS

This research assessed the income losses of disabled employees compared to their occupational compensation for the year 2016. The losses will limit the "compensation" paid under the Workmen's Compensation Act B.E. 2537. The worker who suffered work related loss of organs including fingers, hands and arms in Rayong, PathumThani, Samut Prakan and Samut Sakhon provinces for 2016 was approximately300 people as follows.

1. Social Security Office in Samut Prakarn province 130 people.
2. The Eastern Region Labor Rehabilitation Center in Rayong province 50 people.
3. Pathum Thani Provincial Labor Rehabilitation Center 20 people.
4. Social Security Office 7 in Bangkok 60 people.
5. Social Security Office in Samut Sakhon province 40 people.

4.1 Personal information of the sample

Information of employees who lost their fingers, hands and arms during their work consisted of sex, age, experience, working status, marital status, and education. Most of the workers were male (66%) and female (34%). Most of the respondents were aged 35-39 years (21.3%), followed by those aged 45-49 years, (14.3%). The maximum age was 65 years and the minimum age was 17 years. The average age was 36 years. The majority of workers were employed for 1-5 years (44%), followed by more than 10 years (28%). Most of the employees were married at 52.7% and 36.7% were single. Most of the employees had attained the secondary education level (44.7%) and 42% had only had primary school level. The establishments where the workers had the highest work related accidents were in Samut Prakarn province (44.7%) followed by Bangkok (31%). The major incident was employees who lost their fingers (76.3%) and the second was loss of hand (12.7%). The most common cause of employee loss was finger / hand pump (36.7%) followed by other work-related accidents (23.3%).

4.2 The life style of an employee who loses fingers, hands, and arms from work after receiving compensation

The lifestyle of an employee who losses a finger, hand, or arm from a job after receiving compensation. Employees, who lost most of their fingers, hand or arm from their occupations was 63.8%. Secondly, workers who were undergoing reconstruction were 30%. The new job description was different from the original job was 6.3% and can adapt with colleagues after the incident 98 percent, but still cannot adapt with colleagues after the incident 2.0%.

For workers who lost their fingers, hands or arms during occupation, currently, 62.7% rent their homes. 24.3%. Seventy one point three percent (71.3%) of the injured require care. Most of the care (33.5%) was provided by parents and children, secondly by 1 child/wife (19.7%). The majority of people lived with their husbands / wives (36%) and second, 20.7% were alone. Most of them were family members (99.3%).

Problems/Obstacles of the worker losing their fingers, hands, or arms from a work related injury to their present life sustainability. Workers who lost their fingers, hands, or arms from most occupations did not work as before (41.9%). Secondly, had no problems/obstacles in their current life 23.5% and had difficulties 12.4%.

4.3 Opinions of employees who lost their fingers, hands or arms from occupational related injury

Incorporation of improved safety equipment related to machinery or things that can cause harm. Most of the respondents had opinions on machine improvements or what on other mechanisms that could cause them harm. There should be workplace protection measures (24%), secondly repairs or purchase of new machinery (23 %).

The implementation of safety measures by companies. Employees who lost fingers, hands or arms from occupation injuries, most of the respondents commented on the implementation of the safety measures by the company should be checked before the use of machinery (22%) and secondly the prevention of workplace hazards (15.0%).

4.4. Assessment of socio-economic losses of employees who lose their finger, hand and arm performance after receiving the compensation

The mean score of socio-economic loss of employees who lost fingers, hands or arms from a work related incident after receiving compensation. Performance appraisal of employees who lose their finger, hand or arm performance was lower than for normal fully fit employees. Most employers had the opinion that loss of organs caused limitation during work, the average was 2.95. Secondly, there was a comment that they cannot work normally, as it will be managed by artificial arms or work equipment modification, the mean was 2.89. The evaluation of the acceptance of the loss of external factors. Most employers had the opinion that employers often supported wage increases or promotion to fully fit people rather than those who had lost body organs, the average was 2.93. Secondly, there was a comment that employers often do not accept people who have lost organs. The average was 2.70. There was less evidence of human capital accumulation. Most employers had the opinion that if a worker has not lost a body organ, he can improve his ability to work better with an average of 3.27. Secondly, there was the opinion that with further professional training people who have lost body organs can progress in a career as normal, the average was 2.99. The old human capital assessment cannot be utilized. Most employers have the opinion that a good education for people who have lost organs will help them to earn better jobs than those who lost their organs but did not have a good education level, the average was 2.96. Second, the financial loss before the organ loss and after the organ loss of different organ parts. Employees had the opinion that injured workers can continue to work in the current job without having to study at any kind of school, with an average of 2.91. Analysis of the socio-economic losses of employees who lost their fingers, hand or arm from an employment related incident after receiving compensation. Most employers had the opinion that the old human capital accumulation cannot be utilized, most valuable the mean was 2.87. Secondly, there was a comment. That there was less evidence of human capital accumulation, the mean was 2.79.

4.5 Relationship between variables affecting compensation for type of organ loss

The Multiple Linear Regression Model of Receiving Compensation for Employee Loss Category, The SOS = f (SEX, AGE, STATUS, EXP, EDU, PHY, WB). SOS = Organs, SEX = gender, AGE = age, STATUS = marital status EXP = work experience, EDU = education level, PHY = lost organ, and WB = Career The regression model by Stepwise method, Table 1.

Table 1: Relationship between variables influencing compensation for type of organ failure

ANOVA

[a].Predictors: (Constant), PHY:,
[b].Predictors: (Constant), PHY, AGE:.
p-value < 0.05

Model	df	Sum of Square	Mean Square	F–Statistic	p-value
1. Regression	1	2529	2529	11.646	0.002[a]
Residual	35	7601	2171		
Total	36	10130			

Model	df	Sum of Square	Mean Square	F–Statistic	p-value
2. Regression	2	3455	1727	8.797	0.001[b]
Residual	34	6675	1963		
Total	36	10130			

Coefficients

	Unstandardized		Standardized		
Model	B	Std. Error	Beta	t–Statistic	p-value
1.(Constant)	14558.484	42816.583		0.340	0.736
PHY	42861.808	12559.682	0.500	3.413	0.002
2.(Constant)	190704.100	90783.198		2.101	0.043
PHY	39379.054	12049.570	0.459	3.268	0.002
AGE	-4337.689	1998.186	-0.305	-2.171	0.032

p-value < 0.05

From Table 1, the model for receiving compensation for the loss category of the employee is p-value = 0.001 <0.05. It can be concluded that there is a correlation between the compensation and the compensation for the loss of organs of the employees, at a significance level of 0.05. This is the multiple regression equation for assessing compensation for the type of organ loss of an employee. Equation 5 is

SOS=190704.100+39379.054 PHY- 4337.689 AGE (5)

Receiving compensation for the type of organ loss of an employee is associated with the employee's lost organ and age.

4.6 Relationship between variables affecting earnings of employee wages before an accident

The Multiple Linear Regression Model of Employee Wage Revenue Prior to Accident WB3 = f (SEX, STATUS, EXP, EDU, AGE1, PHY, WB2), where WB3 = wage earners income prior to experiencing Accident, SEX = gender, STATUS = marital status, EXP = experience before an accident EDU = education level, AGE1 = pre-accident age, PHY = lost organ, and WB2 = work or occupation before an accident. The regression model by using the Stepwise as shown in Table 2

Table 2 Relationship between the variables affecting the income of the employees before the accident

ANOVA

Model	df	Sum of Square	Mean Square	F–Statistic	p-value
1.Regression	1	393420017.52	393420017.52	30.210	0.000[a]
Residual	289	3763574358.97	13022748.65		
Total	290	4156994376.50			
2.Regression	2	569192850.14	284596425.10	22.845	0.000[b]
Residual	288	3587801526.36	12457644.19		
Total	290	4156994376.50			
3.Regression	3	659101611.51	219700537.20	18.026	0.000[c]
Residual	287	3497892764.99	12187779.67		
Total	290	4156994376.50			
4.Regression	4	856403516.27	214100879.10	18.552	0.000[d]
Residual	286	3300590864.24	11540527.48		
Total	290	4156994376.50			

[a].Predictors: (Constant), PHY:,
[b].Predictors: (Constant), PHY, SEX:.
[c].Predictors: (Constant), PHY, SEX, EDU:.
[d].Predictors: (Constant), PHY, SEX, EDU, AGE1:. p-value < 0.05

Coefficients

	Unstandardized		Standardized		
Model	B	Std. Error	Beta	t – Statistic	p-value
1.(Constant)	9553.731	303.585		31.47	0.000
EXP	130.389	23.723	0.308	5.49	0.000
2. (Constant)	11770.387	660.610		17.817	0.000
EXP	128.448	23.208	0.303	5.535	0.000
SEX	-1653.397	440.169	-0.206	-3.756	0.000
3. (Constant)	9153.452	1164.171		7.863	0.000
EXP	143.320	23.599	0.338	6.073	0.000
SEX	-1434.142	442.796	-0.178	-3.239	0.001
EDU	821.876	302.599	0.153	2.716	0.007
4. (Constant)	4396.65	1614.565		2.723	0.007
EXP	98.509	25.393	0.232	3.879	0.000
SEX	-1360.06	491.250	-0.169	-3.154	0.002
EDU	1388.477	324.779	0.259	4.275	0.000
AGE1	97.65	23.616	0.273	4.135	0.000

p-value < 0.05

From Table 2, the model of the wage rate of the employee before the accident was p-value = 0.000 <0.05. It can be concluded that the wage earner's income before the accident correlated at a significance level of 0.05. This is a multiple regression equation for getting compensation for the wage rate of the employee before the accident. Equation 6 is WB3=4396.651-1360.064SEX+1388.477EDU+97.65 AGE1+ 98.509EXP (6)

It is the wage earner's income before the accident that correlates with the educational level, gender, age and work experience of the employee before the accident. It has the power to predict the effect of models at 45.4%.

4.7 Relationship between the variables affecting the income of the employee's wages after rehabilitation or compensation

The Multiple Linear Regression Model of Employee Wage Revenue after Rehabilitation or Benefit ECO1 = f (SEX, STATUS, EXP, EDU, AGE, WB2), where ECO1 = wage income of employees after rehabilitation. SEX = gender, STATUS = marital status, EXP=work experience, EDU=education level, AGE=age, PHY= lost organ, and WB2 = occupation by using the Stepwise regression model as shown in Table 3.

Table 3 Relationship between the variables affecting wage income of employees after rehabilitation

ANOVA

Model	df	Sum of Square	Mean Square	F – Statistic	p-value
1. Regression	1	359697960.732	359697960.732	24.454	0.000[a]
Residual	275	4045046333.673	14709259.40		
Total	276	4404744294.404			
2. Regression	2	750786616.826	375393308.40	28.150	0.000[b]
Residual	274	3653957677.578	13335611.96		
Total	276	4404744294.404			
3. Regression	3	926986130.387	308995376.80	24.256	0.000[c]
Residual	273	3477758164.017	12739040.89		
Total	276	4404744294.404			
4. Regression	4	1034756873.299	25869218.30	20.879	0.000[d]
Residual	272	3369987421.106	12389659.64		
Total	276	4404744294.404			
5. Regression	5	1084509070.545	216901814.10	17.704	0.000[d]
Residual	271	3320235223.859	12251790.49		
Total	276	4404744294.404			

[a].Predictors: (Constant), EDU:,
[b].Predictors: (Constant), EDU, AGE:.
[c].Predictors: (Constant), EDU, AGE, SEX:.
[d].Predictors: (Constant), EDU, AGE, SEX, EXP:.
[e].Predictors: (Constant), EDU, AGE, SEX, EXP, WB2:. p-value < 0.05

Coefficients

	Unstandardized		Standardized		
Model	B	Std. Error	Beta	t – Statistic	p-value
1.(Constant)	6499.754	902.875		7.199	0.000
EDU	1644.825	332.818	0.286	4.945	0.000
2.(Constant)	-408.081	1538.241		-0.265	0.791
EDU	2519.259	355.494	0.438	7.087	0.000
AGE	125.393	23.155	0.334	5.415	0.000
3. (Constant)	2634.337	1711.594		1.539	0.125
EDU	2312.984	351.851	0.402	6.574	0.000
AGE	119.531	22.686	0.319	5.269	0.000
SEX	-1697.202	456.352	-0.203	-3.719	0.000
4. (Constant)	3062.744	1694.198		1.808	0.072
EDU	2329.207	347.036	0.405	6.712	0.000
AGE	84.965	25.257	0.227	3.364	0.001
SEX	-1668.285	450.157	-0.199	-3.706	0.000
EXP	83.923	28.455	0.182	2.949	0.003
5. (Constant)	3039.348	1684.786		1.804	0.072
EDU	2318.690	345.139	0.403	6.718	0.000
AGE	72.64	25.849	0.194	2.810	0.005
SEX	-1740.262	449.068	-0.208	-3.875	0.000
EXP	85.892	28.313	0.187	3.034	0.003
WB2	50.56	25.091	0.111	2.015	0.0045

p-value < 0.05

From Table 3, the income model of wage rates of employees after rehabilitation or compensation was p-value = 0.000 <0.05. It can be concluded that the wage rate of the employee after rehabilitation or compensation was at a significance level of 0.05. This was the multiple regression equation for getting compensation as an income of the employee after rehabilitation or compensation. Equation 7 was ECO1= 3039.348-1740.262SEX+2318.69EDU+72.64AGE+85.892EXP + 50.56 WB2 (7)

It was the income of the worker's wage rate after being rehabilitated or receiving compensation related to gender, age, education level and work experience of employees before the accident. It had the ability to predict the effect of the model at 49.6%.

4.8 The relationship between the variables affecting wage income difference between normal and injured person's pay rate

The Multiple Linear Regression Model of the wage rate of the difference between the normal and the injured person's pay rate. The ECO_WB3 = f (SEX, STATUS, EXP, EDU, AGE1, PHY, WB2) were

used for this study. Normal sex and body dysfunction, SEX = gender, STATUS = marital status, EXP = work experience, EDU = Education level, AGE1 = age, PHY = loss of organs, and WB2 = work or career, the regression model was by using the Stepwise method as shown in Table 4.

Table 4 Relationship between the variables affecting income of the difference between normal and physically impaired wage rates

ANOVA

Model	df	Sum of Square	Mean Square	F – Statistic	p-value
1. Regression	1	121333348.266	121333348.266	25.611	0.000[a]
Residual	275	1302848456.066	4737630.749		
Total	276	1424181804.332			
2. Regression	2	169820432.893	8491021645	18.548	0.000[b]
Residual	274	1254361371.440	4577961210		
Total	276	1424181804.332			

[a].Predictors: (Constant), EDU:,
[b].Predictors: (Constant), EDU, WB2:. p-value < 0.05

Coefficients

	Unstandardized		Standardized		
Model	B	Std. Error	Beta	t – Statistic	p-value
1.(Constant)	-2384.061	42816.583		-4.653	0.000
EDU	955.302	12559.682	0.292	5.061	0.000
2.(Constant)	-313.95	90783.198		-5.657	0.000
EDU	1028.24	12049.570	0.314	5.501	0.000
WB2	48.23	1998.186	0.186	3.254	0.001

p-value < 0.05

From Table 4, the revenue model of the difference between the wages of workers and regular employees who lost their physical performance had a p-value = 0.000 <0.05, so it can be concluded as the income difference between the wage rate, normal and employees who had lost fitness, Equation 8 is

ECO_WB3=-3131.95+1028.24EDU+48.23WB2 (8)

It was the income of the difference between the wages of regular workers and those who had lost physical fitness relative to the level of education and occupation of the former employee. It was estimated at 34.5%.

4.9 The relationship between the dependent variables and the independent variables of the multiple linear regression model.

Table 5 Relationship between the dependent variables and the independent variables of the multiple linear regression model.

Model	Dependent variables	Independent variables
1	Model of receiving compensation for type of organ loss	The type of organ lost and Age.
2	The income model of the workers before an accident	Gender, Education level, Age, and Work experience.
3	The income model of the workers after rehabilitation	Gender, Education level, Age, and Work experience before the accident.
4	The income model difference between the wages of regular and degraded workers	Education level and Occupation of the former employee.

From Table 5, the multiple linear regression model had resulted in statistical significance. The independent variables affected by the evaluation of work-related finger, hand or arm; were gender, age, education level, and job. The male workers with a high level of education had the greater economic loss of income than female workers, or with lower education. But the compensation paid to the worker was in accordance to the type of organ lost and the worker's age.

5. CONCLUSIONS AND DISCUSSION

Most worker can help themselves without having to care because most lose their fingers. Living with the family continues to work the same but the behavior is different. Most of the workers are from the provinces, so they live in rented houses with their husbands/wives. There are people who need to be raised/supervised, often a parent who can adapt to the family. Problems /obstacles of working in the present life. Most of the work is not convenient/ not done fully. Most of the employees have physical, mental, intellectual and social health status. Take care of yourself except in the moments after the danger. At first, caregivers were needed because they were not used to helping themselves it also takes time to adjust. Employees have an opinion that they have a worse economic and social status than ever before. There are limits to the use of fingers or hands are not the same. Loss of opportunity to work to earn more or get progress in the work. This is an economic loss and there is a feeling that they are inferior and embarrassed by friends, which is a loss of social value. Most of the respondents were concerned about the economic downturn their income declined / lack of income and do not work as usual. The family of income to raise and the disorganized families are unhappy about the social side, which has a psychological impact. The society is ashamed of friends because the organ is not fully consistent with the opinion of the family or caretaker.

The results showed that the model for receiving compensation for the type of organ loss of the worker was associated with loss of organs and age. This will be the way to get paid. According to the type of loss of organs of the employee. SOS=190704.10+39379.05PHY-4337.69 AGE Income model of wage rate of employees before the accident is WB3=4396.65+1360.06SEX+388.477EDU+97.65AGE1+98.509EXP. It is the income of the employee before the accident. Factors related to work experience before accident, gender, education and age before accident. Revenue model of wage rate of employees after rehabilitation or compensation. ECO1=3039.348+1740.26EX+2318.69EDU+72.64 AGE+85.89EXP+50.56WB2 It is the income of the employee's wage rate after being rehabilitated or receiving compensation related to gender. Education Age and experience of employees before the accident. The income pattern of the difference between the wages of regular employees and the physically impaired employee is ECO_WB3 =-3131.95+1028.24EDU+ 48.23WB2 It is the income of the difference between the wages of regular employees and those with physical impairments that are related to the level of education and occupation of the employee after rehabilitation or compensation.

The suggestion from this research is that the government should set policies and actions to help after the employee suffers a loss of fingerprints. Hands and arms from the occupation include: 1) Review the help after the employee is in danger. Because employees lose opportunities lost revenue in order to choose a career and find a new job. 2) Keep track of the ongoing assistance provided to employees who are no longer able to work. For employees who are in danger of low income. The burden of caring for or caring for a loved one. The cost of medical treatment is much higher than the cost of medical care provided by the agency. Difficult to live. Employees should receive medical expenses as they actually pay. 3) Set up monitoring measures for establishments where employees are in danger to improve their safety. And there are some ways to do it, if there are more workers in the workplace. 4) Review the provision or donation of artificial organs to employees, which is considered to be a government benefit to help employees work normally or almost normally. 5) Encourage the invention of equipment to help the employees to work as normal. 6) The law regulates the establishment of a business policy to accept employees to work in their own establishments to continue to earn. And income should not be lower than the original, even in the new job. Or to find work in other establishments to continue to work. 7) Revise the legal compensation to suit the employee. From the above, it is interesting to continue studying: 1) Impact assessment on industrial work losses. Case study on fingerprints. Hand and arm from the occupation 2) The subject of education, occupational promotion and employment of employees who lose fingerprints. Hands and arms from the occupation. 3) Reasonable legal compensation for employees who lose fingerprints. Hands and arms from the occupation.

REFERENCES

Neamklay, A. (2008). "The legal problems on law enforcement of workmen's compensation: a case study of accidental injury sustained by the employee on working." Master of Laws, Graduate School, Sripatum UniversityChonburi Campus Thesis.

Bleichrodt, H., & Quiggin, J. (1999). Life-cycle preferences over consumption and health: when is cost-effectiveness analysis equivalent to cost–benefit analysis?. Journal of health economics, 18(6), 681-708.

Geweke, J., & Keane, M. (2000). An empirical analysis of earnings dynamics among men in the PSID: 1968–1989. Journal of econometrics, 96(2), 293-356.

Greenberg, D., Bakhai, A., Neumann, P.J., & Cohen, D.J. (2004). Willingness to pay for avoiding coronary restenosis and repeat revascularization: results from a contingent valuation study. Health policy, 70(2), 207-216.

Keeler, E.B. (2001). The value of remaining lifetime is close to estimated values of life. Journal of health economics, 20(1), 141-143.

Arayasajjapong, L. (2007) "Assessing the Employee's Loss of Foregone Earnings under Workers' Compensation Law : In case of a Permanent Leg Disability" Thammasat University Thesis.

Meyer, B.D., Viscusi, W.K., & Durbin, D.L. (1995). Workers' compensation and injury duration: evidence from a natural experiment. The American economic review, 322-340.

Thianprathuangchai, M. (2009). Provision of Thai Labors Welfare: Case Studies on Workmen Compensations Fund, Social Security Fund and Employee Welfare Fund. Quality of Life and Law Journal Vol 5. No1. (January – June 2009)

Ratkul, P. (2005). Thai research and education for Nation Development Institute. "The evaluation of effects from damage at works: the case of work-related eye injury" Research Fund Report from Social Security Office 2005.

Polachek, S.W., & Kim, M.K. (1994). Panel estimates of the gender earnings gap: individual-specific intercept and individual - specific slope models. Journal of Econometrics, 61(1), 23-42.

Sonthonchai, S. (2005).Socio-Economic impact of the workers who had occupational hand loss. Research fund report from social security office 2005.

Krutchaiyant, T. (2002). "The Workmen's Compensation Act: A Comparative Study of Thailand Act and some countries in ASIA." Dhurakijpundit University Thesis.

Viscusi, W.K. (1980). Imperfect job risk information and optimal workmen's compensation benefits. Journal of Public Economics, 14(3), 319-337.

Missing Value Imputation based on K Nearest Neighbor Method with Correlation Coefficient

Manita Kuama*, Wuttichai Srisodaphol and Prem Jansawang

Department of Statistics, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand
*Corresponding email: manita.k36@gmail.com
Email: wuttsr@kku.ac.th
Email: prem@kku.ac.th

ABSTRACT

This study is to propose a method for issing data imputation, namely Correlated K Nearest Neighbor (Corr-KNN) which prepare the data for analysis. The correlation coefficient is used to select the variables having complete data which highly correlate with the variable having missing data. After that, the selected data set in the variables having complete data and the missing data are used in K Nearest Neighbor (KNN) method by replacing missing data with substituted data. The Corr-KNN method is compared with the KNN and K Nearest Neighbor Feature Selection (KNNFS) methods by using the data on UCI Machine Learning Repository database. The performance of each method is measured by the Root Mean Squared Error (RMSE). The results indicate that the Corr-KNN method has powerful imputation with smaller RMSE than the compared methods.

Keywords: correlation coefficient; imputation; K Nearest Neighbor; missing data

1 INTRODUCTION

Missing data problem is a serious problem in data management before analyzing the data for any study or research. Since the missing data can occur in the data set so, the results from the data analysis will be deviate from the truth. However, when missing data occurred, the preliminary management is eliminating cases or variables having missing data. Unfortunately, in case of the dataset have a lot of missing data, the management by this way caused a loss of data which was collected too much and the remained data is insufficient to analyze.

The popular method of dealing with missing data and not make a loss of data which was collected is the method of filling the missing value (Imputation). The imputation could be made in several ways. The method that must use the models to impute missing values, namely Model-Donor Imputation. The methods are in Model-Donor Imputation such as Mean Imputation, Regression Imputation and Multiple Imputation. Moreover, another way for imputation that does not have to use models is called Real-Donor Imputation. Real-Donor Imputation can make by replacing the missing values from the set of remaining observed values in the same dataset, for example, Hot-Deck Imputation and K Nearest Neighbor (KNN) methods.

One of simple and effective method to impute missing values is the KNN method which is a method in the Real-Donor Imputation. KNN method is the method to impute missing values based on remains observed values or complete data that have similar pattern with the cases having missing data. This method can be used to impute quantitative data by using the average of the first *k* cases from the cases having complete data that is close to the case having missing data in the same variable. KNN method is appropriate to use in case of multiple missing values occurring in each case since we can consider randomly impute missing values. Unfortunately, the KNN method also has disadvantages in case of the variables having complete data that does not correlate with the variable having missing data. This case, the estimated values are inclined by those variables. As a result of this drawback, the efficiency of imputation is not work.

Many researchers have proposed missing values imputation by using the concept of KNN method to improve the accuracy of the estimated values. Meesad & Hengpraphrom (2008) proposed the K Nearest Neighbor Feature Selection (KNNFS). The KNNFS method select the variables having complete data that are closest to the variable having missing data based on the distance between the data in the variable having missing data and data in the variable having complete data. Then they use KNNFS to impute the missing value of microarray data by the KNN method. Myneni, Srividya & Dandamudi (2016) proposed a method for missing values imputation, namely the Correlated Cluster Based Imputation. This method considers the correlation between variable having missing data and variables having complete data. Then, only variables having complete data which highly correlate with the variable having missing data were selected for dividing into clusters and impute the missing values with respect to cluster mean value of data in same group of variables having missing data.

Therefore, in this study, we then propose missing value imputation which has more effective, namely the Correlated K Nearest Neighbor (Corr-KNN). This method is developed from KNNFS method of Meesad & Hengpraphrom (2008) by using the correlation coefficient between the variable having missing data with the variables having complete data. After that, the selected data in the variables having complete data is used in K Nearest Neighbor (KNN) method.

2 METHODS

2.1 The proposed missing value imputation method

The Corr-KNN method is developed from KNNFS method of Meesad & Hengpraphrom (2008) by adjusted the selection of variables having complete data which highly correlated with variable having missing data based on the correlation coefficient. After that, the selected data is used in K Nearest Neighbor (KNN) method for replacing missing data with substituted data. The steps of missing value imputation by the Corr-KNN method are explained as follows.

Step 1: The randomly missing value based on Missing Completely at Random (MCAR) is considered for imputing. Other missing values that have not imputed by Corr-KNN method, they are imputed by the average of the data in same variable. The dataset which contain missing values and the first randomly missing value ($y_{m_1 m_2} = NA$) which is imputed is showed in Table 1.

Table 1: The dataset which contain missing values

Variables / Cases	V_1	V_2	V_3	...	V_j	...	V_{m_2}	...	V_p
C_1	y_{11}	y_{12}	NA	...	y_{1j}	...	y_{1m_2}	...	y_{1p}
C_2	y_{21}	y_{22}	y_{23}	...	y_{2j}	...	y_{2m_2}	...	y_{2p}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
C_i	y_{i1}	y_{i2}	y_{i3}		y_{ij}		y_{im_2}	...	y_{ip}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
C_{m_1}	$y_{m_1 1}$	$y_{m_1 2}$	$y_{m_1 3}$	...	$y_{m_1 3}$	...	$y_{m_1 m_2} = NA$	...	$y_{m_1 p}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
C_q	y_{q1}	y_{q2}	y_{q3}	...	y_{qj}	...	y_{qm_2}	...	y_{qp}

Step 2: The correlation coefficient between the variable having missing data and the other variables having complete data is calculated by (1),

$$r_{m_2 j} = \frac{\sum_{i=1}^{q} y_{im_2} y_{ij} - \frac{\left(\sum_{i=1}^{q} y_{im_2}\right)\left(\sum_{i=1}^{q} y_{ij}\right)}{q-1}}{\sqrt{\left[\sum_{i=1}^{q} y_{im_2}^2 - \frac{\left(\sum_{i=1}^{q} y_{im_2}\right)^2}{q-1}\right]\left[\sum_{i=1}^{q} y_{ij}^2 - \frac{\left(\sum_{i=1}^{q} y_{ij}\right)^2}{q-1}\right]}} \quad (1)$$

$$; i = 1,2,...,q, i \neq m_1 \text{ and } j = 1,2,...,p, j \neq m_2$$

where

$r_{m_2 j}$ is the correlation coefficient between variable having missing data and variables j,

C_{m_1} is the case having missing data,

V_{m_2} is the variable having missing data,

y_{im_2} is data of the case i in the variable having missing data,

y_{ij} is data of the case i in the variable having complete data.

Step 3: The c variables having complete data which highly correlated with the variable having missing data are selected.

Step 4: Missing value is imputed by KNN method by using data from Step 3 as follows.

1) The number of k is specified.

2) The Euclidean distance between data in the case having missing data and the other cases having complete data are calculated by

$$dist(C_{m_1}, C_i) = \sqrt{\sum_{j=1}^{p} (y_{m_1 j} - y_{ij})^2} \quad (2)$$

$$; \ i = 1,2,...,q, i \neq m_1 \text{ and } j = 1,2,...,p, j \neq m_2$$

where

$dist(C_{m_1}, C_i)$ is the distance between data in the case having missing data and the case having complete data,

$y_{m_1 j}$ is data of the case having missing data in the variable j,

y_{ij} is data of the case having complete data in the variable j,

p is total number of variables in dataset,

q is total number of cases in dataset,

3) The calculated distances are sorted based on the k distances that is less or close to the case having missing data.

4) The missing value is imputed by the average of the data in the same variable with missing data and k cases from 3). The imputed missing value is calculated by

$$\hat{y}_{m_1 m_2} = \frac{\sum_{a=1}^{k} y_{a,m_2}}{k} \quad (3)$$

where

$\hat{y}_{m_1 m_2}$ is the estimated of missing data,

y_{am_2} is data in the same variable with missing data and the cases that is close to the case having missing data where $a = 1,2,...,k$.

Step 5: Repeat step 1 to step 4 until all missing values have been imputed.

2.2 Performance Measurement

Root Mean Square Error (RMSE) is a measure of performance for considers the accuracy of estimated value from the missing value from the Corr-KNN, KNN and KNNFS methods. RMSE is calculated by (Gajawada & Toshniwal, 2012)

$$RMSE = \sqrt{\frac{1}{M} \sum_{i=1}^{M} (y_i - \hat{y}_i)^2} \quad (4)$$

where

M is total number of missing values,

y_i is an actual value of data,

$\hat{y}_i$ is the estimated value of the missing value.

The scope of this study are as follows.

1) The data which is used in this study is quantitative data from the UCI Machine Learning Repository database include: Breast Cancer Wisconsin (Diagnostic), Forest type mapping, Glass, Image Segmentation, Leaf, Musk (Version 1), Seeds, SPECTF Heart, Statlog (Vehicle Silhoue), Wine, Yeast

2) Missing data are simulated by missing completely at random (MCAR) with 5, 15 and 25 percentage of missing data, respectively. For each missing percentage of missing data are repeated for 30 iterations.

3) Set the number of variables for selecting to impute the missing values based on the correlation coefficient between the variable having missing data and the variable having complete data (c). In this study, we set $c = 3$. (Meesad & Hengpraphrom, 2008)

4) Set the number of cases having complete data for selecting to impute the missing value by the KNN method base on distance of cases that close to the case having missing data (k). In this study, we set $k = 10$.

3 RESULTS

In this study, the efficiency of the missing value imputation methods are evaluated by the RMSE of the imputing missing value of Corr-KNN, KNN, and KNNFS methods. The efficiency is considered based on 10 quantitative datasets with 5%, 15%, and 25% of missing data with missing completely at random (MCAR), respectively. The results are showed in Tables 2-4 and Figure 1.

Table 2 shows the RMSE of the missing value imputation methods for 5% of missing data. The Corr-KNN method has smallest RMSE for all datasets except the Seeds and Yeast dataset. Hence, Corr-KNN is the most efficient method for missing values imputation because it has the powerful estimation ability for actual value more than other methods.

Table 2: RMSE of imputation method for 5 % of missing data

Dataset	Method		
	KNN	**KNNFS**	**Corr-KNN**
Breast Cancer Wisconsin (Diagnostic)	34.626	34.468	**30.662***
Forest type mapping	4.117	6.624	**2.915***
Image Segmentation	21.425	22.974	**22.202***
Leaf	0.406	0.315	**0.242***
Musk (Version 1)	38.181	24.774	**21.92***
Seeds	0.631	**0.603***	0.617
SPECTF Heart	6.91	6.346	**6.167***
Statlog (Vehicle Silhoue)	7.996	12.652	**7.498***
Wine	56.969	62.405	**46.86***
Yeast	**0.097***	0.1	0.098

Remark: * is the smallest RMSE for each data set.

Table 3 shows the RMSE of imputation method for 15% of missing data. The Corr-KNN method has lowest RMSE value for all datasets except the Statlog (Vehicle Silhoue) dataset. Corr-KNN is the most efficient method to impute the missing values because Corr-KNN is the method that approximates actual values more than any other method.

Table 3: RMSE of imputation method for 15 % of missing data

Dataset	Method		
	KNN	KNNFS	Corr-KNN
Breast Cancer Wisconsin (Diagnostic)	38.012	38.389	**35.956***
Forest type mapping	4.751	7.177	**3.899***
Image Segmentation	23.523	24.285	**23.442***
Leaf	0.4*3	0.349	**0.291***
Musk (Version 1)	40.665	30.268	**27.735***
Seeds	0.673	0.628	**0.619***
SPECTF Heart	7.304	6.81	**6.627***
Statlog (Vehicle Silhoue)	**11.097***	14.755	11.461
Wine	66.68	70.408	**62.26***
Yeast	**0.097***	0.098	**0.097***

Remark: * is the smallest RMSE for each data set.

Table 4 shows the RMSE of imputation method for 25% of missing data. The Corr-KNN method has lowest RMSE value for all datasets. Corr-KNN is the most efficient method to impute in missing values because Corr-KNN is the method that approximates actual values more than any other method.

Table 4: RMSE of imputation method for 25% of missing data

Dataset	Method		
	KNN	KNNFS	Corr-KNN
Breast Cancer Wisconsin (Diagnostic)	47.051	46.954	**44.473***
Forest type mapping	5.467	7.64	**4.804***
Image Segmentation	24.37	24.668	**23.919***
Leaf	0.517	0.628	**0.398***
Musk (Version 1)	43.943	36.451	**33.375***
Seeds	0.742	0.709	**0.684***
SPECTF Heart	7.657	7.284	**7.107***
Statlog(Vehicle Silhoue)	14.527	17.293	**13.516***
Wine	72.991	75.249	**68.499***
Yeast	**0.098***	0.099	**0.098***

Remark: * is the smallest RMSE for each data set.

Moreover, to simplify the performance of Corr-KNN, the RMSE of imputation methods for each case of missing data percentage on 10 datasets are illustrated in Figure 1. Figure 1 clearly shows that Corr-KNN outperformed than other methods by the smaller RMSE.

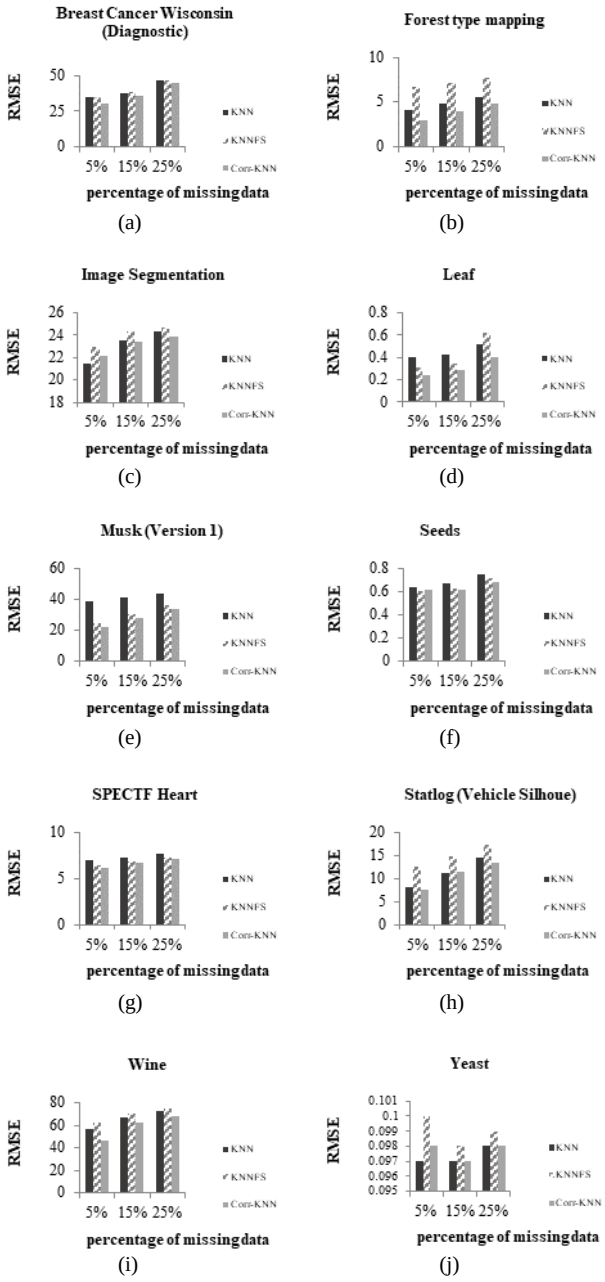


Figure 1: RMSE of imputed data for each imputation methods on 10 datasets.

(a) Breast Cancer Wisconsin (Diagnostic),
(b) Forest type mapping, (c) Image Segmentation, (d) Leaf,
(e) Musk (Version 1), (f) Seeds, (g) SPECTF Heart,
(h) Statlog (Vehicle Silhoue), (i) Wine, (j) Yeast

4 CONCLUSIONS

In this study, we proposed the missing value imputation, namely Corr-KNN. This Corr-KNN method is considering the correlation coefficient between variable having missing data and variables having complete data. After that, the variables having complete data which highly correlated with the variables having missing data are selected for using in K Nearest Neighbor (KNN) method for replacing missing data with substituted data. Based on the results from 10 real datasets, Corr-KNN has the powerful estimation ability for actual value more than KNN and KNNFS methods. Therefore, Corr-KNN can be applied for ensuring the accuracy of missing value imputation for data management before analyzing the data.

ACKNOWLEDGEMENTS

The authors are very thankful to Department of Statistics, Faculty of Science, Khon Kaen University for financial support.

REFERENCES

Gajawada, S., & Toshniwal, D. (2012). Missing value imputation method based on clustering and nearest neighbours. *International Journal of Future Computer and Communication, 1*(2), 206-208.

Meesad, P., & Hengpraprohm, K. (2008, June). Combination of knn-based feature selection and knnbased missing-value imputation of microarray data. In 2008 3rd International Conference on Innovative Computing Information and Control (pp. 341-341). IEEE.

Myneni, M. B., Srividya, Y., & Dandamudi, A. (2017). Correlated Cluster-Based Imputation for Treatment of Missing Values. In Proceedings of the First International Conference on Computational Intelligence and Informatics (pp. 171-178). Springer, Singapore. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Estimation of Population Mean Using a New Compromised Imputation Method for Missing Data in Survey Sampling

Kanisa Chodjuntug and Nuanpan Lawson*

Department of Applied Statistics, Faculty of Applied Science,
King Mongkut's University of Technology North Bangkok, Bangkok, Thailand
Email: kanisa.c@ubu.ac.th
*Corresponding email: nuanpan.n@sci.kmutnb.ac.th

ABSTRACT

Missing data is one of the serious problems that occurs in the business world. Making the decision based on the results analysis from incomplete data may leads to wrong decision. This paper presents the estimation of population mean using a new compromised imputation method for missing data. The bias and mean squared error of a proposed estimator are derived up to first degree of approximation. The efficiency of the proposed estimator is better than several other estimators. Theoretical findings are supported by empirical study to show the efficiency of the proposed estimator.

Keywords: Compromised imputation method; Mean square error; Ratio estimator

1 INTRODUCTION

In economics research, the data collection methods typically rely on survey sampling. It is usually impossible to survey an entire population due to its large size. For this reason, samples collected from the population have been used to infer about the characteristics of the population such as population mean, total, and proportion. One of the most common and serious problems in survey sampling is missing data, which can cause a significant effect to data analysis process. Missing data naturally occurs in data collection when a few sampling units refuse to respond or are unable to participate in the survey. Thus, the imputation is the most popular method used to substitute values for missing data which helps researchers deal with the problems efficiently.

In the statistical literature on missing data, Little and Rubin (1987) defined three mechanism: Missing Completely at Random (MCAR), Missing at Random (MAR) and Not Missing at Random (NMAR). While data are MCAR, if missingness does not depend on observed value and missing value, the data are MAR, if missingness is related to the observed data but not the missing values and the data are NMAR, if missingness is related to observed data but not the missing values. In this study, we implicitly assume MCAR which introduced by Heitjan and Basu (1996).

Let $\bar{Y} = N^{-1}\sum_{i=1}^{N} y_i$ be the population mean of study variable Y. A simple random sample without replacement (SRSWOR), s, of size n is drawn from finite population $\Omega = \{1, 2, \ldots, N\}$ of size N to estimate the population mean Y. Let r be the number of responding units out of sampled n and the number of non-responding units is $(n-r)$. Let the set of responding units be denoted by R and that of non-responding units be denoted by R^c. For sample units $i \in R$, the value y_i is observed. Nevertheless, for the units $i \in R^c$, the y_i values are missing and imputed values are to be derived. The imputation is carried out with the aid of a quantitative auxiliary variable x such that, the value of x for unit i is x_i. The value of x_i is known and positive value for every $i \in s$. Also, the data $x_s = \{x_i : i \in s\}$ are known.

In addition, the imputation technique is also applicable when information on an auxiliary variable is available. The auxiliary information has been used in practice to increase the precision of the estimators. When population parameters of the auxiliary variable are known, several estimators for population mean of study variable have been discussed in previously literature. For example, assume that a researcher wants to estimate the population mean of the household food cost. The researcher collects the household food cost and income data samples. According to the example, the household food cost is the study variable y and the household income is an auxiliary variable x. Over the last decade, several imputation techniques have been published and discussed. Lee et al. (1994) used the information on an auxiliary variable for the purpose of imputation. A few well-known imputation methods such as mean, ratio, product, and regression imputation methods have been used to improve the estimation of population mean with non-response. Singh and Horn (2000) suggested a compromised method under an imputation based estimator of population mean which used the information on an auxiliary variable. Ahmed et al. (2006) suggested several new imputation based estimators that used the information on an auxiliary variable and compared their performances with the mean method of imputation. Moreover, Kadilar and Cingi (2008), Singh et al. (2009) and Diana and Perri (2010) also suggested a few imputation techniques in the case of missing data. Singh et al. (2013) proposed a factor type compromised imputation method in case of missing data. Singh et al. (2014) and Singh and Gogoi (2017) suggested an exponential type compromised imputation method in the case of missing data. More recently, and Pal (2015) and Audu and Adewara (2017) suggested a new power transformation estimator of the population mean.

In this paper, we propose a new estimator for the estimation of population mean using missing data imputation. We suggest to adjust the estimator proposed by Singh et al. (2014) and then compare the bias and mean square error equations of the new proposed estimator with the naïve ones.

2 SOME IMPUTATION METHODS

There are some well-known imputation methods which are as follows:

2.1 Mean Method of imputation

Under this method, the study variable after imputation takes the form

$$y_{.i} = \begin{cases} y_i \ ; i \in R \\ \bar{y}_r \ ; i \in R^c, \end{cases} \quad (1)$$

The point estimator of population mean $\bar{Y}$ is given by

$$\bar{y}_s = \frac{1}{n}\sum_{i \in s} y_{.i} = \bar{y}_r, \quad (2)$$

where $\bar{y}_r = \frac{1}{r}\sum_{i \in R} y_{.i}$.

Lemma 2.1: The bias $Bias(\cdot)$ and the variance $V(\cdot)$ of $\bar{y}_s$ are respectively given by

$$Bias(\bar{y}_s) = 0 . \quad (3)$$

$$V(\bar{y}_s) = \left(\frac{1}{r} - \frac{1}{N}\right) S_Y^2 . \quad (4)$$

2.2 Ratio method of imputation

The ratio method of imputation is applied with the help of information obtained on an auxiliary variable x in the case of single value imputation. The i th unit requires imputation, the value $\hat{b}x_i$ is imputed, where $\hat{b} = \sum_{i \in R} y_i \Big/ \sum_{i \in R} x_i$. The study variable after imputation becomes

$$y_{.i} = \begin{cases} y_i \ ; i \in R \\ \hat{b}x_i \ ; i \in R^c . \end{cases} \quad (5)$$

Under this method of imputation, the point estimator of population mean is given by

$$\bar{y}_{RAT} = \bar{y}_r \frac{\bar{x}_n}{\bar{x}_r} , \quad (6)$$

where $\bar{x}_n = \frac{1}{n}\sum_{i \in s} x_i$ and $\bar{x}_r = \frac{1}{r}\sum_{i \in R} x_i$.

Lemma 2.2: The bias and the mean square error $MSE(\cdot)$ of $\bar{y}_{RAT}$ are respectively given by

$$Bias(\bar{y}_{RAT}) = \left(\frac{1}{r} - \frac{1}{n}\right) \bar{Y} \left[C_X^2 - \rho_{XY} C_Y C_X \right], \quad (7)$$

$$MSE(\bar{y}_{RAT}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_Y^2 + \left(\frac{1}{r} - \frac{1}{n}\right) \left[S_Y^2 + R_1^2 S_X^2 - 2R_1 S_{XY} \right], \quad (8)$$

Where $S_X^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N-1}$, $S_{XY}^2 = \frac{\sum_{i=1}^{N} (Y_i - \bar{Y})(X_i - \bar{X})}{N-1}$

$R_1 = \frac{\bar{Y}}{\bar{X}}$, $\bar{X} = \frac{\sum_{i=1}^{N} x_i}{N}$, $C_Y = \frac{S_Y}{\bar{Y}}$, $C_X = \frac{S_X}{\bar{X}}$, $\rho_{XY} = \frac{S_{XY}}{S_X S_Y}$.

2.3 Compromised method of imputation

Singh and Horn (2000) proposed the compromised imputation procedure, where the study variable after imputation takes the form

$$y_{.i} = \begin{cases} \alpha \frac{n}{r} y_i + (1-\alpha) \hat{b}x_i \ ; i \in R \\ (1-\alpha) \hat{b}x_i \qquad ; i \in R^c . \end{cases} \quad (9)$$

where α are suitably chosen constant, such that the resultant variance of estimator is minimum.

Thus, the point estimator of the population mean under the above imputation method becomes

$$\bar{y}_{COMP} = \alpha \bar{y}_r + (1-\alpha) \bar{y}_r \frac{\bar{x}_n}{\bar{x}_r} . \quad (10)$$

Lemma 2.3: The bias and the mean square error of $\bar{y}_{COMP}$ are respectively given by

$$Bias(\bar{y}_{COMP}) = (1-\alpha) \left(\frac{1}{r} - \frac{1}{n}\right) \bar{Y} \left[C_X^2 - \rho_{XY} C_Y C_X \right], \quad (11)$$

$$MSE(\bar{y}_{COMP}) = \left(\frac{1}{r} - \frac{1}{N}\right) \bar{Y}^2 C_Y^2 + \left(\frac{1}{r} - \frac{1}{n}\right) \bar{Y}^2 \left[(1-\alpha)^2 C_X^2 - 2(1-\alpha) \rho_{XY} C_Y C_X \right]. \quad (12)$$

and

$$MSE(\bar{y}_{COMP})_{\min} = MSE(\bar{y}_{RAT}) - \left(\frac{1}{r} - \frac{1}{n}\right) \left(1 - \rho_{XY} \frac{C_Y}{C_X}\right)^2 \bar{Y}^2 C_X^2 , \quad (13)$$

Where $\alpha_{opt} = 1 - \rho_{XY} \frac{C_Y}{C_X}$.

3 PROPOSED ESTIMATOR

Motivated by Singh et al. (2014), we propose a new compromised imputation method called an exponential type compromised method. The study variable after imputation is given as below.

$$y_{.i} = \begin{cases} \alpha \frac{n}{r} y_i + (1-k) \bar{y}_r \frac{\bar{X}}{\bar{x}_r} \exp\left(\frac{\bar{X} - \bar{x}_r}{\bar{X} + \bar{x}_r} \right) ; i \in R \\ (1-k) \bar{y}_r \frac{\bar{X}}{\bar{x}_r} \exp\left(\frac{\bar{X} - \bar{x}_r}{\bar{X} + \bar{x}_r} \right) \qquad ; i \in R^c . \end{cases} \quad (14)$$

The point estimator of the population mean under the new method of imputation is given as follows.

$$\bar{y}_{ET} = k\bar{y}_r + (1-k) \bar{y}_r \left(\frac{\bar{X}}{\bar{x}_r} \right) \exp\left(\frac{\bar{X} - \bar{x}_r}{\bar{X} + \bar{x}_r} \right) . \quad (15)$$

4 PROPERTIES OF THE PROPOSED ESTIMATOR

The bias and the mean square error of the proposed estimator up to the first degree of approximations are derived under the following transformations:

$\bar{y}_r = \bar{Y}(1+e_1)$, and $\bar{x}_r = \bar{X}(1+e_2)$.

Then, we have

$E(e_i) = 0, \ i = 1, 2$ and

$$E(e_1^2) = \left(\frac{1}{r} - \frac{1}{N}\right) C_y^2, E(e_2^2) = \left(\frac{1}{r} - \frac{1}{N}\right) C_x^2 , E(e_1 e_2) = \left(\frac{1}{r} - \frac{1}{N}\right) \rho C_Y C_X .$$

Under the above transformation, the estimator takes the following form

$$\bar{y}_{ET} = k\bar{Y}(1+e_1) + (1-k)\bar{Y}(1+e_1)(1+e_2)^{-1} \exp\left\{ -\frac{e_2}{2}\left(1 + \frac{e_2}{2}\right)^{-1} \right\}. \quad (16)$$

Now we have the following theorems:

Theorem 4.1:

The bias of the proposed estimator $\bar{y}_{ET}$ to the first degree of approximations is given by

$$Bias(\bar{y}_{ET}) = (1-k)\bar{Y} \frac{3}{8} \left(\frac{1}{r} - \frac{1}{N}\right) C_X \left[5C_X - 4\rho_{XY} C_Y \right] . \quad (17)$$

Expressing the estimator $\bar{y}_{ET}$ in terms of e_i 's, expanding the right hand side of the above expression, taking expectations and collecting the terms up to the first degree of approximations, we get the expression for bias of the estimator as given in (17).

Proof: By the definition of bias

$$Bias(\bar{y}_{ET}) = E(\bar{y}_{ET} - \bar{Y}) . \quad (18)$$

Such that

$$Bias(\bar{y}_{ET}) = E(\bar{y}_{ET} - \bar{Y})$$
$$= E\left[k\bar{Y}(1+e_1) + (1-k)\bar{Y}(1+e_1)(1+e_2)^{-1} \exp\left\{ -\frac{e_2}{2}\left(1+\frac{e_2}{2}\right)^{-1} \right\} - \bar{Y} \right]$$
$$= E\left[\bar{Y}(1+e_1)\left(k + (1-k)\bar{Y}\left(1+e_2+e_2^2\right)\left\{ 1 - \frac{e_2}{2} + \frac{3e_2^2}{8} \right\} \right) - \bar{Y} \right]$$
$$= E\left[\bar{Y}\left(e_1 + (1-k)\left(-\frac{3e_2}{2} + \frac{15e_2^2}{8} - \frac{3e_1e_2}{2} \right) \right) \right].$$

Therefore

$$Bias(\bar{y}_{ET}) = (1-k)\bar{Y}\frac{3}{8}\left(\frac{1}{r} - \frac{1}{N}\right) C_X \left[5C_X - 4\rho_{XY}C_Y \right].$$

Theorem4.2:

The mean square error of the proposed estimator up to the first order of approximations is given by,

$$MSE(\bar{y}_{ET}) = \bar{Y}^2\left(\frac{1}{r} - \frac{1}{N}\right) \left(C_Y^2 + (1-k)\frac{3}{4}C_X\left[(1-k)3C_X - 4\rho_{XY}C_Y \right] \right). \quad (19)$$

Now using the expression given in equation (16) for $\bar{y}_{ET}$, expanding the terms and taking expectations and retaining the terms up to the first degree of approximations we get expression for mean square error as given in equation (19).

Proof: By the definition of mean square error

$$MSE(\bar{y}_{ET}) = E(\bar{y}_{ET} - \bar{Y})^2. \quad (20)$$

Such that

$$MSE(\bar{y}_{ET}) = E\left[(\bar{y}_{ET} - \bar{Y})^2 \right]$$
$$= E\left[\left(k\bar{Y}(1+e_1) + (1-k)\bar{Y}(1+e_1)(1+e_2)^{-1} \exp\left\{ -\frac{e_2}{2}\left(1+\frac{e_2}{2}\right)^{-1} \right\} - \bar{Y} \right)^2 \right]$$
$$= E\left[\left(\bar{Y}(1+e_1)\left(k + (1-k)\bar{Y}\left(1+e_2+e_2^2\right)\left\{ 1 - \frac{e_2}{2} + \frac{3e_2^2}{8} \right\} \right) - \bar{Y} \right)^2 \right]$$
$$= \bar{Y}^2 E\left[\left(e_1^2 - (1-k)3e_1e_2 + (1-k)^2\frac{9}{4}e_2^2 \right) \right].$$

Therefore

$$MSE(\bar{y}_{ET}) = \bar{Y}^2\left(\frac{1}{r} - \frac{1}{N}\right)\left(C_Y^2 + (1-k)\frac{3}{4}C_X\left[(1-k)3C_X - 4\rho_{XY}C_Y \right] \right).$$

Theorem4.3:

The minimum mean square error of the proposed estimator is given by,

$$MSE(\bar{y}_{ET})_{\min} = \left(\frac{1}{r} - \frac{1}{N}\right)\bar{Y}^2\left[C_Y^2\left(1-\rho_{XY}^2\right) \right]. \quad (21)$$

Since the mean square error of $\bar{y}_{ET}$ as given in (19) is a function of unknown constant k . Therefore, it is natural to search for an optimum value of k , such that the mean square error of the proposed estimators becomes minimum. Hence differentiating equation (19) with respect to k and equating to zero, so we get optimum value of k as

$$k = 1 - \frac{2}{3}\rho_{xy}\frac{C_y}{C_x}. \quad (22)$$

Putting the value of k from the equation (22) into the equation (19), we get the minimum mean square error of $\bar{y}_{ET}$ as defined in (21), which completes the proof.

5 EFFICIENCY COMPARISON OF THE ESTIMATOR

On the basis of expressions of mean square errors of the proposed estimator $(\bar{y}_{ET})_{\min}$ with those of estimators $\bar{y}_s$, $\bar{y}_{RAT}$ and $\bar{y}_{COMP}$ then, we can observe the efficiency of the proposed estimator.

5.1 *Comparison of the estimator* $(\bar{y}_{ET})_{\min}$ *and the estimator* $\bar{y}_s$

$$V(\bar{y}_s) - MSE(\bar{y}_{ET})_{\min} = \bar{Y}^2\left(\frac{1}{r} - \frac{1}{n}\right)\rho_{YX}^2 C_Y^2 > 0 \quad (23)$$

Equation (23) is always true. Hence the estimator $(\bar{y}_{ET})_{\min}$ is better than the estimator $\bar{y}_s$ under optimality condition (21).

5.2 *Comparison of the estimator* $(\bar{y}_{ET})_{\min}$ *and the estimator* $\bar{y}_{RAT}$

$$MSE(\bar{y}_{RAT}) - MSE(\bar{y}_{ET})_{\min}$$
$$= \bar{Y}^2\left[\left(\frac{1}{r} - \frac{1}{n}\right)\left(C_X - \rho_{XY}C_Y\right)^2 + \left(\frac{1}{n} - \frac{1}{N}\right)\rho_{YX}^2 C_Y^2 \right] > 0 \quad (24)$$

Equation (24) is always true. Hence the estimator $(\bar{y}_{ET})_{\min}$ is better than the estimator $\bar{y}_{RAT}$.

5.3 *Comparison of the estimator* $(\bar{y}_{ET})_{\min}$ *and the estimator* $(\bar{y}_{COMP})_{\min}$

$$MSE(\bar{y}_{COMP})_{\min} - MSE(\bar{y}_{ET})_{\min} = \bar{Y}^2\left(\frac{1}{n} - \frac{1}{N}\right)\rho_{YX}^2 C_Y^2 > 0 \quad (25)$$

Equation (25) is always true. Thus, it can be concluded that the proposed estimator $(\bar{y}_{ET})_{\min}$ is always preferable over the estimator $(\bar{y}_{COMP})_{\min}$.

6 EMPIRICAL STUDY

For the empirical study of the proposed strategy with other existing imputation strategies we consider the following data from Rachokarn and Lawson (2017). The data are taken from the Government of India for the West Bengal state in 1981. The data belongs to the population census of 96 villages. This study assumed that the number of agricultural labors in the village was taken as study variable Y and the area of the village as taken as auxiliary variable X . The following values are obtained for the considered variables.

$N = 96$, $\bar{Y} = 137.93$, $\bar{X} = 144.87$, $S_Y = 182.50$, $S_X = 117.57$, $C_Y = 1.32$, $C_X = 0.82$, $\rho_{XY} = 0.77$, $n = 23$, $r = 20$

Table 1: Bias and mean square errors for the proposed estimator and other existing estimators

Estimators	Bias	MSE
$\bar{y}_s$	0	1318.37
$\bar{y}_{RAT}$	-0.14	1090.21
$\bar{y}_{COMP}$	-0.18	725.58
$\bar{y}_{ET}$	0.05	534.17

From Table 1, we can see that the proposed estimator $\bar{y}_{ET}$ is more efficient than the other existing estimators in term of minimum mean square error. The estimator using mean imputation method $\bar{y}_s$ performs the worse when compared to other estimators in term of mean square error. In term of bias, we can see that the estimator $\bar{y}_s$ is an unbiased estimator. Nevertheless, the proposed estimator $\bar{y}_{ET}$

has a smaller bias when compared to $\bar{y}_{RAT}$ and $\bar{y}_{COMP}$, and it is closed to zero.

7 CONCLUSIONS

In this study, we proposed a new estimator for the estimation of population mean using exponential type compromised imputation. The proposed estimator was useful when a few observations were missing in the survey sampling and population mean of auxiliary information was known. With support from the proposed estimator, mean square errors of the proposed estimator were less than other existing estimators; therefore, the proposed estimator was more efficient than the other existing estimators. In addition, the empirical study was carried out to further prove the efficiency of the proposed estimator. In conclusion, the new estimator of the population mean for missing data was successful and provided guidelines when selecting the appropriate estimator under missing data conditions.

REFERENCES

Ahmed, M.S., Al-Titi, O., Al-Rawi, Z., & Abu-Dayyeh, W. (2006). Estimation of a population mean using different imputation methods. *Statistics in Transition*, 7(6), 1247-1264.

Audu, A., & Adewara, A.A. (2017). A modifiled factor type estimator in two phase sampling. *Punjab University Journal of Mathematics*, 49(2),59-73.

Diana, G., & Perri, P.F. (2010). Improved estimators of the population mean for missing data. *Communications in Statistics – Theory and Methods*, 39(18), 3245-3251.

Heitjan, D.F., & Basu, S. (1996). Distinguishing missing at random and missing completely at random. *The American Statistician*, 50(3), 207-213.

Kadilar, C., & Cingi, H. (2008). Estimators for the population mean in the case of missing data. *Communications in Statistics – Theory and Methods*, 37(14), 2226-2236.

Little, R.J.A., & Rubin, D.B. (1987). Statistical Analysis with Missing Data, John Wiley and Sons, NY.

Lee, H., Rancourt, E., & Sarndal, C.E. (1994). Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10(3), 231-243.

Rachokarn, T., & Lawson, N. (2017). A class of ratio chain type exponential estimator for population mean in the presence of non-response. *International Journal of Agricultural and Statistical Sciences*. 13(2), 431-437.

Singh, S., & Horn, S. (2000). Compromised imputation in survey sampling. *Metrika*, 51, 266–276.

Singh, D.R., Singh, N.T, & Shukla, D. (2009). Estimation mean under imputation of missing data using factor type estimator in two phase sampling. *An International Journal of the Polish Statistical Association*, 10(3), 397-414.

Singh, D.R., Singh, N.T., & Shukla, D. (2013). Some new aspects on imputation in sampling. *African Journal of Mathematics and Computer Science Research*, 6(1), 5-15.

Singh, A.K., Singh, P., & Singh, V.K. (2014). Exponential - type compromised imputation in survey sampling. *Journal of Statistics Applications & Probability*, 2, 211-217.

Singh, H.P., & Pal, S.K. (2015). A new chain ratio-ratio-type exponential estimator using auxiliary information in sample surveys. *International Journal of Mathematics And its Applications*, 3(4), 37–46.

Singh, B.K., & Gogoi, U. (2017). Estimation of population mean using exponential dual to ratio type compromised imputation for missing data in survey sampling. *Journal of Statistics Applications & Probability*, 6(3), 515-522.

Missing Data Imputation in Multiple Linear Regression Analysis

Supreeya Srasom* and Tidadeaw Mayureesawan

Department of Statistics, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand
*Corresponding Email: bee.preeyaa@gmail.com
Email: tidadeaw@kku.ac.th

ABSTRACT

The objective of this research is to propose a missing data imputation in multiple linear regression analysis with missing at random dependent variable, namely the Mean Regression Imputation (MRI) method was developed from the Mean Imputation and the Regression Imputation (RI) method. The MRI method is compared the efficiency with 4 methods of missing data imputation which are the Regression Imputation (RI) method, the Mean Imputation method, the Stochastic Regression Imputation (SRI) method and the Expectation Maximization Algorithm (EM Algorithm) method by using the Monte Carlo simulation with R program. For the study three independent variables with sample size (n) 30, 50, and 100; standard deviations of error (σ) 5, 10 and 15; missing percentage 5, 10 and 15. The criteria used to compare performance is an Average Mean Square Error (AMSE). The result of this research indicates that the σ is equal to 10, the proposed MRI method has performance better than the other methods in case of n is equal to 30 at missing percentage are equal to 5 and 10. When the σ is equal to 15, the MRI method has the best performance for all levels of n and all missing percentages.

Keywords: Multiple Linear Regression; Missing Data Imputation; Mean Regression Imputation

1 INTRODUCTION

Prediction is a popular technique that used in many research such as finance economics, social sciences, science, medicine and industry. Because of the current economics and social conditions are more complicated, they always fluctuate and change. This may affect planning of work. Therefore, if you can predict results in the future, it will be more beneficial for planning of work than ignoring it. This can reduce the risk of decisions, erroneous operation or can prepare for any problems that may occur in the future.

Multiple linear regression analysis is data analysis by using independent variables that has more than one variable to explain independent variable. The multiple linear regression models will indicate the average linear correlation between the group of independent variables and the dependent variable. This makes our use this relationship to predict the value of the dependent variable in the future.

In the case of data collection, the data often have problem or missing data. The survey research found the missing data, and there are 2 causes. The first caused is the non-response for some sample or "Unit nonresponse". The second cause is the item non-response for some question. If you bring the incomplete analysis of data to use, the analysis may produce inaccurate results. That can be impact to use of decision-making in various tasks. If the incomplete data is used in regression analysis, the efficiency of data analysis will be reduced. Then, results may biased.

The study of the research related to the missing data of the variables according to the multiple regression analysis showed that Saengsuwan (2008) compared 4 methods of missing data imputation: the Loss Imputation method (Loss method), the Mean Imputation method (Mean method), the Regression Imputation method (RI method) and the Multiple Imputation method (MI method). The results indicate that the RI method is the most effective when the missing percentage is increased. Wongarmart (2012) compared 3 methods for multiple linear regression analysis of missing data imputation: the Expectation Maximization Algorithm method (EM Algorithm method), the K Nearest Neighbor Imputation method (KNN method) and the Predictive Mean Matching Imputation method (PMM method). The results indicate that the EM Algorithm method has the best performance when the standard deviations of error are not high (10-30) and the KNN method has the best performance when the standard deviations of error is high (90). And Lamjaisue (2017) compared 6 methods of missing data imputation: the RI method, the Stochastic Regression Imputation method (SRI method), the KNN method, the EM Algorithm method, the K Nearest Regression Imputation with Equivalent Weighted method (KREW method) and the K Nearest Stochastic Regression Imputation with Equivalent Weighted method (KSEW method). The KREW method is missing data imputation method of combining 2 methods: the KNN method and the RI method then bring to be weighted by Equivalent Weight (EW) method. The KSEW method is missing data imputation method of combining 2 methods: the KNN method and the SRI method then bring to be weighted by EW method. The results indicate that the KSEW method has the best performance when the sample sizes are equal to 20 and 30. The SRI method has the best performance when the sample sizes are equal to 50 and 100.

Based on the above research, the researcher proposed the method of missing data imputation, namely the Mean Regression Imputation method (MRI method) was developed from the Mean method and the RI method. So, in this research, we compared the efficiency with 5 methods of missing data imputation of the dependent variable which are the Mean Imputation method, the RI method, the SRI method, the EM Algorithm method and the MRI method. The criteria used to compare performance is an Average Mean Square Error (AMSE).

2 METHODS

The objective of this research is to propose a missing data imputation in multiple linear regression analysis with missing at random dependent variable. Monte Carlo simulations were performed under different circumstances, as follows;

2.1 Determine an independent variable with normal distribution. In this study will use 3 independent variables by $X_k \sim N(0,100); k = 1,2,3$

2.2 Define an error with a normal distribution $(\varepsilon_i \sim N(0,\sigma^2))$ that have an average value equal to 0, the standard deviations of error equal to 5, 10 and 15, respectively.

2.3 Create the dependent variable that has a linear relationship with independent variable. This uses the linear relationship model by the following:

$$y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \quad (1)$$

When i = 1, 2, …, m, m+1, …, n at n are equal to 30, 50 and 100. Determine the regression coefficient $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 1$. The independent variables are non-missing. $y_1, \ldots, y_m$ are the data of the dependent variable as there is non-missing and $y_{m+1}, \ldots, y_n$ are the data of the dependent variable as there is missing.

2.4 Assign the missing percentage 5, 10 and 15 with missing at random dependent variable.

2.5 Estimate 5 missing data methods for imputation as follows:

1) Mean Imputation method (Mean method)

Mean is a method that used to impute in the missing data of the dependent variables by using the average value of the dependent

variable based on only the data is non-missing (Barnett, 2002) as follows:

$$\bar{y}_j = \frac{\sum_{i=1}^{m} y_i}{m} \tag{2}$$

When $\bar{y}_j$ is The estimated value of the missing observation value of the dependent variable, the observation value that $j; j = m+1,\ldots,n$; y_i is observed values of the dependent variable is non-missing; m is number of observed values of the dependent variable is non-missing.

2) Regression Imputation method (RI method)

RI is a method that regression models to estimate the missing data. The data sets are non-missing (x_i, y_i) ; $i = 1,2,\ldots,$m create the regression model with Ordinary Least Squares (OLS) method. The regression model was used to predict the missing values of the dependent variable (Little and Rubin, 2002) as follows:

$$y_j = \beta_o + \beta_1 x_{j1} + \ldots + \beta_k x_{jk} \tag{3}$$

When y_j is The predicted value of the dependent variable of the observation value $j; j = m+1,\ldots,n$; $\beta_0, \beta_1, \ldots, \beta_k$ are Estimates of regression coefficients; $x_{j1}, x_{j2}, \ldots, x_{jk}$ are The observation value of the independent variables that the dependent variable is missing; m is number of observed values of the dependent variable is non-missing.

3) Stochastic Regression Imputation (SRI) method

SRI is a method that uses regression models from non-missing data sets like the RI method. But different from the RI method is the regression model that predicts the missing values of the dependent variable, It adds a residual term to the regression model (Enderers, 2008) as follows:

$$\hat{y}_j = \beta_o + \beta_1 x_{j1} + \ldots + \beta_k x_{jk} + z_i \tag{4}$$

When y_j is The predicted value of the dependent variable of the observation value $j; j = m+1,\ldots,n$; $\beta_0, \beta_1, \ldots, \beta_k$ are Estimates of regression coefficients; $x_{j1}, x_{j2}, \ldots, x_{jk}$ are The observation value of the independent variables that the dependent variable is missing; m is number of observed values of the dependent variable is non-missing; z_i is an estimate of error with the normal distribution is the mean of 0 and the variance equal to the variance of error in the regression model. This value is derived from random, using the Monte Carlo method.

4) Expectation Maximization Algorithm method (EM Algorithm method)

Little and Rubin (2002) proposed the EM Algorithm method used to fill the missing data of the dependent variable in multiple linear regression analysis. The EM algorithm method is a repetitive process for estimating the maximum likelihood of a parameter when some data is missing. The missing of data is divided into 2 main steps. First, the E-step is the process of finding the expected value of missing data under the conditions of the non-missing data set to bring this expectation to the missing data. Second, the M-step is the procedure for estimating the maximum probability of a parameter, the missing of data is estimated by the E-step. The steps are as follows:

Step 1 manage data set $\mathbf{y = X\beta + \varepsilon}$ into 2 parts are non-missing data and missing data.

$$\begin{bmatrix} \mathbf{y_1} \\ \mathbf{y_2} \end{bmatrix} = \begin{bmatrix} \mathbf{X_1} \\ \mathbf{X_2} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{5}$$

When y_1 is vector of dependent variable that is non-missing data, which is the size $m \times 1$; y_2 is vector of dependent variable that is missing data, which is the size $(n-m) \times 1$; X_1 is the matrix of independent variable that data set of the dependent variable that is non-missing data, which is the size $m \times$(k+1) ; X_2 is the matrix of independent variable that data set of the dependent variable that is missing data, which is the size $(n-m) \times$(k+1) ; β is vector of parameter by size (k+1)$\times 1$; ε is vector of error by size $n \times 1$.

Step 2 estimates linear regression coefficients, starting with the OLS method from the complete data set currently available, calculated from the formula:

$$\boldsymbol{\beta}^{(0)} = (\mathbf{X_1'X_1})^{-1}\mathbf{X_1'y_1} \tag{6}$$

Step 3 enter the E-step step by taking the value of $(\boldsymbol{\beta}^{(0)})$ from step 2 to the expected value. For estimate the missing value of the dependent variable that is missing. To repeat the first round calculated from the formula:

$$E(y_i \,|\mathbf{X}, \mathbf{y_1}, \boldsymbol{\beta}^{(0)}) = \begin{cases} \mathbf{y_i} & ; i = 1,2,\ldots,m \\ \hat{\boldsymbol{\beta}}_{\mathbf{0}}^{(0)} + \sum_{k=1}^{p} \hat{\boldsymbol{\beta}}_{\mathbf{k}}^{(0)} \mathbf{x_{jk}} & ; j = m+1,\ldots,n \end{cases} \tag{7}$$

Step 4 enter the M-step procedure by replacing the missing value with the estimated value in step 3 and then calculating the regression coefficient. To repeat the first round calculated from the formula:

$$\boldsymbol{\beta}^{(1)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^{(1)} \tag{8}$$

Step 5 determine the absolute value of the difference between the regression coefficients in step 2 and step 4 with a value less-than or equal to 0.001. If the absolute value of the difference between all regression coefficients is greater than 0.001, continue to step 6.

Step 6 estimates the new missing value, which returns to the E-step. Repeat the cycle at t ; $t = 2,3,\ldots$ with estimate new regression coefficient that calculated from replacing missing data, calculated from the formula:

$$E(y_i \,|\mathbf{X}, \mathbf{y_1}, \hat{\boldsymbol{\beta}}_{\mathbf{0}}^{(\mathbf{t-1})}) = \begin{cases} \mathbf{y_i} & ; i = 1,2,\ldots,m \\ \hat{\boldsymbol{\beta}}_{\mathbf{0}}^{(\mathbf{t-1})} + \sum_{k=1}^{p} \hat{\boldsymbol{\beta}}_{\mathbf{k}}^{(\mathbf{t-1})} \mathbf{x_{jk}} & ; j = m+1,\ldots,n \end{cases} \tag{9}$$

Enter M-step method to repeat the cycle at t to calculate the new regression coefficient, calculated from the formula:

$$\boldsymbol{\beta}^{(\mathbf{t})} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^{(\mathbf{t})} \tag{10}$$

Do this repeatedly, until the absolute value of the difference between the regression coefficients is less than or equal to 0.001, so stop.

5) Mean Regression Imputation method (MRI method)

MRI is a method of missing data developed from Mean method and RI method. The imputation procedure by the MRI method is performed as follows:

Step 1 find the average of the dependent variables based only on the non-missing data, calculated from this formula:

$$\bar{y}_j = \frac{\sum_{i=1}^{m} y_i}{m} \tag{11}$$

Then the average values add to the missing value. Each dependent variable has the same value is $\bar{y}_j$.

When $\bar{y}_j$ is The estimated value of the missing observation value of the dependent variable, the observation value that $j; j = m+1,\ldots,n$; y_i is observed values of the dependent variable is non-missing; m is number of observed values of the dependent variable is non-missing.

Step 2 use the data from step 1 to create the regression model by using OLS method, and calculated from this formula:

$$\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{12}$$

When $\mathbf{Y}$ is The data set is non-missing for the dependent variable; $\mathbf{X}$ is The data set of independent variables as the dependent variable is non-missing.

Step 3 apply the regression model obtained in step 2 to predict dependent variable which is missing value as follows:

$$y_j = \beta_o + \beta_1 x_{j1} + \ldots + \beta_{jk} x_{ik} \tag{13}$$

When y_j is The predicted value of the dependent variable of the observation value $j; j = m+1,\ldots,n$; $\beta_0, \beta_1, \ldots, \beta_k$ are Estimates of regression coefficients; $x_{j1}, x_{j2}, \ldots, x_{jk}$ are The observation value of the independent variables that the dependent variable is missing; *m* is number of observed values of the dependent variable is non-missing.

2.6 Estimate the regression coefficient with the OLS method at complete data from estimates obtained in each method of missing data imputation method to generate multiple linear regression models.

2.7 Calculate the AMSE value from the 5 missing data imputation methods by using the Monte Carlo simulation in each situation, a number of 1,000 cycles to compare the efficiency of the estimated missing data from the AMSE. Estimating any missing data that provides AMSE value the lowest, it will be method that is the most effective way to estimate the missing data.

3 RESULTS

When the standard deviation of error **(σ)** is equal to 5 at all sample sizes (*n*), the method of missing data imputation that give the lowest AMSE values at every situation is the RI method at all levels of missing percentage as shown in Table 1.

Table 1 The AMSE of 5 methods of missing data imputation when the standard deviation of the error is equal to 5.

Sample Size	Missing Percentage	Missing Data Imputation				
		RI	**SRI**	**Mean**	**MRI**	**EM**
30	5	**3.6577**	4.0138	7.5873	3.8572	4.9404
	10	**2.0825**	2.2517	4.5673	2.1742	2.3392
	15	**1.0101**	1.1060	2.5101	1.0256	1.0680
50	5	**3.7757**	4.3673	11.008	4.4027	6.6067
	10	**2.1868**	2.5474	7.2627	2.4491	2.9983
	15	**1.0718**	1.3854	5.5019	1.1926	1.3958
100	5	**4.1804**	5.4166	19.2042	6.2434	14.3268
	10	**2.3594**	3.1965	13.4321	3.3268	5.5245
	15	**1.1413**	1.7922	10.0024	1.5840	2.1867

When the σ is equal to 10 at *n* is equal to 30, the method of missing data imputation that give the lowest AMSE values is the MRI method except *n* are equal to 50 and 100, the method that give the lowest AMSE values is the RI method at missing percentage are equal to 5 and 10. At all *n* is equal to 15, the EM Algorithm method give the lowest AMSE values at missing percentage is equal to 15 as shown in Table 2.

Table 2 The AMSE of 5 methods of missing data imputation when the standard deviation of the error is equal to 10.

Sample Size	Missing Percentage	Missing Data Imputation				
		RI	**SRI**	**Mean**	**MRI**	**EM**
30	5	14.9066	15.9547	17.7243	**14.8699**	16.3748
	10	8.7524	9.5297	10.486	**8.7284**	9.1456
	15	4.0347	4.4583	5.097	4.0217	**2.0714**
50	5	**15.4895**	17.3137	20.3797	15.5313	18.9922
	10	**9.2001**	10.8422	13.3555	9.2256	10.5382
	15	4.2655	5.571	7.8127	4.2859	**2.3298**
100	5	**16.691**	21.0169	28.7106	17.8276	28.9357
	10	**10.0095**	13.5055	19.2341	10.3871	14.4935
	15	4.5878	7.2271	12.0985	4.8178	**2.918**

When the σ is equal to 15 for all *n*, the method of missing data imputation that give the lowest AMSE values at every situation that is the MRI method at all missing percentages, then the MRI method has the best performance as shown in Table 3.

Table 3 The AMSE of 5 methods of missing data imputation when the standard deviation of the error is equal to 15.

Sample Size	Missing Percentage	Missing Data Imputation				
		RI	**SRI**	**Mean**	**MRI**	**EM**
30	5	32.3858	35.4812	32.7930	**31.6848**	35.5117
	10	19.0018	20.9493	19.5622	**18.7109**	19.8278
	15	9.5532	10.3818	10.1501	**9.4952**	9.7067
50	5	33.6241	38.7934	34.8498	**32.654**	39.4330
	10	19.9614	23.885	22.2844	**19.4725**	22.2659
	15	10.1276	12.9711	12.6256	**9.9999**	10.7833
100	5	37.2673	47.6968	42.5054	**35.9077**	55.2362
	10	21.2352	29.4440	27.5184	**20.7114**	27.3367
	15	10.7221	16.5494	16.2682	**10.5421**	12.6189

4 CONCLUSIONS

To the compare the efficiency of the propose Mean Regression Imputation (MRI) method with 4 methods of missing data imputation for multiple linear regression analysis with missing at random dependent variable, based on the AMSE value indicates that for the standard deviation of error (σ) is equal to 15 for all sample size (*n*), the MRI method has the best performance of missing data imputation at all missing percentages. For the σ is equal to 10 and *n* is equal to 30, the MRI method is the most effective of missing data imputation at all missing percentages. For the σ is equal to 5 for all *n*, the MRI method and the RI method are relatively small differences between the AMSE values at all missing percentages.

REFERENCES

Barnett, V. (2002). Sample Survey Principle & Methods. *New York:* Arnold.

Enderers, C.K. (2008). Applied Missing Data Analysis. *New York: The Guilford Press.*

Lamjaisue, R. (2017). Comparison of missing data estimation methods for the multiple regression analysis with missing at random dependent variable. *Thammasat Journal of Science and Technology*, *25*(5), 766-777.

Little, R.J.A., & Rubin, D.B. (2002). Statistical Analysis with Missing Data. *New York: John Wiley & Sons.*

Saengsuwan, C. (2008). A Comparative Study of Missing Data Estimation Methods in Multiple Regression Analysis. *The Journal of Applied Science, 7*(1), 1-7.

Wongarmart, A. (2012). Comparison of the Estimation Methods for Nonignorable Missing Data in Multiple Linear Regression. *Chulalongkorn University Intellectual Repository.*

Missing Data Imputation Based on Accuracy of Binary Classification.

Jumlong Vongprasert*

Applied Statistics Department, Faculty of Science, Ubon Ratchathani Rajabhat University, Ubon Ratchathani
*Corresponding Email: jumlong.v@ubru.ac.th.

ABSTRACT

The purpose of this study was to comparative accuracy of binary classification based on missing data imputations methods namely; Support Vector Machines (SVM); Neural Networks (NN); Random Forests (RF); Multiple Imputation (MI) and Bagged Tree Imputation (BTI). Three data sets which comprises: i) 7 categorical and 9 continuous independent variables, ii) 9 categorical independent variables and iii) 9 continuous independent variables. The comparisons were made with the following conditions: (i) Three data sets; (ii) three types of missing data: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR); (iii) six level of percentage of missing data (5, 10, 15, 20, 25 and 30). We analyze which imputation method influences most the classifiers' accuracy. The best imputations in overall were obtained using RF and SVM, the imputation under MAR and MCAR were obtained using SVM, the imputation under NMAR were obtained using RF. The imputations under percentage of missing were obtained using SVM or RF

Keywords: missing data imputation; binary classification

1. INTRODUCTION

Missing values are unavoidable in real world datasets, there is a variety of causes why data may be missing. Anyone who does statistical data analysis of any kind runs into the problems of missing data. In a characteristic dataset we always land up in some missing values for attributes. The most serious concern is that missing data can introduce bias into estimates derived from a statistical model (Rubin, 1987; Becker & Walstad, 1990; Becker & Powers, 2001). If the responses are not ignorable, however, estimation of the propensity scores is complicated and often requires additional surrogate (Chen et al., 2008) or instrumental variables (Kott & Chang, 2010) to estimate the model parameters consistently. Missing data analysis is importance since an inference based ignoring the missingness may not only misleading conclusions, but also lose efficiency and lead to biased results. (Little & Rubin, 2002)

2. MISSING DATA

Let $\mathbf{y}=\left(y_1,y_2,\cdots y_n\right)'$ denote the complete set of the outcome variables, and $\boldsymbol{\delta}=\left(\delta_1,\delta_2,\cdots\delta_n\right)'$ be the vector of missing data indicators such that $\delta_i=1$ when y_i is observed and $\delta_i=0$ when y_i is missing. We note that each y_i and the corresponding δ_i can also be vectors. Let $\mathbf{y}_{\text{obs}}$ denote the observed and $\mathbf{y}_{\text{mis}}$ missing components of $\mathbf{y}$. With the above notation, the missing data mechanisms are characterized by the conditional distribution of $\boldsymbol{\delta}$ given $\mathbf{y}$, say $f\left(\boldsymbol{\delta}\middle|\mathbf{y},\phi\right)$, where ϕ denotes some unknown parameters.

Three major types of missing data are (Little & Rubin, 2002):

Missing completely at random (MCAR) denotes the mechanism that missingness does not depend on the values of the data $\mathbf{y}$, missing or observed.

$$f\left(\boldsymbol{\delta}\middle|\mathbf{y},\phi\right)=f\left(\boldsymbol{\delta}\middle|\phi\right)\forall\mathbf{y},\phi \qquad (1)$$

Missing at random (MAR) denotes the mechanism that missingness only depends on the components y_{obs} of that are observed, and not on the components that are missing.

$$f\left(\boldsymbol{\delta}\middle|\mathbf{y},\phi\right)=f\left(\boldsymbol{\delta}\middle|\mathbf{y}_{\text{obs}},\phi\right)\forall\mathbf{y}_{\text{mis}},\phi \qquad (2)$$

Not missing at random (NMAR) denotes the one that the distribution of $\mathbf{y}$ does depend on the missing values in the data.

$$f\left(y_i\middle|x_i,\delta_i=0,\phi\right)\neq f\left(y_i\middle|x_i,\delta_i=1,\phi\right) \qquad (3)$$

3. DATA SET

In this section, we introduce and describe the data set. We used three data set from University of California Irvine Machine Learning Repository, i) Bank Marketing data set with 7 categorical and 9 continuous independent variables and 1,000 instances by simple random sampling from 45,211 instances, ii) Wisconsin Breast Cancer Database with 9 categorical independent variables and 700 instances and iii) Breast Cancer Coimbra Data Set with 9 continuous independent variables and 116 instances.

4. RESEARCH METHODOLOGY

Methods

In this section, we introduced and described the methods applied to impute the original incomplete data set. The five imputation techniques applied are: Support Vector Machines (SVM); Neural Networks (NN); Random Forests (RF); Multiple Imputation (MI) and Bagged Tree Imputation (BTI).

Support Vector Machines (SVM)

SVM are learning machines based on the statistical learning theory, which can use linear and nonlinear kernels for the classification. They minimize the structure risk in a higher dimensional feature space, searching for the hyperplane with the largest margin between the classes. (Xin et al., 2010). SVM are useful approach for solving data classification and recognition problems. In this work we used the SVM implementation from the Radial Basis Function (RBF) kernel from R package 'kernlab'.

Neural Networks (NN)

A number of approaches have been investigated and applied to solve the missing data of this research includes NN, because of their flexibility, fault tolerance and capability to handle incomplete data. NN models have previously been applied to solve different tasks of missing data comprised of neural networks as a key classifier (Ibrahim, Abdullah and Saripan, 2009). In this work we used the NN implementation from the R package 'nnet'.

Random Forests (RF)

RF (Breiman, 2001) is a machine learning technique that builds a multitude of weak decisional trees at training time and outputs the class that is the mode of the classes (classification) or average prediction (regression) of the individual trees. Each tree is individually trained on a sample of the training data, and at each node, the algorithm only searches across a random subset of the features to determine a split. The input vector to be classified is submitted to each of the decision trees in the forest and the prediction is then formed using a majority vote. (Jordanov, Petrov and Petrozziello, 2018). In this work we used the NN implementation from the R package 'randomForest.

Multiple Imputation (MI)

MI is a statistical technique for analyzing incomplete data sets, that is, data sets for which some entries are missing. Application of the technique requires three steps: imputation, analysis and pooling. The figure illustrates these steps.

Imputation: Impute the missing entries of the incomplete data sets, not once, but m times. Imputed values are drawn for a distribution. This step results are m complete data sets.

Analysis: Analyze each of the m completed data sets. This step results in m analyses.

Pooling: Integrate the m analysis results into a final result. Simple rules exist for combining the m analyses.

For imputing the missing data, we use MI algorithm (Verboven et al., 2007), implemented in the amelia function from the Amelia package

Bagged Tree Imputation (BTI)

Bagging predictors approach generates manifold versions of a predictor to get an aggregated one. The aggregation function usually is the average value over all predictor estimations for a numerical outcome, or employs a majority vote when the desired output is a categorical one. The multiple predictions are estimated by bootstrapping from the training set and subsequently using these as new learning sets. Tests on real and artificial data sets, using classification and regression trees and subset bootstrap with linear regression, show that bagging can be beneficial for the accuracy. Vital component of this technique is the instability of the prediction model, but if perturbing the learning set can cause significant changes in the constructed predictor, then bagging can improve the accuracy. (Saar-Tsechansky & Provost, 2007; Rahman & Islam, 2011; Jordanov, Petrov and Petrozziello, 2018). In this work we used the BTI implementation from the R package 'caret'.

Model Evaluation

The accuracy of missing data imputation methods is evaluated by accuracy of classification.

$$\text{Accuracy of classification} \quad \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Structural Flow of the Work

Referring to Section 3, each data set follow extraction of the missing data from the incomplete data sets by the MCAR MAR and NMAR method with the percentages of the missing at 5, 10, 15 20 25 and 30. As a result, each data set was completed and ready for simulation using the missing data imputation techniques, namely SVM, NN, RF, MI and BTI. For executing the tests, we wrote the codes in R-programming and retrieved some equations relating to those techniques from CRAN projects, and used 1,000 replicated for each condition. Next, we tested the results from simulations with the estimators by accuracy of classification. The simulations and results are described in the next section.

5. RESULTS

Missing data imputation methods: SVM, NN, RF, MI and BTI were applied to impute missing data. The goal was to analyze the improvements in accuracy of classification when different algorithms were applied to impute missing data values. Table 1-3 indicates the average of accuracy classified by percentage of missing data and missing type for Bank Marketing data set, Wisconsin Breast Cancer database and Breast Cancer Coimbra data set respectively. Table 4-5 indicates the accuracy of classified percentage of missing data and type of missing respectively. Figure 1 - 5 shows the accuracy of SVM, NN, RF, MI and BTI respectively.

Table 1: Average of accuracy classified by percentage of missing data and missing type for Bank Marketing data set.

Type	Missing	SVM	NN	RF	MI	BT
MAR	5	0.7646	0.7759	0.7944	0.7444	0.7415
	10	0.8576	0.8335	0.8565	0.7721	0.7690
	15	0.8291	0.8286	0.8398	0.7660	0.7789
	20	0.8544	0.8319	0.8789	0.7730	0.7604
	25	0.8388	0.8331	0.8544	0.7471	0.7103
	30	0.8424	0.8113	0.8594	0.7524	0.7757
	Average	0.8412	0.8262	**0.8581**	0.7596	0.7563
MCAR	5	0.8897	0.8764	0.8907	0.8207	0.7999
	10	0.8869	0.8821	0.8871	0.8060	0.7814
	15	0.8904	0.8835	0.8933	0.8108	0.7909
	20	0.8894	0.8846	0.8990	0.8013	0.8147
	25	0.8906	0.8748	0.8915	0.7981	0.7896
	30	0.8929	0.8767	0.8988	0.7935	0.7955
	Average	0.8900	0.8803	**0.8939**	0.8019	0.7944
NMAR	5	0.4600	0.4657	0.4800	0.5762	0.5105
	10	0.5400	0.5522	0.5332	0.5848	0.5597
	15	0.6060	0.5848	0.6080	0.6354	0.6214
	20	0.6368	0.6329	0.6561	0.6203	0.5923
	25	0.7120	0.6847	0.7295	0.6474	0.5804
	30	0.7384	0.7224	0.7619	0.6852	0.6631
	Average	0.6155	0.6071	**0.6281**	0.6249	0.5879
Average		0.7672	0.7566	**0.7780**	0.7198	0.7017

Table 2: Average of accuracy classified by percentage of missing data and missing type for Wisconsin Breast Cancer database.

Type	Missing	SVM	NN	RF	MI	BT
MAR	5	1.0000	0.8517	0.9331	0.6168	0.9805
	10	1.0000	0.8316	0.9245	0.6044	0.9818
	15	1.0000	0.8409	0.9231	0.6017	0.9819
	20	1.0000	0.8392	0.9220	0.6003	0.9827
	25	1.0000	0.8461	0.9223	0.6047	0.9823
	30	**1.0000**	0.8553	0.9241	0.5987	0.9789
	Average	1.0000	0.8441	0.9249	0.6044	0.9813
MCAR	5	0.9614	0.9435	0.9695	0.8004	0.9548
	10	0.9628	0.9423	0.9690	0.7845	0.9537
	15	0.9619	0.9420	0.9684	0.7782	0.9520
	20	0.9625	0.9407	0.9686	0.7728	0.9527
	25	0.9627	0.9404	0.9686	0.7717	0.9508
	30	0.9624	0.9402	0.9683	0.7703	0.9503
	Average	0.9623	0.9415	**0.9687**	0.7797	0.9524
NMAR	5	0.9444	0.8291	0.9999	0.8765	0.9715
	10	0.9252	0.7883	0.9467	0.8521	0.9356
	15	0.9252	0.7813	0.9470	0.8519	0.9348
	20	0.9165	0.5766	0.9317	0.8267	0.9356
	25	0.8975	0.6403	0.9246	0.8150	0.9261
	30	0.1846	0.4595	0.9097	0.7866	0.9197
	Average	0.7989	0.6792	**0.9433**	0.8348	0.9372
Average		0.9204	0.8216	**0.9456**	0.7396	0.9570

Table 3: Average of accuracy classified by percentage of missing data and missing type for Breast Cancer Coimbra data set.

Type	Missing	SVM	NN	RF	MI	BT
	5	0.9065	0.6204	0.8948	0.9689	0.7985
	10	0.8946	0.6142	0.8907	0.9535	0.8020
	15	0.8774	0.5992	0.8892	0.9475	0.8000
MAR	20	0.8654	0.5985	0.8898	0.9452	0.7953
	25	0.8597	0.5964	0.8825	0.9390	0.7926
	30	0.8492	0.5939	0.8793	0.9344	0.7906
	Average	0.8629	0.5970	0.8852	**0.9415**	0.7946
	5	0.7365	0.5722	0.7205	0.9667	0.7120
	10	0.7402	0.5785	0.7243	0.9402	0.7228
	15	0.7417	0.5687	0.7329	0.9375	0.7249
MCAR	20	0.7380	0.5750	0.7220	0.9226	0.7151
	25	0.7297	0.5679	0.7181	0.9215	0.7104
	30	0.7310	0.5674	0.7159	0.9165	0.7045
	Average	0.7361	0.5715	0.7226	**0.9277**	0.7155
	5	0.8333	0.7772	0.9947	0.9807	0.7405
	10	0.9167	0.6908	0.9238	0.9480	0.6310
	15	0.8889	0.5749	0.9282	0.9163	0.6639
NMAR	20	0.8737	0.5368	0.7083	0.9181	0.6167
	25	0.8614	0.4642	0.6272	0.9127	0.5254
	30	0.7431	0.4993	0.7310	0.9160	0.5968
	Average	0.8528	0.5905	0.8189	**0.9320**	0.6290
Average		0.8166	0.5859	0.8045	**0.9331**	0.7020

Table 4: Average of accuracy classified by type of missing.

Missing	SVM	NN	RF	MI	BT
5	0.8329	0.7458	*0.8530*	0.8168	0.8011
10	*0.8582*	0.7459	0.8506	0.8051	0.7930
15	0.8578	0.7338	0.8589	0.8050	0.8054
20	0.8596	0.7129	0.8418	0.7978	0.7962
25	0.8614	0.7164	0.8354	0.7953	0.7742
30	0.7716	0.7029	0.8498	0.7948	0.7972
Average	0.8400	0.7264	**0.8493**	0.8007	0.7943

Table 5: Average of accuracy classified by type of missing.

Type	SVM	NN	RF	MI	BT
MAR	*0.9014*	0.7558	0.8894	0.7685	0.8441
MCAR	*0.8628*	0.7978	0.8617	0.8364	0.8208
NMAR	0.7557	0.6256	*0.7968*	0.7972	0.7180
Average	0.8400	0.7264	**0.8493**	0.8007	0.7943

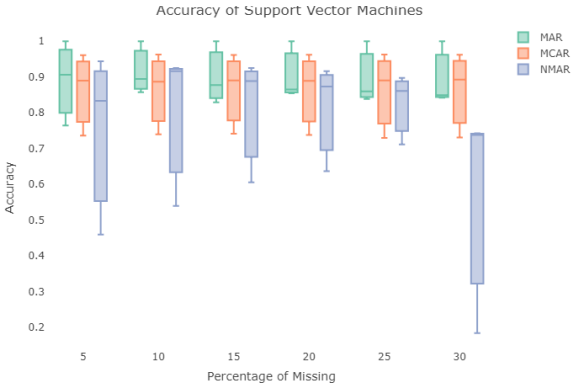


Figure 1: Accuracy of Support Vector Machines

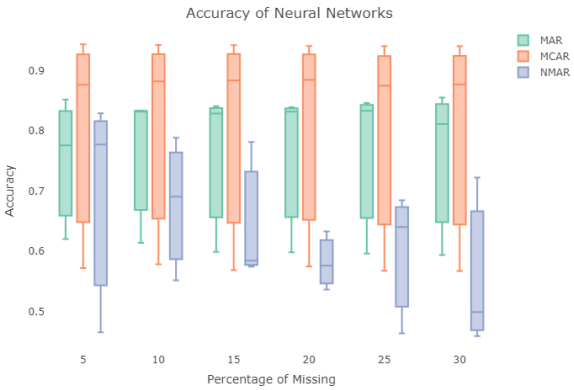


Figure 2: Accuracy of Neural Networks

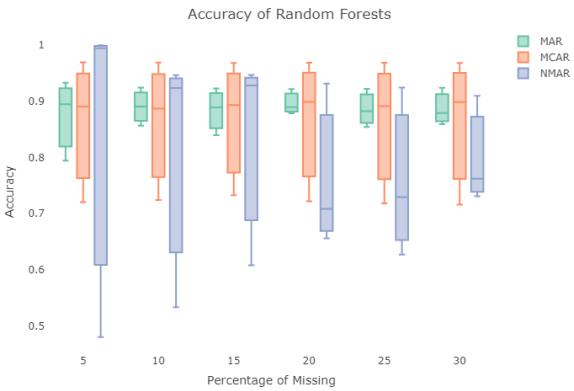


Figure 3: Accuracy of Random Forests

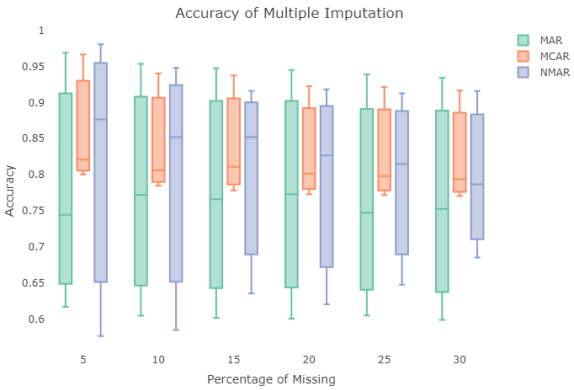


Figure 4: Accuracy of Multiple Imputation

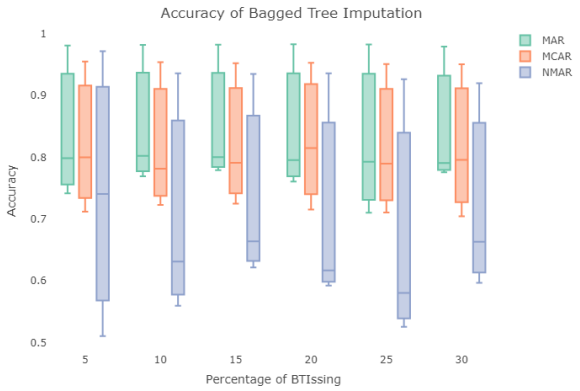


Figure 5: Accuracy of Bagged Tree Imputation

6. CONCLUSIONS AND RECOMMENDATIONS

We applied five imputation methods to treat the problem of missing data. We reviewed and provided technical details of the different methods used included SVM, NN, RF, MI AND BTI. As depicted in Table 1-5, all imputation methods led to an improvement in accuracy prediction, as measured by accuracy of classification. The best imputations in overall were obtained using RF and SVM, the imputation under MAR and MCAR were obtained using SVM, the imputation under NMAR were obtained using RF. The imputations under percentage of missing were obtained using SVM or RF.

ACKOWLEDGMENTS

This work is fully achieved by the collaborations of the Doctor of Philosophy (Educational Research and Evaluation) Department, Faculty of Education, Ubon Ratchathani Rajabhat University, Thailand.

REFERENCES

Chen, S.X., Leung, D.H., & Qin, J. (2008). Improving semiparametric estimation by using surrogate data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(4), 803-823.

Becker, W., & Powers, J. (2001). Student performance, attrition, and class size given missing student data. *Economics of Education Review*. 20, 377-388.

Becker, W.E., & Walstad, W.B. (1990). Data loss from pretest to posttest as a sample selection problem. The Review of Economics and Statistics, 184-188.

Kott, P.S. & Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association.*105, 1265–1275.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

Little, R.J. & D.B. Rubin. (2002). Statistical Analysis with Missing Data. 2nd Edn., John Wiley and Sons, New York, ISBN: 978-0-471-18386-0, pp: 408.

Rahman, G., & Islam, Z. (2011, December). A decision tree-based missing value imputation technique for data pre-processing. In Proceedings of the Ninth Australasian Data Mining Conference-Volume 121 (pp. 41-50). Australian Computer Society, Inc..

Rubin, D.B. (1987) Multiple imputation for nonresponse in surveys, New York: John Wiley & Sons, Inc.

Saar-Tsechansky, M., & Provost, F. (2007). Handling missing values when applying classification models. Journal of machine learning research, 8(Jul), 1623-1657.

Verboven, S., Branden, K.V., & Goos, P. (2007). Sequential imputation for missing values. Computational Biology and Chemistry, 31(5-6), 320-327.

Xin, Z.H.O.U., Ying, W.U., & Bin, Y.A.N.G. (2010). Signal classification method based on support vector machine and high-order cumulants. Wireless Sensor Network, 2(01), 48.

Composite Imputation Method in Logistic Regression Analysis

Sakaowrat Masa* and Uthumporn Domthong

Department of Statistics, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand
*Corresponding Email: Sakaowratmiew@gmail.com
Email: uthudo@kku.ac.th

ABSTRACT

This research aims to propose a new composite imputation method for logistic regression model when missing completely at random (MCAR) occur in the independent variables. The proposed method, namely the Median Multiple Imputation (MED-MI) method, is constructed from combination of two single imputation methods that is Median Imputation (MED) method and Multiple Imputation (MI) method by using Equivalent Weighted (EW) method. This MED-MI method is compared the performance by simulated data with the MI method, K-Nearest Regression Multiple Imputation 2 (KRMI2) method, and K-Nearest Stochastic Regression Imputation with Equivalent Weighted (KSEW) method. The performance of each method is considered by the percentage of accuracy in predicting the value of dependent variables. The results of the simulation study show that the MED-MI method has the higher percentage of accuracy in predicting the value of dependent variables than the MI method, KRMI2 method and KSEW method for all situations.

Keywords: composite imputation; logistic regression model; missing data; multiple imputation

1 INTRODUCTION

Nowadays, data is essential and important in any field such as finance, social science, medical profession, and other researches. The data must be accurate, complete and sufficient before proceed to analyze. Therefore, statistical methods are necessary for describing and analyzing these data in order to obtain valuable and useful information. Linear regression analysis is a simple statistical analysis to consider the relation between the dependent variables and independent variables in the dataset. The dependent variable must be a quantitative variable while independent variable can be either quantitative or qualitative variable. However, some research may use qualitative dependent variable such as medical researchers that the doctor wants to classify the disease and non-disease patients based on the patient's initial information. This research helps to save time of the initial treatment, and choose better patients caring. The linear regression analysis may not be able to analyze the data in this way.

To predict the value of a qualitative dependent variable with a two possible value (dichotomous variable) such as disease and non-disease, died or survived, the method to analyze is the binary logistic regression analysis. The purpose of the binary logistic regression is to consider the relation between dependent variables and independent variables when the dependent variable is a qualitative variable with dichotomous variables. The obtained regression model can be used to predict the probability of interesting event.

Practically, collecting data always has the mistake from recording data, or data providers who are not give the complete answer. The traditional method to manage missing values is the eliminate cases or variables which have missing data. Unfortunately, this method is insufficient data analysis which may causes more errors for obtained results. To reduce the error from missing data, the missing value imputation methods to get a complete data set for further analysis is conducted. However, user should select properly method to use for missing value imputation according to characteristics of the data for efficient imputation.

Multiple Imputation (MI) method is a high efficient imputation method for missing value. It is outstanding when applied to medical study that researchers often study in dichotomous dependent variable and independent variable is both quantitative and qualitative variables (Enders, 2017). Many researchers studied and proposed the missing value imputation for regression analysis. Wilks (1932) proposed the Mean Imputation method which is a simple method for only quantitative independent variables. The concept of Mean Imputation method is imputing by the average of remaining data. Hosmer et al. (2013) proposed a MI method to solve the problem of missing data in case of independent variables is both quantitative and qualitative variable in logistic regression analysis.

Moreover, Sasithorn (2012) proposed and compared two composite missing value imputation methods for linear regression analysis. The first method is K-Nearest Regression Multiple Imputation (KRMI1) which was weighted by the Least Absolute Value (LAV) method and the second method is K-Nearest Regression Multiple Imputation (KRMI2) which was weighted by the Equivalent Weighted (EW) method. They found that the KRMI2 has better performance for imputation than the KRMI1. Pattida (2016) studied and compared the missing value imputation methods when non-ignorable missing values occurred in independent variables for binary logistic regression. The imputation methods in this study are Mean Imputation (MEAN), Median Imputation (MED), K-Nearest Neighbor Method (KNN), and Multiple Imputation (MI). The results showed that MI method is the most effective method among those four method when the regression coefficient is low. In contrast, when the coefficient more is high and the sample size is large, MEAN and MED methods are effective. Rueangluck et al. (2017) proposed and compared composite missing value imputation methods when dependent variables were missing by random. The proposed methods are K-Nearest Regression Imputation with Equivalent Weighted (KREW) and K-Nearest Stochastic Regression Imputation with Equivalent Weighted (KSEW). Both proposed methods are weighted by EW method. The result showed that KSEW method is more effective for missing value imputation than KREW method.

Based on related researches above, we found that the composite missing value imputation methods is more effective than the single imputation methods for linear regression analysis. Therefore, we propose the composite missing value imputation method for logistic regression analysis by developing from MI method, which is more effective than other methods for logistic regression analysis. Moreover, the proposed method is using EW method for weighting because this weighting method is effective and simply to use.

In this study, we propose a new composite imputation method for logistic regression model, which called the Median Multiple Imputation (MED-MI) method. This research is developed under three quantitative independent variables when missing completely at random (MCAR) occur in the independent variables. The MED-MI method will be constructed from combination of two single imputation methods that is Median Imputation (MED) method and Multiple Imputation (MI) method by using EW method. The performance is considered by the percentage of accuracy in predicting the value of dependent variables.

The structure of this paper consists of theories and literature review in section 2. The details of the proposed methodology are given in section 3. The result from simulation study is illustrated in section 4, and conclusion is given in section 5.

2 LITERATURE REVIEWS

2.1 Binary Logistic Regression Analysis

Binary logistic regression analysis used to analyze relation of independent variable and dependent variable in order to predict desired subject. The dependent variable (Y) represents 2 values which are 0

and 1, where $Y=1$ expresses desired situation, and $Y=0$ expresses undesired situation which means the dependent variable is Bernoulli distribution.

$$Y=\begin{cases}1 \text{ when desired situation occurs with probability } \pi(x)\\ 0 \text{ when no desired situation occurs with probability } 1-\pi(x)\end{cases}$$

X is independent variable that can be either quantitative data or qualitative data. The average mean of dependent variable is depended on independent variables which called conditional mean or $E(Y \mid X)$. In case of Y represents 2 values which are 0 and 1, conditional mean is quantified as a number between 0 and 1, expressed as $0 \le E(Y \mid X) \le 1$

Let $E(Y \mid X)=\pi(X)$ be a conditional mean of Y which depends on X with logistic regression. The model of logistic regression is shown as follows:

$$E(Y \mid X)=\pi(X)=\frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_q x_{ij}}}{1+e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_q x_{ij}}} \quad (1)$$

where Y is $n\times 1$ observations vector of $y_1, y_2, y_3, ..., y_n$

$\pi(X)$ is conditional mean of Y which depends on X

x_{ij} is observations i^{th} of independent variable j^{th} for $i=1, 2, ..., n$ and $j=1, 2, ..., p$

β_0 is y-intercept or mean of Y when $X=0$ which is parameter of the regression model

β_q is regression coefficient for $q=1, 2, ..., p$,

n is a total number of observation

p is a number of independent variable

Then, we can use logit transformation convert $\pi(X)$ from [0,1] to $(-\infty,\infty)$. By equation (1), we will receive the logit function that can be substituted by $g(X)$ which has linear regression as shown in equation (2)

$$g(X)=\ln\left(\frac{\pi(X)}{1-\pi(X)}\right)=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_q x_{ij}. \quad (2)$$

2.2 Imputation

2.2.1 Median Imputation (MED) Method

Median imputation method is used to impute missing value with median of observations. Median is the middle value when data is sorted by ascending order, or descending order. The missing imputation by MED method is outstanding when data set has some very large or very small values.

2.2.2 Multiple Imputation (MI) Method

Multiple imputation method is used to impute missing data with at least two sets buildup data set from several methods in order to impute each missing data. The building up data set is a complete data set $X_s^*=\left(X_{obs}, X_{j(miss)}\right)$ including the rest of observations data value X_{obs} and the data that has built to estimate missing value $X_{s(miss)}$. For $s=1, 2, ..., m$ when m is a number of built data set and $1<m<n$ which proper number of data set is m = 3 or 5 in case of missing data less than 30 percent (Haung & Carriere, 2006). In the past, this method has not widely used because it had complicated calculation, and inconvenient. However, computer nowadays had been using as calculation device which make it more convenience, and give accurate result. Therefore, MI method becomes more popular (Sinharay et al., 2001). To apply this method, we first construct several data sets to estimate missing value that occurs on each independent variable (x_1, x_2, x_3). Then, we build up data set to impute missing value which could be done by several ways such as sampling with replacement of complete data, creating regressive linear equation, or decision tree learning. After that to obtain the complete data set, missing values will be replaced by existing values in same position from the data set which were built up. The obtained complete data set is analyzed to find parameter. The parameter is calculated by mean of m parameters from m data set.

2.2.3 K-Nearest Neighbor Imputation (KNN) Method

K-Nearest Neighbor imputation method is used to impute missing values with sample unit that similar to the missing data. The imputed value is imputed by the K units from remaining data (x_i, y_i) that has least distance to the unit which is missing. The distance between sample units and missing value is considered from Euclidean Distance. The Euclidean distance in (3) is used to measure distance between sample unit and replace missing values with average of K sample units, as shown below:

$$D_{ij}=\sqrt{\sum_{p=1}^{3}(x_{ip}-x_{jp})^2+(y_i-y_j)^2}\ ;\ i=1, 2, 3, ..., m\ ,\ j=m+1, ..., n \quad (3)$$

where D_{ij} is Euclidean distance of sample unit i and j

The process of missing value imputation by KNN method are as follows;

1) convert all remains data of independent variable and dependent variable to standard value.

2) calculate D_{ij} $m(3)$ for every possible pair of independent variable and dependent variable, then select K units of D_{ij} that has least value for each sample unit j

3) calculate average of data K units of independent variable from step 2),

4) replace missing value by the value from 3).

2.2.4 Regression Imputation (RI) Method

Regression imputation method is used to impute missing data with regression equation by applying known value data set, then calculate regressive coefficient by the least square in order to impute missing value of independent variables.

2.2.5 Stochastic Regression Imputation (SRI) Method

Stochastic regression imputation method is used to impute missing data with regression equation by applying known value data set, which is different from the RI method that add random error term by the least square in order to impute missing values of independent variables.

2.3 Composite Imputation

Composite imputation is a missing value imputation method by combining the single imputation method with properly weighted method.

In this study, Equivalent Weighted (EW) method is applied. This method is equally weight for all missing value imputation methods that were combined (Jirakarn, 2009). EW method is calculated by

$$W_j=\frac{1}{m} \quad (4)$$

where W_j is the average weight of the imputation method j

m is the total number of imputation methods that were combined.

3 METHODS

The proposed composite missing value method, called the Median Multiple Imputation (MED-MI) method, is constructed from combination of two single imputation methods that is Median Imputation (MED) method, and Multiple Imputation (MI) method by using Equivalent Weighted (EW) method. The MED-MI method is calculated by

$$\hat{x}_{MED\_MI}=\frac{1}{2}(\hat{x}_{MED}+\hat{x}_{MI}) \quad (5)$$

where $\hat{x}_{MED}$ is estimated value by MED method

$\hat{x}_{MI}$ is estimated value by MI method

$\hat{x}_{MED-MI}$ is estimated value by $MED-MI$ method.

The simulation study were organized as following.

1) Let three independent variables, be distributed as follows

$X_1 \sim N(50,10)$ $X_2 \sim Exp(1)$ $X_3 \sim Exp(2)$,

with sample size n = 50, 100, 200, 400 and 500. From the hypothesis testing on the correlation coefficients of the independent variables which the correlation coefficient is zero, then is there are no correlation among independent variables.

2) Predicted dependent variables are performed based on independent variables in logistic regression model. The regression coefficients are set as $\beta_0 = -1.56$, $\beta_1 = 0.05$, $\beta_2 = 0.01$, and $\beta_3 = 0.34$. The model is showed as follows:

$$\ln\left(\frac{\pi(X)}{1-\pi(X)}\right) = -1.56 + 0.05x_{i1} + 0.01x_{i2} + 0.34x_{i3}\ ;\ i = 1,\ 2,\ 3,\ ...,\ n. \tag{6}$$

The probability from model (6) is used to predict the dependent variables (Y), with a Bernoulli distribution. In this study, the cut-off point is set as 0.45 to classify desired groups and non- desired groups. This cut-off point is referred from the proportion of desired group (survive patients) of Meliodosis patients which were treated in Khon Kaen hospital between 1st Jan 2014 to 30th Dec 2017. This data is the experiment real data for this study.

3) The data of independent variable are random by missing completely at random (MCAR) with 5%, 10% and 15% of missing data.

4) The missing values are computed by Median Multiple Imputation with Equivalent Weighted (MED-MI)method.

5) Predicted dependent variables based on independent variables are imputed by MED-MI method for logistic regression. Let the regression coefficient $\beta_0 = -1.56$, $\beta_1 = 0.05$, $\beta_2 = 0.01$, and $\beta_3 = 0.34$. The model is shown as follows.

$$\ln\left(\frac{\pi(X)}{1-\pi(X)}\right) = -1.56 + 0.05x_{i1} + 0.01x_{i2} + 0.34x_{i3}\ ;\ i = 1,\ 2,\ 3,\ ...,\ n. \tag{7}$$

The probability from model (7) is used to predict the dependent variables $(\hat{Y})$, with a Bernoulli distribution. In this study, the cut-off point is set as 0.45 to classify desired groups and non- desired groups.

6) The efficiency of the missing value imputation methods are measured by the percentage of accuracy in predicting the value of the dependent variables. The percentage of accuracy is calculated as follows.

$$Accuracy = \frac{\sum_{i=1}^{L} P_i}{L} \tag{8}$$

$$P_i = \frac{T_i}{n} \times 100 \tag{9}$$

where P_i is the percentage of accuracy in predicting value of dependent variables in the iteration i^{th} ; $i = 1,\ 2,\ 3,\ ...,\ L$

T_i is the number of correctly values that are predicted in the iteration i^{th} ; $i = 1,\ 2,\ 3,\ ...,\ L$

n is the sample size

L is the number of iteration of the simulation

Accuracy is the mean of percentage of accuracy in predicting value of dependent variables**.**

4 RESULTS

In this section, the results of the comparison of MED-MI method with the existing imputation methods by percentage of accuracy in predicting value of dependent variables are considered. The results show that, in all cases of sample size, the MED-MI imputation method has the highest percentage of accuracy in predicting value of dependent variables comparing to other methods with situations of percentage of missing data. This result shows in **Table 1**.

Table 1: The percentage of accuracy in predicting value of dependent variables with%5 , 10% and 15% of missing data.

Sample size	Imputation method	Percentage of missing data 5%	10%	15%
50	MI	99.16	98.58	98.04
	KRMI2	99.18	98.62	98.08
	KSEW	99.09	98.42	97.82
	MED_MI	**99.22**	**98.63**	**98.14**
100	MI	99.23	98.63	97.91
	KRMI2	99.26	98.68	98.01
	KSEW	99.15	98.52	97.74
	MED_MI	**99.28**	**98.69**	**98.04**
200	MI	99.27	98.58	97.85
	KRMI2	99.30	98.63	97.92
	KSEW	99.22	98.45	97.62
	MED_MI	**99.31**	**98.65**	**97.94**
400	MI	99.26	98.58	97.94
	KRMI2	99.28	98.61	97.99
	KSEW	99.20	98.43	97.74
	MED_MI	**99.29**	**98.64**	**98.03**
500	MI	99.28	98.61	97.89
	KRMI2	99.30	98.66	97.94
	KSEW	99.21	98.47	97.70
	MED_MI	**99.31**	**98.68**	**97.98**

To obtain the obviously results, in each case of percentage missing data in Table 1, we illustrated the result in Figures 1 to 3.

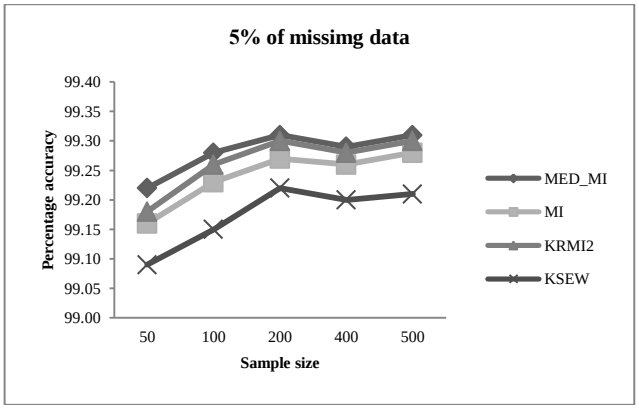


Figure 1: The percentage of accuracy in predicting value of dependent variables with 5% of missing data.

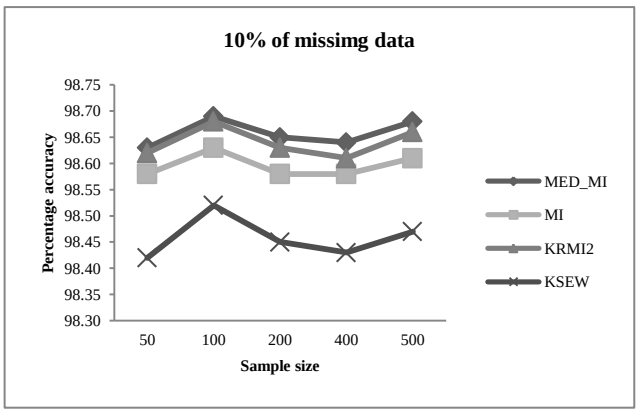


Figure 2: The percentage of accuracy in predicting value of dependent variables with 10% of missing data.

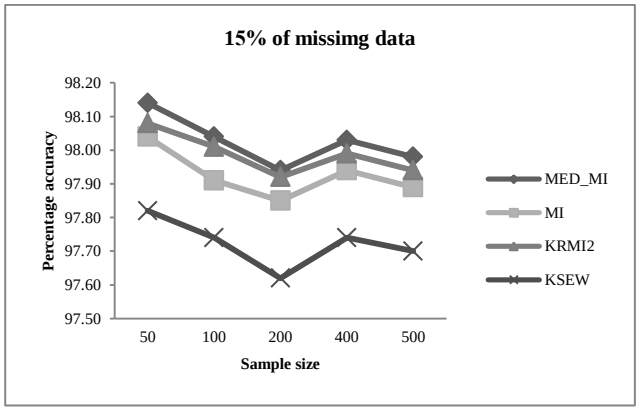


Figure 3: The percentage of accuracy in predicting value of dependent variables with 15% of missing data.

According to figures 1 to 3, it clearly shows that the MED-MI method is more effective than the other missing value imputation methods for imputing independent variables in logistic regression model.

5 CONCLUSIONS

In this study, the MED-MI method, the missing value composite imputation method for logistic regression model when missing completely at random (MCAR) occurs in the independent variables is proposed. The MED-MI method is constructed from combination of two single imputation methods which is MED method and MI method by using EW method. Moreover, the MED-MI method is compared with the existing imputation methods by simulation study. The results show that, the MED-MI has the highest percentage of accuracy in predicting the value of dependent variables comparing to other methods. Therefore, in logistic regression analysis when some value of independent variables are missing, MED-MI method is the effective imputation method to apply to the data set.

ACKNOWLEDGEMENTS

The authors would like to thank professors and classmates for their valuable suggestions, also thank to the Department of Statistics, Faculty of Science, Khon Kaen University for their support in this study.

REFERENCES

Enders, C.K. (2017). Multiple imputation as a flexible tool for missing data handling in clinical research. *Behaviour Research and Therapy*, 98, 4-18.

Haung, R., & Carriere, K.C. (2006) Comparison of methods for incomplete repeated measures data analysis in small samples. *Joural of Statistical Planning and Inference*, (136), 235-247.

Hosmer, Jr, D.W., Lemeshow, S., & Sturdivant, R.X. (2013) *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Sinharay, S., Hal, S.S., & Russell, D. (2001). The use multiple imputation for the analysis of missing data. *Psychological Methods*, 6(4), 317-329.

Wilks, S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. *Annals of Mathematical Statistics, 3 (3)*, 163–195.

Nunlaong, J. (2009). A Comparison of Missing Value Estimation Methods for Forecasting Models. Master of Science (Applied Statistics), Graduate College, King Mongkut's University of Technology North Bangkok (in Thai).

Nilpattarachat, P. (2016). A Comparison of the Estimation Methods for Nonignorable Missing Data in Logistic Regression Analysis. Master of Science (Statistics), Graduate School of Chulalongkorn University (in Thai).

Lamjaisue, R., Thongteeraparp, A., & Sinsomboonthong, J. (2017). Comparison of missing data estimation methods for the multiple regression analysis with missing at random. Thai Journal of Science and Technology, 25(5), 765-777 (in Thai).

Sompomgnawakij, S. (2012). A Comparison of Composite Imputation Methods. Master of Science (Statistics), Graduate School of Kasetsart University (in Thai).

A Time Series Model to Predict the Number of People per Day Calling for an Appointment for HIV Counseling and Testing

Tanarat Muangmool[1,2]*, Anouar Nechba[2], Kanchana Than-in-at[2], Paporn Mongkolwat[2], Niphatta Mungkhala[2], Tanawan Samleerat[3], Wasna Sirirungsi[3], and Patrinee Traisathit[2,4]

[1]Master's Degree Program in Applied Statistics, Department of Statistics, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand.
*Corresponding Email: tanarat_m@cmu.ac.th
[2]Prevention and Treatment of HIV infection and virus-associated cancers in Southeast Asia (PHPT), Chiang Mai, Thailand.
Email: a.nechba@gmail.com
Email: kanchana.than-in-at@phpt.org
Email: paporn.mongkolwat@phpt.org
Email: niphatta.mungkhala@phpt.org
[3]Faculty of Associated Medical Sciences, Chiang Mai University, Chiang Mai, Thailand.
Email: tanawan.s@cmu.ac.th
Email: wasna.s@cmu.ac.th
[4]Center of Excellence in Bioresources for Agriculture, Industry and Medicine, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand.
Email: patrinee.t@cmu.ac.th

ABSTRACT

Napneung is an ongoing research project aimed at evaluating new methods to increase the uptake of HIV, hepatitis B, hepatitis C and syphilis testing for individuals in Northern Thailand. To reach out individuals potentially willing to be tested, one of the approaches was to distribute in different public places vouchers for free of charge testing. Those willing to be tested had to call a hotline for an appointment. Each voucher had a unique identifier along with the date and location of distribution. To optimize the allocation of resources, it would be useful to forecast the number of individuals requesting an appointment. We fitted an autoregressive moving average (ARIMA) model with exogenous variables (ARIMAX). The exogenous variables were minimum, maximum and mean daily temperatures (°C), mean daily rainfall (mm) and the number of distributed vouchers in the previous 30 days. The Granger causality test was used to select the variables in the model. The fitted model was evaluated using the Akaike information criterion and the Ljung-Box test. The analyzed data were from 4,182 persons who made a call on 13 October 2015 to 31 December 2017. The median daily number of calls was 4 [interquartile range: 2 to 7]. Calls were preferentially made on weekdays (seasonal index > 1.00). The number of distributed vouchers was useful to predict the number of calls (Granger test; $p<0.001$). From a seasonal ARIMAX$(1,1,1)(0,0,2)^7$ model, it was estimated that 3 to 5 calls are made per day. Staff allocation should be planned to respond incoming call sufficiently, especially on Monday which the number of calls was highest. However, the small number of calls per day might not cause the resource allocation problems. The prediction would help the staff to work easier.

Keywords: HIV testing and counseling; Napneung; ARIMAX; Exogenous variable.

1 INTRODUCTION

In Thailand, the estimated number of people newly infected with the human immunodeficiency virus (HIV) has decreased by 70% from 21,000 in 2005 to 6,400 in 2016 (UNAIDS, 2017). Although new HIV infections are declining, UNAIDS (2016a) has estimated that 10% of HIV-infected people were unaware of their HIV status. This is associated with high-risk sexual behaviors such as unprotected sex with their partner (Vagenas et al., 2014), thereby increasing the risk of HIV transmission, especially in key populations including men who have sex with men, commercial sex workers, people who inject drugs, and transgender people (UNAIDS, 2016b). According to the 90-90-90 ambitious treatment target, UNAIDS (2014) has provided HIV testing and counseling (HTC) services as a global strategy to decrease the number of people who are unaware of their HIV status (WHO, 2011).

HTC services are a gateway to care, treatment, and essential support for HIV-infected people (Hall et al., 2012; Johnston et al., 2016). Fear, stress, confidential information, and difficulties with the procedures might have resulted in barriers against persuading people to seek HTC services (Conway et al., 2015; Deblonde et al., 2010; Schwarcz et al., 2011), and reducing the difficulties might be advantageous toward them.

In Northern Thailand, a project called "Napneung" (ClinicalTrial.gov: NCT02752152) is offering an easier method to contact and set up appointments for HTC services to persuade more people to seek testing and counseling for four infections (HIV, Hepatitis B and C, and Syphilis). This project is aimed at minimizing traditional obstacles against getting tested such as wasting time, cost, lack of confidentiality, etc. To promote the service, vouchers for free testing are distributed in several areas in Chiang Mai and Chiang Rai. Individuals who receive a voucher are invited to make a call for an appointment for which the place and date of distribution can be tracked by the voucher number. Promoting the HTC services is also made through traditional media (press, radio, and newspapers) as well as social media (Facebook https://www.facebook.com/napneung/) and a dedicated website (https://www.napneung.net/).

To make them efficient, HTC services should be planned in terms of stocking sufficient test kits, human resources, time schedules, and places (WHO, 2012). As the number of appointments might vary depending on the promotion of the services and climate factors (Phithakkitnukoon et al., 2012), predicting the number of appointments is necessary for planning and preparation. The aim of this study was to predict the number of appointments via the Napneung project based on the number of calls per day.

2 METHODS

Data were collected from all of the adults (age > 18 years) who made a call to the Napneung project HTC services in Northern Thailand between October 13, 2015, and December 31, 2017. The information on the voucher number, appointment, and date of the call were recorded by the well-trained staff without requesting personal information from the callers. Each appointment was arranged at one of four facilities located in Chiang Mai and Chiang Rai depending on the person's purpose, time schedule, and the availability of places. Vouchers for free testing were distributed by volunteers at specific places, e.g. pubs, restaurants, associations, and special events (festivals, fairs, etc.).

The protocol of the Napneung project HTC services was reviewed and approved by the Ethics Committee of the Faculty of Associated Medical Sciences, Chiang Mai University.

The outcome of interest was the number of calls to the HTC services per day, either with or without a voucher. The exogenous

variables comprising the minimum, maximum and mean daily temperatures (°C), the mean daily rainfall (mm), and the number of distributed vouchers during the previous 30 days were included in the prediction model. These climate variables were the average values from each province in Northern Thailand recorded by the Thailand Meteorological Department.

Medians and interquartile ranges (IQRs) were used for the continuous variables, while frequencies and percentages are presented for the categorical variables.

A time series analysis was used to predict the number of calls per day. Time series data were preliminarily considered to detect trends and seasonal components using an autocorrelation function (ACF) and a partial autocorrelation function (PACF). The unit root and seasonal unit root were tested for stationarity using the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) and Osborn-Chui-Smith-Birchenhall (OCSB) tests, respectively. According to the assumption that data must be stationary prior to using an autoregressive integrated moving average (ARIMA) model (Shumway & Stoffer, 2017), the number of calls per day was adjusted by a differencing process until they were stationary. The ARIMA models were considered according to the lowest Akaike information criterion (AIC) value. Exogenous variables were tested using the Granger causality test (Lopez & Weber, 2017) prior to inclusion in the autoregressive integrated moving average with exogenous (ARIMAX) model (Andrews et al., 2013) to predict the incoming calls per day from the 1st to 14th January 2018. The assumption of the model that residuals were diagnosed independence its using Box-Ljung test. All test results where $p<0.05$ were considered as statically significant. All analyses were performed using Stata version 14.0 (StataCorp LP, College Station, TX, USA).

3 RESULTS

From October 13, 2015 to December 31, 2017, 4,182 clients made a call to ask for information or to make an appointment for the HTC services. Of these, 64% had received vouchers from Napneung distributors, 94% made a call and an appointment with 14% cancelling after having done so. More than half of the appointment calls were from people who received a voucher. The median number of calls was 4 per days [IQR: 2 to 7] (Table 1).

Table 1: The characteristics of the number of calls to the Napneung HTC services (N=4,182)

Variables	N (%) or Median (IQR)
Number of calls per day	4 (2 to 7)
Number of calls with a voucher	2,680 (64.0)
Number of calls with an appointment	3,928 (94.0)
Without a voucher	1,296 (33.0)
With a voucher	2,632 (67.0)
Cancellation after making an appointment	561 (14.0)

Abbreviations: IQR, interquartile range

Figure 1 shows that the number of calls changed over time. There was an increasing trend and a seasonal pattern each week. The KPSS results show that the number of calls was non-stationary ($p>0.10$), but after adjusting the data, they became stationary ($p<0.001$). Figure 2 shows the ACF and PACF of the number of calls both before (Figure 2a, 2b) and after (Figure 2c, 2d) the differencing process. Significant ACF spike lags were found at 7, 14, 21, and 28, which seem to be weekly seasonal variations (Figure 2c). Moreover, PACF showed that the correlation coefficients declined slightly, which indicates that the data contained a trend component (Figure 2d). The results of the KPSS test show that the trend component was stationary (no unit root, $p<0.001$) as did the OCSB test for the seasonal component (no seasonal unit root, $p<0.001$).

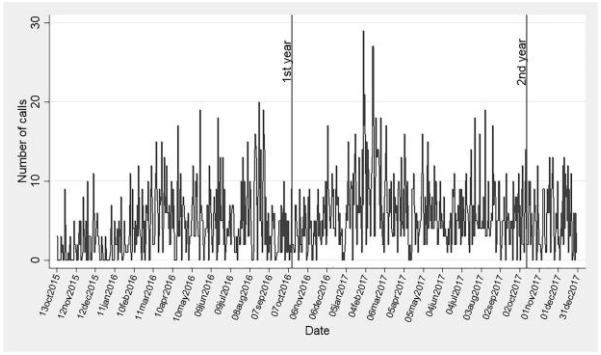


Figure 1: The daily number of calls for an appointment after receiving voucher between October 13, 2015 and December 31, 2017.

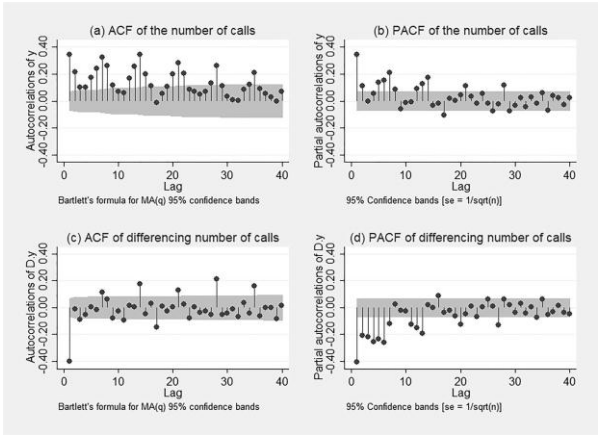


Figure 2: The autocorrelation and partial autocorrelation functions of the number of calls and the subsequent differencing.

In addition, most people made a call on weekdays rather than weekends. The seasonal index of Monday to Sunday was 1.35, 1.22, 1.28, 1.16, 0.96, 0.56, and 0.43, respectively. The highest number of incoming calls occurred on Mondays and the lowest on Sundays. The calls on Mondays comprised 35% over the median number of calls per day.

The ARIMA(p,d,q)(P,D,Q)^{L} model was used to consider all possible orders of the selection of the six models used, resulting finally in the selection of ARIMA(1,1,1)(0,0,2)^{7} based on its minimal AIC value (Table 2).

According to the causality of each exogenous variable with the number of calls, only the number of distributed vouchers in the previous 30 days was useful for predicting the number of calls and thus was included in the model ($p<0.001$) (Table 3). The independence of the residuals agreed with the assumption of the ARIMAX model ($p>0.05$). The daily predicted number of calls is shown in Figure 3 based on the ARIMAX(1,1,1)(0,0,2)^{7} model, which would be 3 to 5 per day from January 1st to 14th, 2018.

Table 2: The ARIMA model selection and Box-Ljung tests of the models' residuals

Models	AIC	Ljung-Box test	
		Statistics	p-value
ARIMA(1,1,0)(0,0,1)^7	4,726.50	82.09	<0.001
ARIMA(1,1,0)(0,0,2)^7	4,705.52	81.12	<0.001
ARIMA(0,1,1)(0,0,1)^7	4,528.28	38.27	<0.001
ARIMA(0,1,1)(0,0,2)^7	4,504.44	26.88	<0.001
ARIMA(1,1,1)(0,0,1)^7	4,505.01	9.87	0.13
ARIMA(1,1,1)(0,0,2)^7	4,482.97	4.73	0.57

Abbreviations: AIC, Akaike information criterion

Table 3: The Granger causality test for the selection of useful exogenous variables into the model

Exogenous variables	Granger causality test		
	Chi-squared	p-value	Interpretation
Minimum daily temperature (ºC)	3.93	0.14	Not useful
Maximum daily temperature (ºC)	5.41	0.07	Not useful
Mean daily temperature (ºC)	4.82	0.09	Not useful
Mean daily rainfall (mm)	1.21	0.54	Not useful
Number of distributed vouchers in the previous 30 days	12.70	<0.001	Useful

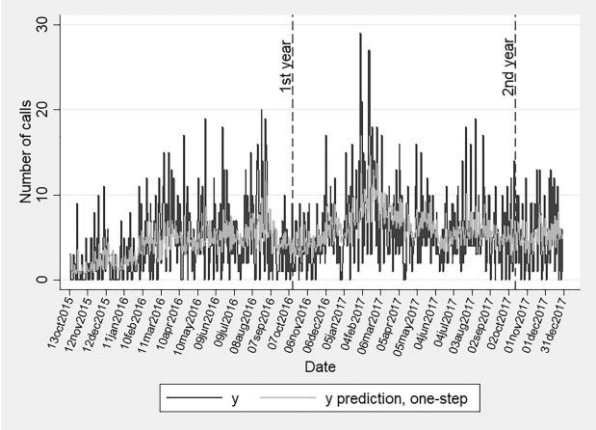


Figure 3: The ARIMAX(1,1,1)(0,0,2)^7 model for the daily incoming calls prediction.

4 DISCUSSION

The number of calls per day were fitted and predicted using an ARIMAX model. There was a trend component, seasonal component by days of the weeks, and the number of vouchers distributed in the previous 30 days influenced the number of calls. ARIMAX's specialization helped to explain the predicted value influenced by the exogenous factors in the ARIMA models, which used only dependent variables. We found that only the number of distributed vouchers affected the number of calls.

The rate of appointments in our study was quite high compared to a previous study by Chhim et al. (2017) (67% versus 25%). They studied increasing HIV case detection in Cambodia among high-risk populations by using peer navigators or using risk-tracing snowball approach. Our study might have been due to fact that information on the HTC services was provided by well-trained distributors, and thus covered more information than the snowball messaging approach among clients only used in the previous study. However, the snowball approach was more effective in detecting new HIV cases in the target group.

Phone contact is sometimes considered as easier than a first face-to-face contact (Schenker et al., 2010). In this study, the number of appointments for the HTC services was 90%, which was higher than another study in Thailand by Khawcharoenporn, Chunloy and Apisarnthanarak (2016). Their study on the response to face-to-face contact among university students showed that only 27% of them made an appointment and continued on to HIV testing and counseling.

A limitation of our study is that the amount of data was too small to consider the seasonality in terms of weeks or months since data were collected over the relatively short time period of two years. This would have requested at least 100 observations (Hanke & Wichern, 2014). In addition, we were concerned that the number of calls per day or the outcome had exceeded of zero counts (11%). When most of the values of the outcome are zero, a prediction can become negative even though this is not possible due to the nature of the count data (Kane et al., 2014). However, using the ARIMAX model in our study was still suitable as other studies have suggested using alternative models if the outcome with zero values exceeded 40% (Charlton, 2015; Hasan et al., 2012).

In conclusion, call behavior of individuals requesting to HTC service could be explained by seasonality including the days of the week, which occurred more on weekdays. Staff allocation should be prepared to respond the number of calls sufficiently, especially on Monday which incoming calls were highest. However, the small predicted number of calls might not cause the resource allocation problems. The prediction could help the staff to work easier.

ACKNOWLEDGMENTS

We thank all the clients who participated in Napneung HIV testing and counseling service. We are also grateful to Napneung project teams; Nirattiya Jaisieng, Natthanidnan Sricharoen, Jiraporn Kamkorn, Ampika Kaewbundit, Chutharat Kasemrat, Pornpimon Moolnoi, Laddawan Laomanit, Warunee Khamjakkaew, Nantawan Wangsaeng, Areerat Kongphonoi, Subenya Jinasa, Nusara Krapunpongsakul and Pongsak Pirom. We also thank Napneung advisory board and collaborating institutions; Nakornping Hospital; Prattana Leenasirimakul, Sanpatong Hospital; Virat Klinbuayaem, CAREMAT, PIMAN Center, STIs center, Maharaj Nakorn Chiang Mai (Suandok) hospital, MAP Foundation, Chiangrai Prachanukroh Hospital; Jullapong Achalapong. We would like to thank Dr. Gonzague Jourdain who are the one of leaders of Napneung project for our manuscript improvement and Nicolas Salvadori for statistical advice. We also thank statistical professors at Chiang Mai University for helpful methods discussion; Asst. Prof. Dr. Watha Minsan, Asst. Prof. Dr. Bandhita Plubin and Asst. Prof. Dr. Sukon Prasitwattanaseree. We thank the Thai Meteorological Department for providing the climate data. The Napneung project is funded by Expertise France (Initiative 5%). Tanarat Muangmool receives a scholarship from the Graduate school, Chiang Mai University.

REFERENCES

Andrews, B. H., Dean, M. D., Swain, R., & Cole, C. (2013). *Building ARIMA and ARIMAX models for predicting long-term disability benefit application rates in the public/private sectors*. Society of Actuaries: University of Southern Maine. Retrieved from https://www.soa.org/research-reports/2013/research-2013-arima-arimax-ben-appl-rates/

Charlton, S. (2015). Zero-inflated count time series. In *Wiley StatsRef: Statistics Reference Online* (pp. 1–5). American Cancer Society.

Chhim, S., Macom, J., Pav, C., Nim, N., Yun, P., Seng, S., … Yi, S. (2017). Using risk-tracing snowball approach to increase HIV case detection among high-risk populations in Cambodia: An intervention study. *BMC Infectious Diseases*, *17*. https://doi.org/10.1186/s12879-017-2790-1

Conway, D. P., Holt, M., Couldwell, D. L., Smith, D. E., Davies, S. C., McNulty, A., … Guy, R. (2015). Barriers to HIV testing and characteristics associated with never testing among gay and bisexual men attending sexual health clinics in Sydney. *Journal of the International AIDS Society*, *18*(1). https://doi.org/10.7448/IAS.18.1.20221

Deblonde, J., De Koker, P., Hamers, F. F., Fontaine, J., Luchters, S., & Temmerman, M. (2010). Barriers to HIV testing in Europe: A systematic review. *European Journal of Public Health*, *20*(4), 422–432.

Hall, H. I., Holtgrave, D. R., & Maulsby, C. (2012). HIV transmission rates from persons living with HIV who are aware and unaware of their infection. *AIDS*, *26*(7), 893–896.

Hanke, J. E., & Wichern, D. W. (2009). *Business forecasting (9th Edition)*, New Jersey: Pearson Prentice-Hall, Inc.

Hasan, M. T., Sneddon, G., & Ma, R. (2012). Regression analysis of zero-inflated time-series counts: Application to air pollution related emergency room visit data. *Journal of Applied Statistics*, *39*(3), 467–476.

Johnston, L. G., Steinhaus, M. C., Sass, J., Sirinirund, P., Lee, C., Benjarattanaporn, P., & Gass, R. (2016). Recent HIV testing among young men who have sex with men in Bangkok and Chiang Mai: HIV testing and prevention strategies must be enhanced in Thailand. *AIDS and Behavior*, *20*(9), 2023–2032.

Kane, M. J., Price, N., Scotch, M., & Rabinowitz, P. (2014). Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. *BMC Bioinformatics*, *15*(1). https://doi.org/10.1186/1471-2105-15-276

Khawcharoenporn, T., Chunloy, K., & Apisarnthanarak, A. (2016). Uptake of HIV testing and counseling, risk perception and linkage to HIV care among Thai university students. *BMC Public Health*, *16*, 556.

Lopez, L., & Weber, S. (2017). Testing for Granger causality in panel data. *IRENE Working Paper 17-03*, IRENE Institute of Economic Research.

Phithakkitnukoon, S., Leong, T. W., Smoreda, Z., & Olivier, P. (2012). Weather effects on mobile social interactions: A case study of mobile phone users in Lisbon, Portugal. *PLoS ONE*, *7*(10). https://doi.org/10.1371/journal.pone.0045745

Schenker, I., Chemtob, D., Yamin, N., Shtechman, N., & Rosenberg, H. (2010). The national HIV/AIDS hotline in Israel: Data on utilization, quality assurance and content. *International Public Health Journal*, (2), 339–344.

Schwarcz, S., Richards, T. A., Frank, H., Wenzel, C., Hsu, L. C., Chin, C.-S. J., … Dilley, J. (2011). Identifying barriers to HIV testing: personal and contextual factors associated with late HIV testing. *AIDS Care*, *23*(7), 892–900.

Shumway, R. H., & Stoffer, D. S. (2017). Time series regression and exploratory data analysis. In R. H. Shumway & D. S. Stoffer, *Time Series Analysis and Its Applications* (pp. 45–74). Cham: Springer International Publishing.

UNAIDS. (2014). 90-90-90: An ambitious treatment target to help end the AIDS epidemic. Joint United Nations Program on HIV/AIDS Geneva, Switzerland. Retrieved from http://www.unaids.org/sites/default/files/media_asset/90-90-90_en.pdf

UNAIDS. (2016a). Country factsheets - HIV and AIDS estimates. Joint United Nations Program on HIV/AIDS Geneva, Switzerland. Retrieved May 21, 2018, from http://www.unaids.org/en/regionscountries/countries/thailand

UNAIDS. (2016b). Prevention gap report 2016. Joint United Nations Program on HIV/AIDS Geneva, Switzerland. Retrieved from http://www.unaids.org/en/resources/documents/2016/prevention-gap

UNAIDS. (2017). UNAIDS data 2017. Joint United Nations Program on HIV/AIDS Geneva, Switzerland. Retrieved from http://www.unaids.org/en/resources/documents/2017/2017_data_book

Vagenas, P., Ludford, K. T., Gonzales, P., Peinado, J., Cabezas, C., Gonzales, F., … Altice, F. L. (2014). Being unaware of being HIV-infected is associated with alcohol use disorders and high risk sexual behaviors among men who have sex with men in Peru. *AIDS and Behavior*, *18*(1). https://doi.org/10.1007/s10461-013-0504-2

WHO. (2011). *Guide for monitoring and evaluating national HIV testing and counselling (HTC) programmes: Field-test version*. Geneva: World Health Organization. Retrieved from http://www.who.int/iris/handle/10665/44558

WHO. (2012). *Service delivery approaches to HIV testing and conselling (HTC): A strategic HTC programme framework*. World Health Organization. Retrieved from http://www.who.int/iris/handle/10665/75206

Numerical Approximation of the Fractional HIV Model

Kunwithree Phramrung[1], Anirut Luadsong[1*] and Nitima Aschariyaphotha[2]

[1]King Mongkut's University of Technology Thonburi/Department of Mathematics/Bang Mod, Thung Khru, Bangkok, Thailand
*Corresponding Email: anirut.lua@kmutt.ac.th
Email: kunwithree.kp@mail.kmutt.ac.th
[2]King Mongkut's University of Technology Thonburi/Ratchabuti Learning Park/Rang Bua, Chom Bueng, Ratchaburi, Thailand
Email: nitima.asc@kmutt.ac.th

ABSTRACT

This paper developed the HIV model based on the fractional differential equation. A numerical approximation was proposed in time and space. The HIV model was presented with the time fractional of the Caputo derivative. The fractional HIV model was used by an implicit finite difference method. This model solved numerically using the Crank-Nicolson technique. The numerical results were compared with the approximation of the HIV model using the integer order differential equations to confirm the accuracy of the proposed method. Finally, numerical simulations were performed to demonstrate consistent results.

Keywords: HIV model; implicit finite difference method; fractional differential equation; Caputo derivative

1 INTRODUCTION

In recent years, the pattern of viral infection has been considered in areas where the virus has diffusion. The models are assumed that the only free viruses diffusion to target cells (Wang, et al., 2016). Hybrid systems of differential equations describe the diffusion of the virus by the following factors: 1) the interaction between the virus and the immune system is localized according to the type of tissue identified. 2) The virus can move freely and their motion is characterized by Fickian diffusion.

In the case of Wang et al. (2007) the density of the uninfected cells, infected cells, and free virus are defined in one dimension. Brauner et al. (2011) extended in two dimensions with periodic boundary conditions, and indicated that the recruitment rate depends on the area. Recent work by Wang et al. (2014) proposed a zero-bound boundary condition in a finite domain, $\Omega \in \mathbb{R}^d$, at the homogeneous Neumann boundary condition, $\partial\Omega$ (Arafa, 2012; Wang et al., 2014; Wang et al., 2016).

Recent studies have found that the efficacy of viral infections is significant in transferring the virus to non-target cells. In this case, the virus particles can be transferred from infected target cells to uninfected ones through the synapses of the virus. To determine the effects of both diffuse and spatial heterogeneity, the cell-to-cell hybridization was transformed into the Wang et al. (2014) model.

$$\begin{aligned}\frac{\partial p(\mathbf{x},t)}{\partial t} &= \lambda(\mathbf{x}) - \beta_1(\mathbf{x})p(\mathbf{x},t)r(\mathbf{x},t) - \beta_2(\mathbf{x})p(\mathbf{x},t)q(\mathbf{x},t) \\ &\quad - a(\mathbf{x})p(\mathbf{x},t)\end{aligned}$$

$$\begin{aligned}\frac{\partial q(\mathbf{x},t)}{\partial t} &= \beta_1(\mathbf{x})p(\mathbf{x},t)r(\mathbf{x},t) + \beta_2(\mathbf{x})p(\mathbf{x},t)q(\mathbf{x},t) \\ &\quad - b(\mathbf{x})q(\mathbf{x},t)\end{aligned}$$

$$\frac{\partial r(\mathbf{x},t)}{\partial t} = \mathcal{D}\Delta r(\mathbf{x},t) + k(\mathbf{x})q(\mathbf{x},t) - m(\mathbf{x})r(\mathbf{x},t) \quad (1)$$

Here the meaning of each symbol is listed for $(\mathbf{x},t) \in \Omega \times (0,\infty)$; $\Omega \in \mathbb{R}^d, d = 1,2,3$ with the homogeneous Neumann boundary condition $\frac{\partial r(\mathbf{x},t)}{\partial \mathbf{n}} = 0$; $\mathbf{x} \in \partial\Omega, t > 0$ where $\mathbf{n}$ denotes an outward unit normal to $\partial\Omega$ and initial conditions $p(\mathbf{x},0) = p^0(\mathbf{x}) \geq 0, q(\mathbf{x},0) = q^0(\mathbf{x}) \geq 0$ and $r(\mathbf{x},0) = r^0(\mathbf{x}) \geq 0$; $\mathbf{x} \in \Omega$. The density of uninfected cells, the density of infected cells and the density of free viruses which are denoted by $p(\mathbf{x},t), q(\mathbf{x},t)$ and $r(\mathbf{x},t)$, respectively. $\lambda(\mathbf{x})$ and $b(\mathbf{x})$ denote the number of newly produced uninfected cells and the death rate of uninfected cells, respectively. The death rate of infected cells, the death rate of free viruses and the transmission coefficient for the virus to cell infection denote by $a(\mathbf{x}), m(\mathbf{x})$ and $\beta_1(\mathbf{x})$, respectively. $\beta_2(\mathbf{x}), k(\mathbf{x})$ and $\mathcal{D}$ denote the transmission coefficient for the cell to cell infection, the rate of virus production due to the lysis of infected cells and the diffusion coefficient, respectively.

However, the amount of work done in modeling HIV infection has been limited to the ordinary integer differential equation. Fractional calculus has recently been widely applied in many fields. Many mathematicians and applied researchers attempt to simulate real processes with fractional calculus. Therefore, we propose a system of fractional differential equations for HIV models using the Crank-Nicolson technique to describe the behavior of HIV infection transformation. Section 2 describes the approximation of the Caputo derivative of fractional order. Section 3 is a discretization numerical approximation of the space variable using by Crank-Nicolson technique. Section 4 is used the numerical experiment to confirm the accuracy of the scheme. Finally, the summary of our article is section 5.

2 APPROXIMATION OF CAPUTO DERIVATIVE OF FRACTIONAL ORDER

This section presents a numerical approximation of the time fractional derivative in the Caputo sense. The first-order approximation formula for computation of the time fractional derivative of order α can be obtained by simple quadrature formula and finite difference method (Atangana & Alqahtani, 2016; Dalir & Bashour, 2010; Ding, 2016; Murio, 2008; Loverro, 2004; Oliveira & Machado, 2014; Ren et al., 2006; Thamareerat et al., 2017). The definition of the Caputo fractional derivative is presented as follows:

Definition 1: The fractional derivative of $f(t)$ in the Caputo sense is defined as equation (2):

$$D_t^\alpha f(t) = \frac{1}{\Gamma(n-\alpha)} \int_0^t (t-\tau)^{n-\alpha-1} \frac{d^n f(\tau)}{d\tau^n} d\tau \quad (2)$$

where $n-1 < \alpha \leq n, n \in N, t > 0$ and $\Gamma(\cdot)$ denotes the gamma function. For $n = 1$, a numerical approximation based upon the Caputo fractional derivative as equation (3):

$$D_t^\alpha f(t) = \frac{1}{\Gamma(1-\alpha)} \int_0^t (t-\tau)^{-\alpha} \frac{df(\tau)}{d\tau} d\tau \quad (3)$$

where $0 < \alpha \leq 1, t > 0$. For some positive integer N, the grid size in time for finite difference method is defined by $\Delta t = \frac{T}{N}$. The grid points in the time interval $[0,T]$ are labeled $t_j = j\Delta t\,; j = 0,1,2,\ldots,N$. The value of the function f at the grid point is $f^j = f(t_j)$. A discrete approximation to the Caputo derivative of fractional order can be obtained by simple quadrature formula as equation (4):

$$D_t^{\alpha} f(t_n) = \frac{1}{\Gamma(1-\alpha)} \int_{t_0}^{t_n} (t_n - \tau)^{-\alpha} \frac{df(\tau)}{d\tau} d\tau \,; 0 < \alpha \leq 1 \qquad (4)$$

By $\int_{t_0}^{t_n} f(y)dy = \sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} f(y)dy$. The equation (4) can be rewritten as equation (5):

$$D_t^{\alpha} f(t_n) = \frac{1}{\Gamma(1-\alpha)} \sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} (t_n - \tau)^{-\alpha} \frac{df(\tau)}{d\tau} d\tau \qquad (5)$$

The approximation formula at the time level n can be obtained as equation (6):

$$D_t^{\alpha} f(t_n) = \frac{\Delta t^{1-\alpha}}{\Gamma(2-\alpha)} \sum_{j=0}^{n-1} \left(\frac{f^{j+1} - f^{j}}{\Delta t} + O(\Delta t) \right) d_j \qquad (6)$$

where $\Gamma(z+1) = z\Gamma(z)$ and $d_j = (n-j)^{1-\alpha} - (n-j-1)^{1-\alpha}$. Then defining the indices j as $n-j$, which $d_{n-j} = j^{1-\alpha} - (j-1)^{1-\alpha}$. Thus equation (6) can be written as equation (7):

$$D_t^{\alpha} f(t_n) = \frac{\Delta t^{-\alpha}}{\Gamma(2-\alpha)} \sum_{j=1}^{n} (f^{n-j+1} - f^{n-j})\, d_{n-j} + \frac{O(\Delta t^{2-\alpha})}{\Gamma(2-\alpha)} \sum_{j=1}^{n} d_{n-j}. \qquad (7)$$

Letting $D_t^{\alpha} f(t_n) = \mathfrak{D}_t^{\alpha} f(t_n) + O(\Delta t)$ and $\sigma_{\alpha} = \frac{\Delta t^{-\alpha}}{\Gamma(2-\alpha)}$. The first-order approximation formula of the time fractional derivative in the Caputo sense is given as equation (8):

$$\mathfrak{D}_t^{\alpha} f(t_n) = \sigma_{\alpha} \sum_{j=1}^{n} (f^{n-j+1} - f^{n-j})\, d_{n-j} \qquad (8)$$

3 THE DISCRETIZATION

This section presents a numerical approximation of the space variable by Crank-Nicolson technique (Sweilam et al., 2012). The system (1) can be transformed into the fractional differential equations of order α as a system (9):

$$\frac{\partial^{\alpha} p(\mathrm{x},t)}{\partial t^{\alpha}} - \lambda(\mathbf{x}) + \beta_1(\mathbf{x}) p(\mathbf{x},t) r(\mathbf{x},t) + \beta_2(\mathbf{x}) p(\mathbf{x},t) q(\mathbf{x},t) + a(\mathbf{x}) p(\mathbf{x},t) = 0$$

$$\frac{\partial^{\alpha} q(\mathrm{x},t)}{\partial t^{\alpha}} - \beta_1(\mathbf{x}) p(\mathbf{x},t) r(\mathbf{x},t) - \beta_2(\mathbf{x}) p(\mathbf{x},t) q(\mathbf{x},t) + b(\mathbf{x}) q(\mathbf{x},t) = 0$$

$$\frac{\partial^{\alpha} r(\mathrm{x},t)}{\partial t^{\alpha}} - \mathcal{D}\Delta r(\mathbf{x},t) - k(\mathbf{x}) q(\mathbf{x},t) + m(\mathbf{x}) r(\mathbf{x},t) = 0 \qquad (9)$$

For some positive integer M, the grid sizes in time for finite difference technique is defined by $\Delta \mathbf{x} = \frac{\boldsymbol{X}}{M}$. The grid points in the space interval $[0, \boldsymbol{X}]$ are labeled $\mathbf{x}_i = i\Delta \mathbf{x}\,; i = 0,1,2,\ldots,M$. The value of the function f at the grid point is $f_i^j = f(\mathbf{x}_i, t_j)$. The system (9) can be present as a system (10):

$$\frac{\partial^{\alpha} p(\mathbf{x}_i,t_j)}{\partial t^{\alpha}} - \lambda(\mathbf{x}_i) + \beta_1(\mathbf{x}_i) p_i^j r_i^j + \beta_2(\mathbf{x}_i) p_i^j q_i^j + a(\mathbf{x}_i) p_i^j = 0$$

$$\frac{\partial^{\alpha} q(\mathbf{x}_i,t_j)}{\partial t^{\alpha}} - \beta_1(\mathbf{x}_i) p_i^j r_i^j - \beta_2(\mathbf{x}_i) p_i^j q_i^j + b(\mathbf{x}_i) q_i^j = 0$$

$$\frac{\partial^{\alpha} r(\mathbf{x}_i,t_j)}{\partial t^{\alpha}} - \mathcal{D}\Delta r_i^j - k(\mathbf{x}_i) q_i^j + m(\mathbf{x}_i) r_i^j = 0 \qquad (10)$$

Now employing the Crank-Nicolson technique, the system (10) is converted to a system (11):

$$\mathfrak{D}_t^{\alpha} p_i(t_n) - \lambda(\mathbf{x}_i) + \beta_1(\mathbf{x}_i) p_i^n r_i^n + \beta_2(\mathbf{x}_i) p_i^n q_i^n + a(\mathbf{x}_i) p_i^n = 0$$

$$\mathfrak{D}_t^{\alpha} q_i(t_n) - \beta_1(\mathbf{x}_i) p_i^n r_i^n - \beta_2(\mathbf{x}_i) p_i^n q_i^n + b(\mathbf{x}_i) q_i^n = 0$$

$$\mathfrak{D}_t^{\alpha} r_i(t_n) - \frac{\mathcal{D}}{2}\left(\frac{r_{i+1}^{n-1} - 2r_i^{n-1} + r_{i-1}^{n-1}}{\Delta \mathbf{x}^2} + \frac{r_{i+1}^{n} - 2r_i^{n} + r_{i-1}^{n}}{\Delta \mathbf{x}^2} \right) - k(\mathbf{x}_i) q_i^n + m(\mathbf{x}_i) r_i^n = 0 \qquad (11)$$

The system (11) is substituted by the approximation formula of time fractional derivative in the Caputo sense, $\mathfrak{D}_t^{\alpha}(\cdot)$. The approximate solution of computing at the time level n as a system (12):

$$\sigma_{\alpha} \sum_{j=1}^{n} \left(p_i^{n-j+1} - p_i^{n-j}\right) d_{n-j} - \lambda(\mathbf{x}) + \beta_1(\mathbf{x}) p_i^n r_i^n + \beta_2(\mathbf{x}) p_i^n q_i^n + a(\mathbf{x}) p_i^n = 0$$

$$\sigma_{\alpha} \sum_{j=1}^{n} \left(q_i^{n-j+1} - q_i^{n-j}\right) d_{n-j} - \beta_1(\mathbf{x}) p_i^n r_i^n - \beta_2(\mathbf{x}) p_i^n q_i^n + b(\mathbf{x}) q_i^n = 0$$

$$\sigma_{\alpha} \sum_{j=1}^{n} \left(r_i^{n-j+1} - r_i^{n-j}\right) d_{n-j} - \frac{\mathcal{D}}{2\Delta x^2} (r_{i+1}^{n-1} - 2r_i^{n-1} + r_{i-1}^{n-1} + r_{i+1}^{n} - 2r_i^{n} + r_{i-1}^{n}) - k(\mathbf{x}) q_i^n + m(\mathbf{x}) r_i^n = 0 \qquad (12)$$

For $n = 1$, at the first-time level of the system (12) can be presented as a system (13):

$$p_i^1 = \frac{\lambda(\mathbf{x}) + \sigma_{\alpha} d_0 p_i^0}{\left(\sigma_{\alpha} d_0 + \beta_1(\mathbf{x}) r_i^1 + \beta_2(\mathbf{x}) q_i^1 + a(\mathbf{x})\right)}$$

$$q_i^1 = \frac{\beta_1(\mathbf{x}) p_i^1 r_i^1 + \sigma_{\alpha} d_0 q_i^0}{\left(\sigma_{\alpha} d_0 - \beta_2(\mathbf{x}) p_i^1 + b(\mathbf{x})\right)}$$

$$r_i^1 = \frac{\left(\begin{matrix} k(\mathbf{x}) q_i^1 + \sigma_{\alpha} d_0 r_i^0 \\ + \frac{\mathcal{D}}{2\Delta x^2} (r_{i+1}^0 - 2r_i^0 + r_{i-1}^0 + r_{i+1}^1 + r_{i-1}^1) \end{matrix} \right)}{\left(\sigma_{\alpha} d_0 + m(\mathbf{x}) + \frac{\mathcal{D}}{\Delta x^2}\right)} \qquad (13)$$

For $n \geq 2$, the time level n of the system (12) can be written as a system (14):

$$p_i^n = \frac{\left(\lambda(\mathbf{x}) + \sigma_{\alpha} d_0 p_i^0 + \sigma_{\alpha} \sum_{j=1}^{n-1} \left(d_{n-j} - d_{n-j-1}\right) p_i^{n-j}\right)}{\sigma_{\alpha} d_{n-1} + \beta_1(\mathbf{x}) r_i^n + \beta_2(\mathbf{x}) q_i^n + a(\mathbf{x})}$$

$$q_i^n = \frac{\left(\begin{matrix} \beta_1(\mathbf{x}) p_i^n r_i^n + \sigma_{\alpha} d_0 q_i^0 \\ + \sigma_{\alpha} \sum_{j=1}^{n-1} \left(d_{n-j} - d_{n-j-1}\right) q_i^{n-j} \end{matrix} \right)}{\sigma_{\alpha} d_{n-1} - \beta_2(\mathbf{x}) p_i^n + b(\mathbf{x})}$$

$$r_i^n = \frac{\left(\begin{matrix} k(\mathbf{x}) q_i^n + \sigma_{\alpha} d_0 r_i^0 \\ + \frac{\mathcal{D}}{2\Delta x^2} (r_{i+1}^{n-1} - 2r_i^{n-1} + r_{i-1}^{n-1} + r_{i+1}^{n} + r_{i-1}^{n}) \\ + \sigma_{\alpha} \sum_{j=1}^{n-1} \left(d_{n-j} - d_{n-j-1}\right) r_i^{n-j} \end{matrix} \right)}{\sigma_{\alpha} d_{n-1} + \frac{\mathcal{D}}{\Delta x^2} + m(\mathbf{x})} \qquad (14)$$

4 NUMERICAL EXPERIMENTS

This section presents examples for β_1 is a constant (Wang et al., 2014). The approximation of the integer order differential equation will be compared with the approximation of the fractional differential equation. To confirm that the formula for the approximation of the proposed scheme corresponds to the approximation of the integer order differential equation.

Example: Let the parameters as follows: $\alpha = 0.99, \lambda = 0.332, a = 1, b = 1, \beta_2 = 2, k = 2, m = 1$ and $D = 0.01$. The initial conditions in this case are $p(\mathbf{x}, 0) = 0.99, q(\mathbf{x}, 0) = 0, r(\mathbf{x}, 0) = e^{-(\mathbf{x}-5)^2} \times 10^{-3}, \mathbf{x} \in \Omega = [0,10]$. For $\beta_1 = 0.47$, it is found that the density of free virus $r(\mathbf{x}, t)$ became to zero. That is free of infection shown in figure 1(a). The absolute error between the approximation solutions of the fractional order with the integer order differential equation shown in figure 1(b). It is found that the numerical results for two scheme agree as well.

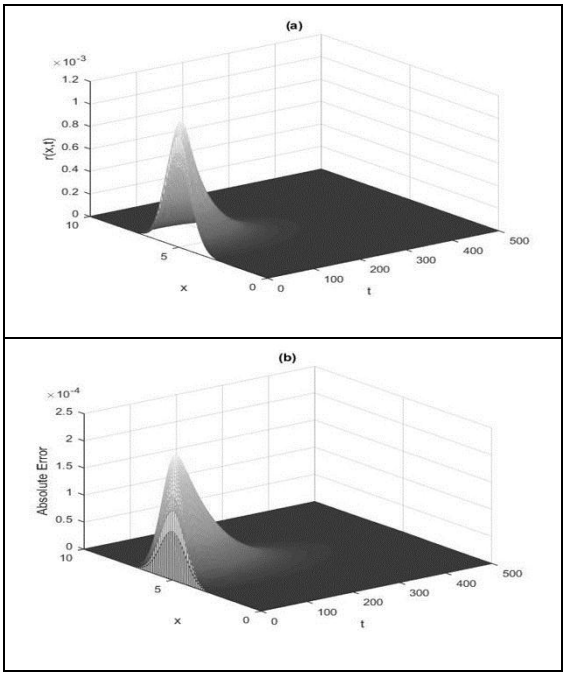
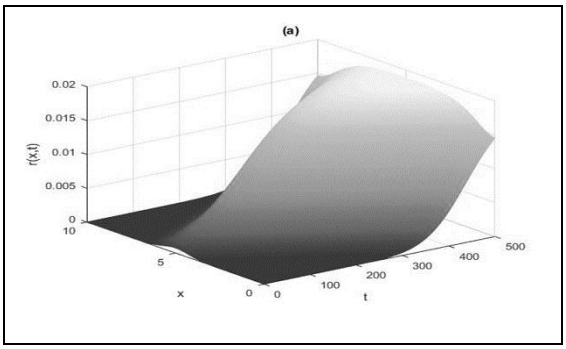


Figure 1. (a) The approximation solution of the fractional order differential equation (b) The absolute error between the approximation solutions of the fractional order with the integer order differential equation for $\beta_1 = 0.47$.

On the other hand, if $\beta_1 = 0.55$, then the density of free virus $r(\mathbf{x}, t)$ is unstable. That is the existence of the infection state remains shown in figure 2(a). The absolute error between the approximation solutions of the fractional order with the integer order differential equation shown in figure 2(b).



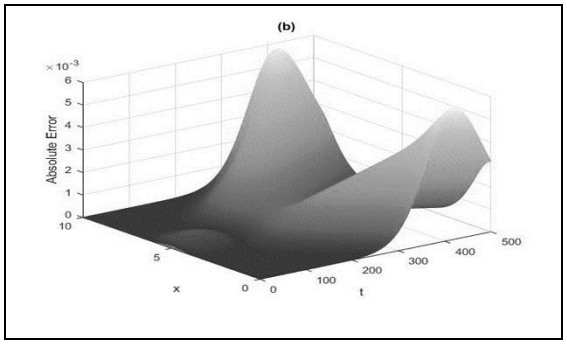


Figure 2. (a) The approximation solution of the fractional order differential equation (b) The absolute error between the approximation solutions of the fractional order with the integer order differential equation for $\beta_1 = 0.55$.

5 CONCLUSIONS

This research investigated mathematical models have been in the understanding of the dynamics of HIV infection. Numerical approximation of the fractional HIV model by using the Crank-Nicolson technique in the Caputo sense. It found that the Crank-Nicolson technique for the fractional differential equations is equally reliable with the approximation solution of the integer order differential equation.

ACKNOWLEDGEMENTS

This research was supported by King Mongkut's University of Technology Thonburi (KMUTT). The researcher would like to thank the advisor for providing advice and taking care of this research and Kanchanaburi Rajabhat University for providing a scholarship.

REFERENCES

Arafa, A. A. M., Rida, S. Z., & Khalil, M. (2012). Fractional modeling dynamics of HIV and $CD4^+$T-cells during primary infection. *Nonlinear biomedical physics*, 6(1), 1.

Atangana, A., & Alqahtani, R. T. (2016). Numerical approximation of the space-time Caputo-Fabrizio fractional derivative and application to groundwater pollution equation. *Advances in Difference Equations*, 2016(1), 156.

Dalir, M., & Bashour, M. (2010). Applications of fractional calculus. *Applied Mathematical Sciences*, 4(21), 1021–1032.

Ding, H. (2016). General Padé approximation method for time–space fractional diffusion equation. *Journal of Computational and Applied Mathematics*, 299, 221–228.

Loverro, A., (2004). Fractional calculus: history, definitions and applications for the engineer. *Rapport technique, University of Notre Dame: Department of Aerospace and Mechanical Engineering*, 1-28.

Murio, D. A. (2008). Implicit finite difference approximation for time fractional diffusion equations. *Computers & Mathematics with Applications*, 56(4), 1138-1145.

De Oliveira, E. C., & Machado, J. A. (2014). A review of definitions for fractional derivatives and integral. *Mathematical Problems in Engineering*, 2014.

Ren, J., Sun, Z. Z., & Dai, W. (2016). New approximations for solving the Caputo-type fractional partial differential equations. *Applied Mathematical Modelling*, 40(4), 2625–2636.

Sweilam, N. H., Khader, M. M., & Mahdy, A. M. S. (2012). Crank-Nicolson finite difference method for solving time-fractional diffusion equation. *Journal of Fractional Calculus and Applications*, 2(2), 1-9.

Thamareerat, N., Luadsong, A., & Aschariyaphotha, N. (2017). Stability results of a fractional model for unsteady-state fluid flow problem. *Advances in Difference Equations*, 2017(1), 74.

Wang, F. B., Huang, Y., & Zou, X. (2014). Global dynamics of a PDE in-host viral model. *Applicable Analysis*, 93(11), 2312-2329.

Wang, J., Yang, J., & Kuniya, T. (2016). Dynamics of a PDE viral infection model incorporating cell-to-cell transmission. *Journal of Mathematical Analysis and Applications*, 444(2), 1452-1564.

Systematic Review and Meta-analysis of Positive Youth Development (PYD) Programmes on Condom and Hormonal Use among Adolescents

Ratu Luke Mudreilagi[1,2], Thammasin Ingviya[3], Rassamee Sangthong[1*]

[1]Prince of Songkla University/Epidemiology Unit, Hat Yai, Thailand
[2]Fiji National University/School of Nursing, Suva, Fiji
Email: luke2016.psu@gmail.com
[3]Prince of Songkla University/Department of Family Medicine and Preventive Medicine, Hat Yai, Thailand
Email: thammasin@gmail.com
*Corresponding Email: rassamee.s@psu.ac.th

ABSTRACT

Sexual health is one of the important health issues among adolescents. Attention to their perceptions and needs is essential, along with development of policies, services, and programmes that address those needs. Positive youth development (PYD) may be a promising strategy for promoting adolescent sexual and reproductive health. This systematic review was conducted to assess the effect of PYD programs to improve sexual health including condom use and hormonal use. The retrieval and inclusion criteria were developed a priori and applied to search the literature. Several databases were searched for articles about PYD programmes published between 2000 and 2013. Two independent reviewers assessed the methodological quality and abstracted data using PRISMA (Preferred Reporting Items for Systematic Reviews and Meta Analyses) guidelines. A meta-analysis was performed on the interventions that had a contemporaneous control group and conceptual similarity in the outcomes of interest.

A total of 6 studies (2,079 female and 1,755 male adolescents) were reviewed. Meta-analysis showed PYD programs did not significantly increase condom use [OR=1.02 (0.78-1.31)] among male adolescents. However, it could increase the use of hormonal contraception by two times [OR=1.8 (1.15-2.82)] among female adolescents in the intervention groups compared to the control groups. Effective PYD programmes can promote adolescent sexual health and should be part of a comprehensive approach especially among female adolescents. Further research should be done to increase the effectiveness of PYD program on sexual health among male adolescents and other sexual outcomes such as pregnancy in female adolescents.

Keywords: positive youth development (PYD); sexual health; adolescents; meta-analysis

1 INTRODUCTION

Adolescents who engaged themselves in sexual behavior put themselves at risk for pregnancy, HIV, and other sexually transmitted infection (Mirzazadeh et al, 2017). Teenage pregnancy rate was from 20/10,000 female adolescents per year in Europe to as high as 143/10,000 female adolescents per year in Sub Saharan Africa (Sedgh et al, 2015). Meanwhile, the proportion of abortion from teenage pregnancies ranged from 17% in Slovakia to 69% of all abortions in Sweden (Shepherd et al, 2010). UNICEF reported that teen pregnancy might result in adverse psychological and socio-economic outcomes for the mother and her child. A teen mother is prone to suffer from depression, leading to suicide. While the child of the teen mother is at risk to be neglected or abused (UNICEF Malaysia Communications, 2008).

Numerous studies have been employed to promote protected sex including condom and hormonal use (Oringanje et al., 2016; Ghobadzadeh et al, 2016). However the protected sex practice is still low as the condom use and the hormonal use were 60% and 41%, respectively (Ghobadzadeh et al., 2016; Marseille et al., 2018) There is also a need to concentrate efforts on sustaining and prolonging the adoption of healthy sexual behaviors and outcomes (Gavin, Catalano, Markham, 2010).

Positive youth development (PYD) is a comprehensive framework outlining the supports to enhance youth's interests, skills, and abilities (Gavin et al, 2010). PYD encompasses psychological, behavioral and social characteristics that reflects the 5 Cs.(Bowers et al., 2010). The 5Cs include competent, confidence, connection, character and caring (Lerner et al, 2005). These 5Cs could lead youth to reach their full potential and contribute to their society (Bowers et al, 2010).

Many individual PYD studies reported that PYD program is vital in promotion of adolescent and young people sexual and reproductive health (Gavin et al , 2010). Adolescents with greater self-respect and honored family connectedness address more steady hormonal contraceptive use (Ghobadzadeh et al, 2016). It is therefore worth to see the combination of previous studies of PYD on sexual health. The purpose of the current study is to use systematic review and meta-analysis to assess the effectiveness of PYD programmes on condom and hormonal use among adolescents.

2 METHODS

The literature search was limited to papers, published in English language between January 2000 and December 2013. The search included the data sources, including PubMed, BMJ Journals, ProQuest, Springer link, and Google scholar. The Search terms included "positive youth development", "sexual health", "adolescents", "systematic review" and "meta-analysis". Additional unpublished papers were also searched.

Articles were then independently screened for the following inclusion criteria: First, randomized controlled trials and quasi-experiment of PYD programs were eligible. Second, the population studied must include either adolescents and/or young adults aged 13 – 19 years. Third, the study outcome must include condom and hormonal uses. Finally, the study must report number of students between the intervention and the control groups.

2.1 Reviewing process:

Two reviewers independently reviewed all eligible articles. We used RoB 2.0, tool to assess the risk of bias in those articles (Higgins & Thompson, 2002). Risk of bias arising from randomization processes, deviations from intended interventions, missing data, measurements of the outcomes and the result reports was assessed. If there was a discrepancy in the risk of bias between the two reviewers, a discussion to find a mutual agreement was done.

2.2 Data extraction:

The following information of each study was extracted: first author's name, year of publication, total numbers of sample included in each study, and number of samples in intervention and control groups by sexual health outcomes including hormonal use and condom use. All data were then recorded in an excel spread sheet.

2.3 Meta-analysis

All data analysis were done with R program version 3.3.3 using the "metafor" package (R Development Core Team, 2017) for meta-analysis (Viechtbauer, 2010). Cochran Q test and I^2 statistic were used to examine heterogeneity across the studies. Funnel plots were used to explore reported bias and heterogeneity. Fixed effect and random effect (DerSimonian and Laird) models (Guolo & Varin, 2017) were used to estimate the pooled odds ratios of two main sexual health outcomes. A Fixed effect model was used when Cochran Q's p-value was > 0.1 and I^2 < 25%, otherwise, a random effect model was used.

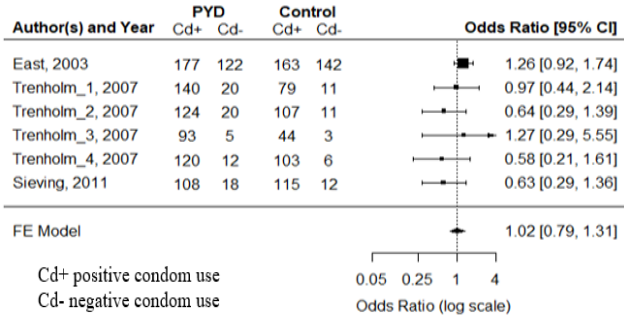
3 RESULTS

We identified 24 articles, 20 published reported from electronic database and 4 unpublished reports. After removal of duplicated articles across database, 13 abstracts were further screened for the eligible criteria and reviewed with the RoB2.0 tool. Eventually, the full text of 6 studies, which met the inclusion criteria were obtained. The 6 studies were summarized in Table 1.

Table 1: Summary of PYD programs in this systematic review and meta-analysis

East, 2003 n = 1467 The intervention programs is made up for services dedicated to improve their psychological skills, sexuality and health education, community services or recreational activities, help with school and employment concern.
Trenholm 1 *(My Choice, My Future Program)*, 2007 n = 498 Intervention: Classroom intervention based on three curriculums. The Reasonable Reasons to Wait: The Keys to Character focus on character development, reasons to wait to engage in sex, peer influencing, dating, avoiding STDs, relationship skills, and the benefits and ingredients of a strong marriage at the eighth year grade. The Art of Loving Well which features short stories, poetry, classic fairy tales and myths that taught about healthy and loving relationship. During the final year of the program, the Wait Training Curriculum which focus on relationship skills and risk avoidance was given. They also showed materials which provide information on STDs and instructed students that abstinence is the only sure way to avoid contracting.
Trenholm 2 (*Recapturing the Vision & Vessels of Honor*), 2007 n = 523 The program was ventured on identifying personal courage and resources, developing approaches for fulfilling personal careers goals, and building critical skills that would help youth accomplish their goals and prevent negative influences. The complimentary Vessels of Honor curriculum comprised of six key area of focus (1) honorable behavior, (2) effective communication for opposing pressure to engage in sex and other high risk behaviors, (3) development of good relationship and fulfilling social needs and emotional feelings through friendship instead of having sex, (4) physical development and its association for changing pressure, (5) sexual abuse and date rape and how to refrain from it, (6) strategies for deciding a mate and the advantages of a committed relationship. Other complimentary services including home visit by social workers, referrals to local services after school tutoring cultural events, a family retreat, an Annual Teen Abstinence Rally, an annual Teen Talk Symposium with celebrity panelists.
Trenholm 3 (*A Life Options Model Curriculum for Youth*), 2007 n = 359 The curriculum contains 10 topic areas, nearly all of which have abstinence as a target: (1) group-building, (2) self-esteem, (3) values and goal setting, (4) decision –making skills, (5) risk-taking behavior, (6) communication skills, (7) relationship and sexuality, (8) adolescent development and anatomy, (9) sexually transmitted disease, and (10) social skills. The unit on relationship and sexuality addressed marriage in addition to abstinence.
Trenholm 4 (*Teens in Control*), 2007 n = 481 Postponing Sexual Involvement curriculum has five topics which focused on the risks of early sexual involvement and the advantages of being abstinence, social and peer pressure to have sex, and the development of a particular skills for resisting peer pressure using extensive practice sessions and reinforcement. The Sex Can Wait curriculum covered several key areas such as: self-concept and self-esteem: physical and psychological changes during puberty; values; communication skills: information on the risk of STDs; skills for resisting social and peer pressure; and the formulation of career goals, planning on how to achieve them.
Sieving, 2011 n = 506 The program had 3 interventions: 1) Case management: one to one visit and discussion focus on emotional skills, health relationship, responsible sexual behaviors, positive family and school and community involvement. 2) Peer Leadership Components intervention provides a hands on skill-building experience to foster development and pro-social interaction skills. The sequence of peer educator training is followed by service learning programming. 3) Peer Educator Training and Employment; Just in Time program. They address communication skills, stress management skills, conflict resolutions skills, expectation and skills for healthy relationship, understanding social influences on sexual behaviors, sexual decision-making and contraceptive use skills.

From the funnel plots (Figure 2 and 4), risk of reporting bias and/or publication bias was low. The individual and pooled odds ratio of PYD programs on condom use showed no statistical significant among male adolescents (Figure 1). On the contrary, the forest plot (Figure 3) of PYD on hormonal used, showed that the proportion of hormonal uses in female adolescents in the intervention group was 2 times significantly higher than those in the control group.



I^2 (total heterogeneity / total variability): 15.01%
Test for Heterogeneity: Q (df = 5) = 5.8832, p-val = 0.3178
FE = Fixed Effect Model

Figure 1: Forest plot of condom use between female adolescents in the PYD and the control group.

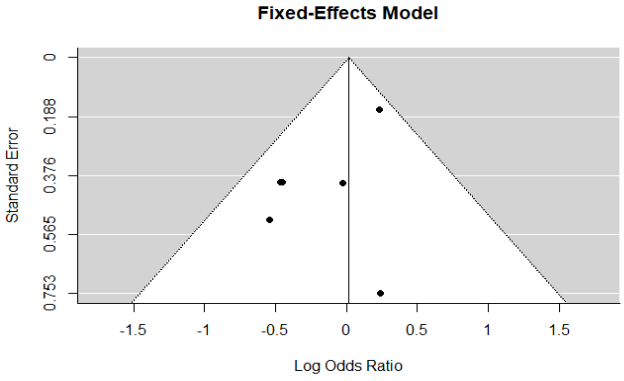
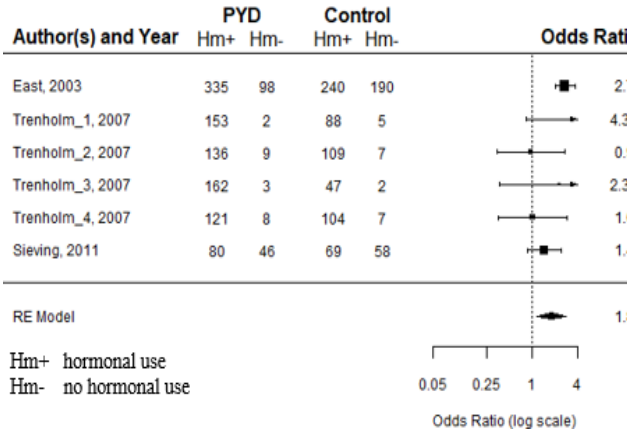


Figure 2: Funnel plot for condom use by fixed-effect model (FE)



I^2 (total heterogeneity / total variability): 47.84%
Test for Heterogeneity: Q (df = 5) = 9.5861, p-val = 0.0878
RE = Random Effect Model

Figure 3: Forest plot of hormonal use between female adolescents i n the PYD and the control group

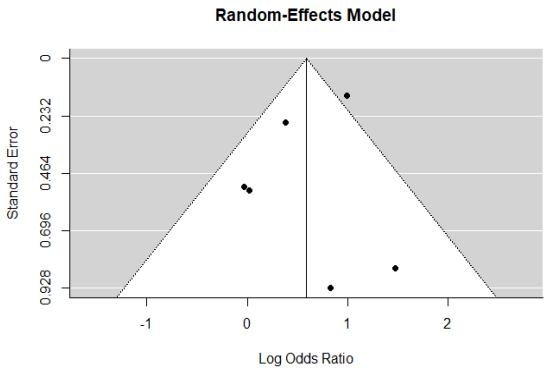


Figure 4: Funnel plot for hormonal use by random-effect model

4 DISCUSSIONS

Our systematic review and meta-analysis showed that PYD programs cannot increase condom use among male adolescents but can increase hormonal use among female adolescents by two times. This hormonal use among female adolescents may prevent them from getting pregnant however they are still at high risk of sexual transmitted diseases.

Adolescents' brain structures and functions have changed substantially compared to when they were children (Steinberg et. al, 2010). This is part of a natural process to transform them towards adulthood. They have developed secondary sex characteristics, sexual attraction, sexual desire and feel exciting to have intimacy relationship. Their executive functions such as higher cognitive function, self-control and appropriate decision making, however, will be maturely developed later on when they reach early adulthood at 20s (Swartz et .al, 2014). These two biological-driven factors strongly lead them to have sexual initiation and sex activity.

In order to unleash the sexual risk behaviour's among adolescents, appropriate PYD program should be further studied. Our study suggested that the effectiveness of PYD programs for male and female adolescents were different. Hence the program should account for gender. The studies included in the study were mainly conducted in the United States and England which may not be applicable to adolescents in different culture such as in Thailand.

A number of study were excluded because the data extraction could not be made. This may lead to some bias towards null hypothesis and may underestimate the results. PYD program interventions were largely different from one study to another. This makes it difficult to draw a conclusion which PYD feature is essential and effective to promote sexual health. The use of condom and hormonal contraception defined in each study was, however, only slightly different and still could be combined.

ACKNOWLEDGEMENTS

The authors would like to thank Miss Teerohah Donroman for her valuable assisting in the data management and analysis.

REFERENCES

Bowers, E. P., Li, Y., Kiely, M. K., Brittian, A., Lerner, J. V., & Lerner, R. M. (2010). The Five Cs model of positive youth development: A longitudinal analysis of confirmatory factor structure and measurement invariance. *Journal of Youth and Adolescence*, *39*(7), 720–735.

Gavin, L. E., Catalano, R. F., & Markham, C. M. (2010). Positive youth development as a strategy to promote adolescent

sexual and reproductive health. *Journal of Adolescent Health*, *46*(3), S1–S6.

Ghobadzadeh, M., Sieving, R. E., & Gloppen, K. (2016). Positive youth development and contraceptive use consistency. *Journal of Pediatric Health Care*, *30*(4), 308–316.

Guolo, A., & Varin, C. (2017). Random-effects meta-analysis: the number of studies matters. *Statistical Methods in Medical Research*, *26*(3), 1500–1518.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558.

Lerner, R. M., Almerigi, J. B., Theokas, C., & Lerner, J. V. (2005). Positive youth development a view of the issues. *The Journal of Early Adolescence*, *25*(1), 10–16.

Marseille, E., Mirzazadeh, A., Biggs, M. A., P. Miller, A., Horvath, H., Lightfoot, M., Kahn, J. G. (2018). Effectiveness of school-based teen pregnancy prevention programs in the USA: a systematic review and meta-analysis. *Prevention Science*, *19*(4), 468–489.

Oringanje,C., Meremikwu, M.M.H., Esu, E., Meremikwu, A., & Ehiri, J.E. (2016). Interventions for preventing unintended pregnancies among adolescents. Cochrane Database of Systematic Review

R Development Core Team. (2017). R: A language and environment for statistical computing (Version 3.3.3). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Sedgh, G., Finer, L.B., Bankole, A., Eilers, M.A., & Singh, S. (2015). Adolescent pregnancy, birth, and abortion rates across countries: levels and recent trends. *Journal of Adolescent Health*, *56*(2), 223–230.

Shepherd, J., Kavanagh, J., Picot, J., Cooper, K., Harden, A., Barnett-Page, E, Price, A. (2010). The effectiveness and cost-effectiveness of behavioural interventions for the prevention of sexually transmitted infections in young people aged 13–19: a systematic review and economic evaluation. *Health Technology Assessment*, *14*(7). https://doi.org/10.3310/hta14070

Steinberg, L. (2010). A behavioral scientist looks at the science of adolescent brain development. *Brain and Cognition*, *72*(1), 160–164.

Swartz, J. R., Carrasco, M., Wiggins, J. L., Thomason, M. E., & Monk, C. S. (2014). Age-related changes in the structure and function of prefrontal cortex–amygdala circuitry in children and adolescents: A multi-modal imaging approach. *NeuroImage*, *86*, 212–220.

UNICEF Malaysia Communications. (2008). *Young people and family planning: Teenage pregnancy*.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor Package. *Journal of Statistical Software*, *36*(3). https://doi.org/10.18637/jss.v036.i03

Predicting TB Death Using Decision Tree Model in Reliable Mortality Data

Muhamad Rifki Taufik[1*], Apiradee Lim[2], Phattrawan Tongkhumchum[3] and Nurin Dureh[4]

[1]Research Methodology, Mathematics and Computer Science Department, PSU, Pattani, Thailand
*Corresponding Email: m.rifki.taufik@gmail.com
[2] Research Methodology, Mathematics and Computer Science Department, PSU, Pattani, Thailand
Email: apiradee.s@psu.ac.th
[3] Research Methodology, Mathematics and Computer Science Department, PSU, Pattani, Thailand
Email: phattrawan.t@psu.ac.th
[4] Research Methodology, Mathematics and Computer Science Department, PSU, Pattani, Thailand
Email: dnurin@gmail.com

ABSTRACT

Medical record faces a big problem that is low-quality data where over 50% of data registered ill-defined category. Verbal autopsy study attempted to fulfill it by providing more accurate data. Tuberculosis (TB) case has been chosen as an interesting variable. Several efforts conducting prediction of TB death has not been done properly where the accuracy of prediction still can be improved. This study aimed to predict TB death using decision tree in verbal autopsy data. The data split into a training set and a testing set in three ratios 60:40, 70:30, 80:20, respectively where both R programming and Python had been used to analyze. The data consisted of four categorical determinants and one dichotomous outcome. Based on R, The result showed that death registry and age-gender group significance in ratio 60:40 and 80:20 and additionally province also significant in ratio 70:30 with the TB death. Python gave specific categories which were significant variables that were Death registry respiratory, female and male aged 50-59 and female with age 80+, and provinces Nakon Nayok and Songkla were associated with TB mortality in ration 60:40 and 80:20. in ratio 70:30, DR respiratory and other cancer groups, agsx 5,9,12, female aged 50-59 and 70-79, male aged 80+, and die outside hospital were associated with TB Death.

Keywords: decision tree, prediction, machine learning, verbal autopsy

1 INTRODUCTION

In Thailand, 35% of all deaths occur in hospitals, and the cause of death is medically certified by attending physicians. About 15% of hospital deaths are registered with nonspecific diagnoses, despite the potential for greater accuracy using information available from medical records. Further, issues arising from transcription of diagnoses from Thai to English at registration create uncertainty about the accuracy of registration data even for specified causes of death. Procedures for death certification and coding of underlying causes of death need to be streamlined to improve the reliability of registration data (Pattaraarchachai et al., 2010). Ascertainment of cause for deaths that occur in the absence of medical attention is a significant problem in many countries, including Thailand, where more than 50% of such deaths are registered with ill-defined causes. Routine implementation of standardized, rigorous verbal autopsy methods is a potential solution (Polprasert et al., 2010).

Verbal autopsy (VA) has become a primary source of information about causes of death in populations lacking vital registration and medical certification (WHO, 2012). VA has the capability to produce more reliable data. Reliable data on the levels and causes of mortality are cornerstones for building a solid evidence base for health policy, planning, monitoring and evaluation (Prasartkul, Porapakham, Vapattanawong and Rittirong, 2007; Yang et al., 2006). In the case where the majority of deaths occur at home and where civil registration systems do not function, there is little chance that deaths occurring away from health facilities will be recorded and certified as to the cause or causes of death (Soleman et al., 2006). The methods enabled health professionals to estimate the specific cause of deaths in countries where the low quality of causes of death in the Death Registry (DR) database and reliable data such as the VA data is available (Pipatjaturon et al., 2017). VA comes as an essential public health tool for obtaining a reasonable direct estimation of the causal structure of mortality at a community or population level, VA is also capable to identify ill-defined categories to specific causes of death such as diabetes, cancer, and TB (Polprasert et al., 2010) and some misclassification from other categories (Pattaraarchachai et al., 2010). Considering an ability to predict the correct caused earns verbal autopsy data to be applied in this research with TB as an interesting case.

In medical decision making (classification, diagnosing, etc.) there are many situations where a decision must be made effectively and reliably. And VA is the most reliable mortality data in Thailand. Conceptual simple decision-making models with the possibility of automatic learning are the most appropriate for performing such tasks. Decision trees are a reliable and effective decision-making technique that provide high classification accuracy with a simple representation of gathered knowledge and they have been used in different areas of medical decision making (Podgorelec et al., 2002).

2 METHODS

This study used secondary data from a 2005 VA survey, which assessed the causes of death based on a sample of 9,644 cases (3,316 in-hospital deaths and 6,328 outside-hospital deaths) from 28 districts in nine provinces (Rao et al., 2010). The data consisted of four determinants which were the location of death (inside or outside the hospital), age-gender group, province, and death registry, and one binary outcome that is dying from TB or other diseases. This study developed predictive models using decision tree where the data were treated in three ratios, 60:40, 70:30, and 80:20. Classification accuracy and area under the ROC curve would measure the goodness of fit in each ratio. R programming and Python had been employed to do the analysis since these both tools are widely used and have strong analysis. Both Python and R are amongst the most popular languages for data analysis.

Table 1 Variable detail

Variable	Number of group	Detail	Coded
Province (pro)	9 province	Bangkok, Nakhon Nayok, Suphan Buri, Ubon Ratchathani, Loei, Phayao, Chiang Rai, Chumphon, and Songkhla	1, 2, …, 9

Age-gender (AgSx)	12 group	ages 5-39, 40-49, 50-59, 60-69, 70-79 and 80+ years for each sex.	1, 2, …, 12
Location of Death (ghos)	2 location	Inside hospital and outside the hospital	1, 2
Death Registry (DRgrp)	21 major cause	TB, Septicemia, HIV, Other Infectious, Liver Cancer, Lung Cancer+, Other Digestive Cancer, Other Cancer, Endocrine, Mental-Nervous, Ischemic, Stroke, Other CVD, Respiratory, Digestive, GenitoUrinary, Ill-defined, Transport Accident, Another injury, Suicide, All other	1:a, 2:b, …, 21:u
TB (tb)	Binary	Death due to TB or others	0, 1

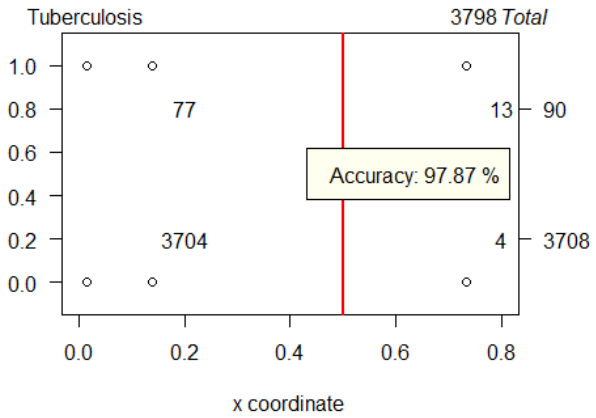
3 RESULTS AND DISCUSSION

Decision tree constructed a tree as a predictive model where the first branch is the most important determinant. All those ratios showed the same chosen significant determinants that were death registry and age-gender group. Those models depicted high accuracy prediction where all ratios got classification accuracy of nearly 98%.

Figure 1 and Figure 2 depicted result from ratio 60:40 in the R program. CA in Figure 1 gave a high accuracy of prediction TB. Unfortunately, after AUC assessed the model, this ratio did not develop a good predictive model since the value of AUC was only 0.608. This value was close to the threshold line where the value was 0.5

In ratio 70:30, R program introduced a similar result from the previous ratio with CA 97.68 and AUC 0.599. Even the values were a bit decreasing, but, these differences were not significant.

The third ratio in Figure 3 also did not improve predictive performance where the decision tree gave CA 97.84 and AUC 0.597 that very closed to both previous ratios. Further figures would depict decision tree performance predicting TB death in Python.



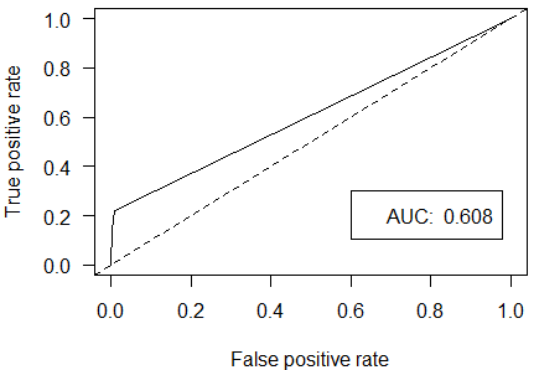
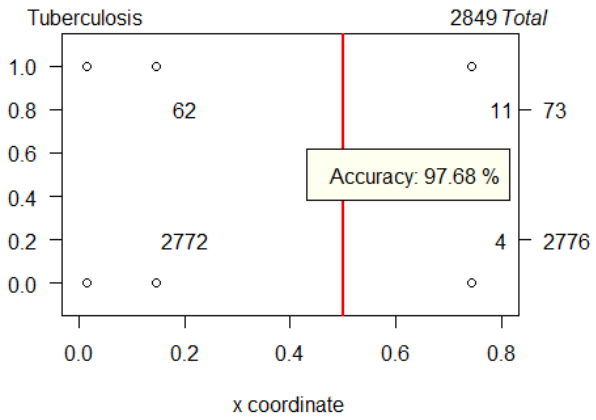


Figure 1 ratio 60:40 in R



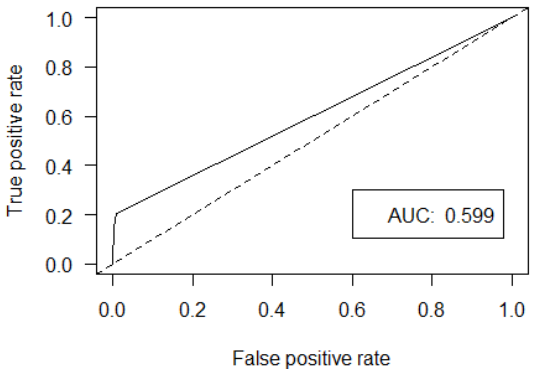


Figure 2 ratio 70:30 in R

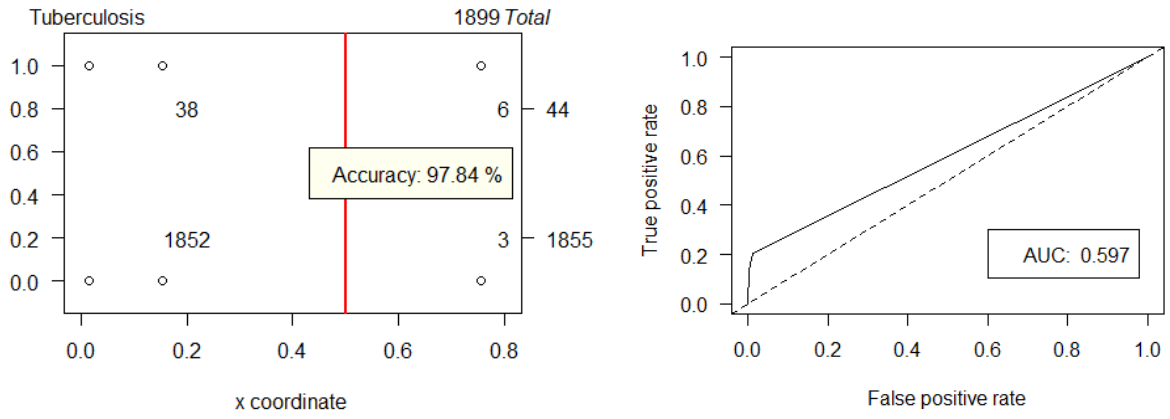


Figure 3 ratio 80:20 in R

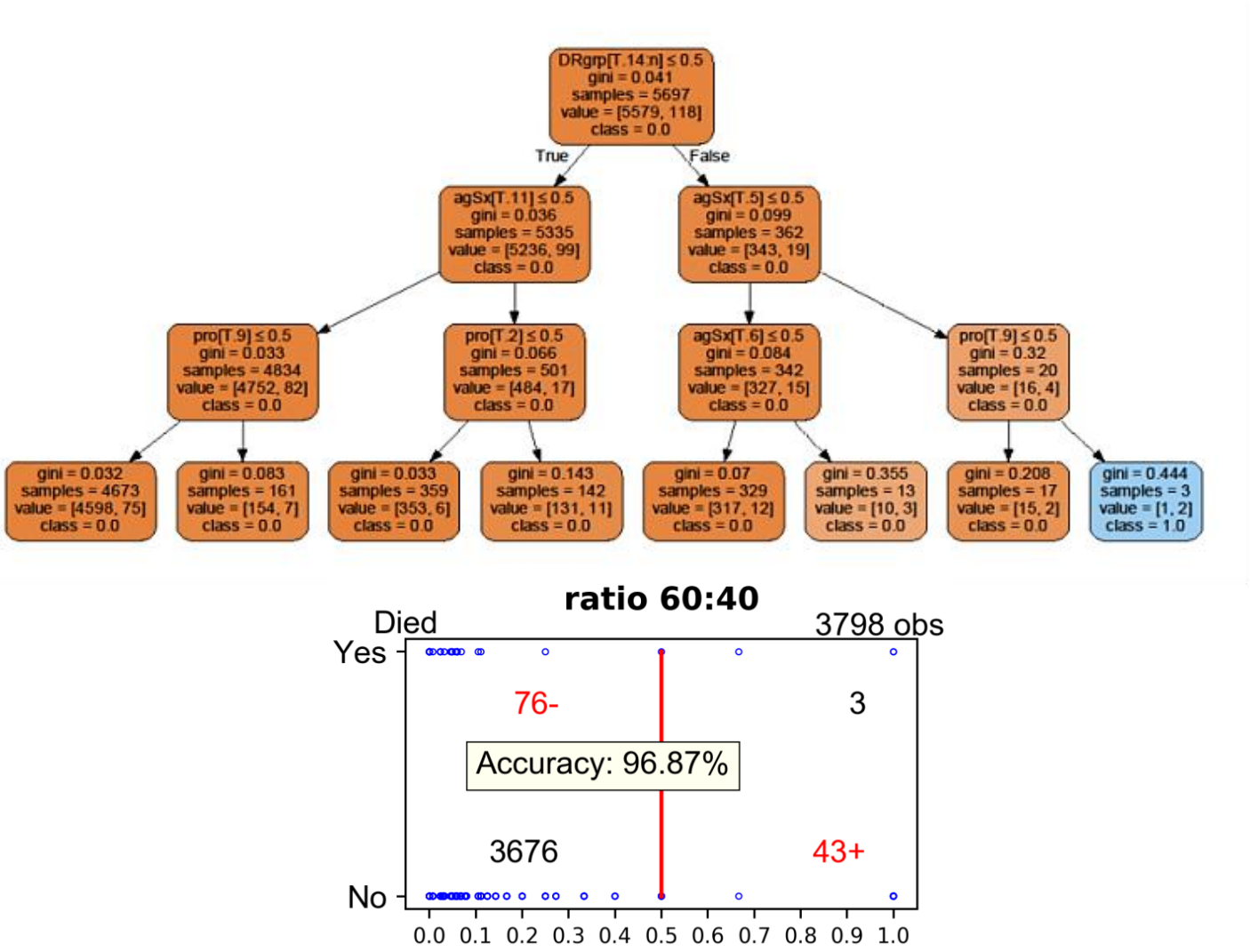


Figure 4 ratio 60:40 in Python

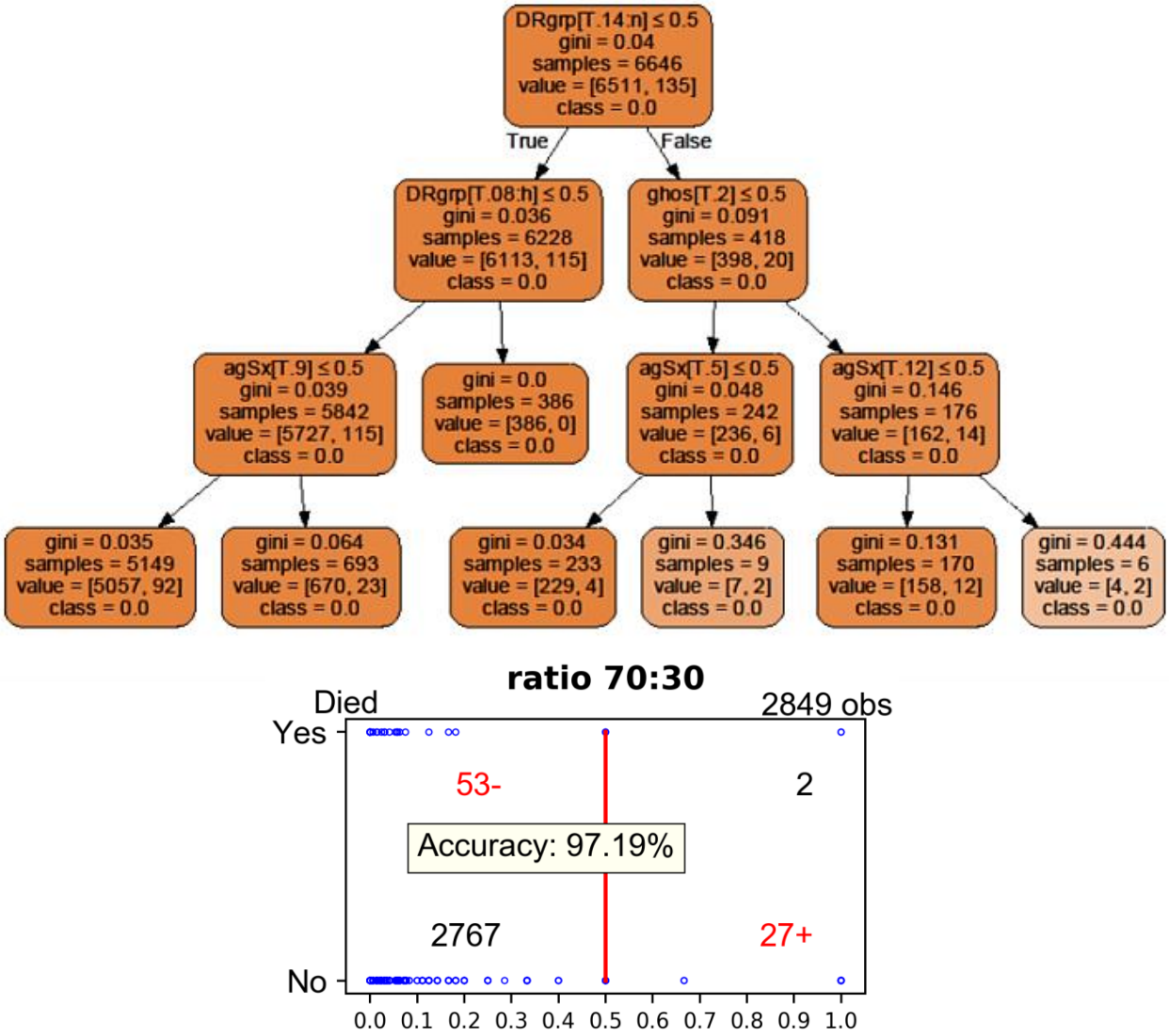


Figure 5 ratio 70:30 in Python

Ratio 60:40 for the training set and the testing set was able to predict TB death with 96.87 showed in Figure 4. The CA was similar to the current ratio in R. Decision tree in Python predicted TB patient more spread out compared to the decision tree in R.

Somehow, ratio 70:30 in Figure 5 gave accuracy 97.19% where this CA improved a bit, even not significant from the previous ratio.

The last ratio 80:20 did not improve the accuracy since a number of the testing set also was decreased. Figure 6 showed that CA in this ratio was 96.52% with only 1899 observation in the testing set.

Figure 7 depicted ROC curve from three different ratios. It showed that the predictive model would give better performance in the larger testing set. Another way, a number of training sets did not improve predictive model even the number of observation had been increased.

ROC curve was able to distinguish whether the model is better. R programming gave not a significant value of AUC. Contrary, AUC in Python could able to gain some movement. The highest AUC was in ratio 60:40, the lower number of the testing set decreasing value of AUC.

TB death patient highly related to death registry record and age-gender group. According to various ratios, R programming gave not significant result for both measurement classification accuracy and ROC curve. In Python, classification accuracy still did not give a significant result, but ROC showed that ratio might give some impact to the result.

4 CONCLUSIONS

Based on R, death registry and age-sex variables were significantly associated with TB mortality in all ratios. Additionally, the province is also significant in ratio 70:30 and 80:20. Different ratios of training and testing sets, the result gave not different result from decision tree showed a similar result in AUC and CA

Based on Python, Death registry respiratory, female and male aged 50-59 and female with age 80+, and provinces Nakon Nayok and Songkla were associated with TB mortality in ration 60:40 and 80:20. in ratio 70:30, DR respiratory and other cancer groups, agsx 5,9,12, female aged 50-59 and 70-79, male aged 80+, and die outside hospital were associated with TB Death. Different ratios of training and testing sets gave similar result also from decision tree shown in CA. Contrary, the larger testing set gave higher AUC result from a decision tree model.

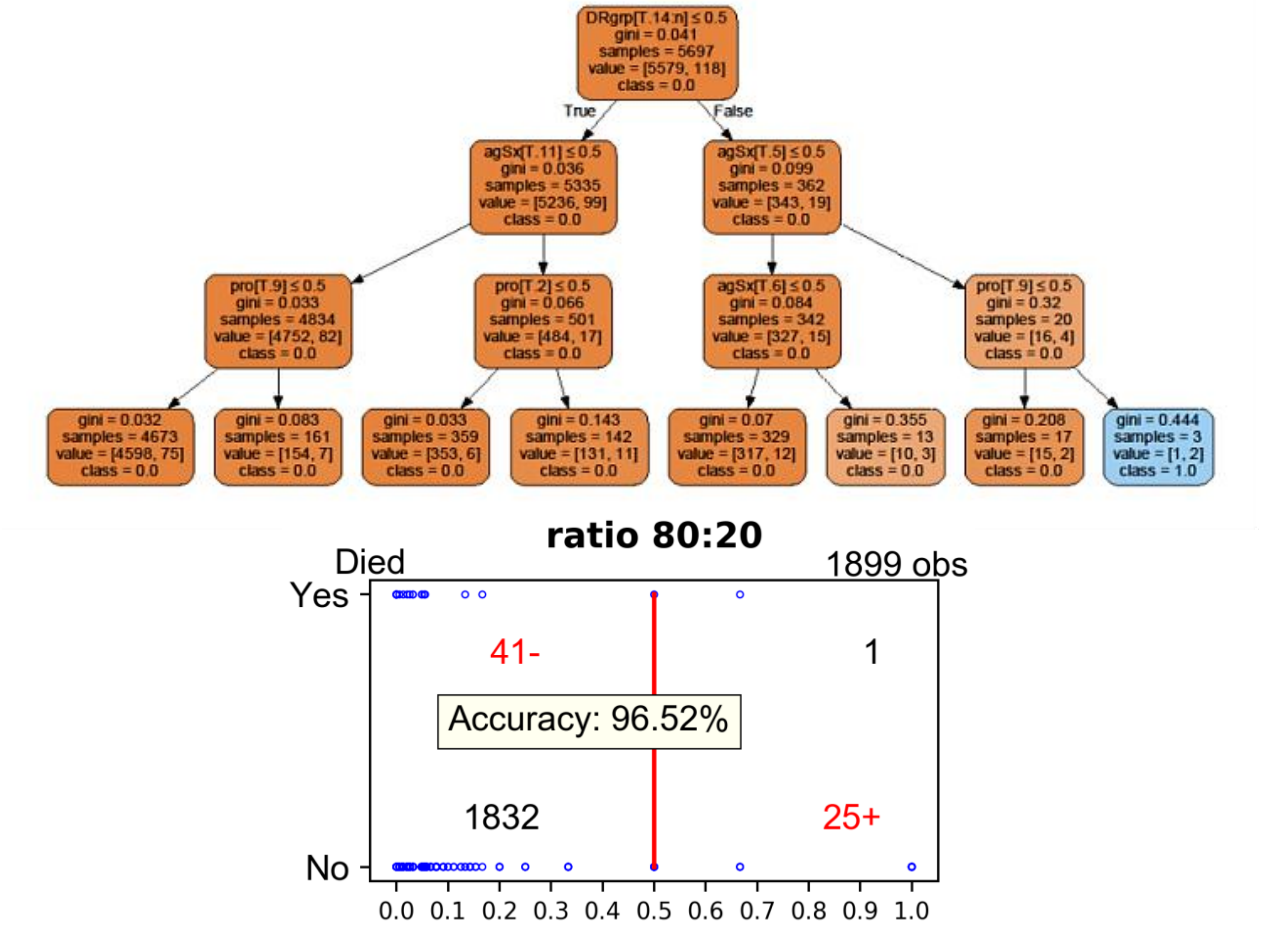


Figure 6 ratio 80:20 in Python

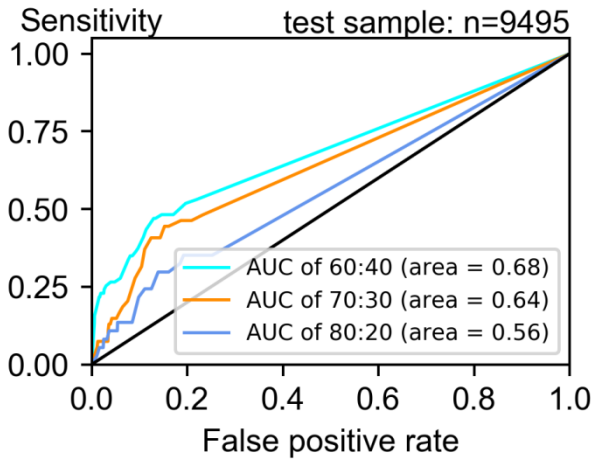


Figure 7 ROC for all ratios in Python

ACKNOWLEDGMENTS

I am grateful to Dr. Kanitta Bundhamcharoen as an officer in the Ministry of Public Health for allowing permission to obtain the data to do a study. This work was supported by the Higher Education Research Promotion and Thailand's Education Hub for Southern Region of ASEAN Countries Project Office of the Higher Education Commission.

REFERENCES

Pattaraarchachai, J., Rao, C., Polprasert, W., Porapakkham, Y., Pao-In, W., Singwerathum, N., & Lopez, A. D. (2010). Cause-specific mortality patterns among hospital deaths in Thailand: validating routine death certification. *Population Health Metrics*, *8*(1), 1–12.

Pipatjaturon, N., Tongkumchum, P., & Ueranantasun, A. (2017). Estimating lung cancer deaths in thailand based on verbal autopsy study in 2005. *Pertanika Science & Technology*, *25*(2), 469–478.

Polprasert, W., Rao, C., Adair, T., Pattaraarchachai, J., Porapakkham, Y., & Lopez, A. D. (2010). Cause-of-death ascertainment for deaths that occur outside hospitals in Thailand: application of verbal autopsy methods. Population Health Metrics, 8(1), 13.

Prasartkul, P., Porapakham, Y., Vapattanawong, P., & Rittirong, J. (2007). Development of a verbal autopsy tool for investigating cause of death : the Kanchanaburi project. *Journal Of Population And Social Studies*, *15*(2), 1–22.

Rao, C., Porapakkham, Y., Pattaraarchachai, J., Polprasert, W., Swampunyalert, N., & Lopez, A. D. (2010). Verifying causes of death in Thailand: rationale and methods for empirical investigation. Population health metrics, 8(1), 11..

Soleman, N., Chandramohan, D., & Shibuya, K. (2006). Verbal autopsy: Current practices and challenges. *Bulletin of the World Health Organization*, *84*(3), 239–245.

Podgorelec, V., Kokol, P., Stiglic, B., & Rozman, I. (2002). Decision trees: an overview and their use in medicine. Journal of medical systems, 26(5), 445-463..

World Health Organization. (2012). *Verbal autopsy standards*. Switzerland: World Health Organization.

Yang, G., Rao, C., Ma, J., Wang, L., Wan, X., Dubrovsky, G., & Lopez, A. D. (2006). Validation of verbal autopsy procedures for adult deaths in China. *International Journal of Epidemiology*, *35*(3), 741–748.

2-periods Coupon Bond Assessment in Regard to the Existence of Jump Diffusion Model on Asset Prices

Di Asih I Maruddani*

Department of Statistics, Diponegoro University, Semarang, Indonesia
*Corresponding Email: maruddani@live.undip.ac.id

ABSTRACT

Credit risk theory for bond valuation is conventionally expressed as zero coupon bond and assumes normal distribution in the logarithmic returns of asset prices. Those assumptions have been widely used in the Black-Scholes-Merton bond framework to model the return of assets. In real bond trading, the most common form of debt is coupon bond and the logarithmic returns of asset prices are not normally distributed. The characteristics of asset prices commonly contain jump processes. This paper study the valuation of 2-periods coupon bond where the asset price process contains a compound Poisson jump component, in addition to a continuous log-normally distributed component. We derive closed-form solution for equity and default probability of the bond which gives coupon for 2 periods using straightforward integration.

Keywords: jump diffusion processes; compound Poisson; compound option; extreme value assets

1 INTRODUCTION

Academic literature on credit risk has grown substantially over the last decades. One way of modeling credit risk is the structural approach which relies on a thorough description of the economics of default. A wide range of theoretical structural models have been proposed and they all contribute to the literature as they focus on the effects of various realistic economic considerations on credit risk valuation and prediction.

Credit risk is the risk resulting from credit events such as changes in credit ratings, restructuring, bankruptcy, etc. Formal definition of credit risk is the distribution of financial losses due to unexpected changes in the credit quality of counterparty in a financial agreement (Giesecke, 2004). The point of view in credit risk is the default event, which happens when the firm cannot fulfill its legal obligations in accordance with bond contract.

Merton (1974) was the seminal paper which builds a model based on the capital structure of the firm, which becomes the first of the structural model. He assumes that firm is financed by equity and a zero coupon bond with face value *K* and maturity date *T*. In this approach, the company defaults at the bond maturity date *T* if its assets have smaller value than the face value of the bond at time *T*.

Up to this time, most corporation tend to issue risky coupon bond rather than a zero coupon bond. At every coupon date until the final payment, the firms have to pay the coupon. At the maturity date, the bondholder receives the face value of the bond plus the last coupon. The bankruptcy of the firm occurs when the firm fails to pay the coupon at the coupon payment and/or the face value of the bond at the maturity date. Geske (1977) has derived formulas for valuing corporate bonds and suggested that when company has coupon bond outstanding, the common stock and coupon bond can be viewed as a compound option (Geske, 1979). KMV Corporation uses Black & Scholes and Merton methodologies by first converting the debt structure into an equivalent zero coupon bond with maturity one year, for a total promised repayment (Vasicek, 2001). Although KMV claims that its methodology can accommodate different classes of debt, they did not explain on how this equivalent amount is distilled from more complex capital structures (Reisz & Perlich, 2004).

It is of great importance for those in charge of managing risk to understand how financial asset returns are distributed. Practitioners often assume for convenience that the distribution is normal. It is now commonly accepted that financial asset returns are, in fact, heavy-tailed.

However, empirical evidence has led many to reject this assumption in favor of various heavy-tailed alternatives. Heavy tailed performance show the existence of skewness and excess kurtosis performance

In this paper we give general valuation of 2 periods coupon bond with classical default time approach. As a result, it produces some important formulas of credit risk valuation. Those formulas give complete valuation in each coupon date and at the maturity date. To construct the formulas, we use straight forward integration.

2 METHODS

2.1 Geometric Brownian Motion with Jump Diffusion

Dmouj (2006) gives Geometric Brownian Motion model with this equation:

$$X(t) = \left(\mu - \frac{1}{2}\sigma^2\right) + \sigma W(t); t \geq 0 \tag{1}$$

And Geometric Brownian Motion with Jump Diffusion is as this equation:

$$X(t) = \exp\left(\left(\mu - \frac{1}{2}\sigma^2\right) + \sigma W(t)\right)\prod_{j=1}^{n} Y_j \tag{2}$$

For geometric Brownian motion model, the stochastic differential equation follows this equation:

$$dX(t) = f\big(X(t)\big)dt + g\big(X(t)\big)dW(t) \tag{3}$$

For geometric Brownian motion with jump diffusion process, the stochastic differential equation follows this equation:

$$dX(t) = f\big(X(t)\big)dt + g\big(X(t)\big)dW(t) + X(t)dJ(t) \tag{4}$$

With $f\big(X(t)\big)$ is a drift, $g\big(X(t)\big)dW(t)$ is diffusion term, $W(t)$ is Brownian motion, and $J(t)$ is jump process (Brigo *et al*, 2008).

$W(t)$ is standard Brownian motion. $J(t)$ is standard jump process which is defined as:

$$J(t) = \sum_{j=1}^{N_T} \big(Y_j - 1\big) \tag{5}$$

$$dJ(t) = \big(Y_{N_T} - 1\big)dN(t) \tag{6}$$

$N(t)$ is Poisson process with intensity λ and $W(t)$, $N(t)$, $Y(t)$ independent each other. $W(t)$ is Brownian motion, μ and σ are parameter of *X* and *t* (Maruddani & Trimono, 2017).

2.2 Coupon Bond

Let's consider the simple set-up examined by Merton, the model considers a corporation financed through a single debt and single equity issue. The debt comprises a zero coupon bond with notional value *K* maturing at time *T*. There are no payments until time *T*, and equity holders will wait until *T*, before they decide whether default or not. If they defaulted before *T*, they would forgo the chance of benefiting from an increase of the asset value.

To build this model, we have to make some assumptions for simplifying the analytic solution. We considers geometric Brownian

motion with volatility σ and no dividend payments, the equity value can be determined with the standard Black–Scholes call option formula. Further, the model makes all the other simplifying assumptions of the Black-Scholes option pricing formula which are constant return and volatility, no transaction costs, no dividends, no riskless arbitrage, scurity trading is continuous, risk free rate is constant for all maturities, and short selling proceeds is permitted (Maruddani et al, 2015).

T*wo* periods coupon bond is a bond which gives coupon for bondholder twice in its bond period. The classical structural models consider that the default occurs if the value of the assets falls below a critical value associated with the firm's liabilities at maturity date or below the coupon payment at coupon date.

The financing arrangements for making or missing the interest payments are specified in the indenture conditions of the bond. In Fig 1 we illustrate the cash flow of common corporate coupon bond that is 2 periods coupon bond.

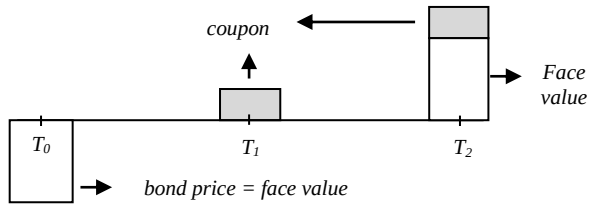


Fig 1: Cash Flow of Multiperiods Coupon Bond

Suppose we look at first coupon date T_1. If the estimated asset value exceeds or equal the coupon payment of the bond, the bondholder will receive their coupon payment, *c*. The firm is not default at time T_1, and the bankruptcy probability for the firm at time T_1, the firm will default if the asset value is not sufficient to pay the coupon payment.

With those assumptions, the equity of bond issuer at coupon date is

$$\begin{aligned}&\xi_{T_0}^{T_1} \\ &= V_0\Phi_2(d_1,u_1,\rho) \\ &- K_1\Phi_2(d_2,u_1,\rho)\exp(-rT_2) \\ &- \mathrm{c}\Phi(u_1)\exp(-rT_1)\end{aligned} \tag{7}$$

Then the equity if bond issuer at maturity date is

$$\xi_{T_1}^{T_2} = V_{T_1}\Phi(d_1) - K_1\Phi(d_2)\exp\big(-r(T_2 - T_1)\big) \tag{8}$$

the default probability of bond issuer at coupon date is

$$P(\tau = T_1) = \Phi(-u_1) \tag{9}$$

the default probability of bond issuer at maturity date is

$$p(\tau = T_2) = 1 - \frac{\Phi_2(u_1, d_3, \rho)}{\Phi(u_1)} \tag{10}$$

with

$$u_1 = \frac{\ln\frac{V_0}{V_1^*} + \left(r - \frac{\sigma^2}{2}\right)T_1}{\sigma\sqrt{T_1}}$$

$$d_1 = \frac{\ln\frac{V_{T_1}}{K_1} + \left(r + \frac{\sigma^2}{2}\right)(T_2 - T_1)}{\sigma\sqrt{(T_2 - T_1)}}$$

$$d_2 = \frac{\ln\frac{V_{T_1}}{K_1} + \left(r - \frac{\sigma^2}{2}\right)(T_2 - T_1)}{\sigma\sqrt{(T_2 - T_1)}}$$

$$d_3 = \frac{\ln\frac{V_0}{K_1} + \left(r - \frac{\sigma^2}{2}\right)T_2}{\sigma\sqrt{T_2}}$$

$$\rho = \sqrt{\frac{T_1}{T_2}}$$

$$K_1 = K + c$$

3 VALUATION OF 2 PERIODS COUPON BOND WITH JUMP DIFFUSION GEOMETRIC BROWNIAN MOTION PROCESSES

In Jump Diffusion Model, changes in the asset price consist of normal (continuous diffusion) component that is modeled by a Brownian motion with drift process and abnormal (discontinuous, i.e. jump) component that is modeled by a compound Poisson process. Asset price jumps are assumed to occur independently and identically. The probability that an asset price jumps during a small time interval can be written using Poisson process (Matsuda, 2004).

Under the assumption of log normality of the jump distribution we analytically solve the valuation problem of an *N*-staged investment opportunity under two different scenarios. Firstly, we consider the case where investment costs are deterministic and perfectly known at the beginning of the project. Secondly, we consider the case where investment costs are stochastic and unknown at the beginning of the project, but where it is known that they follow a jump-diffusion process.

Such investments, including bond, can be modeled commonly as an *N*-stage on the commercialization phase where in each stage the company faces the option of shutting the project down or of continuing its operations, that is, to continue to invest in the project (Andergassen & Sereno, 2010). The arrival of new strategically important information at discrete points in time can be accommodated by modeling the dynamics of the project value as a jump-diffusion process, where the Gaussian diffusion process represents business-as-usual uncertainty and where punctuated jumps at random intervals represent exceptional events such as major project failures or important breakthroughs.

In bond investment, the most common form of debt instrument is a coupon bond. In the U.S and in many other countries, coupon bonds pay coupons every six months and face value at maturity. Suppose the firm has only common stock and coupon bond outstanding. The coupon bond has *2* interest payments of *c* dollars each. The firm is assumed to default at the coupon date, if the total assets value of the firm is not sufficient to pay the coupon payment to bondholder. And at the maturity date, the firm can default if the total asset is below the face value of the bond. For this case, if the firm defaults on a coupon payment, then all subsequent coupon payments (and payments of face value) are also default on.

At every coupon date until the final payment, the firm has the option of paying the coupon or forfeiting the firm to bondholder. The final firm option is to repurchase the claims on the firm from the bondholders by paying off the principal at maturity. The financing arrangements for making or missing the interest payments are specified in the indenture conditions of the bond.

We assume that the firm can neither repurchase shares nor issue new senior debt. The firm only issued single coupon bond that gives *n* times coupon until the maturity date. At the maturity date, firm has to pay coupon *c* and the face value *K* (Maruddani et al., 2015b).

Suppose we look at first coupon date T_1. If the estimated asset value exceeds or equal the coupon payment of the bond, the bondholder will receive their coupon payment, *c*. Consequently, the firm will default if the asset value is not sufficient to pay the coupon payment.

Based on those assumptions, then the equity of bond issuer by geometric Brownian motion with jump diffusion process at coupon date is:

$$\xi_{T_0}^{T_1} = \sum_{i=1}^{N_i}\frac{\exp(-\bar{\lambda}\tau)}{i!}V_{BS}(S_t, \tau = T - t, \sigma_i, r_i) \tag{11}$$

And the equity of bond issuer at maturity date by geometric Brownian motion with jump diffusion process is:

$$\xi_{T_1}^{T_2} = \prod_{j=1}^{N}\left[\sum_{n_j=0}^{\infty}\frac{\exp(-\lambda\tau_j)(\lambda\tau_j)^{n_j}}{n_j!}V_0\exp(-\delta_{s_i}T_1)N_2(a_{s_2}, a_{s_1})\right] \tag{12}$$

$$\prod_{j=1}^{N}\left[\sum_{n_k=0}^{\infty}\frac{\exp(-\lambda\tau_k)(\lambda\tau_k)^{n_k}}{n_k!}\right.$$

$$\left. I_j \exp(-rT_1)\, N_1\left(b_{s_N}, b_{s_j}\right)\right]$$

With

$$\delta_{s_k} = \frac{s_k\left(\mu_J + \frac{1}{2}{\sigma_J}^2\right)}{T_k} + \lambda\left[\exp\left(\mu_J + \frac{1}{2}{\sigma_J}^2\right) - 1\right]$$

$$b_{s_k} = \frac{\ln\left(\frac{V_0}{V_0^*}\right) + \left(r - \delta_{s_k} - \frac{1}{2}{\sigma_{s_k}}^2\right)T_k}{\sigma_{s_k}\sqrt{T_k}}$$

$$a_{s_k} = b_{s_k} + \sigma_{s_k}\sqrt{T_k}$$

ACKNOWLEDGEMENTS

This research is officially funded by The Ministry of Research, Technology, and Higher Education Indonesia by grant of PTUPT in 2018 with contract number 101-112/UN7.P4.3/PP/2018.

REFERENCES

Andergassen, R., & Sereno, L. (2010). Valuation of N-investment under jump diffusion. *Working paper University of Pi*sa, Italy

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of political economy*, 81(3), 637-654

Brigo, D., Dalessandrom, A., Neugebauer, M., & Triki, F. (2008). A stochastic processes toolkit for risk management, *Journal of risk management in financial institutions*, *1*(4), 5-13.

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues, *Quantitative finance*, 1, 223-236.

Dmouj, A. (2006). Stock price modeling: theory and practice. *Amsterdam Faculty of Science*. BMI Paper.

Geske, R. (1977). The Valuation of corporate liabilities as compound options, *Journal of Financial and Quantitative Analysis*, 7, 63-81.

Giesecke, K. (2004). Credit Risk: Models and Management 2nd ed. *Risk books*. London.

Kellezi, E. & Gilli, M. (2000). Extreme value theory for tail-related risk measures. *Preprint* University of Geneva.

Mandelbrot, B.B. (1963). The variation of certain speculative prices. *Journal of business*, 36, 392-417.

Maruddani, D.A.I., Rosadi, D., Gunardi, G., & Abdurakhman, A. (2015). Valuation of one period coupon bond valuation based on default time and empirical study in Indonesian bond data, *Far east journal of mathematical sciences*, *98*(1), 57-73.

Marudani, D.A.I., & Trimono, T. (2017). Stock price prediction of PT Astra Argo Lestari Tbk. with jump diffusion model. *Jurnal riset akuntansi mercu buana*, *3*(1), 57-67.

Merton, R. (1974). On the pricing of corporate debt: the risk structure of interest rate. *Journal of finance*, *29*, 449–470.

Rocco, M. (2011). Extreme value theory for finance: a survey, occasional paper. *Working paper*, Italia. https://www.bancaditalia.it/pubblicazioni/qef/2011-0099/QEF_99.pdf

Tang, Y. 2005. Essay on Credit Risk, *Faculty of the Graduate School*. University of Texas, Austin.

Trimono, T., & Maruddani, D. A. I. (2017). Valuasi harga saham PT Aneka Tambang Tbk sebagai peraih IDX best blue 2016. *STATISTIKA: Journal of theoretical statistics and its application*, *17*(1), 33-43.

Complicated Grief and Posttraumatic Stress Disorder in Bereaved Widows from the Civil Unrest in Thailand's Deep South

Wattana Prohmpetch[1*], Phattrawan Tongkumchum[2], Don McNeil[3], Nittaya McNeil[4], and Sayaporn Detdee[5]

[1]Department of Psychology and Guidance/Faculty of Education/Prince of Songkla University, Pattani Campus, Thailand
*Corresponding Email: wattana.ph@psu.ac.th
[2,3,4]Department of Mathematics and Computer Sciences/Faculty of Science and Technology/ Prince of Songkla University, Pattani Campus, Thailand
[2]Email: Phattrawan.t@psu.ac.th
[5]Songkhla Rajanagarindra Psychiatric Hospital, Songkhla, Thailand
[5]Email: ssayaa4@gmail.com

ABSTRACT

This cross-sectional study aimed to investigate the prevalence and predictors both complicated grief (CG) and posttraumatic stress disorder (PTSD). These are mental disorder which is a severe stress of widows losing their husband from the civil unrest in Thailand's deep south. The sample was 156 volunteered widows who wanted to participate Resilience Enhancing Program of Songkhla Rajanagarindra Psychitric Hospital in 2016**.** CG was assessed with Inventory of Complicated Grief (ICG), PTSD was assessed with PTSD Symptom Scale-Self Report (PSS-I), and other determinants were obtained by using questionnaire. Of the total sample, 74.4% (n=116) were classified as CG and 82.7% (n=129) fulfilled the criteria of PTSD. Age, residence province, relationship with neighbors and resilience were significantly predicted CG score, whereas residence provinces and relationship with relatives significantly predicted PTSD score. Even the same as bereaved widows, there were different predictors between CG and PTSD. Therefore, to effectively prevent, providing treatment, and rehabilitate bereavement-related distress both CG and PTSD of widows losing their husband, related determinants should be considered.

Keywords: civil unrest; complicated grief; bereaved widows; posttraumatic stress disorder

1 INTRODUCTION

Bereavement event is a period of mourning after the death of loved one. Bereavement victims can meet posttraumatic stress disorder (PTSD) and complicated grief (CG). PTSD is a mental disorder that can develop after one month person is exposed to a traumatic event or bereavement event. Criteria of four symptoms are re-experiencing, avoiding situation or trigger memories of the traumatic event, negative beliefs and feelings and hyper arousal. These can lead to a significant disturbance and impairment in social, occupational, or other important areas of functioning (American Psychiatric Association, 2013). CG is also mental disorder that consists of symptoms at least 6 months after the loss of a loved one including a sense of disbelief regarding the death; persistent intense longing, yearning, and preoccupation with the deceased; recurrent intrusive images of the dying person; and avoidance of painful remainders of the death (Simon et al., 2007).

According Lichtenthal et al. (2004) mentioned that "before 2001 CG was claimed as "PTSD like" because CG symptoms were referred to as traumatic grief which reflects symptoms of both separation distress and traumatic distress. Moreover, similar to symptoms of PTSD, symptoms of CG may be effectively treated with selective serotonin reuptake inhibitors. However, it was argued that sharing common features does not permit equating CG with other diagnostic entities like PTSD. Etiology, course, prognosis, and treatment must all be considered. Others have also asserted that PTSD and CG are not isomorphic. Although the traumatic distress symptoms of CG appear to resemble some of the symptoms of PTSD, the separation distress component is unique"

Recent research has investigated prevalence and determinants which predict CG and PTSD in bereaved widows losing their husband from the civil unrest in Thailand's deep south. The civil unrest is violent situation in southernmost provinces of Thailand which has occurred in four border provinces: Pattani, Yala, Narathiwat, and some districts in Songkhla since 2004. Until 2013, there are 12,549 civil unrests such as shooting, arson attacking, or bombing and other acts of violence over decade, with 9,694 injuries and 5,473 deaths. This civil unrest has had bereavement victims, with 2,450 women who became bereaved widows (Tohmeena, 2013).

Hence, this objective research is not only to examine prevalence and predictors of CG and PTSD, but also to be evidence for discussing argument that distinction between CG and PTSD.

2 METHODS

2.1 Participants and procedure

This cross-sectional study aimed to investigate the prevalence and predictors of PTSD and CG focusing on bereaved widows who have lost their husband and living under civil unrest in Thailand's deep south, with 2,450 bereaved widows (Tohmeena, 2013). Target group was widows who had name lists of Deep South Coordination Center in Pattani, Yala and Narwthiwat. In May 2015, we invited those 200 widows to recruit as part of a program of research that was designed to study the effectiveness of resilience enhancing program (Detdee et al., 2018). All 156 participants, were willing and voluntary, who had been interviewed with questionnaires by four psychologists and ten psychology students who were trained to well use the questionnaire and interview.

2.2 Measures

The participants were interviewed with questionnaires comprise three sets as flows: **1) demographic questionnaires** was conducted by researchers that consist of age, residence, religion, marital status, level of education, occupation, income, congenital disease, trauma event occur, directly facing or witnessing bereavement event, received compensation from government, employed by government project, allocated to husband's relative, saving, debt, perception in quality of life, relationship with family member, relative and neighbor or co-worker. **2) Thai resilience quotient screening test** by Department of Mental Health (2009) is a rating scale of 20 items to measure 3 factors of resilience: security and emotion (bear), motivation (resolve), and coping (fight). This rating test has 4 levels (not true=1 and always true=4). The overall score range from 0 to 80. The score was grouped as: lower 55 (low) 30-40 (very low) and lower 30 (extremely low). **3) inventory of complicated grief** (ICG) (Prigerson *et al*., 1995) was used to measure the symptom of complicated grief comprising 19 items. This ICG rating scales have five levels: never, rarely, sometime, often, and always. Respondents with ICG score more than 25 were considered as having complications of bereavement (Prigerson *et al*., 1995). **4) PTSD Symptom Scale-Self Report** (PSS-I) was from Fao and Tolin (2000). This 17-items semi-structured interview assesses the presence and severity of the Diagnostic and Statistical Manual of Mental Disorders-IV (DSM-IV) PTSD symptoms. Rating to reflect a combination of frequency and severity (from 0= "not at all" to 3= "5 or more time per

week/very much") yielding a slightly higher coefficient and sensitivity to PTSD. PSS-I more than 15 score is a cut of point.

2.3 Statistical analysis

Two sample t test and ANOVA were used to investigate relation between outcomes (CG or PTSD) and demographic factors. Multiple linear regression was used to model between outcomes (CG or PTSD) and demographic factors. The model was separately fitted to the data for CG and PTSD using both treatment and sum contrasts. All data analysis was conducted using R Programing (R Core Team, 2015).

3 RESULTS

3.1 The distributions of CG and PTSD

Figure 1 shows the distributions of CG and PTSD. The shapes of the distributions are symmetric.

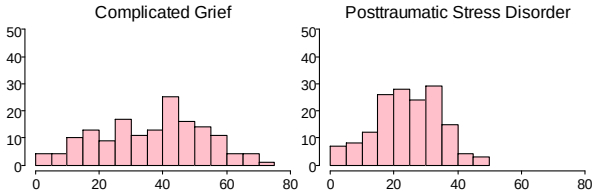


Figure 1: Histograms of CG and PTSD

3.2 The prevalence of CG and PTSD

The prevalence of CG was 74.4% (n=116), with mean 37.2 and standard deviation was 16.1 whereas prevalence of PTSD was 82.7% (n=129), with mean 24.6 and standard deviation was 10.0.

3.3 The characteristics of study sample

Table 1 shows characteristics of study sample based on demographic factors. The number and percent of each factor level is shown which were described the following determinants.

Table 1: Number (n) and percentage (%) of determinants (N=156)

Determinants	n	%
Age		
<40	45	28.85
40-49	61	39.10
50+	50	32.05
Residence		
Pattani	62	39.74
Yala	70	44.87
Narathiwat	24	15.38
Religion		
Islamic	128	82.05
Buddhist	28	17.95
Marital status		
Widow	128	82.05
New married	28	17.95
Level of education		
Uneducated	19	12.18
Primary	85	54.49
Secondary	35	22.44
Diploma& bachelor	17	10.89
Occupation		
Employed	33	21.15
Agriculture	46	29.49
Employee of Private or Gov.	58	37.18
Private business and contractor	19	12.18
Incomes		
<4,500	88	56.41
4,500-5,000	26	16.67
5,000+	42	26.92
Congenital disease		
Yes*	64	41.03
No	92	58.97
Trauma event occur		
2012-2015	45	28.85
2008-2011	49	31.41
2004-2007	62	39.74
Facing or witnessing bereavement event		
Directly	31	19.87
Indirectly	125	80.13
Compensation		
Yes	143	92.86
In process	11	7.14
Compensation budget		
<300,0000	51	32.69
300,000–499,999	88	56.41
500,000+	17	10.90
Enployed by government		
Yes	128	82.05
No	28	17.95
Allocated to husband's relative		
Yes	94	60.26
No	62	39.74
Saving		
Yes	41	26.28
No	115	73.72
Debt		
Yes	125	80.13
No	31	19.87
Quality of life		
Low	114	73.08
Normal	42	26.92
Relationship with family		
Positive	106	67.95
Negative	50	32.05
Relationship with relatives		
Positive	104	66.67
Negative	52	33.33
Relationship with neighbor & co-worker		
Positive	108	69.23
Negative	48	30.77
Resilience Quotient		
Low	36	23.08
Very low	39	25.00
Extremely low	81	51.92

*Diseases: blood pressure (38.7%), diabetes (12.9%), gastritis (9.7%), heart disease (6.5%), allergy (6.5%), and other diseases such as pain, and wound injury (25.8%).

3.4 Multiple linear regression predicting CG and PTSD

Table 2 and Figure 2 show that the result of multiple linear regressions indicated four predictors: age, residence, relationship with neighbor and co-worker and resilience can predict CG in bereaved widows.

Meanwhile in Table 3 and Figure 3 show that only two predictors: residence and relationship with relatives can predict PTSD.

Table 2: Multiple linear regression predicting CG using treatment contrasts

Predictors	Coef.	SE	p-value
Constant	26.96	7.23	<0.001
Age:			
<40	0		
40-49	10.80	7.15	0.127
50+	17.86	7.08	0.012

Predictors	Coef.	SE	p-value
Residence:			
Pattani	0		
Yala	-5.74	2.69	0.034
Narathiwat	3.15	3.76	0.404
Relationship with neighbor& co-worker:			
Positive	0		
Negative	7.56	2.71	0.005
Resilience:			
Low	0		
Very low	7.90	7.51	0.295
Extremely low	-7.93	3.29	0.017

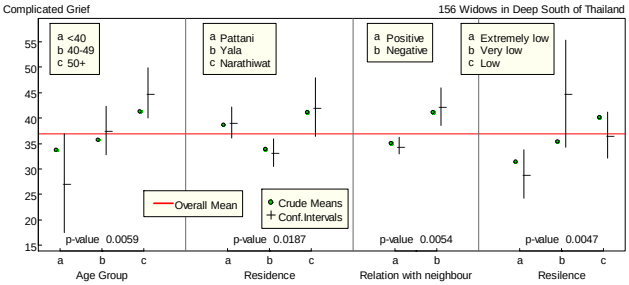


Figure 2: Multiple linear regression predicting CG using sum contrasts

Table 3: Multiple linear regression predicting PTSD using treatment contrasts

Predictors	Coef.	SE	p-value
Constant	23.84	1.31	0.000
Residence			
Pattani	0		
Yala	-3.22	1.74	0.034
Narathiwat	3.15	3.76	0.404
Relationship with relatives			
Positive	0		
Negative	7.56	2.71	0.005

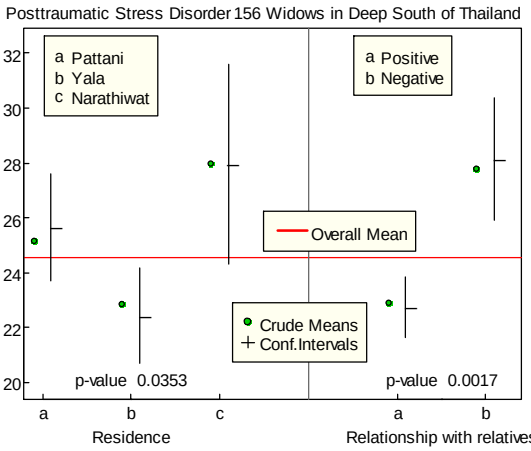


Figure 3: Multiple linear regression predicting PTSD using sum contrasts

DISCUSSIONS

The prevalence of PTSD (82.7%) in bereaved widows losing their husband from the civil unrest in Thailand's deep south had higher than prevalence of CG (74.4%). Compare to lifetime prevalence of PTSD (6.8%) in Thailand (Tantirangsee, 2018), PTSD (24.2%) at least one trauma experience in German (Eichhorn et al., 2014), CG (49.4%) in patients (Simon et al., 2007), both PTSD and CG were very high prevalence in bereaved widows. Because of this target was focusing on bereaved windows and the context where diversity has. As Tantirangsee et al. (2017) found that the highest of rate to relate with PTSD is facing extremely traumatic experience and see someone being badly or killed.

Even though CG and PTSD had the same determinant which was residence, could predict both CG and PTSD particularly in Yala. Because of in Yala had the highest percent of CG and PTSD. However, only age, relationship with neighbor or co-worker, and resilience could predict CG meanwhile relationship with relatives could predict only PTSD in bereaved widows. As Tantirangsee et al. (2017) reported PTSD in Thailand related from extremely traumatic experiences, domestic violence, and see someone being badly or killed. Even though Simon e al. (2007) found that PTSD conditions share symptoms with CG but most predictors were different between CG and PTSD. This result can support argument that sharing common features does not permit equating CG with other diagnostic entities like PTSD, therefore, etiology, course, prognosis, and treatment must all be considered.

ACKNOWLEDGEMENTS

The authors would like to thank Asst. Prof. Dr.Metta Kuning and Staff of Deep South Coordination Center in Pattani for supporting data and helping contact the subjects. This study is funded by Songkhla Rajanagarindra Psychiatric Hospital.

REFERENCES

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed., Washington, DC., U.S.A., pp. 271-290.

Deep South Incident Database. (2017). *Thirteen Years of Conflict in Southernmost Provinces. Center for Conflict Studies and Cultural Diversity (CSCD)*, Prince of Songkla University, Pattani Campus. Available online: ps://www.deepsouthwatch. org/node/11053 [January 25, 2018].

Department of Mental Health. (2009). *Manual for Mental Health Strength in Employee Crisis Program*. Nonthaburi: The Agricultural Cooperative Federation of Thailand Limited.

Detdee, S., Prohmpetch, W., & Tantirangsee, N. (2018). The effectiveness of resilience enhancing program on resilience level and depression among widows affected by south Thailand insurgency. *Journal of Mental Health of Thailand*, *26*(2), 103-116.

Eichhorn, S., Brähler, E., Franz, M., Friedrich, M., & Glaesmer, H. (2014). Traumatic experiences, alexithymia, and posttraumatic symptomatology: a cross-sectional population-based study in Germany. *European Journal of Psychotraumatology*, 5(1), 23870.

Foa, E.B., & Tolin, D.F. (2000). Comparison of the PTSD symptom scale–interview version and the clinician- administered PTSD scale. *Journal of Traumatic Stress*, 13(2), 181-191.

Lichtenthal, W.G., Cruess, D.G., & Prigerson, H.G. (2004). A case for establishing complicated grief as a distinct mental disorder in DSM-V. *Clinical Psychology Review*, 24,637-662.

Prigerson, H.G., Maciejewski, P.K., Reynolds, C.F., Bierhals, A.J., Newsom, J.T., Fasiczka, A., Frank, E., Doman, J., & Miller, M. (1995). Inventory of complicated grief: a scale to measure maladaptive symptoms of loss. *Psychiatry Research*, 59(1), 65-79.

R Core Team. (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available online: http://www.R-project.org/. [June 16, 2018]

Silverman, G.K., Johnson, J.G., & Prigerson, H.G. (2001). Preliminary explorations of the effects of prior trauma and loss on risk for psychiatric disorders in recently widowed people. *The Israel Journal of Psychiatry and Related Sciences*, *38*(3/4), 202.

Simon, N.M., Shear, K.M., Thompson, E.H., Zalta, A.K., Perlman, C., Reynolds, C.F., Frank, E., Melhem, N.M., & Silowash, R. (2007). The prevalence and correlates of psychiatric comorbidity in individuals with complicated grief. *Comprehensive Psychiatry*, 48(5), 395-399.

Tantirangsee, N., Makwichit, K., Likhasithdamrongkul, W., Boonratt-ana, A., & Lertkiatratchata, M. (2017). Factors related with post-traumatic stress disorder: Thai national mental health survey 2013. *Journal of Mental Health of Thailand*, *25*(2), 122-135.

Tohmeena P. (2013). Mental remedy for people affected by unrest in the southernmost provinces of Thailand. *Journal of Mental Health of Thailand*, 21(3), 171-84.

Estimating the Population Coefficient of Quartile Variation for Bootstrap Confidence Intervals

Pot Somboon* and Tidadeaw Mayureesawan

[1]Department of Statistics, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand
*Corresponding Email: potsomboon@gmail.com
Email: tidadeaw@kku.ac.th

ABSTRACT

This research aims to study bootstrap confidence intervals for Coefficient of Quartile Variation (CQV) in 2 methods namely, Nonparametric Bootstrap (NB) and Parametric Bootstrap (PB) method when estimate the population coefficient of quartile variation from 5 quartiles method, there are Tukey, Moore and McCabe (M&M), Mendenhall and Sincich (M&S), Freund and Perles (F&P) and Minitab. The data case have Normal(4, 1), Lognormal(0, 1.5), Lognormal(0, 1), Lognormal(0, 0.75), Lognormal(0, 0.5) and Lognormal(0, 0.25) distribution and the sample sizes (n) are 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 19, 20, 25, 26, 29, 30, 35, 36, 39, 40, 45, 46, 49 and 50. The consideration performance of confidence interval is considered from coverage probability (CP) at confidence level of 0.95 and the average width of confidence interval (AW) by using Monte Carlo simulation with R program. The result found that confidence interval of NB method when estimate the population coefficient of quartile variation by F&P quartile method performs better than the other methods when the distributions are Normal(4, 1), Lognormal(0, 1), Lognormal(0, 0.75) and Lognormal(0, 0.25) distribution in various sizes of sample. For the confidence interval of PB method, it can confirm that the F&P method for estimate quartile is more effective than the other methods for Normal(4, 1), Lognormal(0, 0.5) and Lognormal(0, 0.25) distribution in many levels of n. And the performance of PB confidence interval based on Minitab quartile method has higher efficiency than the other methods for Lognormal(0, 1.5), Lognormal(0, 1) and Lognormal(0, 0.75) distribution in various sizes of sample.

Keywords: bootstrap confidence interval, coefficient of quartile variation, coverage probability

1. INTRODUCTION

Relative dispersion is the measure of dispersion calculated from the ratio between absolute variation and the average. Relative dispersion used to compare the variation of data starting from 2. When the data has the average or the difference unit of data, Coefficient of Quartile Variation (CQV) is the appropriate measure of dispersion for the data with non-normal distribution and avoids one-tailed or two-tailed outliers data. CQV can be calculated from:

$$CQV = \frac{Q_3 - Q_1}{Q_3 + Q_1} \quad (1)$$

When Q_1 and Q_3 are Quartile 1 and Quartile 3, respectively, from the CQV formula, $Q_3 - Q_1$ means Interquartile Range (IQR). Several researchers have developed confidence intervals for coefficient of quartile variation as follows:

Bonett (2006) presented the 95% confidence interval for coefficient of quartile variation by using variance of difference and studied the confidence interval for the case of normal distribution and non-normal distribution with several sample sizes, it was found that the confidence interval will be efficient when the sample size is small; this is for the case of non-normal distribution.

Tongkaw and Pongsakchat (2014) indicated confidence intervals for coefficient of quartile variation in 2 methods namely, Bootstrap and Bonett Bootstrap by adjusting the Bonette method and applying resampling technique before comparing the efficiency of the two confidence intervals with Bonett method. The data case has Normal(4, 1), Lognormal(0, 1) and Gamma (1.5, 1). It was discovered that Bootstrap confidence internval and Bonett Bootstrap confidence interval have better efficiency than Bonett method when the sample size of every distribution is small.

Altunkaynak and Gamgam (2017) demonstrated confidence intervals for coefficient of quartile variation in 2 methods namely, Nonparametric Bootstrap and Parametric Bootstrap, using resampling technique of bootstrap. They found that the confidence intervals of the two methods will have better efficiency than Bonett method when the distribution of data has the small sample size.

Langford (2006) studied the estimation quartile from 5 quartiles methods, there are Tukey, Moore and McCabe (M&M), Mendenhall and Sincich (M&S), Freund and Perles (F&P) and Minitab towards IQR. As a result of the estimation, it affects IQR to contain the differences due to each estimate Q_1 and Q_3 differently. However, it depends on the size of the samples.

Cangur, Pasin, and Ankarali (2015) studied the estimation of quartiles through Microsoft Excel (Freund and Perles) and Minitab for the distribtions of Standard Normal, Chi Square and small extreme value. It was found that the esimation of quartiles via Microsoft Excel is best for asymmetrical distributions while the esimation of quartiles via Minitab is best for symmetrical distributions.

According to the estimation of quartiles with these methods, it was revealed that each provides different results which depend on the size of the sample and the type of data distribution that affect the result of CQV in the equation (1) differently. This research interests in studying the bootstrap confidence intervals for coefficient of quartile variation in 2 methods namely, Nonparametric Bootstrap (NB) and Parametric Bootstrap (PB) methods when estimate the population coefficient of quartile variation from 5 quartiles method, there are Tukey, Moore and McCabe (M&M), Mendenhall and Sincich (M&S), Freund and Perles (F&P) and Minitab.

2. LITERATURE REVIEWS

2.1 Quartile calculations

Cangur, Pasin, and Ankarali (2015) stated the method to find the position of quartile for unclassified data ($x_1, x_2, ..., x_n$) which consists of Quartile 1 (Q_1), Quartile 2 (Q_2) and Quartile 3 (Q_3) as follows:

1. Tukey method

Tukey method has the step of searching Q_1 and Q_3 as follows:

- Sort data, from least to greatest ($x_{(1)}, x_{(2)}, ..., x_{(n)}$).
- Search Q_2 with this formula $(n+1)/2$.
- Search the median of the lower data from data 1 ($x_{(1)}$) to Q_2. This position is Q_1.
- Search the median of the top data from Q_2 to data n ($x_{(n)}$). This position is Q_3.

2. Moore and McCabe method (M&M method)

M&M method can be considered into 2 cases as follows:

- When the total number of data is the odd number, the position of Q_1 and Q_3 are as follows:

$$Q_1 = \frac{(n+1)}{4}, \; Q_3 = \frac{(3n+3)}{4}.$$

- When the total number of data is the even number, the position of Q_1 and Q_3 as follows:

$$Q_1 = \frac{(n+2)}{4}, \; Q_3 = \frac{(3n+2)}{4}.$$

3. Mendenhall and Sincich method (M&S method)

M&S method is to find Q_1 and Q_3 as follows:

$$Q_1 = \frac{(n+1)}{4}, \; Q_3 = \frac{(3n+2)}{4}.$$

4. Freund and Perles method (F&P method)

F&P method is to find the quartiles via Microsoft Excel; the formula to find Q_1 and Q_3 as follows:

$$Q_1 = \frac{(n+3)}{4}, \; Q_3 = \frac{(3n+1)}{4}.$$

5. Minitab method

Minitab is a popular method to find Q_1 and Q_3 as follows:

$$Q_1 = \frac{(n+1)}{4}, \; Q_3 = \frac{3(n+1)}{4}.$$

2.2 Bootstrap method

Bootstrap method is the resampling with replacement, by means of selecting the sample size n which are $x_1, x_2, ..., x_n$ that are independent of each other, from the populations distributed by many types. θ is the parameter and $\hat{\theta}_B$ is the estimated parameter. Perform the sampling once for n times, selecting from the sample group $x_1, x_2, ..., x_n$. The received result will be added to the previous sample for the next sampling. $x_1^*, x_2^*, ..., x_n^*$ is the n size sample group received from $x_1, x_2, ..., x_n$, called Bootstrap sample. Estimate θ and then receive the estimated B; there are $\hat{\theta}_1^*, \hat{\theta}_2^*, \; ... \; , \hat{\theta}_B^*$. Repeat the process according to B. This method was presented by Efron and Tibshirani (1993).

2.3 Confidence intervals for coefficient of quartile variation using Nonparametric Bootstrap (NB method)

Set $\hat{Q}_{1i}^*$ and $\hat{Q}_{3i}^*$ as the estimated quartile 1 and the quartile 3, following whit bootstrap method, then receive Coefficient of Quartile Variation of Bootstrap (CQV_i^*) as follows:

$$CQV_i^* = \frac{\hat{Q}_{3i}^* - \hat{Q}_{1i}^*}{\hat{Q}_{3i}^* + \hat{Q}_{1i}^*} \quad (2)$$

Coefficient of Quartile Variation of Bootstrap (CQV_i^*) are

$$CQV_1^*, CQV_2^*, \; ... \; , CQV_B^*.$$

Sort CQV_i^*, from least to greatest represented by

$$CQV_{(1)}^*, CQV_{(2)}^*, \; ... \; , CQV_{(B)}^*.$$

When B is round of resampling.

The confidence interval for coefficient of quartile variation using Nonparametric Bootstrap applies percentile bootstrap technqiue as follows:

$$\left[CQV_{(BL)}^*, CQV_{(BU)}^* \right] \quad (3)$$

When $CQV_{(BL)}^*$ and $CQV_{(BU)}^*$ becomes the percentile at $(\alpha/2)B$ and the percentile at $(1-\alpha/2)B$ or the lower class limit and the upper class limit of the confidence intervals for coefficient of quartile variation using Nonparametric Bootstrap, respectively.

2.4 Confidence intervals for coefficient of quartile variation using Parametric Bootstrap (PB method)

Set M_{CQV_B}, $S_{CQV_B}^2$ and T_i^* as the point estimation, the variation of the point estimation, and Bootstrap – t for coefficient of quartile variation of parametric bootstrap, respectively, as follows:

The point estimation for coefficient of quartile variation of Parametric Bootstrap is

$$M_{CQV_B} = \frac{\sum_{i=1}^{B} CQV_i^*}{B} \quad ; \; i = 1, 2, ..., B \quad (4)$$

The variance of the point estimation for coefficient of quartile variation of Parametric Bootstrap is

$$S_{CQV_B}^2 = \frac{\sum_{i=1}^{B} \left(CQV_i^* - M_{CQV_B} \right)^2}{B} \quad ; \; i = 1, 2, ..., B \quad (5)$$

The Bootstrap - t for coefficient of quartile variation of Parametric Bootstrap is

$$T_i^* = \frac{\left(CQV_i^* - M_{CQV_B} \right)}{S_{CQV_B}} \quad ; \; i = 1, 2, ..., B \quad (6)$$

The confidence interval for coefficient of quartile variation of Parametric Bootstrap can be calculated as follows:

$$\left[M_{CQV_B} - T_{(1-\alpha/2)B}^* S_{CQV_B}, \; M_{CQV_B} + T_{(1-\alpha/2)B}^* S_{CQV_B} \right] \quad (7)$$

When $T_{(1-\alpha/2)B}^*$ is T_i^* sort from least to greatest represented by $T_{(1)}^*, T_{(2)}^*, \; ... \; , T_{(B)}^*$ thus $T_{(1-\alpha/2)B}^*$ is $T_{(i)}^*$ at $(1-\alpha/2)B$ of Bootstrap - t

3. METHODS

In this topic, it states the scope and the methodology of the research with details as follows:

3.1 Scope of research

1. The sample sizes (n) are 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 19, 20, 25, 26, 29, 30, 35, 36, 39, 40, 45, 46, 49 and 50.

2. Regarding case study of the symmetrical distributions, Normal(4,1) and the positively skew distribution include Lognormal(0, 1.5), Lognormal(0, 1), Lognormal(0, 0.75), Lognormal(0, 0.5) and Lognormal(0, 0.25).

3. Set the confidence level $(1-\alpha)$ at 0.95

4. Specify the resampling with bootstrap method for 1,000 times.

5. Simulate the different events for 5,000 iterations with R program.

6. Analyze the efficiency of the bootstrap confidence interval regarding Coverage Probability (CP) with reliability at 0.95. If CP of the confidence interval is ranged in [0.9444, 0.9561] and Average Width of Confidence Interval (AW) refers to the narrowest, it is regarded as the best efficient confidence internval.

3.2 Processes

1. Create the data of population in accordance with the distribution as regulated at 5,000 observed values.
2. Assess CQV of the populations.
3. Do the sampling of the sample size n from data in No. 1 and examine the distribution data if it corresponds to the setting.
4. Calculate the confidence interval at 95% for coefficient of quartile variation by NB and PB according to the equation (3) and (7), respectively.
5. Examine the confidence interval at 95% for coefficient of quartile variation if it covers coefficient of quartile variation of the populations or not, including width of confidence interval.
6. Redo the process in No. 3 to No. 5 for 5,000 iterations.
7. Calculate CP and AW.
8. Redo No.1 to No. 7 for every event that has been set under the scope of the research.

4. RESULTS

From Table 1 assessment of quartile estimate method (QEM) in bootstrap confidence interval (BCI), there is the efficiency of distribution and the sample sizes (n). It was found that the Nonparametric Bootstrap (NB method) confidence interval has the efficiency when estimating coefficient of quartile variation of the populations by Freund and Perles method (F&P method) for the case of these distributions: Normal(4, 1) at $9 \le n \le 20$, Lognormal(0, 0.75) at $9 \le n \le 15$, Lognormal(0, 1) at $10 \le n \le 19$ and Lognormal(0, 0.25) at $10 \le n \le 12$. For NB method when estimating coefficient of quartile variation by Tukey method there is the efficiency for the case of these distributions and odd number of n are Normal(4, 1) and Lognormal(0, 1) at $9 \le n \le 19$ and Lognormal(0, 0.75) at $9 \le n \le 15$.

For Parametric Bootstrap (PB method) confidence interval, it will have the efficiency when estimating coefficient of quartile variation of the populations by Freund and Perles (F&P method) and Minitab methods. For F&P method, there is the efficiency when the distributions are Normal(4, 1) at $10 \le n \le 20$, Lognormal(0, 0.5) at $10 \le n \le 15$ and Lognormal(0, 0.25) at $10 \le n \le 12$. For Minitab method, there is the efficiency when the distributions are Lognormal(0, 1.5) $29 \le n \le 50$, Lognormal(0, 1) at $11 \le n \le 49$ and Lognormal(0, 0.75) at $7 \le n \le 13$. For PB method has the efficiency for odd number of n when estimating coefficient of quartile variation by Tukey method and Moore and McCabe method (M&M method). For Tukey and M&M method there are the efficiency for the case of these distributions: Normal(4, 1) Lognormal(0, 0.5) and Lognormal(0, 1) at $11 \le n \le 15$. And the M&M method there is the efficiency when the distribution is Lognormal(0, 1.5) at $13 \le n \le 49$.

Table 1 Quartile estimate method (QEM) that makes the bootstrap confidence interval (BCI) having the efficiency in each distribution and sample size (n) with the confidence level at 0.95.

BCI	QEM	Ditribution	n
NB	Tukey	Normal(4, 1) and Lognormal(0, 1)	$9 \le n \le 19$; for odd number of n
		Lognormal(0, 0.75)	$9 \le n \le 15$; for odd number of n
	M&M	-	-
	M&S	Lognormal(0, 1), Lognormal(0, 0.75) and Lognormal(0, 0.5)	7
	F&P	Normal(4, 1)	$9 \le n \le 20$
		Lognormal(0, 0.75)	$9 \le n \le 15$
		Lognormal(0, 1)	$10 \le n \le 19$
		Lognormal(0, 0.25)	$10 \le n \le 12$
	Minitab	-	-
PB	Tukey	Normal(4, 1) and Lognormal(0, 0.5)	$11 \le n \le 15$; for odd number of n
	M&M	Lognormal(0, 1.5)	$13 \le n \le 49$; for odd number of n
		Lognormal(0, 1)	$11 \le n \le 15$; for odd number of n
	M&S	Normal(4, 1) and Lognormal(0, 0.5)	7 and 8
		Lognormal(0, 0.75)	7
	F&P	Normal(4, 1)	$10 \le n \le 20$
		Lognormal(0, 0.5)	$10 \le n \le 15$
		Lognormal(0, 0.25)	$10 \le n \le 12$
	Minitab	Lognormal(0, 1.5)	$29 \le n \le 50$
		Lognormal(0, 1)	$11 \le n \le 49$
		Lognormal(0, 0.75)	$7 \le n \le 13$

5. CONCLUSIONS

Regarding the Nonparametric Bootstrap confidence interval at 95%, there is the efficiency when estimating coefficient of quartile variation of the populations by Freund and Perles; the distributions are Normal(4, 1), Lognormal(0, 0.1), Lognormal(0, 0.75), and Lognormal(0, 0.25). And the Nonparametric Bootstrap confidence interval at 95%, there is the efficiency when estimating coefficient of quartile variation by Tukey method for the case of these distributions and odd number of n are Normal(4, 1), Lognormal(0, 1) and Lognormal(0, 0.75). Additionally, it depends on the sample size and the distribution.

Regarding the Parametric Bootstrap confidence interval at 95%, there is the efficiency when estimating coefficient of quartile variation of the populations by Freund and Perles and Minitab. For the Freund and Perles method, the efficiency of distributions are Normal (4, 1), Lognormal(0, 0.5), Lognormal(0, 0.5) and Lognormal(0, 0.25). For the Minitab method, the efficiency of distrubtions are Lognormal(0, 1.5), Lognormal(0, 1) and Lognormal(0, 0.75). And the Parametric Bootstrap confidence interval at 95%, there is the efficiency for odd number of sample sizes when estimating coefficient of quartile by Tukey method and Moore and McCabe method. For Tukey and Moore and McCabe method there are the efficiency for the case of these distributions: Normal(4, 1) Lognormal(0, 0.5) and Lognormal(0, 1). For the Moore and McCabe method there is the efficiency when the

distribution is Lognormal(0, 1.5). Additionally, it depends on the sample size and the distribution.

REFERENCES

Altunkaynak, B., & Gamgam, H. (2017). Boostrap confidence intervals for the coefficient of quartile variation. *Communications in Statistics–Simulation and Computation*. doi:10.1080/03610918.2018.1435800

Bonett, D.G. (2006). Confidence interval for a coefficient of quartile variation. *Computational Statistics & DataAnalysis*, *50*(11), 2953-2957.

Cangur, S., Pasin, O., & Ankarali, H. (2015). Comparison of sampling distributions and performances of minitab and freund & perles Quartile. *Pakistan Journal of Statistics*, *31*(1), 1-20.

Efron, B., & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. U.K.: Chapman & Hall.

Langford, E. (2006). Quartiles in Elementary Statistics. *Journal of Statistics Education*, *14*(3). doi:10.1080/10691898.2006.11910589

Tongkaw, A., & Pongsakchat, V. (2014). Confidence intervals for a coefficient of quartile variation with bootstrap method. *ICAS 2014*, Khon Kaen, 19-21.

A Comparison of Regression and Artificial Neural Network Models for Predicting Thai Gold Bullion Price in Thailand

Passadee Manketkorn[1], Jaratsri Rungrattanaubol[1], Nupian Thepmong[2] and Anamai Na-udom[2*]

[1]Naresuan University, Faculty of Science, Department of Computer Science and IT, Phitsanulok, Thailand
Email: passadeem56@email.nu.ac.th
Email: jaratsrir@nu.ac.th
[2]Naresuan University, Faculty of Science, Department of Mathematics, Phitsanulok, Thailand
Email: nupiant@nu.ac.th
[*]Corresponding Email: anamain@nu.ac.th

ABSTRACT

The objective of this paper is to develop the predictive model for Thai gold bullion price. Gold is well-known for its precious quality and used as essential material in many industries such as jewelry, financial gold, electronics, computers, dentistry and for the gold trader who gains benefits from buying and selling gold. The paper focuses on a comparison of models developed based on multiple linear regression (MLR) and artificial neural networks (ANN). It consists of two main parts 1) the development of the model for predicting weekly gold price and 2) the model for predicting daily gold price. The dataset for weekly gold price model is taken from the gold price for 10 years (2006-2015) from website, while the daily gold price model uses the data based on four influenced factors: Thai gold price, Exchange rate, International gold price and SET50 for 3 years (2014-2017). The results reveal that ANN slightly outperforms MLR for weekly gold price predictive models, while MLR performs better than ANN for daily gold price predictive model. The efficient models can be further developed as a program in which will benefit the user in order to predict both the weekly and daily gold price in Thailand.

Keywords: Gold Bullion Price Predictive Model; Multiple Linear Regression; Artificial Neural Network

1 INTRODUCTION

Gold has played a major role in various industries and investments of many countries around the world. The precious quality of gold leads to its pricy use in many industries such as in jewelry, electronics, computers and dentistry (World Gold Council, 2016). In addition, gold has been used as financial assets in international currency reserves. Hence, the effective way of foreseeing gold prices will reduce risk value in gold selling, buying and investing. However, gold price is unstable and sometime very sensitive till obtaining extreme price fluctuations. This causes difficulty for the investors to capture the trend and foreseeing gold prices.

Gold predictive models have recently been one of the most interesting topics for researchers so various proposed techniques and methods have been presented including auto regression integrated moving average (ARIMA) (Guha, 2016; Tripathy, 2017) based on the time series method, Artificial Neural Network (ANN) (Mombeini, 2015) and Multiple Linear Regression (MLR) (Ismail, 2009). Different datasets of gold prices were also proposed and discussed. With the time series approach the dataset is typically based on the gold prices with a desired interval e.g. daily, weekly, and monthly. With other approaches, the dataset is set up from possible influential factors affecting the gold prices.

The gold prices in Thailand is updated daily from Monday to Saturday. There are gold bullion price and gold ornament price both for buying and selling recorded every day, which can be retrieved from the website (Gold Trader Association). Since we focused on the model to support the industrial investors who want to buy or stock the gold bullion as supplied materials and for the investors who sell and buy gold for profits. In this paper, we designed and developed two types of gold bullion price predictive models 1) the weekly predictive model and 2) the daily predictive model.

The two techniques considered here are Multiple Linear Regression (MLR) and Artificial Neural Network (ANN). Hence the aim of this study is to make a comparison between these two techniques, then the accuracy prediction of the gold bullion price predictive models from each technique are calculated and compared.

In section 2 we present the research methods including MLR and ANN, then section 3 describes the details of the construction of the weekly gold bullion price models and prediction accuracy, then the daily gold bullion price models discussed in section 4. In section 5, the conclusion is summarized and discussed.

2 METHODS

In order to compare the prediction accuracy of the MLR and ANN models, the dataset of the gold bullion price from 2006 to 2015, for 10 years, was collected from the website and to be used for the weekly predictive gold price model. While the daily predictive gold price model used the dataset based on four influential factors, which are Thai gold price, Exchange rate, International gold price and SET50 from 2014 to 2017. The factors introduced here were proposed in (Mombeini, 2015). The MLR and ANN techniques are presented in the next section.

2.1 Multiple Linear Regression (MLR)

Regression analysis is one of the most popular techniques and has been widely used in the context of yield prediction on many applications and domains, since it is simple to construct and there are various software that support the model construction (Montgomery, 2012). The method is relied on the assumption of random error arising from a large number of insignificant input factors. By giving an output response, *y*, and input variables = $(x_1, \dots, x_d)$, the relationship between *y* and *x* can be shown in (1).

$$y = f(x) + \varepsilon \tag{1}$$

ε is a random error assumed to be normally distributed with zero mean and variance σ^2. The true response surface function $f(x)$ is unknown, then a response surface $g(x)$ is invented to approximate $f(x)$ and the predicted values are obtained by using $\hat{y} = g(x)$. The $g(x)$ can be a polynomial function of $(X_1, X_2, \dots, X_d)$ and the observed dataset can be written with a matrix form of X in (2)

$$y_0 = X\beta + \varepsilon \tag{2}$$

where $y_0 = (y_1, y_2, \dots, y_n)^T$, X is a $n \times \alpha$ matrix, β is a $\alpha \times 1$ vector of the regression coefficients, and ε is $n \times 1$ vector of random error. The number of unknown parameters, which is defined as $\alpha = 2d + \binom{d}{2} + 1$. The vector of least squares estimators $\hat{\beta}$ is determined subject to the minimization of

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = (y_0 - X\beta)^T (y_0 - X\beta) \tag{3}$$

Then the minimization of (3) becomes

$$X^T X\hat{\beta} = X^T y_0 \tag{4}$$

Hence, the least squares estimator of β is

$$\hat{\beta} = (X^T X)^{-1} X^T y_0 \quad (5)$$

, while $(X^T X)$ is invertible. After β is estimated, the equation (5) can be used to predict the gold price at any untried settings of input variables.

2.2 Aritificial Neural Network (ANN)

Artificial neural network (ANN) is designed based on the learning system of the human brain, where millions of neurons are closely interconnected. ANN has been used in many applications such as image recognition, predictive models and complex decision making systems. ANN consists of a set of connected nodes, which is simplified as an artificial neuron. The node has a basic architecture as shown in Figure 1. A set of inputs (p_i) is connected to the node, where a summation function $(\sum)$ performs. A set of input is summed with a set of weight (w_i) corresponding to each input (p_i) and the bias (i.e. b_1) assigned to the node. Then, the summation of this is passed to the activation function (f), which is nonlinear, in this paper the sigmoid function was used. Eventually an output response (y) is obtained as shown in Figure 1.

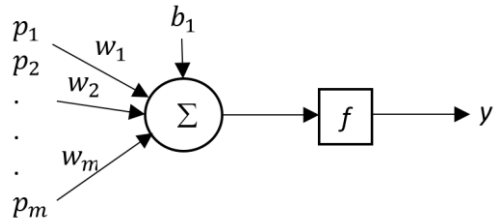


Figure1: A basic node model (Na-udom, A, 2015)

The process of a basic node model in Figure 1 can be formulated and written as (6).

$$y = f(w_i p_i + b_1) \quad (6)$$

A set of weight (w_i) is calculated and assigned during the learning process, in which backpropagation method is used in this paper. The backpropagation learning process is fine-tuned by some parameters; learning rate, momentum rate, training time and random seed number. Learning rate influences how large the weight adjustment should be and momentum rate influences the current adjustment to move in the same direction as previous.

ANN typically consists of an input layer, a hidden layer and an output layer. In theory, ANN can have more than one hidden layers; however one hidden layer is often sufficient enough. In this study we only used one hidden layer and each node in each layer is interconnected to every node in the next layer as displayed in Figure 2.

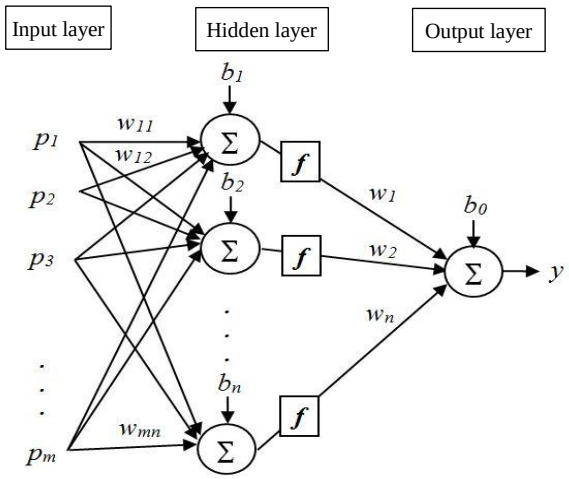


Figure 2: A structure of ANN (Na-udom, A, 2015)

Figure 2 displays an ANN model, which has *m* inputs, one hidden layer with *n* number of nodes and one output y.

The entire processing unit of ANN can be rewritten as,

$$y = \left[\sum_{i=1}^{n} w_i \times \left(f\left(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} p_i + b_i\right)\right)\right] + b_0 \quad (7)$$

,where n is a number of node in a hidden layer and m is a number of inputs, b_i is a bias of each node and b_0 is a bias at the output node.

3 WEEKLY GOLD PRICE PREDITIVE MODELS

The dataset for the weekly gold price predictive model is taken from the Gold Trader Association website, which is a daily gold price for 10 years, a total of 3,023 records. Then the daily buying gold bullion price is transformed to weekly gold price by averaging the gold price in a week. The dataset reduces to 522 records (weekly), with maximum, minimum, average and standard deviation values of 26466.67, 10160, 17641.51, and 4497.08, respectively. The plot graph of average weekly gold price is in Figure 3.

Figure 3: A graph of average weekly gold price

3.1 The design of weekly gold price dataset

In this study, based on the concept of autoregressive models, we used the average weekly gold price of the previous 4 week to predict the gold price in week 5. Hence, examples of the dataset is shown in Table 1, where W_t is the current week and W_{t+1} is the week we want to predict the gold price. W_{t-3} is referred as the gold price of the previous 3 week before current week.

In fact, with this study, we also combined the concept of moving average models to the weekly gold price models. However, the prediction accuracy of such models were not good enough. Then, we did not present such models in this paper.

Table 1: Example of weekly gold price dataset

W_{t-3}	W_{t-2}	W_{t-1}	W_t	W_{t+1}
18966.67	18758.33	18508.33	18600.00	18875.00
18758.33	18508.33	18600.00	18875.00	19291.67
18508.33	18600.00	18875.00	19291.67	19941.67
18600.00	18875.00	19291.67	19941.67	19791.67
…	…	…	…	…

With this design, the total number of records in the dataset is 518. Then, the dataset is split into two sets; the first 466 records are served as the training set, while the others 52 as the test set.

3.2 The development of weekly gold price models

The predictive models were developed with WEKA by applying the training set for building predictive models, then validating them with the test set. The prediction accuracy is measured by Root Mean Square Error (RMSE), its formula is defined in (8).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{k}(y_i - \hat{y}_i)^2}{k}} \quad (8)$$

, where k is a number of test points, y_i and $\hat{y}_i$ is actual and predicted value respectively.

The result of the predictive model from MLR is defined in (9) with the RMSE on the training set and the test set are **311.71** and **235.71** respectively.

$$\hat{y} = 0.1153 * w_{t-2} - 0.447 * w_{t-1} + 1.3257 * w_t + 97.7531 \quad (9)$$

To develop the ANN predictive model, we varied parameters by:
1) Different setting of (learning rate, momentum rate): (0.1,0.1) (0.1,0.3) (0.3,0.1) (0.2,0.3) (0.3,0.2) (0.2,0.2) (0.3,0.3)
2) Number of hidden nodes: 2, 3, 4, 5
3) trainingTime: 1000

The ANN models that minimizes RMSE on the test set is shown in Figure 4. The least RMSE on the test set is 231.9379 from the ANN with 2 hidden nodes and learning and momentum rate of 0.3 and 0.3.

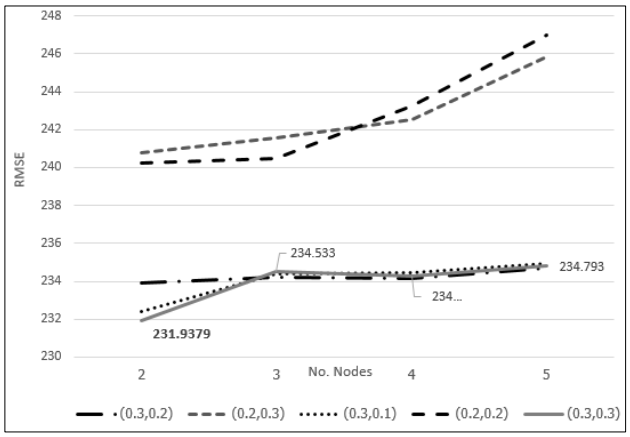


Figure 4: The accuracy prediction (RMSE) of ANN models on test set

3.3 The result of weekly gold price models

The accuracy prediction of MLR and ANN models can be summarized in Table 2. The result displays that the optimal ANN model is slightly better than MLR in terms of the accuracy prediction on test set, while the MLR training set is a little better than ANN training set. However, in this paper we judges the best models based on the accuracy prediction of test set.

Table 2: RMSE of MLR and optimal ANN for weekly models

Method	RMSE	
	Training Set	Test set
MLR	311.71	235.71
Optimal ANN	315.21	**231.94**

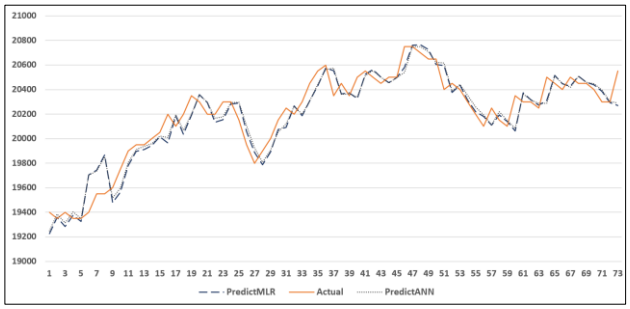


Figure 5: Graph for actual, MLR and ANN predicted value for test set

4 DAILY GOLD PRICE PREDITIVE MODELS

The dataset for the daily gold predictive model contains four sets of data from influential factors; Thai gold price, Exchange rate, International gold price and SET50 as input variables to predict the gold price of the next day. The total of 740 data taken from 24 March 2014 to 4 April 2017.

4.1 The design of daily gold price dataset

The design of daily gold price dataset is different from the weekly since we want to use the factors affecting the gold price, proposed by Mombeini (2015) and Lin (2015). In this paper, four input variables are today gold price (Gold), Exchange rate (Exchange), International gold price (Gold Spot) and index of SET50 (SET50) in order to predict the tomorrow gold price (Gold+1). The dataset is statistically analyzed with the values as shown in Table 3. The three input variables, which are significantly related to Gold price (Gold+1), and tested with p-value < 0.05, are Gold, Exchange and Gold Spot. Gold has the highest correlation with Gold+1 value (r) 0.202.

Table 3: The input variables and their statistic values

input variables	min	max	mean	SD	r	P-value
Gold	19550	22800	19599.93	2106.49	.202**	< .001
Exchange	31.92	36.71	34.367	1.44	.099**	.007
Gold Spot	1151.00	1366.25	1181.81	223.71	.114**	.002
SET50	761.75	1074.80	956.661	67.83	-.032	.377
Gold+1	**19550**	**22800**	**19012.16**	**3935.71**	**1**	-

4.2 The development of daily gold price models

With the design of daily gold price as shown in Table 4, the total number of records in the dataset is 739. The dataset is then split into two sets; the first 665 records are served as the training set, while the others 74 as the test set.

Table 4: Example of daily gold price dataset

Gold	Exchange	Gold Spot	SET50	Gold+1
20400	32.57	1310.75	912.35	20300
20300	32.63	1313.50	917.34	20300
20300	32.74	1304.00	921.49	20000
20000	32.71	1296.00	919.06	20000
20000	32.66	1294.75	929.84	19900
…	…	…	…	…

The result of the predictive model from MLR is defined in (10) with the RMSE on the training set and the test set are **152.17** and **118.02** respectively.

$$\hat{y} = 0.7554 * Gold + 138.2461 * Exchange + 3.9593 * Gold_{spot} - 4746.81 \quad (10)$$

To develop the ANN predictive daily gold price model, we varied parameters in a similar way to the weekly gold price models. The optimal ANN with the least RMSE on the test data is 121.67, with 2 hidden nodes and learning and momentum rate of 0.1 and 0.1.

We, then, tried to construct ANN without SET50 input, since its *r* value was quite low, and from the result of MLR model, SET50 is not included as shown in (10). The optimal ANN, in this case, has the least RMSE on the test data is 117.72, with 4 hidden nodes and learning and momentum rate of 0.1 and 0.3, which is slightly better than MLR (118.02).

4.3 The result of daily gold price models

In Table 3, the MLR daily gold price model slightly outperforms the optimal ANN model on both training and test set. Figure 6 shows the accuracy prediction for each test point.

Table 3: RMSE of MLR and optimal ANN for daily models

Method	RMSE	
	Training Set	Test set
MLR	152.17	**118.02**
Optimal ANN	153.16	121.67
Optimal ANN (3 input)	152.23	**117.72**

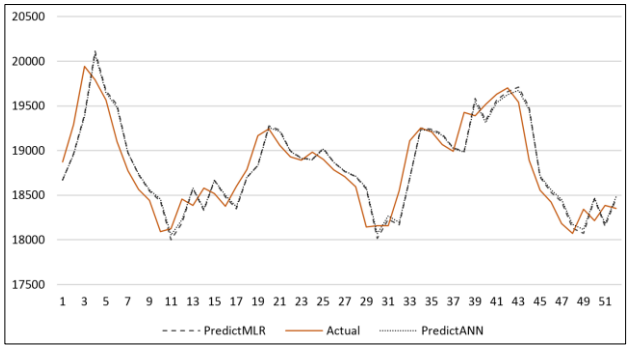


Figure 6: Graph for actual, MLR and ANN predicted value for test set

Finally, we have developed the web-based application, using WEKA engine, for users to predict the daily gold price.

Figure 7: A snapshot of the daily gold bullion predictive program

5 CONCLUSIONS AND DISCUSSIONS

This paper presents the construction of the weekly gold price predictive model from only gold price data. The design of the dataset is based on the concept of an autoregressive model. Whereas, the construction of the daily gold price predictive model is based on the factors affecting the gold price. From the study, it shows that both ANN and MLR have a similar accuracy prediction. ANN is slightly better for the weekly gold price predictive model, while MLR is a little better for the daily model. However, in terms of the program development, MLR is much easier to implement than ANN.

REFERENCES

Gold Trader Association. (2016). Retrieved July 4, 2016, from https://www.goldtraders.or.th/DailyPrices.aspx

Guha, B. & Bandyopadhyay, G. (2016). Gold price forecasting using ARIMA model. *Journal of Advanced Management Science*, 4(2), 117-121.

Ismail, Z. Yahya, A. & Shabri, A. (2009). Forecasting gold prices using multiple linear regression method. *American Journal of Applied Sciences*, 6(8), 1509-1514.

Lin, C. (2015). Building Prediction Models for gold prices based on back-propagation neural network. *International Conference on Modelling, Simulation and Applied Mathematics (MSAM 2015)*, 155-158.

Mombeini, H. & Yazdani-Chamzini, A. (2015). Modeling gold price via artificial neural network. *Journal of Economics, Business and Management*, 3(7), 699-703.

Montgomery, D. C., Peak, R. A. & Vining, G. G. (2012). Introduction to Linear Regression Analysis. Fifth Edition, John Wiley & Sons, New Jersey.

Na-udom, A. & Rungrattanaubol, J. (2015). A comparison of artificial neural network and regression model for predicting the rice production in lower Northern Thailand. *Information Science and Applications*, *Lecture Notes in Electrical Engineering*. 339. 745-752.

Shafiee, S. & Topal, E. (2010). An overview of global gold market and gold forecasting. *Resources Police*, 35(3), 178-189.

Tripathy, N. (2017). Forecasting gold price with auto regressive integrated moving average model. *International Journal of Economics and Financial Issues*, 7(4), 324-329.

World Gold Council. (2016). Gold Demand Trends Q2 2016. Retrieved August 4, 2016, from http://www.gold.org/supply-and-demand/gold-demand-trends/back-issues/gold-demand-trends-q2-2016

Superskewness Adjusted Black Scholes Option Pricing Model

Abdurakhman

Departemen Matematika,Universitas Gadjah Mada,
Yogyakarta, Indonesia
Corresponding Email: rachmanstat@ugm.ac.id

ABSTRACT

Black-Scholes (B-S) in 1973 succeeded in formulating an option price pricing formula. They use the assumption of stock returns following a normal distribution and the share price is logically distributed. Some stock returns are not normally distributed, so the option pricing model should consider a higher moment. The model in this paper inspired by the Gram-Charlier expansion to high-moment adjustment on the Black-Scholes formula. The approximation method used is the approach with the Hermite polynomial. Finally, we have a formulation for European type call option prices using fifth moment.

Keywords: Superskewness, Gram-Charlier, Hermite polyinomial, option

1 INTRODUCTION

Black and Scholes (1973) developed a European-type option pricing model based on normal distributed stock returns. The B-S model is a simple model widely used by practitioners in the options market. It may be said that the option practitioner makes the B-S formula an option price reference with slight modifications to market conditions. Of course the option price based on the B-S formula is not exactly 100% predict the option price in the market, there is still an error.

There is still some evidence in the market, the B-S model's option price valuation deviated in predicting market prices. This can be caused by many factors, such as ignore higher normal distribution moments. A study conducted by Jarrow and Rudd (1982) showed that a given probability distribution can be approximated by the expansion of distributions of two or more moments. Another is that, in options trading, the exchanges often limit the movement of daily price changes of an asset. As a result, asset returns are not perfectly normal distributed. The inclusion of higher moments is considered very important to get a better level of precision, given that not all data is perfectly normal distribution. Corrado and Su (1996) developed an option pricing formula using the Gram-Charlier expansion. Lin et al. (2015) discusses options pricing with several models including the Black-Scholes model and the Gram-Charlier Expansion model. The comparison of the option pricing model using Relative Price Error (RPE) and Squared Relative Price Error (SRPE) in some cases shows that the G-C approach still contains fluctuating errors. Based on the facts mentioned above, in this paper we formulate option price of European B-S model using a higher moment, until the fifth moment. Furthermore, the new model will be symbolized by GC_{SKS}.

The Black-Scholes model for option price valuation is a model used by many researchers and financial practitioners. This model was developed by Fischer Black and Myron Scholes (1976), and give the famous formula

$$C_{BS} = S_0 N(d_1) - Ke^{-rT} N(d_2) \tag{1}$$

Where d_1 and d_2

$$d_1 = \frac{\ln \frac{S_0}{K} + \left(r + \frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}} \quad ; \quad d_2 = d_1 - \sigma\sqrt{T} \tag{2}$$

Hermite Polynomial was defined by Laplace in 1810 and studied in detail by Chebyshev in 1859, was ignored and only recognized after Charles Hermite wrote this polynomial in 1864. The form of this polynomial is

$$H_n(z) = \sum_{k=0}^{n/2} \frac{(-1)^k n^{[2k]}}{2^k k!} z^{n-2k} \tag{3}$$

From equation (3), we get $H_0(z) = 1$, $H_1(z) = z$, $H_2(z) = (z^2 - 1)$, ..., $H_4(z) = (z^4 - 6z^2 + 3)$. This polynomial has orthogonal properties :

$$\int_{-\infty}^{\infty} H_m(z) H_n(z) n(z) dz = \begin{cases} 0 & , m \neq n \\ m! & , m = n \end{cases} \tag{4}$$

2. GRAM-CHARLIER EXPANSION

Gram-Charlier expansion is a very strong expansion to estimate normal density. In the last two decades, the use of this expansion has been introduced in the field of finance for leptokurtik return model, skewness, group volatility and so on. Hermite polynomials can be used to obtain expansion of probability functions in a series of derivatives n (z). One of the advantages of this polynomial is the fact that the density function can be formally expanded as :

$$g(z) = \sum_{n=0}^{n/2} c_n H_n(z) n(z) \tag{5}$$

Where n(z) is the standard normal density function, $H_n(z)$ is the hermite-order nth polynomial of equation (3). The coefficient c_n of (4) is derived from the Hermite Polynomial. If the equation (5) of the two segments multiplied by $H_m(z)$, and it is integrated from $-\infty$ to ∞, then :

$$\begin{aligned}
\int_{-\infty}^{\infty} g(z) H_m(z) dz &= \int_{-\infty}^{\infty} \sum_{n=0}^{n/2} c_n H_n(z) H_m(z) n(z) dz = \sum_{n=0}^{n/2} c_n \int_{-\infty}^{\infty} H_n(z) H_m(z) n(z) dz \\
&= c_0 \int_{-\infty}^{\infty} H_0(z) H_m(z) n(z) dz + c_1 \int_{-\infty}^{\infty} H_1(z) H_m(z) n(z) dz \\
&\quad + \ldots + c_{m-1} \int_{-\infty}^{\infty} H_{m-1}(z) H_m(z) n(z) dz \\
&\quad + c_m \int_{-\infty}^{\infty} H_m(z) H_m(z) n(z) dz + c_{m+1} \int_{-\infty}^{\infty} H_{m+1}(z) H_m(z) n(z) dz + \ldots
\end{aligned}$$

Using the orthogonal properties on equation (4) we obtained

$$\begin{aligned}
\int_{-\infty}^{\infty} g(z) H_m(z) dz &= 0 + 0 + \ldots + c_m \int_{-\infty}^{\infty} H_m(z) H_m(z) n(z) dz + 0 + \ldots \\
&= c_m \int_{-\infty}^{\infty} H_m(z) H_m(z) n(z) dz \\
&= c_m m!
\end{aligned}$$

Finally, we have:

$$c_m = \frac{1}{m!} \int_{-\infty}^{\infty} g(z) H_m(z) dz \tag{6}$$

From equation (6) we get the values : $c_0 = 1$, $c_1 = \mu_1$, $c_2 = (1/2!)(\mu_2 - 1)$, $c_3 = (1/3!)(\mu_3 - 3\mu_1)$, $c_4 = (1/4!)(\mu_4 - 6\mu_2 + 3)$, $c_5 = (1/5!)(\mu_5 - 10\mu_3 + 15$

μ_1) with μ_3, μ_4, μ_5 are skewness, kurtosis, and superskewness respectively. The expansion for normal standard pdf is :

$$g(z)=\sum_{n=0}^{\infty} c_n H_n(z)n(z)$$
$$=n(z)\left[\underbrace{\begin{bmatrix} H_0(z)+\mu_1 H_1(z)+\frac{1}{2!}(\mu_2-1)H_2(z)+ \\ \frac{1}{3!}(\mu_3-3\mu_1)H_3(z)+\frac{1}{4!}(\mu_4-6\mu_2+3)H_4(z) \end{bmatrix}}_{g(z)_{1-4}} + \underbrace{\frac{1}{5!}(\mu_5-10\mu_3+15\mu_1)H_5(z)}_{g(z)_5}\right] \quad (7)$$
$$=g(z)_{1-4}+g(z)_5$$

Corrado and Su (1996) develop the third and fourth moments for B-S option pricing valuation and give the formula for Call option price.

$$\begin{aligned} C_{GC_4} &= e^{-rT}E\left[\max(S_T-K,0)\right] \\ &= C_{BS}+\frac{\mu_3}{3!}I_1+\frac{(\mu_4-3)}{4!}I_2 \\ &= C_{BS}+\mu_3 Q_3+(\mu_4-3)Q_4 \end{aligned} \quad (8)$$

Where

$$Q_3=\frac{1}{3!}S_0\sigma\sqrt{T}\left(n(d_1)\left(2\sigma\sqrt{T}-d_1\right)+\sigma^2 T N(d_1)\right)$$
$$Q_4=\frac{1}{4!}S_0\sigma\sqrt{T}\left(n(d_1)\left({d_1}^2-3\sigma\sqrt{T}\left(d_1-\sigma\sqrt{T}\right)-1\right)+\left(\sigma\sqrt{T}\right)^3 N(d_1)\right)$$

Substituting equation (7) to the general formula in equation (1) give the option pricing formula based on superskewness moment as follows

$$\begin{aligned} C_{GC_{SKS}} &= e^{-rT}\int_K^{\infty}(S_T-K)g(z)dS_T \\ &= e^{-rT}\int_K^{\infty}(S_T-K)\left[g(z)_{1-4}+g(z)_5\right]dS_T \\ &= C_{GC_4}+C_{GC_5} \end{aligned}$$

Where

$$C_{GC_5}=e^{-rT}\int_K^{\infty}(S_T-K)\left[\frac{(\mu_5-10\mu_3)}{5!}H_5(z)n(z)\right]dS_T$$
$$=e^{-rT}\int_{-d_2}^{\infty}\left(S_0e^{mT+z\sigma\sqrt{T}}-K\right)\left[\frac{(\mu_5-10\mu_3)}{5!}H_5(z)n(z)dz\right]$$

Use partial integral

$u=S_0e^{mT+z\sigma\sqrt{T}}-K$ and $v=\int\frac{d}{dz}H_4(z)n(z)dz$ we gets that

$$\begin{aligned} C_{GC_{SKS}} &= \frac{1}{120}\left[\mu_5-10\mu_3\right]\left(\sigma\sqrt{T}\right)^5 S_0N(d_1)+ \\ &\quad S_0\sum_{j=2}^{5}\left(\sigma\sqrt{T}\right)^{j-1}H_{5-j}(-d_2)n(d_1) \\ &= \frac{1}{120}\left[\mu_5-10\mu_3\right] \\ &\quad \left[S_0\sigma\sqrt{T}\begin{pmatrix}\sigma^4T^2N(d_1)+ \\ n(d_1)\begin{bmatrix}4\sigma^3T^{3/2}-6d_1\sigma^2T+3d_1^2\sigma\sqrt{T}+ \\ d_1\sigma\sqrt{T}-3\sigma\sqrt{T}-d_1^3+3d_1\end{bmatrix}\end{pmatrix}\right] \end{aligned}$$

Finally we have that the formula for call option price using fifth moment

$$C_{GC_{SKS}}=C_{BS}+\mu_3Q_3+(\mu_4-3)Q_3+(\mu_5-10\mu_3)Q_5 \quad (9)$$

From the latest model option pricing formula in equation (9), we can see some analysis as follows :

1. The option price formula in equation (9) is an extension of the B-S option pricing formula and Gram-Charlier.
2. Theoretically, if the data is normally distributed (skewness = 0, kurtosis = 3 and the superskewness = 0), then the option price formula in equation (9) equals the B-S's formula..

ACKNOWLEDGEMENTS

The author would like to thank to the anonymous referees for their valuable suggestions. This study is funded by Universitas Gadjah Mada research project.

REFERENCES

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of political economy*, *81*(3), 637-654.

Corrado, C. J., & Su, T. (1996). Skewness and kurtosis in S&P 500 index returns implied by option prices. *Journal of Financial research*, *19*(2), 175-192.

Jarrow, R., & Rudd, A. (1982). Approximate option valuation for arbitrary stochastic processes. *Journal of financial Economics*, *10*(3), 347-369.

Lin, S. H., Huang, H. H., & Li, S. H. (2015). Option pricing under truncated Gram–Charlier expansion. *The North American Journal of Economics and Finance*, *32*, 77-97.

Can Bagging Improve Forecasting Accuracy of Decomposition Method? A Case Study of Thai Direct non-Life Insurance Premium

Pawanee Koprasertthaworn and Naowarut Meejun*

Silpakorn University/Department of Statistics /NakornPathom, Thailand
Email: koprasertthawor_p@silpakorn.edu
*Corresponding Email: Meejun_n@silpakorn.edu

ABSTRACT

Bootstrap aggregating (bagging) is a popular ensemble technique used to improve prediction performance. An ensemble consists of a number of training sets generated via resampling with replacement. Then a particular model is trained on each training set. The method gives multiple predictions before combining into one single estimator. Combining forecast has shown to perform better than a single forecast. In this study, bagging is introduced in forecasting Thai direct non-Life insurance premium. The aim is to improve the forecast accuracy of classical decomposition method. The methodology involves decomposing the time series into three main components: trend, seasonality and remainder. The remainder part is bootstrapped and then added back to the trend and seasonal part to obtain a bootstrapped time series. The modified decomposition method is applied to modelling each bootstrapped time series and the final forecast is obtained by averaging the forecast set. The method purposed is evaluated on a series of Thai direct non-Life insurance premium. The empirical results obtained with bagging decomposition method outperform the classical one.

Keywords:Bagging; Decomposition method; insurance premium

1 INTRODUCTION

The classical decomposition method is one of the oldest common forecasting techniques. The method is simple and easy to understand. The methodology is based on an analysis of individual components of the time series. A time series is decomposed into three main components: trend, seasonality and remainder. The information in each component can be extracted separately. The technique is useful for understanding and exploring the variation in each part of the time series. Also, it can be used in forecasting. The seasonality part is normally expressed with seasonal indices. It is an essential part in the business and economic time series as it shows periodic fluctuations of constant length. This makes the technique becomes wildly used. However, every method has drawbacks. A large random variation or outlier can distort the estimation of seasonal indices and trend.

Bootstrap aggregating (bagging) is a popular ensemble technique used to improve prediction performance (Fotios et al., 2018). It has been applied to combine with the other techniques in order to increase forecasting accuracy (Bergmeir et al., 2016, Dantas et al., 2017). An ensemble consists of a number of training sets generated via resampling with replacement. Then a particular model is trained on each training set. The method gives multiple predictions before combining into one single estimator. Combining forecast has shown to perform better than a single forecast.

However, adding bagging to the statistical forecasting techniques is not always promising (Bergmeir et al., 2016).There are several important things need to be considered in order to make bagging work efficiently. The bagging will perform well if it is well designed to tackle the sources of uncertainty. Previous work has demonstrated that bagging is very effective when combining with several forecasting techniques (Yang & Dong, 2018, Dantas et al.,2017, Zhang & Wang, 2018); but, there has been no empirical testing with the classical decomposition method.

In this study, bagging is introduced to work with the classical decomposition method. The aim is to find out if bagging can improve the forecast accuracy of classical decomposition method. We aim to make it simple and easy to implement as if it works this could be an alternative to the original one. The proposed method is applied to forecast Thai direct non-Life insurance premium.

The remainder of this paper is organized as follows. Section 2 introduces the proposed techniques including data description. In Section 3, the empirical results are presented. Finally, the conclusions of this study are presented in Section 4

2 METHODS

2.1 Data

In this study, a time series of Thai direct non-Life insurance premium was used as a case study. The data set consist of quarterly Thai non-Life insurance premium at national scale in total of 72 observations. The period considered runs from the first quarter in 2000 to the fourth quarter in 2017. The data from the first quarter in 2000 to the fourth quarter in 2016 were arranged as training set (in-sample) and the observations from the first quarter in 2017 to the fourth quarter in 2017 were arranged as test set (out-of-sample). The datasets were obtained online from the Office of Insurance Commission (OIC). A time series of direct non-Life insurance premium was adjusted before sending to the process. It was divided by Thai GDP.
Figure 1 shows a plot of time series.

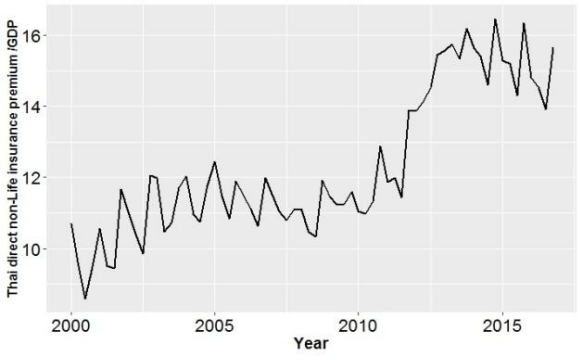


Figure 1: A time series of ratio of Thai direct non-Life insurance premium to Thai GDP

2.2 Decomposition Method

In decomposition, a time series is descried via two forms of model: an additive decomposition and multiplicative decomposition. The model is given in general form as follow:

$$Y_t = f(S_t, T_t, C_t, E_t) \quad (1)$$

where$Y_t$is the time series value at time t, S_t is a seasonal component, T_t is a trend component at time t, C_t is a cyclical component at time t and E_t is irregular or remainder component at time t. The cyclical component is often treated together with the trend component. Mathematically, time series decomposition is a mathematical procedure

that transforms a time series in to multiple time series, mainly 3 series: season, trend and remainder. To decompose the series, the first step is to estimate the seasonal indices using moving average. After that the seasonal component and deseasonalised series can be obtained. This depends on the type of the model. Finally, the remainder component is calculated by removing the estimated seasonal and trend component from the original series. Each part of the series can be modeled and forecasted individually.

2.3 Bagging

Bootstrap aggregating (Bagging) as proposed by Breiman (1996) is a well-known machine learning technique. In bagging, predictors are built on bootstrapped versions of the original data. For each training set, a specific model or technique is applied in order to make a prediction. The idea is to apply the same technique or model to different samples and then the predictions resulted from each data set are combined into one single prediction. The goal is to improve the accuracy of one data set by using multiple copies.

2.4 Decomposition with Bagging

In this study, bagging is introduced to work with the classical decomposition method. The aim is to find out if bagging can improve the forecast accuracy of classical decomposition method. The proposed method is applied to forecast Thai direct non-Life insurance premium in order to validate the method. The methodology is described as follows:

2.4.1 Step1-Decomposition

The time series is decomposed into three components: seasonal, trend and remainder. For this work, additive decomposition method was used, so the result can be written as in (2)

$$Y_t = \hat{S}_t + \hat{T}_t + e_t \qquad (2)$$

where $\hat{S}_t$ is estimated seasonal index, $\hat{T}_t$ is estimated trend and e_t is the remainder or residual . The seasonal part and the trend part are kept together and treated as constants. The remainder is a variation part. This part play important role in the next step.

2.4.2 Step 2- Bootstrapping the remainder

New versions of the remainders of time series are generated by bootstrapping the remainders part kept in the step 1. After that, the seasonal and the trend treated as constants in step 1 are added back to new version of remainder to obtain bootstrap time series. At this stage, it is very important to take into account that there is no autocorrelation in the remainder component. Otherwise, different techniques of bootstrapping will be the options such as Block Bootstrap. In this work, the remainders part was checked and there is no structure found in the series.

2.4.3 Step3- Bagging

The same model applied to original series in step1 is applied to bootstrap time series to generate forecasts. In this work, according step 2 there were 600 bootstrap samples generated and in each sample the classical decomposition was applied to generate forecasts.

2.4.4 Step 4- Aggregation

The final stage is to aggregate the forecasts in order to get the final result. There are various ways to combine forecasts together such as mean, median. In this work, the average of forecasts is used as the final forecast.

3 RESULTS

3.1 Decomposition of Time series

The time series of Thai direct non-Life insurance premium was used as a case study. For an analysis purpose the series was adjusted to series of ratio of Thai direct non-Life insurance premium to Thai GDP denoted by Y_t.

A plot of ratio of this series to the Thai GDP is illustrated in Figure1. As we can see, there was the effect of season in the series. Also, the insurance premium went up sharply in 2011-2013. In order to investigate each part of time series, the additive decomposition method was applied. The result is shown in Figure 2.

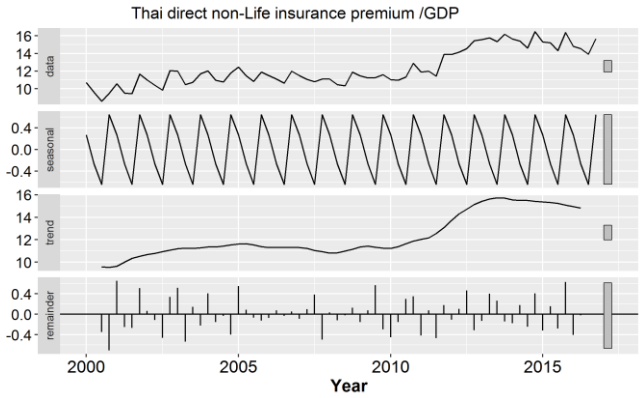


Figure 2: components of time series

3.1.1 The seasonal indices

Figure 2 shows the original time series on the top and another 3 times series extracted from the original one: seasonal variation, trend and remainder. The series is obviously influenced by the season as seen on a plot of seasonal series (the second from the top). The same pattern repeated every year. The seasonal indices were calculated and revealed in table1.

Table 1: Seasonal index

Quarter	Season index
1	0.2664
2	-0.264
3	-0.6418
4	0.6398

3.1.2 Trend analysis

Having obtained the seasonal indices, the trend-cycle of the series can be estimated. First, the deseasonalised data is estimated as follow:

Deseasonalised data= $Y_t - \hat{S}_t = \hat{T}_t + e_t$ (3)

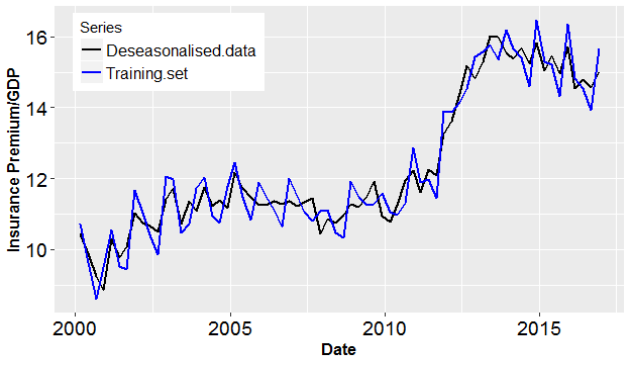


Figure 3: Deseasonalised data and Training set

Figure 3 shows plots of deseasonalised data comparing with training set. As seen in the figure, a linear trend is not a good choice for this data set. Polynomial regression models can be used to estimate the trend for this data set. An estimated trend equation was calculated as shown in (4) and the fitted trend line is shown in Figure 4.

$$\hat{T}_t = 10.7016 - 0.6060t + 0.1106t^2 - 0.0066t^3 + 0.0002t^4 - 2 \times 10^{-6}6t^5 + 9 \times 10^{-6}t^6 \quad (4)$$

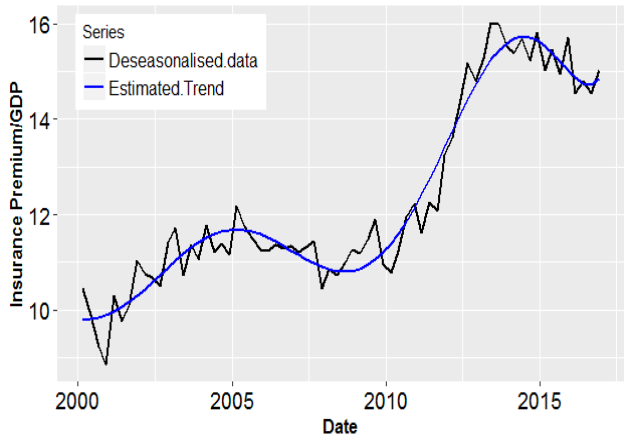


Figure 4: Deseasonalised data and Estimated Trend

3.1.3 *The remainders*

The variation part, remainders are obtained as in (5)

$$e_t = Y_t - \hat{S}_t - \hat{T}_t \quad (5)$$

At this stage, a series of remainderse_t ; t= 1, 2, 3, … 72 was obtained. This series is treated as an original series for the process of bootstrapping. The remainder series was checked if there is any structure, especially autocorrelation. Figure5 and figure 6 show a time series of remainder and the autocorrelation respectively. There is no structure found in this series, so the next step can be carry on.

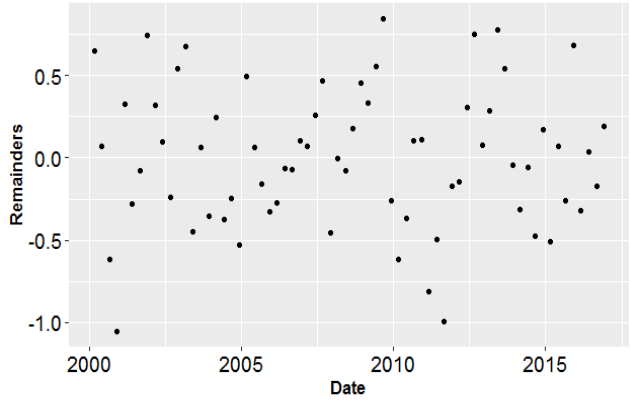


Figure 5: Remainders

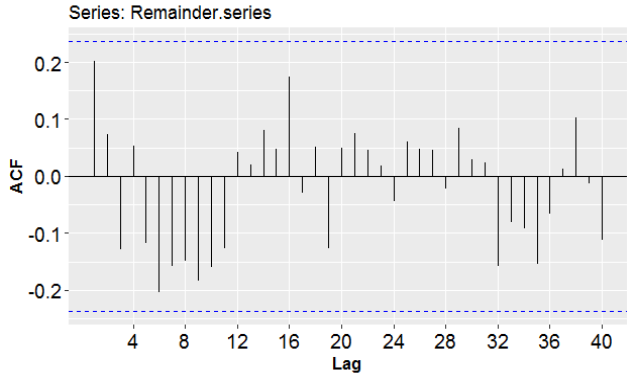


Figure 6: Autocorrelation function of remainder series

3.2 *Bootstrapping the remainder*

To generate a new time series, first bootstrapped remainder series were generated by bootstrapping the remainders and then these bootstrapped remainder series were added back to the trend and seasonal components as shown in (6)

$$Y_t^* = \hat{S}_t + \hat{T}_t + e_t^* \quad (6)$$

where e_t^* (t=1,2,3,…,72) are bootstrapped remainders and Y_t^* (t=1,2,3,…,72) are bootstrapped series.In this work, a typical bootstrap method was used while other types of bootstrap should be used if there is autocorrelation presented in the remainders. There were 600 new bootstrapped series generated.

3.3 *Bagging*

Both the original series and 600 bootstrapped series were used. The same technique, additive decomposition method with polynomial trend was performed on each series. A set of forecasts for training set and test set was calculated from the original series and bootstrapped series.

3.4 *Aggregation*

To obtain final forecast at each time t, an average of a bag of 600 forecasts was calculated. Table 2 shows forecasting results of two methods for the test set: decomposition method and decomposition method with bagging. The purposed methodology was evaluated using Mean Absolute Percentage Errors (MAPE) calculated as follow:

$$\left(\frac{1}{n}\sum_{t=1}^{n}\frac{|Y_t - \hat{Y}_t|}{|Y_t|}\right) \times 100 \quad (7)$$

. whereY_t is time series at time t and $\hat{Y}_t$ is fitted value at time t.

Table 2: Forecasting for test set by decomposition method and decomposition method with bagging.

Quarter	Y_t	decomposition method	decomposition method with bagging
1	55,101.15	55,578.06	55,634.07
2	53,123.05	51,521.04	51,638.33
3	51,731.28	50,804.05	50,965.63
4	59.693.57	57,818.59	58,017.49
	MAPE	2.2036%	2.0125%

As shown in Table 2 there is slightly improvement in forecasting with bagging.

4 CONCLUSIONS

This work applies a combination of Bootstrap aggregating method with the classical decomposition method to the Thai direct non-Life insurance premium in order to predict the future of insurance business. The combined methodology named as Decomposition with Bagging is applied for the first time as long as we are aware. The preliminary result obtained confirms that the methodology potentially can improve forecast accuracy even the series is not big and especially it is a quarterly data. As stated in the work of Bergmeir et al. (2016), the bagging may not always work well for the yearly and quarterly data. However the forecasting resulted from methodology proposed suggest that Decomposition with Bagging can be a choice for purposed of prediction.

REFERENCES

Bergmeir, C., Hyndman, R.J., & Benitez, J.M (2016). Bagging exponential smoothing method using STL decomposition and Box-Cox transformations. *International Journal of Forecasting* 32 (2) 303-312.

Bremen, L. (1996) Bagging predictors. *Mach. Learn*ing, 24(2),123-140.

Dantas, T., Oliveira, F., Repolho, H., (2017) Air transportation demand forecast through Bagging Holt Winters methods. *Journal of Air Transport Management*, 59,116-123.

Petropoulos, F., Hyndman, Rob J., & Bergmeir, C (2018). Exploring the sources of uncertainty: Why does Bagging for time series forecasting work*?*. *European Journal of Operational Research*, 268(2), 545-554.

Yang, D., & Dong, Z. (2018). Operational photovoltaics power forecasting using seasonal time series ensemble. *Solar Energy*, 166, 529-541.

Zhang, X., & Wang, J. (2018). A novel decomposition - ensemble model for forecasting short-term load time series with multiple seasonal patterns. *Applied Soft Computing*, 65, 478-494.

Logistic Regression Versus Linear and One Way ANOVA on the Lecturer Performance Index (LPI) IAIN Purwokerto

Mutijah[1*]

[1]IAIN Purwokerto, Banyumas, Indonesia

*mutijah1972@gmail.com

ABSTRACT

This paper compares the statistical analysis on the LPI data of IAIN Purwokerto. We analyze them based on the type of the LPI data and analyze the significance of the relationship. We do the analysis as follows; compares the analysis results of the significance using the linear regression and One Way ANOVA with the logistic regression. We compared significance of the relationship analysis between linear and logistic regression in independent variables of either age, the department assessment, the student assessment with the dependent variable of the LPI. We also compare the analysis results of the significance using One Way ANOVA with the logistic regression. We determine that there is a relationship if the average of the LPI gives the same results. They include the analysis of relationship between either the lecturer category, the assessment component category and the LPI. The analysis of linear regression showed the relationship between age and the LPI was not significant, the relationship between the department assessment and the LPI was significant, and the relationship between the student assessment and the LPI was significant. This showed the same results as previous research using the logistic regression. The analysis results using One Way ANOVA that the relationship between the lecturer category and the LPI was significant. This means that the average of the LPI was not the same for all the lecturer categories. In other words, the lecturer category has no relationship with the LPI. This is contrary as previous research using the logistic regression. The analysis result by One Way ANOVA for the assessment component category was not significant, this is also contrary using the logistic regression that the assessment component category has no the relationship with the LPI.

Keywords: logistic; linear; one way ANOVA; lpi

1 INTRODUCTION

We can use the statistical tools to analyze the data. The statistical analysis tools can use the logistic regression, the linear regression or the one-way ANOVA. The logistic regression analysis is used to analyze data if the data for the dependent variable is a category data, while the linear regression analysis can be used if the data for the dependent variable is continuous data. The one-way ANOVA is used if the variable is divided into two or more category and the data values for every category are continuous. The regression analysis aims to analyze the relationship between the independent variables and the dependent variable. It can be done if the dependent variable is in the form of a category. The linear regression analysis aims to analyze the relationship between the independent variables and the dependent variable. It can also be done if the data of the dependent variable is continuous data. The one-way ANOVA aims to analyze the average similarity of a variable dividing in the category if the data values of the variable is continuous data. In this study, the one-way ANOVA also be used to analyze the relationship between the independent variables and the dependent variable of the LPI by provisions if the average among the independent variable categories is the same, so that it is stated to have a relationship, but if the average among the independent variable categories is different so it is declared to have no a relationship.

This study will analyze data taken from the IAIN Purwokerto. The data includes age, the lecturer category, the assessment component of the LPI, and the main data from this study, namely the data of the LPI. The data of ages, the lecturer category, and the assessment component are the data for the independent variables and the data of the LPI is the data for the dependent variable. We find the data are the lecturer age of IAIN Purwokerto were 26,00-65,58 years, the lecturers of IAIN Purwokerto included State Civil Aparatus (SCA), the internal lecturer of Non State Civil Aparatus (NSCA), and External Lecturers (EL), the assessment component of the LPI consisted the department and student assessment, while the data values of the LPI was in intervals 0-4. The IAIN Purwokerto has made the category of the LPI in two categories. They are the LPI $\geq$ 3.00 called fine and the LPI < 3.00 called no fine. In this case, the difference among the lecturer category appears in its salary which it greatly affects the lecturer performance Index for the lecturers of IAIN Purwokerto. Whereas the assessment components of the LPI are important in determining whether fine or not fine in the LPI of IAIN Purwokerto. The department assessment is determined by the factors themselves but for the student assessment depends on the others.

Based on the exposure of the LPI results of the IAIN Purwokerto in 2017, we find the low LPI or the no fine LPI in the young lecturers and the lecturers in the category of External Lecturers (EL). Based on this problem, we also suspect that there is a significant of the relationship between each age, the lecturer category, and the assessment component and the LPI of IAIN Purwokerto. Therefore, to prove the hypothesis, we try to analyze the data by using the logistic and linear regression, and also the one-way ANOVA. In the case of the one-way ANOVA, we determine if the average is the same so there is a relationship, but if the average is not the same, then there is no relationship.

In detail, based on the LPI category then we analyze the relationship between each the independent variable age, the lecturer category, the department assessment, the student assessment and the LPI of IAIN Purwokerto by using the logistic regression. This data analysis has been carried out by Mutijah (2018). An idea that the LPI value is only an interval among 0-4, the small number value will very

determine the results of the analysis, therefore is important to compare the results of the analysis if the data type of the LPI of IAIN Purwokerto is continuous. Based on this condition, we perform a regression analysis for each the independent variable age, the lecturer category, and the assessment component and the LPI of IAIN Purwokerto. We propose an idea, it comes from the lecturer category data and the assessment component of the LPI which consists of two types, namely the department and student assessment. The idea is "if the average of the LPI in the category is the same then we declare there is a relationship and vice versa".

We compare the significance is the main part of using the logistic and linear regression, and also one-way ANOVA in this paper. We compare the significance of the data analysis between the linear and logistics regression in each the independent variable age, the department and student assessment. We also compare between one-way ANOVA and the logistic regression in each the independent variable in either the lecturer category, the assessment component and the LPI of IAIN Purwokerto as a dependent variable.

2 LOGISTIC REGRESSION VERSUS LINEAR ON THE LPI OF IAIN PURWOKERTO

The modeling of the relationship between a dependent variable and the independent variables is one of the most widely used of all statistical techniques. We refer to this type of modeling as regression analysis. A regression model provides the user with a functional relationship between the dependent variables and the independent variable that allows the user to determine which of the independent variables have an effect on the dependent variable. The regression model allows the user to explore what happen to the dependent variable for specified changes in the independent variables.

The basic idea of the regression analysis is to obtain a model for the functional relationship between a dependent variable and one or more the independent variables. Regression model have a number of uses that the model provides a description of the major features of the data set, the equation relating the dependent variable to the independent variables produced from the regression analysis provides estimates of the dependent variable for values of the independent variables are not observed in the study, and in some applications of the regression analysis, the researcher is seeking a model which can accurately estimate the values of a variable that is difficult or expensive to measure the independent variables that inexpensive to measure and obtain.

We would like to write an equation in the linear regression as shown in (1)

$$y = \mathrm{a} + \mathrm{bx} \tag{1}$$

The equation in (1) stated that y as a linear function of x and a, b are constant.0

To begin the linear regression analysis, we present description of variables and types of data as in table 1.

Table 1. Variables and types of data to analyze linear regression

Variables	Number of Lecturer	Type of Data	
		Logistic	Linear
Lecturer Age	220	Continuous	Continuous
Department Assessment	201	Continuous	Continuous
Student Assessment	201	Continuous	Continuous
LPI	220	Category	Continuous

Furthermore, we analyze the relationship between the lecturer age and the LPI, the department assessment and the LPI, the student assessment and the LPI resulting the linear regression model respectively are

$$y = 3{,}182 + 0{,}001\mathrm{X} \tag{2}$$

$$y = 6{,}623 + 0{,}250\mathrm{X} \tag{3}$$

$$y = 6{,}623 + 0{,}250\mathrm{X} \tag{4}$$

The significance analysis of the linear regression results as in equation (2), (3), and (4) showed that the regression model in equation (2) is not significant by p-value 0.765, equation (3) is significant by p-value <0.001, and equation (4) is also significant by p-value <0.001. The significance results are the same for the logistic regression analysis as previous research by Mutijah (2018).

3 ONE-WAY ANOVA ON THE LPI OF IAIN PURWOKERTO

In this research, analysis of variance (ANOVA) concern the interest of comparison of treatment means where the focus is on the evaluation of the effects of two or more independent variables on a dependent variable rather than on comparison of treatment means as in designs. Particular attention is given to focus on either comparison of treatment means or examination of the effects of the dependent variable.

Beginner to analyze the data using one-way ANOVA, we present types of data of the lecturer category and the assessment component as in table 2.

Table 2. Variables the lecturer category, the assessment component, and the data types of the LPI

Variables	Category	Coding	Number of Lecturer	The Data Type of the LPI
Lecturer Category	SCA	1	131	Continuous
	NSCA	2	50	Continuous
	EL	3	39	Continuous
Assessment Component	Department Assessment	1	201	Continuous
	Student Assessment	2	201	Continuous

Based on table 2, we analyze the average or mean among the lecturer category and the assessment component by one-way ANOVA. Relating with this, we determine that the mean among the lecturer category and among the assessment component is the same so we stated that there is relationship and vice versa.

The analysis of the lecturer category by one-way ANOVA result that it is significant by p-value 0,002. It means that the mean of the LPI among the category of SCA, NSCA, and EL are not the same, it also means between the lecturer category and LPI is no the relationship. This is contrary in previous research using the analysis of the logistic regression by Mutijah (2018).

The analysis of the assessment component by one-way ANOVA result that it is not significant by p-value 1,00. It means that the mean of the LPI between the department and student assessment are the same so this means that there is relationship. This result will be compared with the analysis using the logistic regression which the department and student assessment are analyzed together related by the LPI in fine and no fine category. This analysis has been done by Mutijah (2018).

As for the specific form of the logistic regression model use as below

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \quad (5)$$

with $\pi(x) = E(Y|x)$ represent the conditional mean of Y given x when the logistic distribution is used.

A transformation of $\pi(x)$ that is central to study the logistic regression is the logit transformation. This transformation is defined in terms of $\pi(x)$ as

$$g(x) = \ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x \quad (6)$$

The importance of this transformation is that g(x) has many of the desirable properties of a linear regression model as in (6).

To begin the logistic regression analysis, we present the data as in table 3.

Table 3. The data to analyze the logistic regression

Variables	Category	Coding
Lecturer Category	SCA	1
	NSCA	2
	EL	3
Assessment Component	Department Assessment	1
	Student Assessment	2
LPI	No Fine	0
	Fine	1

The results of analysis of the logistic regression from the data in table 3 are obtained the analysis of result is not significance by p-value 0,872. It means that there is no relationship between assessment component and LPI. This is also contrary with one-way ANOVA.

4 CONCLUSIONS

Based on the above description, it can be concluded that

1. The significance of the analysis results using the linear regression equal to the logistic regression. The results of this analysis illustrate that there is not a relationship between the lecturer ages and the Lecturer Performance Index (LPI), there is relationship between the department assessment and the Lecturer Performance Index (LPI), and there is also the relationship between the student assessment and the Lecturer Performance Index (LPI) in the IAIN Purwokerto.
2. The significance of the analysis results by one-way ANOVA shows that there is no relationship between the lecturer category and the Lecturer Performance Index (LPI) and this is contrary with the analysis by the logistic regression in Mutijah (2018), and there is a relationship between the assessment components and the Lecturer Performance Index (LPI) with a note that in one-way ANOVA if among the treatment means are the same then it is determined that there is a relationship. The two of the significance analyzed by one-way ANOVA are contrary to the results of the analysis by the logistic regression.

ACKNOWLEDGEMENTS

First, author would like to thank for Committee of International Conference on Applied Statistics (ICAS) 2018 which they have given opportunity to present my research. Second, author also would like to thank for Committee of International Conference on Applied Statistics (ICAS) 2018 which they have served me patiently and they have given suggestion to improve in my full paper. Third, author thank to Kementerian Agama Republik Indonesia and IAIN Purwokerto which have provided the facilities for my research. Part of this research was funded by Kementerian Agama Republik Indonesia and IAIN Purwokerto.

REFERENCES

Hosmer, D. W, & Stanley, L. (2000). Applied Logistic Regression. 2nd edition. John Willey & Sons Inc.

Mutijah. (2018). Logistic regression on the data of lecturer performance index (LPI) of Purwokerto State Islamic Institute. *Proceeding in Process*

Ott, R. L, & Longnecker, M. (2001). An Introduction to Statistical Methods and Data Analysis.

TIM. (2016). Panduan Indeks Kinerja Dosen (IKD) (translate)

The New Exact Solutions of the Fourth Order Nonlinear Estevez-Mansfield-Clarkson Equation by the Simple Equation Method with Riccati Equation

Sirasrete Phoosree and Settapat Chinviriyasit*

Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi (KMUTT) 126 Pracha Uthit Road, Bangmod, Thungkhru, Bangkok 10140, Thailand
Email: sirasrete.pho@mail.kmutt.ac.th
*Corresponding Email: settapat.chi@mail.kmutt.ac.th

ABSTRACT

One of the attractively of nonlinear partial differential equations (NPDEs) is exploration the solutions of them. In this work, the simple equation method and the Riccati equation combine to predicate the exact traveling wave solutions of Estevez-Mansfield-Clarkson (EMC) equation which studies the pattern of liquid drops. This method transforms the EMC equation to the ordinary differential equation and defines the solutions in the Riccati equation form. Finally, the solutions of EMC equation are explored.

Keywords: Estevez-Mansfield-Clarkson equation; Nonlinear partial differential equations; Traveling wave solution; Simple equation method

1 INTRODUCTION

The one of special field which has very important in physics is NPDEs. The NPDEs were applied in optical fibers, plasma waves, plasma physics, chemical kinetics, capillary-gravity waves and geochemistry phenomena (Shakeel & Mohyud-Din, 2015; Krisnangkura et al., 2012).

We can solve the NPDEs for finding the exact solutions by many strong methods such as exp-function method (Biazar & Ayati, 2012; Ravi et al. , 2017), G′/G-expansion method (Wang et al., 2008; Taha & Noorani, 2014; Shakeel & Mohyud-Din, 2015; Islam et al., 2017), F-expansion method (Islam et al., 2014; Islam et al., 2017), tanh method (Wazwaz, 2005; Krisnangkura et al., 2012; Bibi & Mohyud-Din, 2014), modified extended tanh method (Elwakil et al., 2002; Yang & Hon, 2006; Taghizadeh & Mirzazadeh, 2010; Taghizadeh & Mirzazadeh, 2011), tanh-coth method (Bekir & Cevikel, 2011; Kudryashov & Shilnikov, 2012; Kumar & Pankaj, 2015), sine-cosine method (Bibi & Mohyud-Din, 2014; Raslan et al., 2017), Jacobi elliptic function method (Ali, 2011; Wang & Xiang, 2013), homogeneous balance method (Eslami et al., 2014), modified simple equation method (Khater et al., 2017), and so on. In 2005, the simple equation (SE) method was introduced to solve the exact solutions of NPDEs by Kudryashov (2005). In recent year, this method has been used to explore the exact solution of NPDEs for many researches as finding the exact solution of the Sharma-Tasso-Olver and the Burgers-Huxley equations by Kudryashov and Loguinova (2008) and solving for the exact solutions of Kodomtsev-Petviashvili (KP) equation, the (2+1)-dimensional breaking soliton equation and the modified generalized Vakhnenko equation by Nofal (2016).

Estevez, Mansfield and Clarkson (1997) introduced one of NPDEs which consider the patterns of liquid drop, Estevez-Mansfield-Clarkson (EMC) equation. The EMC equation is the fourth order NPDEs, may be shown as

$$u_{xxxt} + \lambda u_x u_{xt} + \lambda u_{xx} u_t + u_{tt} = 0, \tag{1}$$

where $u = u(x, y, t)$ and λ is constant. Therefore, we want to find the exact traveling wave solutions of the EMC equation by the SE method.

The research objective is applied the simple equation method with Riccati equation to solve the new exact traveling wave solutions of the forth order nonlinear EMC equation (1).

The structure of this paper can be shown as the following. In section 2, the simple equation method with Riccati equation is used to explore the new exact traveling wave solutions of the fourth order nonlinear EMC equation. Section 3, the 3D exact traveling wave solutions graphs of EMC equation is discussed. The conclusion is stated in the last section.

2 SOLVING THE EMC EQUATION BY THE SIMPLE EQUATION METHOD WITH RICCATI EQUATION

Given the EMC equation (1):

$$u_{xxxt} + \lambda u_x u_{xt} + \lambda u_{xx} u_t + u_{tt} = 0, \tag{2}$$

where $u = u(x, y, t)$ and λ is constant.

Step1: Reducing equation (2) by using the wave transformation (Nofal, 2016) $u(x, y, t) = U(\zeta)$, $\zeta = x + y - ct$, where c is wave velocity constant. We get

$$-\frac{d^4U}{d\zeta^4} - 2\lambda \frac{dU}{d\zeta} \cdot \frac{d^2U}{d\zeta^2} + c\frac{d^2U}{d\zeta^2} = 0. \tag{3}$$

Integrating equation (3) with zero constant yields,

$$-\frac{d^3U}{d\zeta^3} - \lambda\left(\frac{dU}{d\zeta}\right)^2 + c\frac{dU}{d\zeta} = 0. \tag{4}$$

Step2: Applying the SE method (Nofal, 2016), the solution of equation (4) may be defined as,

$$U(\zeta) = \sum_{i=0}^{N} a_i S^i(\zeta). \tag{5}$$

The term $S(\zeta)$ satisfied the Riccati equation,

$$S'(\zeta) = \alpha S^2(\zeta) + \beta, \tag{6}$$

where a_i are real constants, $a_N \neq 0$ and α, β are nonzero constants.

Thus solutions of the equation (6) are described in two cases (Nofal, 2016):

Case 1: $\alpha\beta < 0$,

$$S(\zeta) = -\frac{\sqrt{-\alpha\beta}}{\alpha}\tanh\left(\sqrt{-\alpha\beta}\zeta - \frac{\upsilon \ln(\zeta_0)}{2}\right), \zeta_0 > 0, \upsilon = \pm 1. \tag{7}$$

Case 2: $\alpha\beta > 0$,

$$S(\zeta) = \frac{\sqrt{\alpha\beta}}{\alpha}\tan\left(\sqrt{\alpha\beta}(\zeta + \zeta_0)\right), \zeta_0 \text{ is a constant.} \tag{8}$$

Step3: Balancing (Nofal, 2016) the highest derivative term and the nonlinear term in equation (4) as $\frac{d^3U}{d\zeta^3}$ and $\left(\frac{dU}{d\zeta}\right)^2$, respectively. We get $N+3=2N+2$ so $N=1$. Thus, equation (5) permuted to be

$$U(\zeta)=a_0+a_1S(\zeta). \tag{9}$$

Differentiating equation (9), this yields

$$U'(\zeta)=a_1\alpha S^2(\zeta)+a_1\beta, \tag{10}$$

$$U''(\zeta)=2a_1\alpha^2S^3(\zeta)+2a_1\alpha\beta S(\zeta), \tag{11}$$

$$U'''(\zeta)=6a_1\alpha^3S^4(\zeta)+8a_1\alpha^2\beta S^2(\zeta)+2a_1\alpha\beta^2. \tag{12}$$

Step4: Substituting equations (10) and (12) into equation (4),

$$\begin{aligned}&-6a_1\alpha^3S^4(\zeta)-8a_1\alpha^2\beta S^2(\zeta)-2a_1\alpha\beta^2-\lambda a_1{}^2\alpha^2S^4(\zeta)\\&-2\lambda a_1{}^2\alpha\beta S^2(\zeta)-\lambda a_1{}^2\beta^2+ca_1\alpha S^2(\zeta)+ca_1\beta=0.\end{aligned} \tag{13}$$

The terms of the power of $S^i(\zeta)$ are collected, setting the coefficient to be zero (Nofal, 2016), we obtain

$$S^0(\zeta): \quad -2a_1\alpha\beta^2-\lambda a_1{}^2\beta^2+ca_1\beta=0, \tag{14}$$

$$S^2(\zeta): \quad -8a_1\alpha^2\beta-2\lambda a_1{}^2\alpha\beta+ca_1\alpha=0, \tag{15}$$

$$S^4(\zeta): \quad -6a_1\alpha^3-\lambda a_1{}^2\alpha^2=0. \tag{16}$$

Solving the system of equations (14) - (16), we gets

$$a_1=\frac{-6\alpha}{\lambda}, c=-4\alpha\beta. \tag{17}$$

Substituting equations (7), (8) and (17) into equation (9), the solutions of the EMC equation may be considered as,

Case 1: $\alpha\beta<0$,

$$u(x,y,t)=a_0+\frac{6\sqrt{-\alpha\beta}}{\lambda}\tanh\left(\sqrt{-\alpha\beta}(x+y+4\alpha\beta t)-\frac{\upsilon\ln(\zeta_0)}{2}\right), \tag{18}$$

where $\zeta_0>0, \upsilon=\pm1$.

Case 2: $\alpha\beta>0$,

$$u(x,y,t)=a_0-\frac{6\sqrt{\alpha\beta}}{\lambda}\tan\left(\sqrt{\alpha\beta}(x+y+4\alpha\beta t+\zeta_0)\right), \tag{19}$$

where ζ_0 is a constant.

3. RESULTS

Consider the first exact traveling wave solutions case of the EMC equation,

$$u(x,y,t)=a_0+\frac{6\sqrt{-\alpha\beta}}{\lambda}\tanh\left(\sqrt{-\alpha\beta}(x+y+4\alpha\beta t)-\frac{\upsilon\ln(\zeta_0)}{2}\right),$$

where $\zeta_0>0, \upsilon=\pm1$.

Setting parameters $a_0=0, \alpha=1, \beta=-1, \lambda=1, \upsilon=-1, \zeta_0=2$, $-20\le x\le20, -20\le y\le20$ and $t=0,2,4,6$, the graph shows the kink wave solutions as Figure 1.

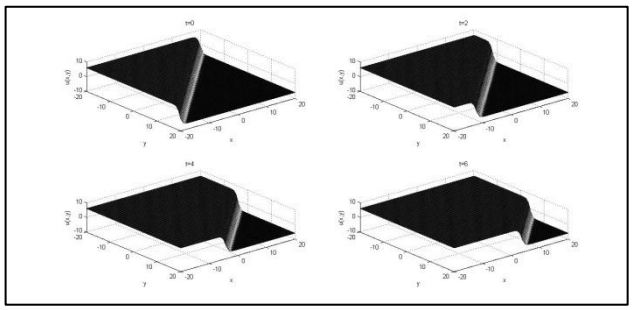


Figure 1: The kink wave solutions of the EMC equation in case 1

For the second exact traveling wave solutions case of the EMC equation,

$$u(x,y,t)=a_0-\frac{6\sqrt{\alpha\beta}}{\lambda}\tan\left(\sqrt{\alpha\beta}(x+y+4\alpha\beta t+\zeta_0)\right),$$

where ζ_0 is a constant.

The traveling wave solutions is periodic wave when we set $a_0=0$, $\alpha=1, \beta=1, \lambda=1, \zeta_0=2, -20\le x\le20, -20\le y\le20$ and $t=0,2,4,6$ and the 3D graph shows as Figure 2.

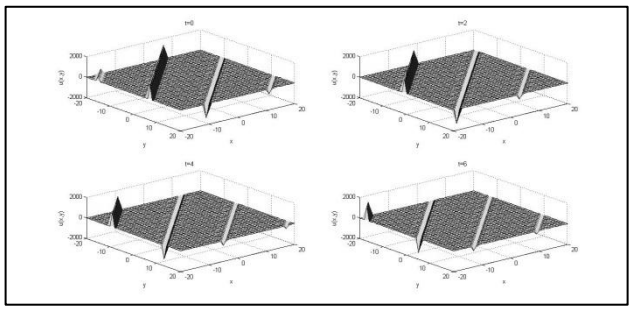


Figure 2: The periodic wave solutions of the EMC equation in case 2

In summary, the EMC equation was solved by the simple equation method with Riccati equation, we got the exact traveling wave solutions in the form of kink and periodic wave.

4. CONCLUSIONS

The simple equation method is applied to solve the EMC equation. By suppose the transformation variable ζ, this yields the EMC equation reduced to be an ODE. The solution may be defined in equations (5) and (6). Taking balancing and solving some algebraic system, the solutions of the EMC equation are achieved in equations (18) and (19). The graphs of equations (18) and (19) with parameters are shown in Figure 1 and Figure 2, respectively.

REFERENCES

Ali, A.T. (2011). New generalized Jacobi elliptic function rational expansion method. *Journal of Computational and Applied Mathematics*, 235(14), 4117-4127.

Bekir, A., & Cevikel, A.C. (2011). The tanh-coth method combined with the Riccati equation for solving nonlinear coupled equation in mathematical physics. *Journal of King Saud University Science*, 23(2), 127-132.

Biazar, J., & Ayati, Z. (2012). Exp and Modified Exp function methods for nonlinear Drinfeld-Sokolov system. *Journal of King Saud University-Science*, 24(4), 315-318.

Bibi, S., & Mohyud-Din, S.T. (2014). New traveling Wave Solutions of Drinefel'd-Sokolov-Wilson equation using Tanh and Extended Tanh methods. *Journal of the Egyptian Mathematical Society*, 22(3), 517-523.

Bibi, S., & Mohyud-Din, S.T. (2014). Traveling wave solutions of KdVs using sine-cosine method. *Journal of the Association of Arab Universities for Basic and Applied Sciences*, 15, 90-93.

Elwakil, S.A., El-labany, S.K., Zahran, M.A., & Sabry, R. (2002). Modified extended Tanh-Function method for solving nonlinear partial differential equations. *Physics Letters A*, 299(2-3), 179-188.

Eslami, M., Fathi vajargah, B., & Mirzazadeh, M. (2014). Exact solutions of modified Zakharov-Kuznetsov equation by the homogeneous balance method. *Ain Sham Engineering Journal*, 5(1), 221-225.

Islam, Md.T., Akbar, M.A., & Azad, Md.A.K. (2017). Multiple closed form wave solutions to the KdV and modified KdV equations through the rational (G'/G)-expansion method. *Journal of the Association of Arab Universities for Basic and Applied Sciences*, 24, 160-168.

Islam, Md.S., Khan, K., & Akbar, M.A. (2017). Application of the improved F -expansion method with Riccati equation to find the exact solution of the nonlinear evolution equations. *Journal of the Egyptian Mathematical Society*, 25(1), 13-18.

Islam, M.S., Khan, K., Akbar, M.A., & Mastroberardino, A. (2014). A note on improved F-expansion method combined with Riccati equation applied to nonlinear evolution equations. *Royal Society Open Science*, 1(2), 140038(1-8).

Khater, M.M.A., Zahran, E.H.M., & Shehata, M.S.M. (2017). Solitary wave solution of the generalized Hirota-Satsuma coupled KdV system. *Journal of the Egyptian Mathematical Society*, 25(1), 8-12.

Krisnangkura, M., Chinviriyasit, S., & Chinviriyasit, W. (2012). Analytic study of the generalized Burger's-Huxley equation by hyperbolic tangent method. *Applied Mathematics and Computation*, 218(22), 10843-10847.

Kudryashov, N.A. (2005). Simplest equation method to look for exact solutions of nonlinear differential equations. *Chaos, Solitons & Fractals*, 24(5), 1217-1231.

Kudryashov, N.A., & Loguinova, N.B. (2008). Extended simplest equation method for nonlinear differential equations. *Apply Mathematics and Computation*, 205(1), 396-402.

Kudryashov, N.A., & Shilnikov, K.E. (2012). A note on the Tanh-Coth method combined with the Riccati equation for solving nonlinear coupled equation in mathematical physics. *Journal of King Saud University-Science*, 24(4), 379-381.

Kumar, A., & Pankaj, R.D. (2015). Tanh-Coth scheme for traveling wave solutions for nonlinear wave interaction model. *Journal of the Egyptian Mathematical Society*, 23(2), 282-285.

Mansfield, E.L., & Clarkson, P.A. (1997). Symmetries and exact solutions for a 2+1-dimensional shallow water wave equation. *Mathematics and Computers in Simulation*, 43(1), 39-55.

Nofal, T.A. (2016). Simple equation method for nonlinear partial differential equations and its applications. *Journal of the Egyptian Mathematical Society*, 24(2), 204-209.

Raslan, K.R., EL-Danaf, T.S., & Ali, K.K. (2017). New exact solution of coupled general equal width wave equation using sine-cosine function method. *Journal of the Egyptian Mathematical Society*, 25(3), 350-354.

Ravi, L.K., Ray, S.S., & Sahoo, S. (2017). New exact solutions of coupled Boussinesq-Burgers equations by Exp-function method. *Journal of Ocean Engineering and Science*, 2(1), 34-46.

Shakeel, M., & Mohyud-Din, S.T. (2015). Improved (G'/G)-expansion and extended Tanh methods for (2+1)-dimensional Calogero-Bogoyavlenskii-Schiff equation. *Alexandria Engineering Journal*, 54(1), 27-33.

Taha, W.M., & Noorani, M.S.M. (2014). Application of the (G'/G)-expansion method for the generalized fisher's equation and modified equal width equation. *Journal of the Association of Arab Universities for Basic and Applied Sciences*, 15, 82-89.

Taghizadeh, N., & Mirzazadeh, M. (2010). The modified Tanh method for solving the improved Eckhaus equation and the (2+1)-dimensional improved Eckhaus equation. *Australian Journal of Basic and Applied Sciences*, 12(4), 6373-6379.

Taghizadeh, N., & Mirzazadeh, M. (2011). The modified extended Tanh method with the Riccati equation for solving (3+1)-dimensional Kadomtsev-Petviashvili (KP) equation. *International Journal of Applied Mathematics and Computation*, 3(4), 238-241.

Wang, M., Li, X., & Zhang, J. (2008). The G'/G -expansion method and travelling wave solutions of nonlinear evolution equations in mathematical physics. *Physics Letters A*, 372(4), 417-423.

Wazwaz, A. M. (2005). Exact solutions to the double Sinh-Gordon equation by the Tanh method and a variable separated ODE method. *Computer and Mathematics with Applications*, 50(10-12), 1685-1696.

Wang, H., & Xiang, C. (2013). Jacobi elliptic function solutions for the modified Korteweg-de Vries equation. *Journal of King Saud University-Science*, 25(3), 271-274.

Yang, Z., & Hon, B.Y.C. (2006). An improved modified extended Tanh-function method. *Zeitschrift fur Naturforschung A*, 61(3-4), 103-115.

Zhen-Ya, Y. (2002). Abundant symmetries and exact Compacton-Like structures in the two-parameter family of the Estevez Mansfield Clarkson equations. *Communications in Theoretical Physics*, 37(1), 27-34.

The Numerical Solution of Fractional Black-Scholes-Schrodinger Equation Using the MLPG Method

Naravadee Nualsaard[1], Anirut Luadsong [1*] and Nitima Aschariyaphotha[2]

[1]King Mongkut's University of Technology Thonburi/Department of Mathematics/Bang Mod, Thung Khru, Bangkok, Thailand
[2]King Mongkut's University of Technology Thonburi/Ratchaburi Learning Park/Rang Bua, Chom Bueng, Ratchaburi, Thailand
*Corresponding Email: anirut.lua@kmutt.ac.th
Email: naravadee401@gmail.com
Email: nitima.asc@kmutt.ac.th

ABSTRACT

This paper, mainly focuses on the fractional Black-Scholes-Schrodinger equation for the option price in financial problems which is solved by using a numerical technique. The meshless local Petrov-Galerkin (MLPG) method is applied for spatial discretization. The approximation of time fractional derivative is interpreted in the Caputo's sense by a simple quadrature formula. In MLPG method, the moving kriging interpolation is applied to constructing shape function. The Kronecker delta function is chosen to be the test function for simplifying the equation. A numerical solution is compared with the semi-classical solution in case of fractional order approach to 1 to verify the results. The results can be concluded that the option prices from MLPG method satisfy with the semi-classical solution.

Keywords: Black-Scholes-Schrodinger equation; fractional model; meshless local Petrov-Galerkin (MLPG) method; option pricing

1 INTRODUCTION

The Black–Scholes equation (Black & Scholes, 1973) is the famous financial model for analyzing of option pricing. One of key assumption of Black–Scholes model has no-arbitrage. In the fact, arbitrage exists in real financial market. The classical Black-Scholes model is extended for arbitrage possibilities by Contreras et al. (2010). The Black–Scholes equation can be interpreted from quantum mechanics's view point in senses of imaginary time Schrodinger equation of a free particle. The Black-Scholes-Schrodinger equation based on arbitrage possibilities is proposed by Contreras, Pellicer et al. (2010). The semi-classical method is applied for the solution of this Black-Scholes-Schrodinger equation.

In recent decades, fractional derivatives have more interested from researchers because it's appearance and various application in engineering and practical science. Many problems in biology, chemistry, physics, mechanics and engineering are transformed in term of fractional differential and integral equations such as diffusion-reaction processes, electrochemistry, financial, and biological system etc. Fractional derivative can be described some occurrence that integer order cannot including financial market. There are several definitions for transforming the notation of differentiation to fractional order. The famous definition such as Grunwald-Letnikov's, Riemann-Liouville's, Caputo's definitions and etc. Each of these definitions also has both advantages and disadvantages. During the last decades, there are many researchers studied the numerical method for solving the fractional Black-Scholes model (Cen et al., 2018; Song & Wang, 2013; Zhang et al., 2016). As mentioned previously, all researches used finite difference method (FDM) for solving the fractional Black-Scholes equation. Solving FDEs by FDM requires the mesh generation, which appears to be computationally costly (Thamareerat et al., 2017). In this paper, the meshless local Petrov-Galerkin (MLPG) method is introduced for solving the fractional Black-Scholes-Schrodinger equation. Some advantages of MLPG method are the flexibility and simplicity of placing nodes at arbitrary locations compared to the FDM.

The purpose of this paper is to develop numerical algorithm for approximating numerical solutions of fractional Black-Scholes-Schrodinger equation. The meshless local Petrov-Galerkin (MLPG) method is applied for space approximation. The approximation of time fractional derivative is interpreted in the Caputo's senses by a simple quadrature formula. The numerical solution is compared with the semi-classical solution in case of fractional order approach to 1 to verify the results.

2 PROBLEM FORMULATION

The Black-Scholes-Schrodinger equation is financial model used for analyzing option price in real financial market. This equation is interpreting the Black-Scholes equation with arbitrage possibilities in quantum mechanic's view point in the senses of Schrodinger equation. Fractional derivative is used in financial market to describe the probability of log-price, which is a useful to specify the variability in price (Phaochoo et al., 2016). Some occurrence of the Black-Scholes-Schrodinger equation is not satisfied with real financial market. Thus, the fractional Black-Scholes-Schrodinger equation is one of choice for description an occurrence that exists in financial market. The fractional Black-Scholes-Schrodinger equation is following:

$$\frac{\partial^\alpha \psi(x,t)}{\partial t^\alpha} + \frac{1}{2}\sigma^2 \frac{\partial^2 \psi(x,t)}{\partial x^2} + v(x,t)\left(\frac{\partial \psi(x,t)}{\partial x} - \psi(x,t)\right) = 0, \quad (1)$$

with $(x,t) \in \mathbb{R} \times [0,T]$, where $\psi(x,t)$ is wave function which represents the movement of particle at time t, x is space variable which represents position of particle, t is time variable, σ^2 is the variation, $v(x,t)$ represent velocity of particle at time t and α is fractional order, $0 < \alpha < 1$. The time fractional derivative presented herein from Caputo's viewpoint is defined by

$$\frac{\partial^\alpha \psi(x,t)}{\partial t^\alpha} = \frac{1}{\Gamma(1-\alpha)} \int_0^t \frac{\partial \psi(x,\tau)}{\partial \tau} (t-\tau)^{-\alpha}\, d\tau, \quad (2)$$

where $\Gamma(\cdot)$denotes the gamma function. In the case of $\alpha = 1$, Eq. (1) reduce to the original Black-Scholes-Schrodinger equation.

3 SPATIAL DISCRETIZATION

The meshless local Petrov-Galerkin (MLPG) method is used for space approximation. The procedures in MLPG method are described in detail. First, there are creating the local weak form over local sub-domain, Ω_s^i which is a small region taken for each node x_i in the global domain, $\Omega \subseteq \mathbb{R}$. Multiplying test functions, $w_i(x)$ for both side of the fractional Black-Scholes-Schrodinger equation and integrating over a local sub-domain which associate with the point x_i; $i = 1,2,\dots,N$, where N is the number of nodes surrounding point x.

$$\int_{\Omega_s^i} \left(\frac{\partial^\alpha \psi(x,t)}{\partial t^\alpha}\right) w_i(x)\, d\Omega$$

$$= \int_{\Omega_s^i} \left[-\frac{1}{2}\sigma^2 \frac{\partial^2 \psi(x,t)}{\partial x^2} - v(x,t)\left(\frac{\partial \psi(x,t)}{\partial x} - \psi(x,t) \right) \right] w_i(x)\, d\Omega \qquad (3)$$

where Ω_s^i is a local sub-domain which associates with the point x_i, and $w_i(x)$ is a test function that has significant for each node. Second, substituting the trial function, $\psi^h(x,t) = \sum_{j=1}^{N} \phi_j(x)\hat{\psi}_j(t)$ for $\psi(x,t)$ into the local weak form Eq. (3) and rearranging as following

$$\sum_{j=1}^{N} \int_{\Omega_s^i} \phi_j(x) w_i(x) \frac{\partial^\alpha \hat{\psi}_j(t)}{\partial t^\alpha} d\Omega$$

$$= -\frac{1}{2}\sigma^2 \sum_{j=1}^{N} \int_{\Omega_s^i} \phi_{j,xx}(x) w_i(x)\, \hat{\psi}_j(t) d\Omega$$

$$- \sum_{j=1}^{N} \int_{\Omega_s^i} v(x,t) \phi_{j,x}(x)\, w_i(x)\, \hat{\psi}_j(t) d\Omega$$

$$+ \sum_{j=1}^{N} \int_{\Omega_s^i} v(x,t) \phi_j(x)\, w_i(x)\, \hat{\psi}_j(t) d\Omega \quad ; i = 1, \dots, N \qquad (4)$$

where $\phi_j(x)$ is the shape function which is constructed by moving kriging interpolation (Yimnak & Luadsong, 2014), $\hat{\psi}_j(t)$ is value of ψ for position x_j at time, t , $\phi_{j,x}(\cdot) = \frac{\partial \phi_j(\cdot)}{\partial x}$ and $\phi_{j,xx}(\cdot) = \frac{\partial^2 \phi_j(\cdot)}{\partial x^2}$. Finally, the Kronecker delta function is chosen as the test function in each sub-domain:

$$w_i(x) = \begin{cases} 0, & x \neq x_i \\ 1, & x = x_i \end{cases} ; \; i = 1,2, \dots, N.$$

Substituting the test function into Eq. (4) and integrate over sub-domain as following

$$\sum_{j=1}^{N} \phi_j(x_i) \frac{d^\alpha \hat{\psi}_j(t)}{\partial t^\alpha}$$

$$= \sum_{j=1}^{N} \left[-\frac{1}{2}\sigma^2 \phi_{j,xx}(x_i) - v(x_i,t)\left(\phi_{j,x}(x_i) - \phi_j(x_i) \right) \right] \hat{\psi}_j(t) \qquad (5)$$

for i =1,2,…, N. Equation (5) can be written in the matrix form as following

$$\boldsymbol{A} \frac{d^\alpha \boldsymbol{\Psi}}{dt^\alpha} = \boldsymbol{B}\boldsymbol{\Psi} \qquad (6)$$

where

$\boldsymbol{A} = [A_{ij}]_{N\times N}$, $A_{ij} = \phi_j(x_i) = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}$,

$\boldsymbol{B} = [B_{ij}]_{N\times N}$; $B_{ij} = -\frac{1}{2}\sigma^2 \phi_{j,xx}(x_i) - v(x_i,t)\left(\phi_{j,x}(x_i) - \phi_j(x_i) \right)$,

$\boldsymbol{\Psi} = [\hat{\psi}_i]^T = [\hat{\psi}_1, \hat{\psi}_2, \dots, \hat{\psi}_N]^T$,

for $i = 1,2, \dots, N$ and $j = 1,2, \dots, N$. Since the moving kriging interpolation has the delta function property. Therefore, $\boldsymbol{A}$ is the identity matrix. Equation (6) can be written as

$$\frac{d^\alpha \boldsymbol{\Psi}}{dt^\alpha} = \boldsymbol{B}\boldsymbol{\Psi}. \qquad (7)$$

From spatial discretization of fractional Black-Scholes-Schrodinger equation we obtained the system of ordinary differential equation each point on the space. Therefore these systems of ODE are still continuous on time.

4 TEMPORAL DISCRETIZATION

For positive integer N, let $\Delta t = \frac{T}{N}$ be the step size of time variable. The nodal points in the time interval $[0,T]$ are given by $t_n = n\Delta t, n = 0,1,2, \dots, N$. Time fractional derivative with notation $\frac{\partial^\alpha \psi(x_i,t_n)}{\partial t^\alpha}$ is the approximation solution of point $\psi(x_i,t_n)$ at time level n. A simple quadrature formula to obtain the discrete approximation of the time fractional derivative in Caputo's sense (Murio, 2008) is applied in here,

$$\frac{d^\alpha \boldsymbol{\Psi}}{dt^\alpha} = \bar{\sigma}_{\alpha,\Delta t} \sum_{j=1}^{n} \omega_j^{(\alpha)} (\boldsymbol{\Psi}^{n-j+1} - \boldsymbol{\Psi}^{n-j}) + O(\Delta t), \qquad (8)$$

where $\omega_j^{(\alpha)} = j^{1-\alpha} - (j-1)^{1-\alpha}$ and $\bar{\sigma}_{\alpha,\Delta t} = \frac{1}{\Gamma(1-\alpha)} \frac{1}{1-\alpha} \frac{1}{\Delta t^\alpha}$. Hence, $\frac{d^\alpha \boldsymbol{\Psi}}{dt^\alpha} = D_t^{(\alpha)} \boldsymbol{\Psi}^n + O(\Delta t)$, and the first-order approximation method for the computation of Caputo's fractional derivative is given by

$$D_t^{(\alpha)} \boldsymbol{\Psi}^n = \bar{\sigma}_{\alpha,\Delta t} \sum_{j=1}^{n} \omega_j^{(\alpha)} (\boldsymbol{\Psi}^{n-j+1} - \boldsymbol{\Psi}^{n-j}). \qquad (9)$$

Next, substituting Eq. (9) into Eq. (7), it follows that

$$\bar{\sigma}_{\alpha,\Delta t} \sum_{j=1}^{n} \omega_j^{(\alpha)} (\boldsymbol{\Psi}^{n-j+1} - \boldsymbol{\Psi}^{n-j}) = \boldsymbol{B}\boldsymbol{\Psi}^n$$

$$\bar{\sigma}_{\alpha,\Delta t} \omega_1^{(\alpha)} (\boldsymbol{\Psi}^n - \boldsymbol{\Psi}^{n-1})$$

$$= -\bar{\sigma}_{\alpha,\Delta t} \sum_{j=2}^{n} \omega_j^{(\alpha)} (\boldsymbol{\Psi}^{n-j+1} - \boldsymbol{\Psi}^{n-j}) + \boldsymbol{B}\boldsymbol{\Psi}^n. \qquad (10)$$

For $n = 1$, we get

$$\left(\bar{\sigma}_{\alpha,\Delta t} \omega_1^{(\alpha)} \boldsymbol{I} - \boldsymbol{B} \right) \boldsymbol{\Psi}^1 = \bar{\sigma}_{\alpha,\Delta t} \omega_1^{(\alpha)} \boldsymbol{\Psi}^0 \qquad (11)$$

and for $n \geq 2$,

$$\left(\bar{\sigma}_{\alpha,\Delta t} \omega_1^{(\alpha)} \boldsymbol{I} - \boldsymbol{B} \right) \boldsymbol{\Psi}^n$$

$$= \bar{\sigma}_{\alpha,\Delta t} \omega_1^{(\alpha)} \boldsymbol{\Psi}^{n-1} - \bar{\sigma}_{\alpha,\Delta t} \sum_{j=2}^{n} \omega_j^{(\alpha)} (\boldsymbol{\Psi}^{n-j+1} - \boldsymbol{\Psi}^{n-j}). \qquad (12)$$

This paper therefore used the formula in Eq. (11) to approximate at the time level $n = 1$ and also Eq. (12) for $n \geq 2$.

5 NUMERICAL EXPERIMENT AND RESULTS

Since the fractional Black-Scholes-Schrodinger equation have no the exact solution, we take an example of Black-Scholes-Schrodinger equation which has the exact solution. The exact solution of Black-Scholes-Schrodinger equation is called the semi-classical solution. Therefore, the numerical solution of proposed method is compared with the semi-classical solution in case of fractional order, α approach to 1. In this paper, the example case is obtained from Contreras et al. (2010). The semi-classical solution in presence of a time dependent arbitrage bubble $f = f(t)$ can be computed as

$$\pi_{sc}(S,t) = \frac{1}{e^{\rho(t,T)}} \pi_{BS}\left(e^{\rho(t,T)} S, t \right) \qquad (13)$$

where $\pi_{BS}(S,t)$ is the arbitrage-free Black–Scholes solution for the specific option with contract $\Phi(S)$ and $\rho(t,T)$ is the ρ factor. The pure Black–Scholes solution $\pi_{BS}(S,t)$ is given by

$$\pi_{BS}(S,t) = e^{-r(T-t)}\left[1 - N\left(d_2(S,t) \right) \right] \qquad (14)$$

where $N(x)$ is the normal distribution function and

$$d_2(S,t) = \frac{\ln \frac{S}{K} + \left(r - \frac{\sigma^2}{2} \right)(T-t)}{\sigma\sqrt{(T-t)}},$$

K is strike price. Contract function, $\Phi(S)$ is given by

$$\Phi(S) = \begin{cases} 1, & 0 < S < K \\ 0, & K < S \end{cases} . \qquad (15)$$

Again, the fractional Black-Scholes-Schrodinger equation is considered

$$\frac{\partial^{\alpha}\psi(x,t)}{\partial t^{\alpha}}+\frac{1}{2}\sigma^{2}\frac{\partial^{2}\psi(x,t)}{\partial x^{2}}+v(x,t)\left(\frac{\partial\psi(x,t)}{\partial x}-\psi(x,t)\right)=0 \quad (16)$$

$(x,t)\in\mathbb{R}\times[0,T]$ with initial and boundary condition as following

$$\psi(x,T)=\begin{cases}1, & -\infty<x<\ln K-\left(r-\frac{1}{2}\sigma^{2}\right)T\\ 0, & \ln K-\left(r-\frac{1}{2}\sigma^{2}\right)T<x\end{cases} \quad (17)$$

$$\psi(x_1,t)=e^{r(T-t)}\pi_{sc}\left(e^{x_1+\left(r-\frac{1}{2}\sigma^{2}\right)t},t\right)$$
$$\psi(x_N,t)=e^{r(T-t)}\pi_{sc}\left(e^{x_N+\left(r-\frac{1}{2}\sigma^{2}\right)t},t\right). \quad (18)$$

Example
Time step arbitrage bubble case is given by

$$f(t)=\begin{cases}0, & 0<t<T_1\\ H, & T_1<t<T_2\\ 0, & T_2<t<T\end{cases} \quad (19)$$

and ρ factor is given by

$$\rho(t,T)=\begin{cases}(T_2-T_1)\dfrac{(r-\bar{\alpha})H}{\sigma-H}, & 0<t<T_1\\ (T_2-t)\dfrac{(r-\bar{\alpha})H}{\sigma-H}, & T_1<t<T_2\\ 0, & T_2<t<T\end{cases}. \quad (20)$$

In this paper, we analyze a binary put option for $\alpha=0.99$, $\sigma=0.5$, $r=0.01$, $\bar{\alpha}=-0.6$, $T_1=0.3$, $T_2=0.6$, $T=1$ and $H=0.1\sigma$. The numerical results are shown in Figures 1 and 2. Figure 1 shows wave function solved by MLPG method. From Figure 1, dots on space-x represent computational nodes. The computational nodes are non-uniform nodes which is the advantage of MLPG method. In Figures 2, option price from MLPG method is shown by dots while the semi-classical solutions are generated by mesh. From Figure 2, we found that dots are almost on the mesh. It is rather overlapping with mesh. Therefore, we can conclude that the option prices of fractional Black-Scholes-Schrodinger equation (in case of α approach to 1) from MLPG method satisfy with the semi-classical solution.

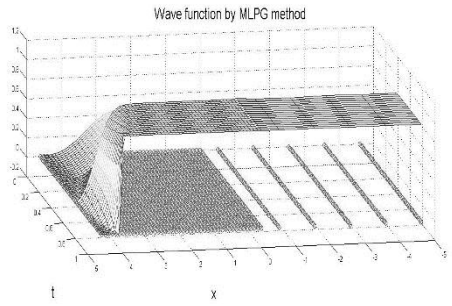


Figure 1: The wave function is solved by MLPG method.

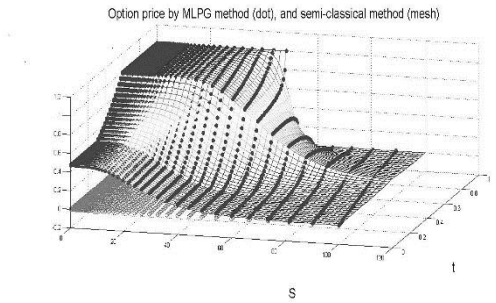


Figure 2: The option price is solved by MLPG method and semi-classical method.

6 CONCLUSIONS

In this paper, the meshless local Petrov-Galerkin (MLPG) method is employed to approximate the solution of fractional Black-Scholes-Schrodinger equation. The Kronecker delta function is chosen as the test function in each sub-domain for simplifying the equation. The effective moving kriging interpolation is applied for constructing nodal shape function. The approximation of time fractional derivative is interpreted in the Caputo's sense by quadrature formula. The numerical solution is compared with the semi-classical solution in case of fractional order approach to 1. The results show that the option prices from MLPG method satisfy with the semi-classical solution. In conclusion, the MLPG method is a new numerical technique for solving the fractional Black-Scholes-Schrodinger equation.

ACKNOWLEDGEMENTS

This research was supported by Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi (KMUTT). The authors would like to thank their advisors for providing advice and taking care of this research. Finally, the authors would like to thank Uttaradit Rajabhat University for providing a scholarship.

REFERENCES

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3), 637-654.

Cen, Z., Huang, J., Xu, A., & Le, A. (2018). Numerical approximation of a time-fractional Black-Scholes equation. *Computers & Mathematics with Applications*, 75, 2874-2887.

Contreras, M., Montalva, R., Pellicer, R., & Villena, M. (2010). Dynamic option pricing with endogenous stochastic arbitrage. *Physica A*, 389, 3552-3564.

Contreras, M., Pellicer, R.,Villena, M., & Ruiz, A. (2010). A quantum model of option pricing: when Black-Scholes meets Schrodinger and its semi-classical limit. *Physica A*, 389, 5447-5459.

Kumar, S., Kumar, D., & Singh, J. (2014). Numerical computation of fractional Black-Scholes equation arising in financial market. *Egyptian Journal of Basic and Applied Sciences*, 1, 177-183.

Murio, D.A. (2008). Implicit finite difference approximation for time fractional diffusion equations. *Computers & Mathematics with Applications*, 56, 1138-1145.

Phaochoo, P., Luadsong, A., & Aschariyaphotha, N. (2016). The meshless local Petrov-Galerkin based on moving Kriging interpolation for solving fractional Black-Scholes model. *Journal of King Saud University- Science*, 28, 111-117.

Song, L., & Wang, W. (2013). Solution of the fractional Black-Scholes option pricing model by finite difference method. *Hindawi Publishing Corporation Abstract and Applied analysis*, 2013, 1-10.

Thamareerat, N., Luadsong, A., & Aschariyaphotha, N. (2017). Stability results of a fractional model for unsteady-state fluid flow problem. *Advances in Difference Equations*, 2017(74), 1-17.

Yimnak, K., & Luadsong, A. (2014). A local integral equation formulation based on moving Kriging Interpolation for Solving Coupled Nonlinear Reaction-Diffusion Equations. *Advance in Mathematical Physics*, 2014, 1-7.

Zhang, H., Liu, F., Turner, I., & Yang, Q. (2016). Numerical solution of the time fractional Black-Scholes model governing European options. *Computers & Mathematics with Applications*, 71, 1772-1783.

New Solitary Wave Solutions for (2+1) Dimensional Chaffee-Infante Equation Using Modified Simple Equation Method

Chutipong Dechanubeksa and Settapat Chinviriyasit*

Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi (KMUTT) 126 Pracha Uthit Road, Bangmod, Thungkhru, Bangkok 10140, Thailand
Email: chutipong2634@gmail.com
*Corresponding Email: settapat.chi@mail.kmutt.ac.th

ABSTRACT

The solutions of nonlinear partial differential equations are challenged and charmed. In this research, the new solitary wave solutions of (2 + 1) dimensional Chaffee-Infante equation are demonstrated by the Modified Simple Equation (MSE) method. This equation is transformed to the series solution form and solved some algebraic system. The new solitary wave solutions of Chaffee-Infante equation are achieved.

Keywords: Chaffee-Infante equation; Modified Simple Equation method; Nonlinear partial differential equation; Solitary wave solutions

1 INTRODUCTION

The nonlinear evolution equations have the important role in nonlinear mathematical and physical science. The traveling wave solutions may well describe the physical phenomena such as vibration, solitons and propogation. There are many several approaches for considering the solitary wave solutions of Nonlinear Partial Differential Equations (NPDEs) such as sine-cosine method (Wazwaz, 2003), Hirota's method (Hirota, 1971), exp-function method (He & Wu, 2006), Kudryashov method (Kudryashov, 1988), F-expansion method (Abdel et al., 2011), hyperbolic tangent method (Malfliet, 1992), (G'/G) - expansion method (Wang et al., 2008), homogeneous balance (HB) method (Wang, 1995), modified simple equation (MSE) method (Jawad et al., 2010), canonical-like transformation method (Cheng, 2009) and trial equation method (Li, 2014).

In 1974, the Chaffee-Infante equation was proposed by Chaffee & Infante (1974) which is well-known arising in mathematical physics (Constantin, 1989). The analytical solutions of this equation were established using Exp-function method by Sakthivel and Chun (2010). Yuanyuan (2018) applied the canonical-like transformation and trial equation methods to discover the traveling wave solutions of this equation. The Geometric structure of the attractor and its bifurcations under perturbation were studied by this equation (see in Carvalho et al. (2012)). Hgele (2011) studied this equation for discovering the metastability. Ricardo (2003) used exact finite-dimensional feedback control via inertial manifold to apply this equation for studying behavior of the solutions of dissipative dynamical systems. In this paper, we consider the (2 + 1) dimensional Chaffee-Infante equation in the following form

$$u_{xt}+\left(-u_{xx}+\alpha u^3-\alpha u\right)_x+\sigma u_{yy}=0 \tag{1}$$

where $u(x,y,t)$ is function of variables x, y and t. α and σ are arbitrary constants.

In 2010, (Jawad et al., 2010) proposed MSE method and applied to find the traveling wave solutions of Fitzhugh-Nagumo equation and Sharma-Tasso-Olver equation. This method was applied in various works such as (Khan, & Akbar, 2013; Khan el al., 2013; Khan & Akbar, 2014a, 2014b). In this paper, the new traveling wave solutions of (2+1) dimensional Chaffee-Infante equation are established using this method. The examples for 3D surfaces of solitary wave solutions are shown by assuming some parameter values.

This work consists of four sections. First section is introduction. The procedure of MSE method is described in Section 2. In Section 3, the analytical solutions and examples of 3D surface for (1) is demonstrated by MSE method. Some conclusions are conferred in last section.

In Section 2, the process of MSE method is briefly described for using construct the exact solutions of NPDEs.

2 MSE METHOD

In this section, the process of MSE method (Jawad et al., 2010; Khan, & Akbar, 2013; Khan el al., 2013; Khan & Akbar, 2014a, 2014b) for constructing the exact solutions of NPDEs is described.

Step 1 : We start with the general form of NPDEs

$$F\left(u,u_x,u_y,u_t,u_{xx},u_{xy},u_{xt},\ldots\right)=0 \tag{2}$$

where u is the function of x, y and t.

Step 2 : The transformation $u(x,y,t)=U(\xi)$ where $\xi=x+y-ct$ is used to transform (2) to ODE

$$G(U,U_\xi,U_{\xi\xi},U_{\xi\xi\xi},\ldots)=0 \tag{3}$$

where c is nonzero constant of wave velocity.

Step 3 : The solution of (3) may be written in the form

$$U(\xi)=\sum_{i=0}^{N}A_i\left(\frac{F'}{F}\right)^i, \tag{4}$$

where $F=F(\xi)$ and A_i are arbitrary constants to be determined.

Step 4 : In order to determine the value N, we replace (3) by (4) for balancing between the highest order derivative and the nonlinear terms.

Step 5 : Substituting N in previous step into (4) to collect all terms which have the same power of F and set its to zero, we construct the system of equations. The values $F(\xi)$, $F'(\xi)$ and $A_i;(i=0,1,2,\ldots N)$ are obtained by solving this system.

Step 6 : Substituting all results in step 5 into (4) where $\xi=x+y-ct$ to find the analytical solutions of NPDEs.

The MSE method is applied to solve the (2+1) dimensional Chaffee-Infante equation (1).

3 SOLUTIONS OF (2+1) DIMENSIONAL CHAFFEE-INFANTE EQUATION

In Section 3, we start with (1) to transform into the following form of ODE by using the wave variable $\xi=x+y-ct$ where c is nonzero constant of wave velocity

$$-cU_{\xi\xi}-U_{\xi\xi\xi\xi}+3\alpha U^2U_\xi-\alpha U_\xi+\sigma U_{\xi\xi}=0. \tag{5}$$

Integrating (5) with zero constant, yields

$$-cU_\xi-U_{\xi\xi\xi}+\alpha U^3-\alpha U+\sigma U_\xi=0. \tag{6}$$

The solution of (6) is defined by (4). To determine the value N, substituting (4) in (6) in order to balancing between the highest order derivative and the nonlinear terms, this gives $N=1$. Therefore,

$$U(\xi)=A_0+A_1\left(\frac{F'}{F}\right) \tag{7}$$

Differentiating

$$U_{\xi} = A_1\left[\frac{F''}{F} - \left(\frac{F'}{F}\right)^2\right], \tag{8}$$

$$U_{\xi\xi\xi} = A_1\left[\frac{F'''}{F} - 3\left(\frac{F'F''}{F}\right)^2 + 2\left(\frac{F'}{F}\right)^3\right], \tag{9}$$

$$U^3 = A_0^3 + 3A_0^2\left(A_1\frac{F'}{F}\right) + 3A_0\left(A_1\frac{F'}{F}\right)^2 + \left(A_1\frac{F'}{F}\right)^3. \tag{10}$$

Substituting (7)-(10) into (6), we obtain

$$\begin{aligned} &-cA_1\left[\frac{F''}{F} - \left(\frac{F'}{F}\right)^2\right] + A_1\left[\frac{F'''}{F} + 3\frac{F'F''}{F^2} + 2\left(\frac{F'}{F}\right)^3\right] \\ &+\alpha\left[A_0^3 + 3A_0^2\left(A_1\frac{F'}{F}\right) + 3A_0\left(A_1\frac{F'}{F}\right)^2 + \left(A_1\frac{F'}{F}\right)^3\right] \\ &-\alpha\left[A_0 + A_1\frac{F'}{F}\right] + \sigma A_1\left[\frac{F''}{F} - \left(\frac{F'}{F}\right)^2\right] = 0. \end{aligned} \tag{11}$$

Collecting the coefficient of all term of F and set to be zero, this gives

$$F^0 : \alpha A_0^3 - \alpha A_0 = 0, \tag{12}$$

$$F^{-1} : -cA_1F'' - A_1F''' + 3\alpha A_0^2A_1F' - \alpha A_1F' + \alpha A_1F'' = 0, \tag{13}$$

$$F^{-2} : cA_1(F')^2 + 3A_1F'F'' + 3\alpha A_0A_1^2(F')^2 - \sigma A_1(F')^2 = 0, \tag{14}$$

$$F^{-3} : -2A_1(F')^3 + \alpha A_1^3(F')^3 = 0. \tag{15}$$

Solving the system (12)-(15), we get three cases

Case 1 : $A_0 = 0, \quad A_1 = \pm\sqrt{\frac{2}{\alpha}}, \quad F(\xi) = c_1 + c_2\xi + c_3e^{\left(\alpha+\frac{3}{c-\sigma}-c\right)\xi}$, and $F'(\xi) = c_2 + \left(\alpha + \frac{3}{c-\sigma} - c\right)c_3e^{\left(\alpha+\frac{3}{c-\sigma}-c\right)\xi}$.

Case 2 : $A_0 = \pm 1, \quad A_1 = \pm\sqrt{\frac{2}{\alpha}}, \quad F(\xi) = c_1 + c_2\xi + c_3e^{\left(\alpha-c-\frac{6\alpha}{c+3\sqrt{2\alpha}-\sigma}\right)\xi}$, and $F'(\xi) = c_2 + \left(\alpha - c - \frac{6\alpha}{c+3\sqrt{2\alpha}-\sigma}\right)c_3e^{\left(\alpha-c-\frac{6\alpha}{c+3\sqrt{2\alpha}-\sigma}\right)\xi}$.

Case 3 : $A_0 = \pm 1, \quad A_1 = \mp\sqrt{\frac{2}{\alpha}}, \quad F(\xi) = c_1 + c_2\xi + c_3e^{\left(\alpha-c-\frac{6\alpha}{c-3\sqrt{2\alpha}-\sigma}\right)\xi}$, and $F'(\xi) = c_2 + \left(\alpha - c - \frac{6\alpha}{c-3\sqrt{2\alpha}-\sigma}\right)c_3e^{\left(\alpha-c-\frac{6\alpha}{c-3\sqrt{2\alpha}-\sigma}\right)\xi}$.

Therefore the exact solutions of (1) are obtained by substituting all results of Cases 1, 2, 3 and $\xi = x + y - ct$ into (7)

Case 1 :

$$u_1(x,y,t) = \gamma\left(\sqrt{\frac{2}{\alpha}}\left[\frac{c_2 + \left(\alpha + \frac{3}{c-\sigma} - c\right)c_3e^{\left(\alpha+\frac{3}{c-\sigma}-c\right)(x+y-ct)}}{c_1 + c_2(x+y-ct) + c_3e^{\left(\alpha+\frac{3}{c-\sigma}-c\right)(x+y-ct)}}\right]\right), \tag{16}$$

Case 2 :

$$u_2(x,y,t) = \gamma\left(1 + \sqrt{\frac{2}{\alpha}}\left[\frac{c_2 + \left(\alpha - c - \frac{6\alpha}{c+3\sqrt{2\alpha}-\sigma}\right)c_3e^{\left(\alpha-c-\frac{6\alpha}{c+3\sqrt{2\alpha}-\sigma}\right)(x+y-ct)}}{c_1 + c_2(x+y-ct) + c_3e^{\left(\alpha-c-\frac{6\alpha}{c+3\sqrt{2\alpha}-\sigma}\right)(x+y-ct)}}\right]\right), \tag{17}$$

Case 3 :

$$u_3(x,y,t) = \gamma\left(1 - \sqrt{\frac{2}{\alpha}}\left[\frac{c_2 + \left(\alpha - c - \frac{6\alpha}{c-3\sqrt{2\alpha}-\sigma}\right)c_3e^{\left(\alpha-c-\frac{6\alpha}{c-3\sqrt{2\alpha}-\sigma}\right)(x+y-ct)}}{c_1 + c_2(x+y-ct) + c_3e^{\left(\alpha-c-\frac{6\alpha}{c-3\sqrt{2\alpha}-\sigma}\right)(x+y-ct)}}\right]\right), \tag{18}$$

where $\alpha \neq 0$ and σ are arbitrary constants and $\gamma = \pm 1$.

The graph of exact solutions for case 1, 2 and 3 with some parameters are shown in the Figure 1, 2, and 3 respectively.

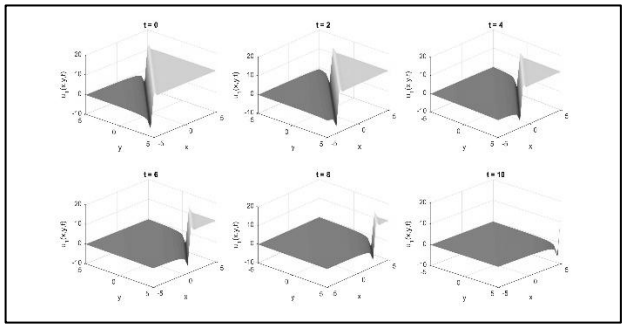


Figure 1 : Example of solitary wave solution u_1 of (2+1) dimensional Chaffee-Infante equation with $\gamma = 1$, $\alpha = 0.5$, $\sigma = 0.5$, $c = 1$, $c_1 = 0$, $c_2 = 1$, $c_3 = 1$, $-5 \le x \le 5$, $-5 \le y \le 5$ and $t = 0, 2, 4, 6, 8, 10$.

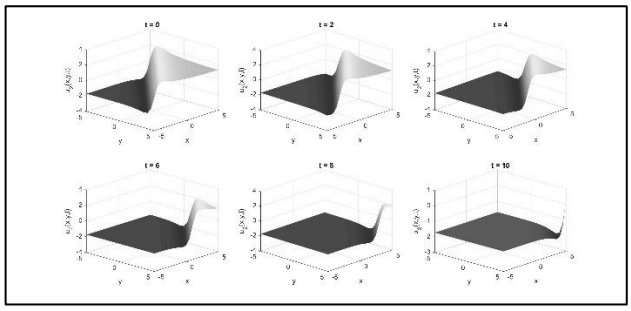


Figure 2 : Example of solitary wave solution u_2 of (2+1) dimensional Chaffee-Infante equation with $\gamma = 1$, $\alpha = 0.5$, $\sigma = 0.5$, $c = 1$, $c_1 = 0$, $c_2 = 1$, $c_3 = 1$, $-5 \le x \le 5$, $-5 \le y \le 5$ and $t = 0, 2, 4, 6, 8, 10$.

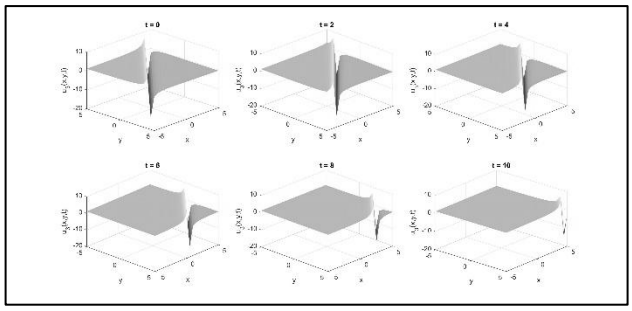


Figure 3 : Example of solitary wave solution u_3 of (2+1) dimensional Chaffee-Infante equation with $\gamma = 1$, $\alpha = 0.5$, $\sigma = 0.5$, $c = 1$, $c_1 = 0$, $c_2 = 1$, $c_3 = 1$, $-5 \le x \le 5$, $-5 \le y \le 5$ and $t = 0, 2, 4, 6, 8, 10$.

4 CONCLUSION AND FUTURE WORK

The solution of (2+1) dimensional Chaffee-Infante equation (1) are explored by the MSE method. These change equation to an ODE (6). Suppose the solution of (6) in the form (4) and talking balancing to obtain (7). Replacing (7) and its derivatives, the system of (12)-(15) are reached. Then, the solutions of (1) are (16)-(18). The MSE method shows that this method is effective for solving the (2+1) dimensional Chaffee-Infante equation.

REFERENCES

Abdel-Razek, M.A., Seddeek, A.K., & Abdel, N.H. (2011). New exact Jacobi elliptic function solution for nonlinear equations using F-expansion method. *Studies in Mathematical Sciences*, 2(1), 88-95.

Carvalho, A., Langa, J., & Robinson, J. (2012). Structure and Bifurcation of Pullback Attractors in a Non-Autonomous Chafee-Infante Equation. *Proceedings of the American Mathematical Society*, 140(7), 2357-2373.

Chaffee, N., & Infante, E.F. (1974). A Bifurcation Problem for a Nonlinear Partial Differential Equation of Parabolic Type. *Applicable analysis*, 4(1), 17-37

Cheng-shi, L. (2009). Canonical-Like Transformation Method and Exact Solutions to a Class of Diffusion Equations. *Chaos Solitons Fractals*, 42(1), 441-446.

Constantin, P. (1989). *Integral Monifold and Intertial Manifolds for Disspasive Partial Equation*. New-York: Springer-Verlag.

He, J., & Wu, X. (2006). Exp-Function Method for Nonlinear Wave Equations. *Chaos, Solitons and Fractals*, 30, 700-708.

Hgele, M.A. (2011). *Metastability of the Chafee-Infante Equation with Small Heavy-Tailed Lvy Noise*. Diss. Humboldt-Universitt zu Berlin, Mathematisch- Naturwissenschaftliche Fakultt II.

Hirota, R. (1971), Exact Solutions of the Korteweg-De Vries Equation for Multiple Collisions of Solitons. *Physical Review Letters*, 27(18), 1192-1194.

Jawad, A.J.M., Petkovic, M.D., & Biswas, A. (2010). Modified Simple Equation Method for Nonlinear Evolution Equations. *Applied Mathematics and Computation*, 217(2), 869-877.

Khan, K., & Akbar, M. (2013). Exact and Solitary Wave Solutions for the Tzitzeica-Dodd-Bullough and the Modified KdV-Zakharov-Kuznetsov Equations Using the Modified Simple Equation Method. *Ain Shams Engineering Journal*, 4(4), 903-909.

Khan, K., Akbar, M., & Alam, M. (2013). Traveling Wave Solutions of the Non-Linear Drinfel'd-Sokolov-Wilson Equation and Modified Benjamin-Bona-Mahony Equations. *Journal of Egyptian Mathematical Society*, 21(3), 233-240.

Khan, K., & Akbar, M. (2014a). Traveling Wave Solutions of the (2+1)-Dimensional Zoomeron Equation and the Burgers Equations via the MSE Method and the Exp-Function Method. *Ain Shams Engineering Journal*, 5(1), 247-256.

Kudryashov, N.A. (1988). Exact Soliton Solutions of the Generalized Evolution Equation of Wave Dynamics. *Journal of Applied Mathematics and Mechanics*, 52, 361-365.

Li, Y. (2014). Trial Equation Method for Solving the Improved Boussinesq Equation. *Advances in Pure Mathematics*, 4, 47-52.

Malfliet, W. (1992). Solitary Wave Solutions of Nonlinear Wave Equation. *American Journal of Physics*, 60, 650-654.

Ricardo, R. (2003). Exact Finite Dimensional Feedback Control via Inertial Manifold Theory with Application to the Chafee-Infante Equation. *Journal of Dynamics and Differential Equations*, 15(1), 61-86.

Sakthivel, R., & Chun, C. (2010). New Soliton Solutions of Chaffee-Infante Equations Using the Exp-Function Method. *Zeitschrift fr Naturforschung A*, 65, 197-202.

Wang, M. (1995). Solitary Wave Solution for Variant Boussinesq Equation. *Physics Letters A*, 199(3-4), 169-172.

Wang, M., Li, X., & Zhang, J. (2008). The (G'/G) -Expansion Method and Traveling Wave Solutions of Nonlinear Evolution Equations in Mathematical Physics. *Physics Letter A*, 372(4), 417-423.

Wazwaz, A.M. (2003). Compacton Solutions of the Kawahara-Type Nonlinear Dispersive Equation. *Applied Mathematics and Computation*, 145(1), 133-150.

Yuanyuan, M. (2018). Exact Solutions to (2+1)-Dimensional Chaffee-Infante Equation. *Pramana*, 91:9.

The Modified Boxplot for Outlier Detection

Mintra Promwongsa*, Wuttichai Srisodaphol and Prem Jansawang

Department of Statistics, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand
*Corresponding Email: mintaa.pws@gmail.com
Email: wuttsr@kku.ac.th
Email: prem@kku.ac.th

ABSTRACT

The aim of this study is to propose a new modified boxplot for outlier detection based on symmetry and skewed data which is called the MK boxplot. This MK boxplot is modified from Kimber's boxplot by using the ratio of lower split interquartile range and upper split interquartile range into the fences of the boxplot. The performance of the boxplot is evaluated by the mean percentage of detected outliers in three cases of simulated data (truncated, uncontaminated and contaminated data) and real data. Furthermore, the existing boxplots for outlier detection are used to make a comparison with the MK boxplot as well. The results from simulated data show that the MK boxplot performs well for symmetric and skewed data when sample size is greater than 30. However, the MK boxplot has better performance than the others for skewed data. Moreover, when the MK boxplot is applied to the real data, it efficiently detects outliers as the shape of real data.

Keywords: boxplot; detection outlier; skewed distribution; split interquartile range

1 INTRODUCTION

Outlier is an observation which is inconsistent with or deviate from the majority of the data. It is very large or very small when it is compared with the other observations in the data set. Outlier might occur from incorrect measurements, including data entry errors, or different population. Outlier may has a negative influence on the data analysis, e.g. outlier goes against the normality of data, increases in variance and reduces the power of statistical tests. Therefore, outlier detection method is considered as an important step of management data before analyzing the data. The traditional method for outlier detection is the boxplot, which was proposed by Tukey (1977). The outliers are labeled by the observations outside the interval called the fence, $[Q_1 - 1.5\text{IQR}, Q_3 + 1.5\text{IQR}]$ where Q_1 and Q_3 stands for the first and the third quartiles, respectively and IQR stands for the interquartile range.

Unfortunately, several researchers have reported that the Tukey's boxplot is only effective for symmetric data. In contrast, the number of observations was marked as potential outliers were immoderate for skewed data. (Hubert & Vandervieren, 2008) Generally, some of the marked observations were presumed to occur naturally in skewed data rather than the real outliers. For instance, when considering Tukey's boxplot with standard normal distribution, the lower and upper fence covers 99.3% of all observations. Hence, only 0.7% or 0.35% in each side of the data were outside the upper or lower fences and were labeled as outliers. When applying Tukey's boxplot to data for right skewed distribution such as a chi-squared distribution with one degree of freedom, percentage of data which were outside the upper fence was up to 7.56% of the data. In order to improve the performance of boxplot, many researchers have attempted to modify boxplot for applying with skewed data. Kimber (1990) proposed the fences of boxplot for skewed data by using the lower and upper parts of the interquartile range that was split at the median, called the split interquartile range (SIQR) instead of using IQR. Carling (2000) replaced the first and third quartiles in fences of Tukey (1977) by the median and suggested that the constant 1.5 in fences was not fixed since it depended on sample size. He also proposed the reasonable constant was 2.3 for obtaining a percentage of an outlier which was less affected by the sample size. Barnett & Cohen (2000) proposed the modified boxplot based on lognormal distribution to solve problems of right censoring and high skewness in lifetime data. Hubert & Vandervieren (2008) proposed the adjusted boxplot by using a robust measure of skewness, namely a medcouple (MC) which was introduced by Brys et al. (2004). They also used the families of skewed distributions for choosing the appropriate constant to insert into exponential terms of fences for efficient applying with skewed data. Walker & Chakraborti (2013) extended the fences of Tukey (1977) and used the concept of splitting interquartile range of Kimber (1990) to insert the ratios of split interquartile range into the fences for applying to skewed data. Adil & Irshad (2015) proposed the modified boxplot by adjusting the boxplot of Hubert & Vandervieren (2008) for solving extreme fences problem by incorporating a moment coefficient of skewness to construct lower and upper fence instead of using a constant. Babura et al. (2016) extended the adjusted boxplot of Hubert & Vandervieren (2008) by using the Bowley coefficient which is a robust measure of skewness and estimated the constant on lower and upper fences which were found by simulating an extreme data from Generalized Extreme Value Distribution (GEV).

In this study, we propose the modified boxplot by using the ratio of split interquartile range for constructing the proper fences, which is called the MK boxplot. The MK boxplot improve the performance of detection outliers with any data regardless of the distributions. For evaluation, simulated and real data are used in various situations and then the three existing boxplots for outlier detection, namely Tukey's boxplot, Kimber's boxplot and Hubert's boxplot are used to make a comparison with the MK boxplot as well.

2 METHODS

In this study, the MK boxplot is modified from Kimber's boxplot (1990) by using the ratio of lower split interquartile range and upper split interquartile range into the fences of the boxplot. The performance measure and the processes of evaluation the MK boxplot based on simulation study and real data are obtained. The details are explained as follows.

2.1 Fences of boxplot modification

The MK boxplot is modified from boxplot which was proposed by Kimber (1990), the lower and upper fences for detecting outliers of Kimber's boxplot was defined as

$$[Q_1 - 1.5(2\text{SIQR}_L),\ Q_3 + 1.5(2\text{SIQR}_U)] \quad (1)$$

where Q_1 and Q_3 are the first and third quartile, respectively,

SIQR_L is a lower split interquartile range $(Q_2 - Q_1)$,

SIQR_U is an upper split interquartile range $(Q_3 - Q_2)$.

The proposed fences of a boxplot are constructed by modifying SIQR. Generally, SIQR_L and SIQR_U are equal for symmetric distribution but not true for skewed distribution. SIQR_U is larger than SIQR_L for right skewed distribution. In contrast, SIQR_L is larger than SIQR_U for left skewed distribution. So, in order to receive the flexible fences as the skewness of the data, we use the ratio of SIQR_U and SIQR_L that could be more of a measured spread of the data than the

distance like IQR or SIQR for inserting into the both fences. The proposed fence of the MK boxplot is given by

$$\left[Q_1 - 1.5(2\text{SIQR}_L)\frac{\text{SIQR}_L}{\text{SIQR}_U},\ Q_3 + 1.5(2\text{SIQR}_U)\frac{\text{SIQR}_U}{\text{SIQR}_L} \right] \quad (2)$$

where Q_1 and Q_3 are the first and third quartile, respectively,

SIQR_L is a lower split interquartile range $(Q_2 - Q_1)$,

SIQR_U is an upper split interquartile range $(Q_3 - Q_2)$.

2.2 Evaluation

2.2.1 Simulated data

To compare the MK boxplot with the existing boxplots, namely Tukey's boxplot, Kimber's boxplot and Hubert's boxplot, we use the mean percentage of detected outliers for three cases of simulated data, i.e. truncated, uncontaminated and contaminated data. In the comparison study, standard normal distribution is selected as symmetric data, chi-squared distribution and F distribution are selected as skewed data. χ_1^2, F(10,10) and F(90,10) are moderate skewed, since the coefficient of skewness is less than 0.6, and χ_5^2, χ_{10}^2, χ_{20}^2 F(10,90) and F(90,90) are mildly skewed, since the coefficient of skewness is around 0. The simulation study of each case has the following steps.

Case I: Truncated data

Step 1 The data from N(0,1), χ_1^2, χ_5^2, χ_{20}^2, F(90,90) and F(10,10) with sample size 10(10)200, 250(50)500 and 600(100)1000 are generated.

Step 2 In each distribution, the lower and upper fences of each boxplot are computed.

Step 3 The data from step 1 are trimmed by 40% (i.e. 20% on both sides).

Step 4 In each boxplot, a number of outliers which are observations from data in step 3 that fall outside the fences are recorded.

We repeat step 1 to 4 for 1000 iterations.

Step 5 The mean of detected outliers from step 4 is computed by

$$\text{mean of detected outliers} = \frac{\text{number of detected outliers}}{1000}.$$

Step 6 The mean percentage of detected outliers of each sample size is computed by

$$\text{mean percentage of detected outliers} = \frac{\text{mean of detected outliers}}{n} \times 100.$$

Case II: Uncontaminated data

Step 1 The data from N(0,1), χ_1^2, χ_5^2, χ_{10}^2, χ_{20}^2, F(90,10), F(10,90), F(90,90) and F (10,10) with sample size 10(10)200, 250(50)500 and 600(100)1000 are generated.

Step 2 In each distribution, the lower and upper fences of each boxplot are computed.

Step 3 In each boxplot, a number of outliers which are observations from data in step 1 that fall outside the fences are recorded.

We repeat step 1 to 3 for 1000 iterations.

Step 4 The mean of detected outliers from step 3 is computed by

$$\text{mean of detected outliers} = \frac{\text{number of detected outliers}}{1000}.$$

Step 5 The mean percentage of detected outliers of each sample size is computed by

$$\text{mean percentage of detected outliers} = \frac{\text{mean of detected outliers}}{n} \times 100.$$

Case III: Contaminated data

Step 1 The data from N(0,1), χ_1^2, χ_5^2, χ_{20}^2, F(90,90) and F(10,10) with sample size 10, 30, 50, 100, 200, 300, 400, 500 and 1000 are generated.

Step 2 In each distribution, the lower and upper fences of each boxplot are computed.

Step 3 The 5% upper and lower tails of the data for standard normal distribution and 5% only upper tail of the data for chi-squared and F distributions from step 1, are replaced by the various contaminators which are multiplied the 5% of tail data by 2, 5 and 10, respectively.

Step 4 In each boxplot, a number of outliers which are observations from data in step 3 that fall outside the fences are recorded.

We repeat step 1 to 4 for 1000 iterations.

Step 5 The mean of detected outliers from step 4 is computed by

$$\text{mean of detected outliers} = \frac{\text{number of detected outliers}}{1000}.$$

Step 6 The mean percentage of detected outliers of each sample size is computed by

$$\text{mean percentage of detected outliers} = \frac{\text{mean of detected outliers}}{n} \times 100.$$

2.2.2 Real data

In order to show the performance of the MK boxplot and compare with existing boxplots, we construct histogram and scatter plot for considering the number of observations which is far from the majority of the data that may be the potential outliers. After that, we compute the lower and upper fences of each boxplot for detecting outliers. The number of detected outliers from each boxplot is compared with the potential outliers in histogram and scatter plot as well. We use two real data sets as following.

1) Indian Liver Patient data set (Bendi et al., 2012) contains Alamine Aminotransferase of 416 liver patients and 167 non liver patients that was collected from north east of Andhra Pradesh, India.

2) Facebook metrics data set (Moro et al., 2016) contains the number of people who clicked anywhere in all posts published in the Facebook's page of 500 worldwide renowned cosmetic brand between the 1 January 2014 to 31 December of 2014.

2.3 Performance measure

The criterion for considering the acceptable mean percentage of detected outliers is as follows. For truncated data, the mean percentage of detected outliers should be about 0% of sample size. For uncontaminated data, the mean percentage of detected outliers should be less than 5% of sample size. For contaminated data, the mean percentage of detected outliers should be about 5% of sample size (the number of contaminators) which are inserted into data. This criterion is referred from Hubert & Vandervieren (2008) and Babura et al. (2016).

3 RESULTS

3.1 Truncated data

In this case, 40% of the data are trimmed (i.e. 20% on both side) so the efficient boxplot should detect 0% outlier. The mean percentages of detected outliers of truncated symmetric and skewed data were obtained. The results show that all boxplots detect 0% outlier when sample size is greater than 30. Hence, all boxplots are the same efficiency when sample size is greater than 30. Therefore, we can use any boxplots for outlier detection.

3.2 Uncontaminated data

In this case, the efficient boxplot should detect outlier less than 5% of sample size. (Babura et.al, 2016) The mean percentages of detected outliers of symmetric and skewed data were obtained. The results show that when data is symmetric or mildly skewed (N(0,1), χ_5^2, χ_{10}^2, χ_{20}^2, F(10,90) and F(90,90)), all boxplots detect outliers less than 5% of sample size when sample size is greater than 30. However, when data is moderated skewed (χ_1^2, F(10,10) and F(90,10)), MK boxplot detect outliers less than 5% of sample size when sample size is greater than 30 while Tukey's boxplot and Kimber's boxplot immoderately detect outliers over 5% for all sample size. The Mean percentages of detected outliers from different distributions with sample size 20, 30, 50, 100, 500 and 1000 are illustrated in Figure 1.

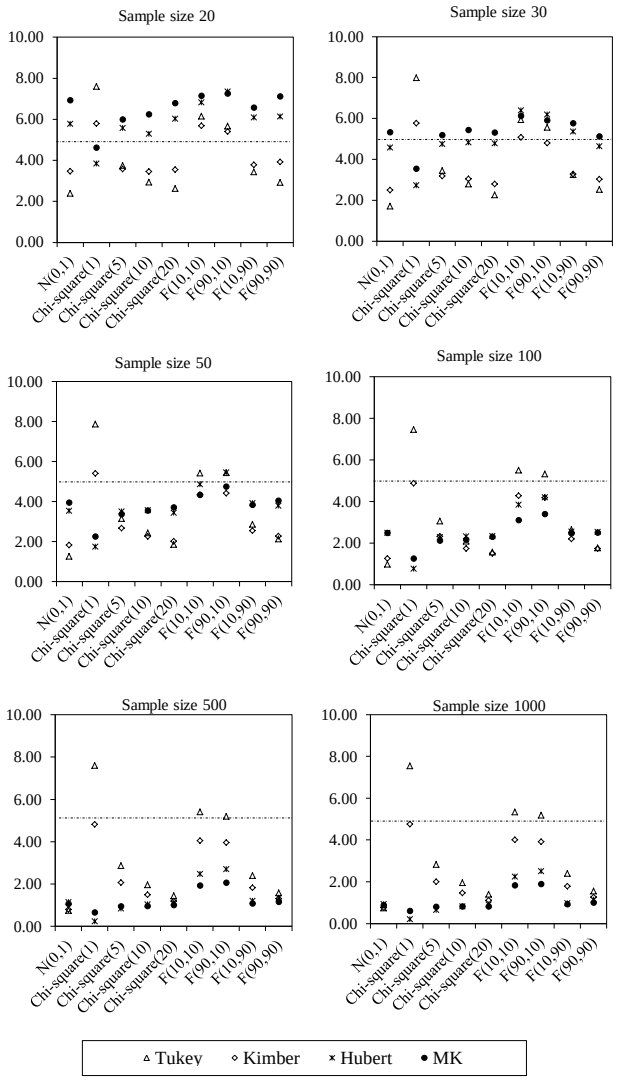
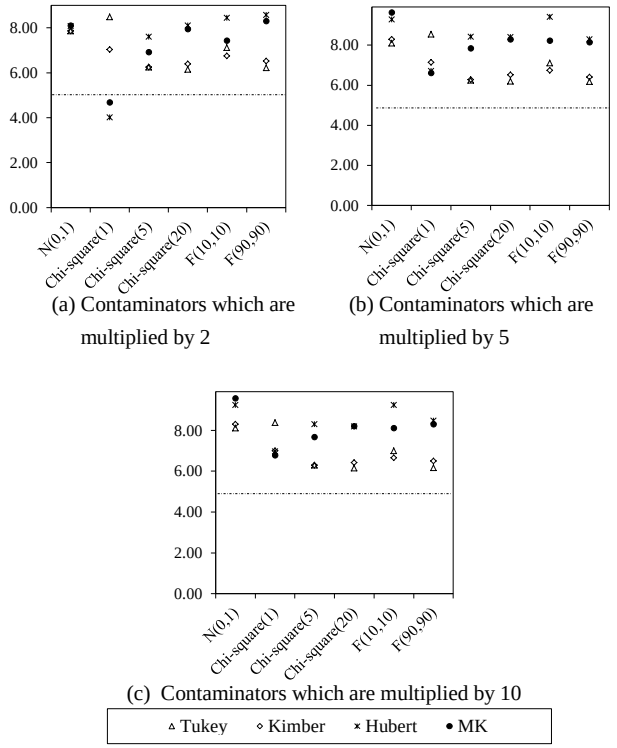


Figure 1: Mean percentages of detected outliers from different distributions.

Figure 1 clearly shows that when data is symmetric or mildly skewed, we can use any boxplots for outlier detection. However, when data is skewed, the MK boxplot shows better performance than the existing boxplots. We give a preference to MK boxplot when sample size is greater than 30.

3.3 Contaminated data

In this case, the 5% lower and upper tails of the standard normal distribution, and only 5% upper tail of the chi-squared and F distributions are replaced by the various contaminators which are multiplied the 5% of tail data by 2, 5 and 10, respectively. The efficient boxplot should detect outliers is about 5% of sample size (the number of contaminators) which are inserted into data set. The mean percentages of detected outliers of different symmetric and skewed data were obtained. The results show that when data is symmetric or mildly skewed (N(0,1), χ^2_5, χ^2_{20} and F(90,90)) for all cases of the contaminators, all boxplots detect outliers close to 5% of sample size as the number of contaminators when sample size 200, 300, 400, 500 and 1000. Hence, boxplots are the same efficiency for symmetric data. Therefore, we can use any boxplots for outlier detection. However, when data is moderated skewed (χ^2_1 and F (10,10)) for all cases of contaminators. The MK boxplot detect outliers close to 5% of sample size as the number of contaminators than the other boxplots when sample size is greater than 30. Moreover, the MK boxplot succeeds in detecting 5% outliers for all distributions when sample size 200, 300, 400, 500 and 1000 while other boxplots fail in detecting outliers for all sample size. The results with sample size 50, 100 and 500 are illustrated in Figures 2-4.

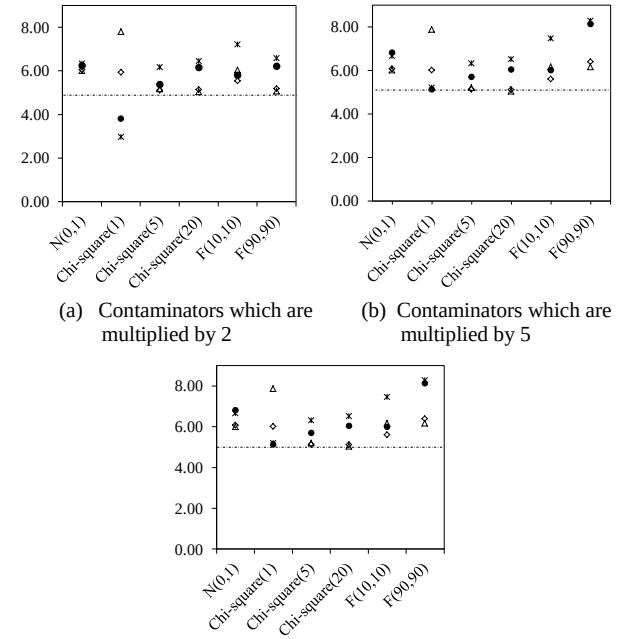


(a) Contaminators which are multiplied by 2

(b) Contaminators which are multiplied by 5

(c) Contaminators which are multiplied by 10

Figure 2: Mean percentage of detected outliers for 5% of upper tail of different distributions multiply by various contaminators, when sample size $n = 50$.

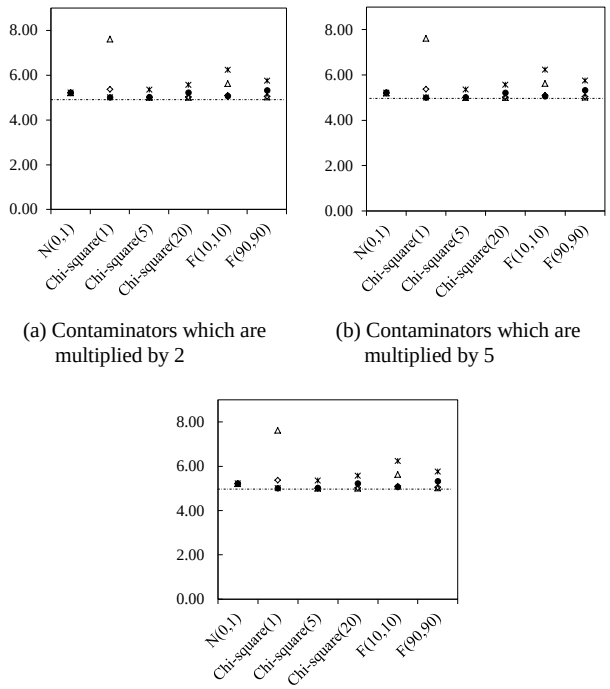


(a) Contaminators which are multiplied by 2

(b) Contaminators which are multiplied by 5

(c) Contaminators which are multiplied by 10

Figure 3: Mean percentage of detected outliers for 5% of upper tail of different distributions multiply by various contaminators, when sample size $n = 100$.



(a) Contaminators which are multiplied by 2

(b) Contaminators which are multiplied by 5

(c) Contaminators which are multiplied by 10

Figure 4: Mean percentage of detected outliers for 5% of upper tail of different distributions multiply by various contaminators, when sample size $n = 500$

From Figures 2-4, we give a preference to MK boxplot for detecting outlier in moderate skewed data. For symmetric and mildly skewed data, we can use any boxplots for outlier detection when sample size is greater than 30.

3.4 Real data

In this section, the MK boxplot and three existing boxplots are applied to real data sets. The descriptive statistics such as minimum, maximum, mean, the first quartile (Q_1), the third quartile (Q_3), median, medcouple (MC) which is robust measure of skewness, histogram and scatter plot are showed for consider the number of detected outliers from each boxplot is compared with the potential outliers in histogram and scatter plot as well.

1) Indian Liver Patient data set: we consider the Alamine Aminotransferase of 416 liver patients and 167 non liver patients from north east of Andhra Pradesh, India. The descriptive statistics of this data are showed as follows.

min = 10 max = 4929 mean = 109.91 median = 42

$Q_1 = 25$ $Q_3 = 87$ MC = 0.541

The histogram and the scatter plot of this data are showed in Figure 5.

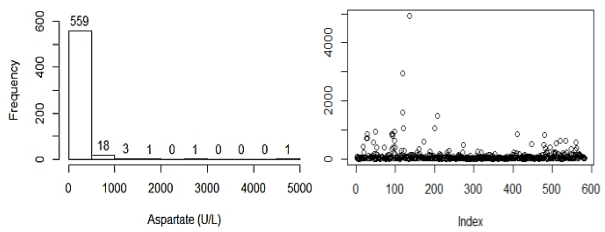


Figure 5: Histogram (a) and scatter plot (b) of Indian Liver Patient data

From histogram in Figure 5 (a) and MC value, we obtain that this data is right skewed distribution and the majority of the data is between 0 and 500. When considering in the histogram, the number of observations which is far from the majority of the data is about 24 values. They are may be the potential outliers. After that we compute the lower and upper fences of each boxplot for detecting outlier, the results are showed in Table 1 and Figure 6.

Table 1: The lower fences, upper fences of Indian Liver Patient data according to the four boxplots.

Method	[lower fence, upper fence]
Tukey	[-68.000, 180.000]
Kimber	[-26.000, 222.000]
Hubert	[14.317, 558.327]
MK	[5.733, 444.353]

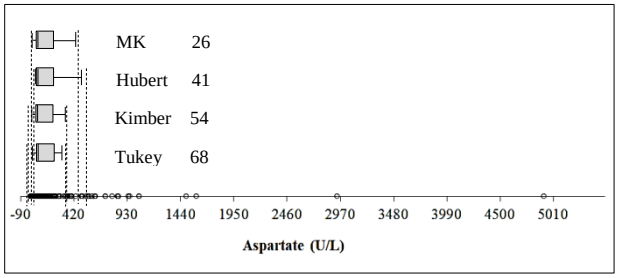


Figure 6: The detection outliers of Tukey, Kimber, Hubert and MK boxplots of Indian Liver Patient data

From Figure 6, we can see that Tukey's boxplot, Kimber's boxplot and Hubert's boxplot immoderately detect outliers. Meanwhile, MK boxplot detect total 26 outliers. When we consider the number of the potential outliers in histogram, we give a preference to MK boxplot for properly detect outliers as the shape of real data.

2) Facebook metrics data set: we consider the number of people who clicked anywhere in all posts in the Facebook's page of 500 worldwide renowned cosmetic brand. The descriptive statistics of this data are showed as follows.

min = 9 max = 11328 mean = 798.78 median = 551.50

$Q_1 = 332.50$ $Q_3 = 955.50$ MC = 0.357

The histogram and the scatter plot of this data are showed in Figure 7.

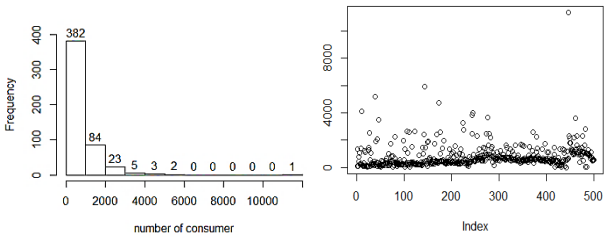


Figure 7: Histogram (a) and scatter plot (b) of Facebook metrics data

From the histogram in Figure 7 (a) and MC value, we obtain that this data is right skewed distribution. When considering in the scatter plot in Figure 7 (b), we found that the number of observations which is far from the majority of the data is about 11 values. They are may be the potential outliers. After that we compute the lower and upper fences of each boxplot, for detecting outlier, the results are showed in Table 2 and Figure 8.

Table 2: The lower fences, upper fences of Facebook metrics data according to the four boxplots.

Method	[lower fence, upper fence]
Tukey	[-602.000, 1890.000]
Kimber	[-324.500, 2167.500]
Hubert	[108.457, 3683.002]
MK	[-23.646, 3191.336]

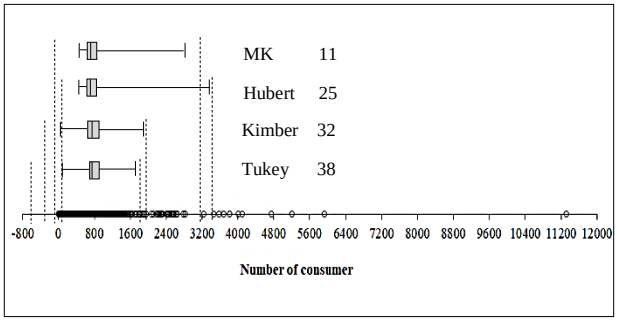


Figure 8: The detection outliers of Tukey, Kimber, Hubert and MK boxplots of Facebook metrics data

From Figure.8, we can see that Tukey's boxplot, Kimber's boxplot and Hubert's boxplot immoderately detect outliers. MK boxplot detect 11 outliers. When we considering the number of potential outliers in histogram and scatterplot, we obtain that MK boxplot properly detect outliers as the shape of real data than other boxplots.

4 CONCLUSIONS AND DISCUSSIONS

In this study, we propose the modified boxplot, namely MK boxplot for detecting outliers in symmetric and skewed data. The performance of the MK boxplot is compared with existing boxplots by the mean percentage of detected outliers based on simulated and real data. The efficient boxplots for any sample size and any shape of the data are showed in Table 3.

Table 3: The efficient boxplots for any shape of the data.

Sample size *(n)*	Symmetric data	Skewed data	
		Moderate skewed	Mildly skewed
$n \leq 30$	Tukey Kimber	MK	Tukey Kimber
$n > 30$	Tukey Kimber Hubert MK	MK Hubert	Tukey Kimber Hubert MK

From Table 3, all boxplots are the same efficiency when sample size is greater than 30 for symmetric and mildly skewed data. We can use any boxplots for outlier detection. In case of sample size of data is less than or equal to 30, we give a preference to Tukey's boxplot and Kimber's boxplot. When data is moderated skewed and sample size is greater than 30, Hubert's boxplot and MK boxplot are effective than the others boxplots. However, MK boxplot is the best choice for outlier detection. Hence, it is effective and easier to compute the fences than the Hubert's boxplot as well.

ACKNOWLEDGEMENTS

The authors would like to express my gratitude to Graduate School and Department of Statistics, Faculty of Science, Khon Kaen University for financial support.

REFERENCES

Adil, I.H., & Ieshad, A.R. (2015). A modified approach for detection of outliers. *Pakistan Journal of Statistics and Operation Research*, 11, 91-102.

Babura, B.I., Adam, M.B., Fitrianto, A., & Rahim, A. (2016). Modified boxplot for extreme data. *Proceedings of The 3rd ISM International Statistical Conference 2016: Bringing Professionalism and Prestige in Statistics.*(pp. 1-9). Kuala Lumper: American Institute of Physics (API).

Barnett, O., & Cohen, A. (2000). A. The histogram and boxplot for the display of lifetime data. *Journal of Computational and Graphical Statistics*, 9(3), 759-778.

Bendi, V. R., Babu, P., & and Venkateswarlu, N. B. (2012). Critical comparative study of liver patients from USA and INDIA: An exploratory analysis. *International Journal of Computer Science*, May, 1694-0784.

Brys, G., Hubert, M., & Struyf, A. (2004). A robust measure of skewness. *Journal of Computational and Graphical Statistics*, 13, 996-1017.

Carling, K. (2000). Resistant outlier rules and the non-Gaussian case. *Computational Statistics and Data Analysis*, 33, 249-258.

Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics and Data Analysis*, 52, 5186-5201.

Kimber, A.C. (1990). Exploratory data Analysis for possibly censored data from skewed distributions. *Applied Statistics*, 39, 21-30.

Moro, S., Rita, P., & Vala, B. (2016). Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. *Journal of Business Research*, 69(9), 3341-3351.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Massachusetts: Addison-Wesley.

Walker, S., & Chakraborti, M. (2013, June). *An Asymmetrically Modified Boxplot for Exploratory Data Analysis*. Paper presented at the 2013 Southern Regional Council on Statistics Summer Research Conference, Burns, USA.

A Comparison of Statistical Models for Predicting Output Responses from Computer Simulated Experiments

Totsaporn Muangngam[1], Anamai Na-udom[1*] and Jaratsri Rungrattanaubol[2]

[1]Department of Mathematics, Faculty of Science, Naresuan University, Phitsanulok, Thailand
Email: totsapornm58@email.nu.ac.th
*Corresponding Email: anamain@nu.ac.th
[2]Department of Computer Science and Information Technology, Faculty of Science, Naresuan University, Phitsanulok, Thailand
Email: jaratsrir@nu.ac.th

ABSTRACT

Computer simulated experiments (CSE) have been widely used to investigate complicated physical phenomena, particularly when physical experiments are not feasible due to limitations of experimental materials. The natures of CSE are time-consuming and the computer codes are expensive. Therefore, experimental designs and statistical models approaches play a major role in the context of CSE in order to overcome these problems. Many researchers have attempted to develop various surrogate models to fit the output responses from CSE. The purpose of this paper is to compare the prediction accuracy of three statistical models namely Kriging model, Radial basis function (RBF) model and Artificial neural network (ANN) model, respectively. These three models are constructed by using the optimal Latin hypercube designs (OLHD). The prediction accuracy of each model is validated though non-linear test problems ranging from 2 to 10 input variables and evaluated by the root mean square of error (RMSE) values. The results show that Guassian RBF model performs well when small dimension of problem with non-complex feature of output response is considered. Further, Guassian RBF model also provides high prediction accuracy for complex feature of output response with small design runs while Kriging models are the most accurate model when the design runs become larger. For medium dimensions of problem, Kriging models are suitable for small design runs while ANN model performs superior over the other models when the design runs are larger. In the case of large dimensions of problem, the results reveal that Multiquadric RBF model is the best choice to construct a surrogate model for CSE.

Keywords: computer simulated experiments; kriging model; radial basis function; artificial neural network

1 INTRODUCTION

Computer simulated experiments (CSE) have been practiced in various fields such as petroleum engineering, mining industrial and applied science to explore the complicated phenomena, especially when physical experiments are impossible due to time, cost or the limitations of experimental units. The examples of CSE are the use of computational fluid dynamics (CFD) model to study the oil mist separator system in the internal combustion engine, use of finite element model for simulating frontal crashes to develop vehicle structure, a study of environmental pollutants that affects human health (Fang et al., 2006) and so on. The nature of output response from CSE is deterministic as the same setting of input variables will always provide the same value of output response. Hence, replication, blocking and randomization that are necessary in physical experiment are irrelevant to CSE (Simpson et al., 2002). The space filling designs such as Latin hypercube design or uniform design that aim to spread the design points over a region of interests are normally practiced in CSE. Moreover, running computer codes are time-consuming and expensive so the construction of approximation model have been developed by many researchers in order to overcome these problems.

Kriging model seems to be the most popular method in modeling output response from CSE due to its interpolation property which is completely accurate when the untried point is close to the design point (Welch et al., 1992). The drawback of Kriging model is that the estimation of all unknown parameters is based on the optimization method and sometimes fails to get the best set of parameters and hence the prediction accuracy of Kriging model becomes worse. Some researchers have focused on enhancement of the parameter estimation method for estimating unknown parameters in Kriging model. For instance, Welch et al. (1992) proposed an efficient algorithm to estimate the parameters using the maximum likelihood method.

While Kriging model has received attention in developing the surrogate models, there are many statistical models such as response surface methodology (RSM), Multivariate adaptive regression splines (MARS), Radial basis function (RBF) and Artificial neural network (ANN) have been adopted to use in the context of CSE. Various researchers have attempted to compare the prediction accuracy of statistical models for predicting the output responses from CSE. For instance, Simpson et al. (2001) compared the performance of Kriging models and RSM in aerospace engineering application. The results showed that Kriging models were slightly more accurate than RSM. Jin et al. (2001) studied the prediction accuracy of four different models, polynomial regression, Kriging, MARS and RBF using Latin Hypercube designs. The authors concluded that RBF model was the most accurate model. Simpson et al. (2002) investigated the prediction accuracy of RSM, RBF and MARS under different types of experimental design. The results revealed that Kriging and RBF were superior over a wide range of designs and sample sizes. Hussain et al. (2002) compared the prediction accuracy between polynomial models and various forms of basis function of RBF using factorial designs and Latin hypercube designs. The results revealed that RBF models provide higher accuracy than polynomial models for all test problems. Fang and Horstemeyer (2006) studied the performance of RBF and RSM and the result showed that RBF models performed better than RSM. Mullur and Messac (2006) investigated the prediction accuracy among various types of RBF, RSM and Kriging models using different classes of designs. The results indicated that RBF model was comparable to Kriging model. Yosboonruang et al. (2013) compared the prediction accuracy among Kriging models, RSM and RBF using optimal Latin hypercube design (OLHD), and two classical designs, Central composite design (CCD) and Fractional factorial design (FFD). The results indicated that RSM performed best when the optimal Latin hypercube design was used with non-complex problem. In the case of complex problem, Kriging models and RBF models were superior over RSM. Furthermore, when the classical designs were considered, RBF model seemed to be the best choice to use. Na-udom and Rungrattanaubol (2013) compared the performance of Kriging and ANN models. The results showed that ANN performed well and can be used as an alternative to Kriging model in some features of problem. Vicario et al. (2016) studied the performance of Kriging model and ANN in predicting the response of four-dimensional computational fluid dynamics experiments. The results revealed that Kriging model was the most accurate model while ANN provided an acceptable prediction accuracy.

According to the results from the published works, there is no certain conclusion on which statistical model is the best for any specific problems in the context of CSE. Therefore, this paper aims to compare the prediction accuracy between the three modeling methods namely Kriging, RBF and ANN models based on two sizes of design runs. The optimal Latin hypercube designs (OLHD) are used in this study. The prediction accuracy of each model is validated though non-linear test problems ranging from 2 to 10 input variables and evaluated by the root

mean square of error (RMSE) values. In section 2, we will present the research method which consists of details of the three statistical models and test problems. The results will be described in section 3 and the conclusions will be presented in the section 4, respectively.

2 METHODS

In this section, we describe details of three statistical models, Kriging, RBF and ANN, and model validation.

2.1 Kriging Model

Kriging model has been originally used in mining engineering and geostatistics (Cressie, 1993). The model has received attention in applications of computer simulated experiments due to its interpolation property. Sacks et al. (1989) adopted Kriging model for CSE and the mathematical form of this model can be expressed as,

$$y = \sum_{j=1}^{d} \beta_j f_j(\boldsymbol{x}) + Z(\boldsymbol{x}) \tag{1}$$

where y is the response, β_j is the parameter of polynomial function, $f_j(\boldsymbol{x})$ is the polynomial function of input variable $j = 1, \dots, d$ and $Z(\boldsymbol{x})$ is stochastic process.

The form of Kriging model is based on the idea that the output response can be modeled as the combination of a polynomial function of input variables and a realization of stochastic process, $Z(\boldsymbol{x})$, with zero mean and a form of correlation function given by

$$Cov\big[Z(\boldsymbol{x_i}), Z(\boldsymbol{x_j})\big] = \sigma^2 R(\boldsymbol{x_i}, \boldsymbol{x_j}) \tag{2}$$

where σ^2 is the process variance and R is the correlation between two design points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

A variety of correlation functions were proposed for Kriging model. The Guassian form is the most popular form and can be written as,

$$R(\boldsymbol{x_i}, \boldsymbol{x_j}) = \prod_{l=1}^{d} exp\left(-\theta_l \left|\boldsymbol{x}_i^{(l)} - \boldsymbol{x}_j^{(l)}\right|^p\right) \tag{3}$$

where $\theta_l > 0$ and $0 < p \leq 2$. In this study, we set $p = 2$.

The polynomial function part in equation (1) can be replaced by a constant vector of 1 as the prediction accuracy of Kriging model will not be affected (Welch et al.,1992, Sacks et al., 1989). Therefore, the subsequent Kriging model is written as,

$$y = \beta + Z(\boldsymbol{x}) \tag{4}$$

All unknown parameters of the correlation function, θ can be estimated by an algorithm based on maximum likelihood estimation method (MLE) proposed by Welch et al. (1992). The estimators of parameter, β and process variance, σ^2 can be obtained by maximizing the log likelihood function as follows,

$$l(\beta, \sigma^2, \theta) = -\frac{1}{2}\left[n\ln\sigma^2 + \ln|R| + \frac{(y - 1\beta)^T R^{-1}(y - 1\beta)}{\sigma^2}\right] \tag{5}$$

where y is the column vector of length n that contains the true response at each design point, R is the $n \times n$ symmetric matrix of correlation $R(\boldsymbol{x_i}, \boldsymbol{x_j})$ for the design points $(1 \leq i, j \leq n)$ with ones along the diagonal.

Given the correlation parameter θ in equation (3), the generalized least square estimator of β is,

$$\hat{\beta} = (1^T R^{-1} 1)^{-1} 1^T R^{-1} y \tag{6}$$

and the MLE of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}(y - 1\hat{\beta})^T R^{-1}(y - 1\hat{\beta}) \tag{7}$$

Substituting $\hat{\beta}$ and $\hat{\sigma}^2$ into the likelihood function in equation (5), so the problem is to numerically maximize

$$-\frac{1}{2}(n\ln\hat{\sigma}^2 + \ln|R|) \tag{8}$$

which is a function of only the correlation parameters and the collected data.

After all parameters are estimated, the next process is to construct a predictor, $\hat{y}(\boldsymbol{x})$, of $y(\boldsymbol{x})$ to act as an approximate model for the output response from computer code. The best linear unbiased predictor (BLUP) at an untried input $\boldsymbol{x}$ is

$$\hat{y}(x) = \hat{\beta} + r^T(x) R^{-1}(y - 1\hat{\beta}) \tag{9}$$

where $r^T(\boldsymbol{x}) = [R(\boldsymbol{x}_1, \boldsymbol{x}) \quad \dots \quad R(\boldsymbol{x}_n, \boldsymbol{x})]$ is the vector of correlation function between n design points and untried input $\boldsymbol{x}$.

In this study, the unknown correlation parameters are estimated using the method that enhanced by Na-udom and Rungrattanaubol (2008).

2.2 Radial Basis Function

Radial basis functions (RBF) were originally developed by Hardy (1971) to fit irregular topographic contours of geographical data. The method is normally used for the exact interpolation of data in multi-dimension problems (Fang and Horstemeyer, 2006), especially when the number of the observations is large. The method uses linear combinations of a radially symmetric function based on Euclidean distance or other distanced such metric to approximate response functions (Jin et al., 2001). The general form of radial basis functions model can be expressed as,

$$y(\boldsymbol{x}) = \sum_{i=1}^{n} \beta_i \phi(\|\boldsymbol{x} - \boldsymbol{x}_i\|) \tag{10}$$

where n is the number of design runs, $\boldsymbol{x}$ is a vector of input variables, $\boldsymbol{x_i}$ is a vector of input variables at the i^{th} design run, $\|\boldsymbol{x} - \boldsymbol{x}_i\|$ is the Euclidean norm, ϕ is a basis function, and β_i is the coefficient for the i^{th} basis function. In this study, Gaussian and Multiquadric basis functions (Fang and Horstemeyer, 2006) presented in Table 1 are used in order to construct the RBF model.

Table 1: Basis functions for RBF model

Name	Symbol	Basis function ($r = \|\boldsymbol{x} - \boldsymbol{x}_i\|$)
Guassian	RBFG	$\phi(r) = e^{-cr^2}, 0 \leq c \leq 1$
Multiquadric	RBFM	$\phi(r) = \sqrt{r^2 + c^2}, 0 \leq c \leq 1$

Replacing $\boldsymbol{x}$ and $y(\boldsymbol{x})$ in equation (10) with the n vectors of input variables and corresponding function values leads to the following equations,

$$\begin{aligned} y(\boldsymbol{x}_1) &= \sum_{j=1}^{n} \beta_j \phi(\|\boldsymbol{x}_1 - \boldsymbol{x}_j\|) \\ y(\boldsymbol{x}_2) &= \sum_{j=1}^{n} \beta_j \phi(\|\boldsymbol{x}_2 - \boldsymbol{x}_j\|) \\ &\vdots \\ y(\boldsymbol{x}_n) &= \sum_{j=1}^{n} \beta_j \phi(\|\boldsymbol{x}_n - \boldsymbol{x}_j\|) \end{aligned} \tag{11}$$

The above equations can be written in matrix form as,

$$y = \boldsymbol{F}\beta \tag{12}$$

where $\boldsymbol{F}_{i,j} = \phi(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|)(i, j = 1,2, \dots, n), \beta = [\beta_1\ \beta_2\ \dots\ \beta_n]^T$ and $y = [y(\boldsymbol{x}_1)\ y(\boldsymbol{x}_2)\ \dots\ y(\boldsymbol{x}_n)]^T$

The least squares estimator of β is

$$\hat{\beta} = (\boldsymbol{F}^T \boldsymbol{F})^{-1} \boldsymbol{F} y \tag{13}$$

In order to estimate parameter c in basis functions, we use the method proposed by Carlson and Foley (1991) and Rippa (1999).

2.3 Artificial Neural Network

Artificial neural network (ANN) is non-parametric approach which any assumption is not required prior to fit the model. The method has been applied successfully in various fields such as data mining, engineering, medicine, finance (Sibanda and Pretorius, 2012) and so on. ANN is inspired by structure, information processing and learning

ability of a biological brain which consists of interconnected sets of neurons.

Figure 1 shows the structure and processing of a real neuron. Dendrites collect inputs from other neurons to cell body, then combines the input information and generates a nonlinear response which sent to other neurons by axon.

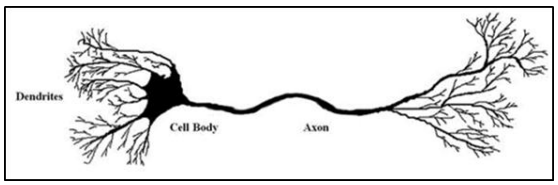


Figure 1: A real neuron (Larous, 2005)

Similarly, the process of a simple artificial neuron model is, input variable (x_i) multiply with weight (w_i) associated with the i^{th} input variable, then the products are combined by a summation function ($\sum$). The summation point is normally referred as a node, and bias (b) is given to each node. The output from a node is passed to a transfer function or an activation function (f) to produce the output response (y) as shown in Figure 2.

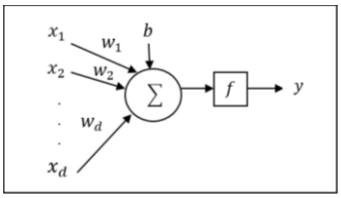


Figure 2: A simple artificial neuron model

The process of artificial neuron model can be rewritten as

$$y = f\left(\sum_{i=1}^{d} x_i w_i + b\right) \tag{14}$$

where f is an activation function, w_i is the weight of each input variable, d is the number of input variables and b is the bias of the summation point or node.

The most frequently used of activation function is sigmoid function that can be expressed as,

$$f(a) = \frac{1}{1 + e^{-a}} \tag{15}$$

where a is the output from a node.

A structure of ANN has three main layers: input layer, hidden layer and output layer as shown in Figure 3. The number of hidden layers and the number of nodes in each hidden layer are both configurable by user. ANN is completely connected network, every node in a previous layer is connected to every node in the next layer by associated weight (Larose, 2005).

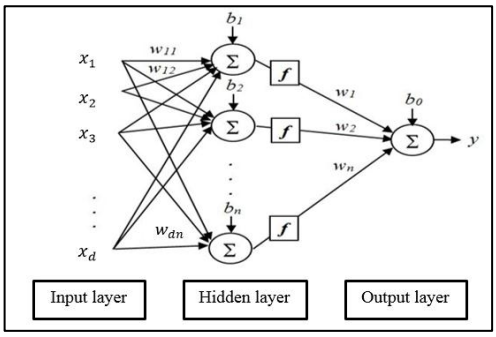


Figure 3: A structure of ANN

The most popular learning algorithm given to training data is back propagation. Prediction errors, the difference between the actual values and output values, are fed back through the network. The weights on all connections are adjusted in order to reduce the error, using gradient descent method. The method involves with learning rate and momentum rate, constant values ranging between zero and 1. Learning rate affects how large the weight adjustment should be and the momentum rate influences the adjustment in the current weight to move along the same direction as previous adjustments.

In this study, we set learning rate, momentum rate, training time, the number of hidden layers and their nodes in Weka 3.8 to construct the ANN models.

Table 2: The detail of test problems

Problen	d	Function
Branin function	2	$y = \left(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)\cos(x_1) + 10, -5 \le x_1 \le 10, 0 \le x_2 \le 15$
Welch function	2	$y = [30 + x_1\sin(x_1)](4 + e^{-x_2}), 0 \le x_1, x_2 \le 5$
2Dfunction	2	$y = \sin(x_1 + x_2) + (x_1 + x_2)^2 - 1.5x_1 + 2.5x_2 + 1, -1.5 \le x_1 \le 4, -3 \le x_2 \le 3$
3Dfunction	3	$y = (x_1 - x_2)^2 + (x_2 + x_3)^4, -10 \le x_1, x_2, x_2 \le 10$
Pressure vessel function	4	$y = 0.6244x_1x_2x_3 + 1.7781x_4x_1^2 + 3.1661x_3^2x_2 + 19.84x_3^2x_1$ $25 \le x_1 \le 150, -1.5 \le x_2 \le 240, 1 \le x_3 \le 1.375, 0.625 \le x_4 \le 1$
Cyclone model	7	$y = 174.42\left(\frac{x_1}{x_5}\right)\left(\frac{x_3}{x_2 - x_1}\right)^{0.85}\sqrt{\frac{1 - 2.62\{1 - 0.36(x_4/x_2)^{-0.56}\}^{3/2}(x_4/x_2)^{1.16}}{x_6x_7}}$ $0.09 \le x_1, x_3, x_4 \le 0.11, 0.27 \le x_2 \le 0.33, 1.35 \le x_5 \le 1.65, 14.4 \le x_6 \le 17.6,$ $0.675 \le x_7 \le 0.825$
Borehole function	8	$y = \frac{2\pi x_3(x_4 - x_6)}{\ln\left(\frac{x_2}{x_1}\right)\left[1 + \frac{2x_7x_3}{\ln\left(\frac{x_2}{x_1}\right)x_1^2x_8} + \frac{x_3}{x_5}\right]}$ $0.05 \le x_1 \le 0.15, 100 \le x_2 \le 50000, 63070 \le x_3 \le 115600, 990 \le x_4 \le 1110,$ $63.1 \le x_5 \le 116, 700 \le x_6 \le 820, 1120 \le x_7 \le 1680, 9855 \le x_8 \le 12045$
9Dfunction	9	$y = 0.28285 + \sum_{i=1}^{9}\left[\frac{3}{10} + \sin\left(\frac{16}{15}x_i - 1\right) + \sin^2\left(\frac{16}{15}x_i - 1\right)\right], -1 \le x_i \le 1, i = 1,2,\dots,9$

Problen	d	Function
10Dfunction	10	$y = \sum_{i=1}^{10}\left[\frac{3}{10} + \sin\left(\frac{16}{15}x_i - 1\right) + \sin^2\left(\frac{16}{15}x_i - 1\right)\right], -1 \le x_i \le 1, i = 1,2,\dots,10$

2.4 Model Validation

To implement the prediction accuracy of the models, we generate OLHD designs using simulated annealing algorithm (SA) under $\phi_p$criteria (Chantarawong, et al, 2010). The number of dimensions or input variables (d) is equal to 2,3,4,7,8,9 and 10. Each dimension has two sizes of design runs (n), small design runs are the number of parameters in quadratic polynomial model and large design runs are calculated by equation (16) (Na-udom, 2007) as follow,

$$n = 2d + 4\binom{d}{2} + 1 \quad (16)$$

Nine non-linear test problems obtained from Hock and Schittkowski (1981) are used to compare the prediction accuracy among the three statistical models. The two-dimension problems are classified into two types, complex problems namely Branin function and Welch function, and non-complex problem namely 2Dfunction. The details of all test problems are given in Table 2

After fitting Kriging, RBFG and RBFM model in R program version 3.4.2 and ANN model in Weka 3.8, the 81, 216 and 625 grid points are test points for 2, 3 and 4 dimension problems, respectively, and the 1,000 random test points are used for larger dimensions of problem. The prediction accuracy of the models is evaluated by RMSE values as follow,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{m}} \quad (17)$$

where m is the number of test points, y_i is the real response of the i^{th} test point and $\hat{y}_i$ is the predicting output response from the models for the i^{th} test point.

Moreover, we calculate percentage improvement (PI) over Kriging model that defined as,

$$PI = \frac{RMSE(Kriging) - RMSE(RBF, ANN)}{RMSE(Kriging)} \times 100\% \quad (18)$$

to compare the performance of the models**.**

3. RESULTS AND DISCUSSIONS

In this section, the prediction accuracy of Kriging, two types of basis function of RBF and ANN model are compared by RMSE values. The different OLHD designs are generated for each dimension and number of design runs. The means, standard deviations of RMSE, and percentage improvement values over Kriging model are shown in Table 3, 4 and 5 for small, medium and large dimensions of problems, respectively.

Table 3: RMSE values for small dimensions (2 to 4 input variables)

Test problem	Model	RMSE					
		Small design runs			Large design runs		
		Mean	S.D.	PI	Mean	S.D.	PI
Branin function	Kriging	47.1271	5.2903		**31.3229**	9.9980	
	RBFG	**45.6944**	5.5329	3.0401	31.6707	10.4940	-1.1104
	RBFM	47.5957	5.1250	-0.9943	36.0136	8.3319	-14.9753
	ANN	60.4779	1.7443	-28.3393	57.7944	0.6217	-84.5117
Welch function	Kriging	5.4923	0.8081		**3.8542**	0.4973	
	RBFG	**5.1697**	1.3390	5.8737	4.2097	0.1754	-9.2237
	RBFM	7.7868	1.2069	-41.7767	5.2123	0.1833	-35.2369
	ANN	9.0204	0.4977	-64.2372	8.4931	0.1942	-120.3696
2Dfunction	Kriging	2.5953	0.1537		2.0243	0.0118	
	RBFG	**1.0280**	0.1132	60.3899	**1.0470**	0.0353	48.2784
	RBFM	3.6987	0.1901	-42.5152	2.6945	0.1185	-33.1077
	ANN	6.8755	0.8220	-164.9212	7.1676	1.0979	-254.0780
3Dfunction	Kriging	26534.58	3374.39		26383.05	2028.55	
	RBFG	**21999.53**	3115.03	17.0911	22924.04	1472.21	13.1107
	RBFM	33836.69	2091.95	-27.5192	29550.92	1844.84	-12.0072
	ANN	29558.61	12579.78	-11.3966	**22565.18**	12012.91	14.4709
Pressure vessel function	Kriging	**944.196**	252.276		**59.657**	17.063	
	RBFG	7938.058	2435.184	-740.7214	12474.46	7191.75	-20810.304
	RBFM	5131.494	425.947	-443.4776	4693.200	192.404	-7766.9729
	ANN	1146.302	393.592	-21.4051	1093.946	70.960	-1733.7261

Table 3 presents the RMSE values from statistical models and PI over Kriging model for small dimensions with two sizes of design runs. For complex two-dimension test problems, Branin and Welch function, the results from small design runs show that average RMSE values obtained from RBFG models are lower than Kriging, RBFM and ANN models, respectively. Moreover, the PI values of RBFG are 3.0401% and 5.8737% for Branin function and Welch function means that RBFG performs slightly better than Kriging. In the case of large design runs, Kriging models are seems to perform best which slightly different from BRBG, PI of RBFG (PI = -1.1104% and -9.2237%). When considering non-complex two-dimension test problem, 2Dfunction, it can be clearly seen that BRFG models are suitable for both sizes of design runs. Hence, RBF model with Guassian basis function would be recommended to use for two-dimension problems with non-complex feature of output response.

For three-dimension test problem, RMSE value from RBFG model is lower than other statistical models when the number of design runs is small. For large design runs, ANN model is the most accurate model while RBFG provides slightly higher RMSE value.

For four-dimension test problem with two sizes of design run, the results show that RMSE values obtained from Kriging model are lower than that of ANN, RBFM and RBFG models. Moreover, the least standard deviation from Kriging model indicates the consistent performance of Kriging model. According to the results obtained from table 3, it could be concluded that RBFG and Kriging are appropriate for small dimensions problems.

Table 4: RMSE values for medium dimensions (7 and 8 input variables)

Test problem	Model	RMSE					
		Small design runs			Large design runs		
		Mean	S.D.	PI	Mean	S.D.	PI
Cyclone model	Kriging	0.**0047**	0.0006		0.0043	0.0003	
	RBFG	0.0590	0.0102	-1155.3191	0.0118	0.0021	-174.4186
	RBFM	0.0630	0.0072	-1240.4255	0.0528	0.0037	-1127.9070
	ANN	0.0103	0.0014	-119.1489	**0.0041**	0.0006	4.6512
Borehole function	Kriging	0.**9489**	0.1057		0.8395	0.1398	
	RBFG	5.0671	0.3419	-433.9973	2.8965	0.2859	-245.0268
	RBFM	5.5991	0.4146	-490.0622	3.2099	0.1541	-282.3585
	ANN	2.0457	0.6200	-115.5865	0.**7845**	0.0801	6.5515

Table 5: RMSE values for large dimensions (9 and 10 input variables)

Test problem	Model	RMSE					
		Small design runs			Large design runs		
		Mean	S.D.	PI	Mean	S.D.	PI
9Dfunction	Kriging	0.2745	0.0140		0.2416	0.0058	
	RBFG	0.3588	0.0265	-30.7104	0.3021	0.0167	-25.0414
	RBFM	0.**2365**	0.0042	13.8434	0.**2237**	0.0040	7.4089
	ANN	0.2517	0.0100	8.3060	0.2425	0.0055	-0.3725
10Dfunction	Kriging	0.2966	0.0130		0.2548	0.0073	
	RBFG	0.3803	0.0151	-28.2198	0.3198	0.0195	-25.5102
	RBFM	0.**2498**	0.0055	15.7788	0.**2346**	0.0033	7.9278
	ANN	0.2741	0.0082	7.5860	0.2612	0.0105	-2.5118

Table 4 presents the RMSE and PI values for medium dimensions test problems (seven-dimension and eight-dimension test problems). The results indicate that Kriging models are suitable for Cyclone model and Borehole function when the number of design run is small. PI values over Kriging of other models are the large negative number. This indicates that Kriging is far better than other models. For large design runs, RMSE values from ANN model are lower than other models. Hence ANN is recommended for use to develop a prediction model. It should be noticed from PI values that the performance of Kriging model is very close to ANN.

Table 5 shows the results obtained from 9 and 10 dimensions test problems. It can be clearly seen that RBFM model performs better than other models as it provides the lowest RMSE values in both small and large number of design runs.

4. CONCLUSIONS

This paper aims to compare the prediction accuracy of Kriging, RBF with Gaussian and Multiquadric basis functions and ANN models. According to the results presented in the previous section, it indicates that RBFG and Kriging models are suitable for small dimensions problems while RBFG model performs best for non-complex problems. Further, RBFG model also provides high prediction accuracy for complex feature of output response with small design runs while Kriging models are the most accurate model when the design runs become larger. For medium dimensions, Kriging model seems to be the best choice to use when the design runs are small. When considering larger design runs, ANN model turns to be the most accurate model. In the case of large dimensions, the results reveal that RBFM model is the best choice to construct a surrogate model for CSE. It can be concluded that RBF model with different basis functions are comparable or better than Kriging model in some cases especially when the dimensions of problems are large. Moreover, ANN model are suitable for medium dimensions with complex feature of output response. Therefore RBF and ANN models are recommended as an alternative choice for constructing the surrogate models of the output response from CSE.

ACKNOWLEDGEMENTS

The authors would like to thanks Department of Mathematics, Department of Computer Science and Information Technology, Faculty of Science, Naresuan University and Science Achievement Scholarship of Thailand for supporting research materials and master degree scholarship.

REFERENCES

Carlson, R.E., & Foley, T.A. (1991). The parameter R2 in multiquadric interpolation. *Computers Math. Applic*, 21, 29-42.

Chantarawong, T. Rungrattanaubol, J. & Na-udom, A. (2010). Enhancement of Stochastic evolutionary algorithm for computer simulated experiments. *Information Technology Journal*, 6(12), 65-69.

Cressie, Noel A.C. (1993). *Statistics for spatial data*. America: John Wiley & Sons.

Fang, H. & Horstemeyer, M.F. (2006). Global response approximation with radial basis functions. *Engineering Optimiztion*, 38(4), 407-424.

Fang, K.T., Li, R., & Sudjianto, A. (2006). *Design and modeling for computer experiments*. London: Chapman & Hall/CRC.

Hardy, R.L. (1971). Multiquadratic equations of topography and other irregular surfaces. *Journal of Geophysical Research*, 76(8), 1905-1915.

Hock, W. & Schittkowski, K. (1981). *Test examples for nonlinear programming codes*. New York: Springer, Berlin Heidelberg.

Hussian, M.F., Barton, R.R., & Joshi, S.B. (2002). Metamodeling: Radial basis functions, versus polynomials. *European Journal of Operation Research*, 138, 142-154.

Jin, R., Chen, W. & Simpson, T.W. (2001). Comparative studies of metamodeling techniques under multiple modeling criteria. *Struct Multidisc Optim*, 23, 1-13.

Larous, D.T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*. New York: John Wileys & Sons.

Morris, M.D. & Mitchell, T.J. (1995). Exploratory design for computational experiments. *Journal of Statistical Planning and Inference*, 43, 381-402.

Muller, A.A. & Messac, A. (2006). Metamodeling using extended radial basis functions: a comparative approach. *Engineering with Computers*, 21, 203-217.

Na-udom, A. (2007). *Experimental design methodology for modeling response from computer simulated experiments*. Doctoral dissertation, Ph.D., Curtin University of Technology, Australia

Na-udom, A. & Rungrattanaubol, J. (2008). Optimization of correlation parameter for kriging approximation model. *International Joint Conference on Computer Science and Software Engineering*, 1, 159-164.

Na-udom, A. & Rungrattanaubol, J. (2013). A Comparison of artificial neural network and Kriging model for predicting the deterministic output Response. *NU Science Journal*, 10(1), 1 – 9.

Rippa, S. (1999). An algorithm for selecting a good value for the parameter c in radial basis function interpolation. *Advances in Computational Mathematics*, 11, 193-210.

Sack, J., Welch, W.J., Mitchell, T.J. & Wynn, H.P. (1989). Design and analysis of computer experiments. *Statistical Science*, 4(4), 409-435.

Sibanda, W. & Pretorius, P. (2012). Artificial neural networks-a review of applications of neural networks in the modeling of HIV epidemic. *International Journal of Computer Applications*, 44(16), 1-4.

Simpson, T.W, Mauery, T.M., Korte, J.J. & Mistree, F. (2001). Kriging models for global approximation in simulation-based multidisciplinary design optimization. *AIAA Journal*, 39(12), 2233-2241.

Simpson, T.W., Lin, D.K.J. & Chen, W. (2001). Sampling strategies for computer experiments: Design and analysis. *International Journal of Reliability and applications*, 2(3), 209-240.

Vicario, G., Craparotta, G., & Pistone, G. (2016). Meta-models in computer experiments: Kriging versus artificial neural networks. *Quality and Reliablity Engineering International*, 32, 2055-2065.

Welch, W.J., Buck, R.J., Sacks, J., Wynn, H.P., Mitchell, T.J. & Morris, M.D. (1992). Screening, predicting, and computer experiments. *Technometrics*, 34(1), 15-25.

Yosboonruang, N., Na-udom, A. & Rungrattanaubol, J. (2013). A comparative study on predicting accuracy of statistical models for modeling deterministic output responses. *Thailand Statistician*, 11(1), 1-15.

Ratio Estimator for The Population Mean using General Ranked Set Sampling with Perfect Ranking

Klairoong Suchon*, Supunnee Ungpansattawong

Department of Statistics, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand
*Corresponding Email: klairoongsuchon@gmail.com
Email: supunnee@kku.ac.th

ABSTRACT

The aim of research is to propose the estimator of population mean using ratio for general ranked set sampling with perfect ranking. This proposed estimator is modified from the estimator of Cochran (1940). Mean squared error and bias of the proposed estimator is obtained

Keywords: general ranked set sampling, population mean, ranked set sampling, ratio estimator**.**

1 INTRODUCTION

The ranked set sampling method was suggested by McIntyre (1952). After that, theories and applications of this method were reviewed by Patil et al. (1994) In 2004 Wang et al. proposed the method of selecting τ units from each set where $\tau > 1$ to measure the interesting features, namely general ranked set sampling (GRSS). Plukpluem (2008) studied and compared the performance of the estimator for the population mean under general ranked set sampling with perfect ranking to other sampling methods. However, all sampling technique above used unbiased estimator but in some cases the study variables may be difficult to observe. So that we can be using auxiliary variable x where x must correlate with the study variables y . Cochran (1940) proposed the ratio estimator of the population mean using $\bar{y} / \bar{x}$ ratio for simple random sampling ($\bar{y}_r$) . And also, Samawi and Muttlak (1996) proposed the ratio estimator of the population mean based on ranked set sampling ($\bar{y}_{rRSS}$) .

In this study we proposed ratio estimator for the population mean using $\bar{y} / \bar{x}$ ratio for general ranked set sampling with perfect ranking "GRSS-WPR" which is adopted from Wang et al. (2004) and Plukpluem (2008), where the proposed estimator is established by modifying the method of Cochran (1940) and Samawi and Muttlak (1996). This estimator is proposed to be used as a guideline for further researcher.

2 SAMPLING METHODS

2.1 Ranked set sampling (RSS). The steps of sampling by RSS can be describes as follows:

Step 1: Randomly select m^2 units from the target population.

Step 2: Allocate the m^2 selected units as randomly as possible into m sets, each of size m .

Step 3: Without yet knowing any values for the variable of interest, rank the units within each set with respect to variable of interest. This may be based on personal professional judgment or done with concomitant variable correlated with the variable of interest.

Step 4: Choose a sample for actual quantification by including the smallest ranked unit in the first set, the second smallest ranked unit in the second set, the process is continues in this way until the largest ranked unit is selected from the last set.

Step 5: Repeat Steps 1 through 4 for r cycles to obtain a sample of size mr .

Table 1: Show ranked set sampling. When $n=8$, $m=4$ and $r=2$

Cycles (r)	Set (m)	Rank (m)			
		1	2	3	4
1	1	☑	*	*	*
	2	*	☑	*	*
	3	*	*	☑	*
	4	*	*	*	☑
2	1	☑	*	*	*
	2	*	☑	*	*
	3	*	*	☑	*
	4	*	*	*	☑

2.2 General ranked set sampling (GRSS) The steps of sampling by GRSS can be describes as follows:

Step 1: Impose sample size n units and τ units to be selected for from each set. Then $\tau > 1$ therefrom m is the number of samples and the number of units in each set when $\tau\binom{m}{\tau} = n$ which allocate the m^2 selected units as randomly as possible into m sets, each of size m. randomly select m^2 units from the target population.

Step 2: Rank the units within each set visually with respect to the study variable or by any inexpensive method.

Step 3: Select the sample for step 2 τ units in each set. Then selection is as follows $\{(Y_{[a]i}, Y_{[b]i}) : 1 \le a \le b \le m; i = 1, 2, ...r\}$.

Step 4: Repeat Steps 1 through 4 for r cycles to obtain a sample of size $r\tau\binom{m}{\tau}$.

Table 2: Show general ranked set sampling. When $n=8$, $\tau=2$, $m=3$ and $r=2$.

Cycles (r)	Set (m)	Rank (m)		
		1	2	3
1	1	☑	☑	*
	2	☑	*	☑
	3	*	☑	☑
2	1	☑	☑	*
	2	☑	*	☑
	3	*	☑	☑

2.3 General ranked set sampling (GRSS) with perfect ranking. The steps of sampling by GRSS-WPR can be describes as follows:

General ranked set sampling (GRSS-WPR) with perfect ranking is same with general ranked set sampling (GRSS) But (GRSS-WPR) is ranking without error of giving order to sample.

Step 1: Impose sample size n units and τ units to be selected for from each set. Then $\tau > 1$ therefrom m is the number of

samples and the number of units in each set when $\tau\binom{m}{\tau}=n$ which allocate the m^2 selected units as randomly as possible into m sets, each of size m. randomly select m^2 units from the target population.

Step 2: Rank the units within each set visually with respect to the study variable or by any inexpensive method.

Step 3: Select the sample for step 2 τ units in each set. Then selection is as follows $\{(Y_{[a]i}, Y_{[b]i}): 1 \le a \le b \le m; i = 1, 2, ...r\}$.

Step 4: Repeat Steps 1 through 4 for r cycles to obtain a sample of size $r\tau\binom{m}{\tau}$.

3 ESTIMATION FOR THE POPULATION MEAN

The traditional ratio estimator for the population mean $\bar{Y}$ of the study variable y is defined by

$$\bar{y}_r = \frac{\bar{y}}{\bar{x}}\bar{X} = \hat{R}\bar{X} \tag{1}$$

In which it is assumed that the population mean $\bar{X}$ of the auxiliary variable x is known. Here $\bar{y}$ is the sample mean of the study variable and $\bar{x}$ is the sample mean of the auxiliary variable

Mean square error (MSE) of the traditional ratio estimator is as follows:

$$MSE(\bar{y}_r) = \gamma(R^2 S_x^2 - 2RS_{yx} + S_y^2) \tag{2}$$

where $\gamma = \frac{1}{n}$; n is the sample size; $R = \frac{\bar{Y}}{\bar{X}}$ is the population ratio; S_x^2 is the population variance of auxiliary variable; S_y^2 is the population variance of study variable and S_{yx} is the population covariance between auxiliary variable and study variable (Cochoran 1977). Note that $f = \frac{n}{N}$ is omitted in (2), Here N is the population size.

Samawi and Muttlak (1996) defined the estimator of population ratio using ranked set sampling as

$$\hat{R}_{RSS} = \frac{\bar{y}}{\bar{x}} \tag{3}$$

Where $\bar{y} = \frac{1}{mr}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{mr}\sum_{i=1}^{n} x_i$. As Samawi and Muttlak (1996) remind that estimator can also be used for the population total and mean, we can write following estimator for the population mean as

$$\bar{y}_{rRSS} = \frac{\bar{y}}{\bar{x}}\bar{X} \tag{4}$$

The MSE and the bias equations of this estimator can be given by

$$MSE(\bar{y}_{rRSS}) = \frac{1}{mr}(S_y^2 - 2RS_{yx} + R^2 S_x^2) - \frac{1}{m^2 r}\left(\sum_{i=1}^{m}\tau_{y[i]}^2 - 2R\sum_{i=1}^{m}\tau_{yx} + R^2\sum\tau_{x(i)}^2\right) \tag{5}$$

$$Bias(\bar{y}_{rRSS}) \cong \bar{Y}\left[\gamma\left(C_x^2 - C_{yx}\right) - \left(W_{x(i)}^2 - W_{yx(i)}\right)\right]$$

Where $\gamma = \frac{1}{mr}$, $C_{yx} = \rho C_y C_x$, $W_{yx} = \frac{1}{m^2 r\bar{X}\bar{Y}}\sum_{i=1}^{m}\tau_{yx(i)}$, $W_{x(i)}^2 = \frac{1}{m^2 r\bar{X}^2}\sum_{i=1}^{m}\tau_{x(i)}^2$, $W_{y(i)}^2 = \frac{1}{m^2 r\bar{Y}^2}\sum_{i=1}^{m}\tau_{y(i)}^2$, $\rho = \frac{S_{yx}}{S_x S_y}$, $C_y = \frac{S_y}{\bar{y}}$ and $C_x = \frac{S_x}{\bar{x}}$

Here S_y and S_x are the sample standard deviations of study and auxiliary variables, respectively.

Here we would also like to remind that $\tau_{x(i)} = \mu_{x(i)} - \bar{X}$, $\tau_{y(i)} = \mu_{y(i)} - \bar{Y}$ and $\tau_{yx(i)} = (\mu_{y(i)} - \bar{Y})(\mu_{x(i)} - \bar{X})$. Note that the values of $\mu_{x(i)}$ and $\mu_{y(i)}$ depend on order statistics from some specific distributions and these values can be found in Arnold et al. (1993)

Adapting the estimator in (1) to the ratio estimator for the population mean suggested by Cochoran (1977) adjusting the method.is general ranked set sampling by Wang et al. (2004) We develop the following estimator:

$$\hat{\bar{Y}}_{GRSSr} = \bar{y}_{GRSS}\left(\frac{\bar{X}}{\bar{x}_{GRSS}}\right) \tag{6}$$

where $\bar{y}_{GRSS} = \frac{1}{r\tau\binom{m}{\tau}}\sum_{i=1}^{m}\sum_{1\le\eta\le\ldots\le r_\tau\le m}\sum_{j=1}^{\tau} Y_{(r_\tau)i}$, $\bar{x}_{GRSS} = \frac{1}{r\tau\binom{m}{\tau}}\sum_{i=1}^{m}\sum_{1\le\eta\le\ldots\le r_\tau\le m}\sum_{j=1}^{\tau} X_{(r_\tau)i}$.

The MSE of $\hat{\bar{Y}}_{GRSSr}$ can be found as follows:

For the first degree of approximation by using Taylor series method as

$$h(\bar{x},\bar{y}) \cong h(\bar{X},\bar{Y}) + \left.\frac{\partial h(\bar{x},\bar{y})}{\partial\bar{x}}\right|_{\bar{x}=\bar{X},\bar{y}=\bar{Y}}(\bar{x}-\bar{X}) + \left.\frac{\partial h(\bar{x},\bar{y})}{\partial\bar{y}}\right|_{\bar{x}=\bar{X},\bar{y}=\bar{Y}}(\bar{y}-\bar{Y})$$

Where $h(\bar{x},\bar{y}) = \hat{\bar{Y}}_{GRSSr}$; $\hat{\bar{Y}}_{GRSSr} = \bar{y}_{GRSS}\left(\frac{\bar{X}}{\bar{x}_{GRSS}}\right)$

MSE of $\hat{\bar{Y}}_{GRSSr}$ is obtained as follows:

$$\hat{\bar{Y}}_{GRSSr} \cong \bar{Y} + \left.\frac{\partial\,\bar{y}_{GRSS}\left(\frac{\bar{X}}{\bar{x}_{GRSS}}\right)}{\partial\,\bar{x}_{GRSS}}\right|_{\bar{x}GRSS=\bar{X},\bar{y}GRSS=\bar{Y}}(\bar{x}_{GRSS}-\bar{X}) + \left.\frac{\partial\,\bar{y}_{GRSS}\left(\frac{\bar{X}}{\bar{x}_{GRSS}}\right)}{\partial\,\bar{y}_{GRSS}}\right|_{\bar{x}GRSS=\bar{X},\bar{y}GRSS=\bar{Y}}(\bar{y}_{GRSS}-\bar{Y})$$

$$\hat{\bar{Y}}_{GRSS} \cong \bar{Y} + \frac{-\bar{Y}}{\bar{X}}(\bar{x}_{GRSS}-\bar{X}) + (\bar{y}_{GRSS}-\bar{Y})$$

$$\hat{\bar{Y}}_{GRSSr} - \bar{Y} \cong \bar{Y} - \bar{Y} + \frac{-\bar{Y}}{\bar{X}}(\bar{x}_{GRSS}-\bar{X}) + (\bar{y}_{GRSS}-\bar{Y})$$

$$\cong \frac{-\bar{Y}}{\bar{X}}(\bar{x}_{GRSS}-\bar{X}) + (\bar{y}_{GRSS}-\bar{Y})$$

$$(\hat{\bar{Y}}_{GRSSr} - \bar{Y})^2 \cong \left[\frac{-\bar{Y}}{\bar{X}}(\bar{x}_{GRSS}-\bar{X}) + (\bar{y}_{GRSS}-\bar{Y})\right]^2$$

$$(\hat{\bar{Y}}_{GRSSr} - \bar{Y})^2 \cong \left[(\bar{y}_{GRSS}-\bar{Y}) - W(\bar{x}_{GRSS}-\bar{X})\right]^2$$

Where $W = \frac{\bar{Y}}{\bar{X}}$

$$(\hat{\bar{Y}}_{GRSSr} - \bar{Y})^2 \cong (\bar{y}_{GRSS}-\bar{Y})^2 - 2(\bar{y}_{GRSS}-\bar{Y})W(\bar{x}_{GRSS}-\bar{X}) + [W(\bar{x}_{GRSS}-\bar{X})]^2 \tag{7}$$

Take the expectation in (7) is given by

$$E(\hat{\bar{Y}}_{GRSSr} - \bar{Y})^2 \cong E\left[(\bar{y}_{GRSS}-\bar{Y})^2 - 2W(\bar{x}_{GRSS}-\bar{X})(\bar{y}_{GRSS}-\bar{Y}) + [W^2(\bar{x}_{GRSS}-\bar{X})]^2\right]$$

$$E(\hat{\bar{Y}}_{GRSSr} - \bar{Y})^2 \cong E(\bar{y}_{GRSS}-\bar{Y})^2 - 2W[E(\bar{x}_{GRSS}-\bar{X})(\bar{y}_{GRSS}-\bar{Y})] + W^2 E(\bar{x}_{GRSS}-\bar{X})^2$$

where

$$MSE(\hat{\bar{Y}}_{GRSSr}) = E(\hat{\bar{Y}}_{GRSSr} - \bar{Y})^2 \cong Var(\bar{y}_{GRSS}) - 2W\,Cov(\bar{x}_{GRSS},\bar{y}_{GRSS}) + W^2\,Var(\bar{x}_{GRSS})$$

$$\cong [\, W^2 Var(\bar{x}_{GRSS}) - 2TW Var(\bar{x}_{GRSS}) + Var(\bar{y}_{GRSS})] \quad (8)$$

and $T = \rho \frac{\sigma_Y}{\sigma_X}$

$$Var(\bar{y}_{GRSS}) = \frac{1}{r\tau \binom{m}{\tau}} \sum_{1 \le r_1 \le \ldots \le r_\tau \le m} Var(\sum_{j=1}^{\tau} y_{[r_j]})$$

$$Var(\bar{x}_{GRSS}) = \frac{1}{r\tau \binom{m}{\tau}} \sum_{1 \le r_1 \le \ldots \le r_\tau \le m} Var(\sum_{j=1}^{\tau} x_{[r_j]})$$

$$Cov(\bar{x}_{GRSS}, \bar{y}_{GRSS}) = \rho_{xy} \frac{\sigma_Y}{\sigma_X} Var(\bar{x}_{GRSS})$$

Finally, we obtain the following MSE equation for the proposed estimator using $Var(\bar{y}_{GRSS})$, $Var(\bar{x}_{GRSS})$ and $Cov(\bar{x}_{GRSS}, \bar{y}_{GRSS})$ in (8)

$$MSE(\hat{\bar{Y}}_{GRSSr}) \cong W^2 S_x^2 - 2WTS_x^2 + S_y^2 \quad (9)$$

The bias using Taylor series expansion of $\hat{\bar{Y}}_{GRSSr}$ about variable $\bar{x}$ and $\bar{y}$ can be approximated as:

$$\hat{\bar{Y}}_{GRSSr} = \bar{y}_{GRSS} - [\, \frac{\bar{Y}}{\bar{X}} (\bar{x}_{GRSS} - \bar{X})] + [\, \frac{\bar{Y}}{\bar{X}} \frac{(\bar{x}_{GRSS} - \bar{X})^2}{\bar{X}}] - [\, \frac{(\bar{y}_{GRSS} - \bar{Y})(\bar{x}_{GRSS} - \bar{X})}{\bar{X}}] \quad (10)$$

where $W = \frac{\bar{Y}}{\bar{X}}$. Take the expectation in (10) is given by

$$E(\hat{\bar{Y}}_{GRSSr}) = E(\bar{y}_{GRSS}) - [\, W E(\bar{x}_{GRSS} - \bar{X})] + [\, W \frac{E(\bar{x}_{GRSS} - \bar{X})^2}{\bar{X}}] - [\, \frac{E(\bar{y}_{GRSS} - \bar{Y})(\bar{x}_{GRSS} - \bar{X})}{\bar{X}}] \quad (11)$$

where

$$Bias(\hat{\bar{Y}}_{GRSSr}) = E(\hat{\bar{Y}}_{GRSSr}) - \bar{Y}$$

and

$$Bias(\hat{\bar{Y}}_{GRSSr}) = W \frac{Var(\bar{x}_{GRSS})}{\bar{X}} - \frac{Cov(\bar{x}_{GRSS}, \bar{y}_{GRSS})}{\bar{X}} \quad (12)$$

We obtain the following bias equation for the proposed estimator using $Var(\bar{y}_{GRSS})$, $Var(\bar{x}_{GRSS})$ and $Cov(\bar{x}_{GRSS}, \bar{y}_{GRSS})$ in (12)

$$Bias(\hat{\bar{Y}}_{GRSSr}) = W \frac{\frac{1}{r\tau \binom{m}{\tau}} \sum_{1 \le r_1 \le \ldots \le r_\tau \le m} Var(\sum_{j=1}^{\tau} x_{[r_j]})}{\bar{X}} - \frac{T \frac{1}{r\tau \binom{m}{\tau}} \sum_{1 \le r_1 \le \ldots \le r_\tau \le m} Var(\sum_{j=1}^{\tau} x_{[r_j]})}{\bar{X}}$$

Finally, the bias of $\hat{\bar{Y}}_{GRSSr}$ as

$$Bias(\hat{\bar{Y}}_{GRSSr}) = W \frac{S_x^2}{\bar{X}} - \frac{T S_x^2}{\bar{X}} \quad (13)$$

4 CONCLUSIONS

The proposed estimator of this research ($\hat{\bar{Y}}_{GRSSr}$) is denoted by

$$\hat{\bar{Y}}_{GRSSr} = \bar{y}_{GRSS} \left(\frac{\bar{X}}{\bar{x}_{GRSS}} \right)$$

The mean squared errors of the proposed estimator can be obtained by the following equation:

$$MSE(\hat{\bar{Y}}_{GRSSr}) \cong W^2 S_x^2 - 2WTS_x^2 + S_y^2$$

Bias of the proposed estimators can be approximated as:

$$Bias(\hat{\bar{Y}}_{GRSSr}) = W \frac{S_x^2}{\bar{X}} - \frac{T S_x^2}{\bar{X}}$$

Therefore, in the future study, we hope to real data studies using the proposed estimator will be conducted in order to compare their efficiencies with those of previously proposed estimators. And develop estimator using GRSS with perfect ranking in the product estimator or Ratio-cum-product estimation utilizing the methods in the papers Singh (1967) and Nitu Mehta (Ranka) & Mandowarab (2016).

ACKNOWLEDGEMENTS

Researchers want to thank their parents and family, which provides the opportunity to receive an education, as well as to assist and encourage research to complete. This thesis was well done. With the help of Associate Prof. Dr.Supunnee Ungpansattawong advisor. This is a great way to do research. It also helps solve the problems that occur during operation.

REFERENCES

Cochran, W.G. (1977). Sampling Techniques. 3rd edition, *Wiley and Sons, New York*.

McIntyre, G.A. (1952). A method for unbiased selective sampling using ranked set. *Journal of Agricultural Research*, 3, 385-390.

Plukpluem, N. (2006). Efficiency of general ranked set sampling with perfect ranking. Master of Science in applied statistics, Graduate School of Khan Kaen Univercity. (in Thai).

Patil, G.P., Sinha, A.K. and Taillie, C. (1994). Handbook of statistics. *Environmental Statistics*, 12, 167-200.

Samawi, H.M, & Muttlak, H.A. (1996). Estimation of ratio using rank set sampling. *Biometrical Journal*, 6, 753-764.

Kheranan, S. (1995). Sampling Theory. Bangkok: *Chulalongkorn Univercity*. (in Thai).

Ungpansattawong, S. (2011). Sampling Techniques. Khonkaen: *Department of Statistics, Faculty of Science, Khon Kaen University*. (in Thai).

Singh, M. P. (1967). Multivariate product method of estimation for finite populations.

Wang, Y.G., Chen, Z., and Liu, J. (2004). General set sampling with cost considerations. *Biometrics*, 60, 556-561.

Discrete-Time Risk Model based on NBMA(1) models

Kodchapown Laphudomsakda[1*] and Jiraphan Suntornchost[2]

[1]Chulalongkorn University/ Mathematics and Computer Science / Faculty of Science, Bangkok, Thailand
*Corresponding Email: kodchapown.l@gmail.com
[2]Chulalongkorn University/ Mathematics and Computer Science / Faculty of Science, Bangkok, Thailand
Email: jiraphan.s@chula.ac.th

ABSTRACT

Risk model is important in the measurements of the aggregate net loss of an insurance company's portfolio. It also indicates the probability of losing of the company's portfolio. The sufficiently precise models are highly necessary to the firms to decide their reserve amounts. There have been many studies attempting to develop new models which are more suitable for real data. Recently, the studies of the risk models based on time series claim counts process have gained attention from researchers. In classical risk models, the claim counts are usually assumed to follow the Poisson distributions. However, the Poisson distributions have a drawback that theirs mean and variance are the same. This assumption of Poisson distribution is rarely found in practice since observations are usually overdispersed. Consequently, extensions of Poisson to other distributions has brought interests from researchers to accommodate overdispersed data in many fields of interest. In this paper, we introduce a new class of discrete-time risk models based on the negative binomial time series process. In our study, we derive some probabilistic properties such as generating function, an expression of the adjustment coefficient and an approximation to ruin probabilities. Moreover, numerical examples to calculate ruin probability and the value-at-risk measure for exponential claim sizes are also discussed.

Keywords: risk model; negative binomial; adjustment coefficient; ruin probability; value-at-risk

1 INTRODUCTION

The study of risk model based on the times series claim counts process is one of the rapidly developing field in insurance risk management. For example, Cossette et al. (2010) considered risk models based on several discrete-claim counts time series, such as Poisson Moving Average processes and Poisson Auto Regressive processes. The marginal distributions for claim counts considered is Poisson family having a property that its mean and variance are the same. This assumption of Poisson distribution is rarely found in practice. Therefore, there are many alternative distributions considered in literature for counts data, for instance, generalized Poisson, zero truncated Poisson, zero-inflated Poisson, Poisson-geometric, geometric and negative binomial were applied in univariate integer-valued time series models. Some of those distributions were introduced in the context of risk models. For example, Hu et al. (2018) considered a discrete-time risk model based on the first-order integer-valued moving average (INMA(1)) process with compound Poisson distributed innovations. In this paper, we are interested in constructing a risk model whose the distribution of the number of claim has overdispersion. The distribution of interest is the negative binomial distribution. This distribution was applied to analyze count data with dispersion in different applications. For example, Byers et al. (2003) used the negative binomial model to examine a discrete outcome, and Chin and Quddus (2003) used the negative binomial model to analyze traffic accident occurrence. However, based on our best knowledge, there exist no model that use the stationary negative binomial moving average with claim counts in discrete-time risk model available in literature. Therefore, in this study, we construct a discrete-time risk model based on negative binomial claim counts process and study some of its properties and also the ruin probability. The organization of this paper is as follows. In section 2, we introduce a discrete-time risk model based on negative binomial claim counts process. In section 3, we find the adjustment coefficient function and an approximation to ruin probability of the negative binomial INMA(1) risk model. An approximation to value at risk of the net loss process is given in section 4. In section 5, we provide some calculation examples of the adjustment coefficient and the value at risk. Finally, section 6 concludes our work.

2 DISCRETE-TIME RISK MODEL BASED ON NEGATIVE BINOMIAL MOVING AVERAGE MODEL

In this section, we introduce the risk model and the negative binomial moving average process (NBMA). In our discussion, we include the probabilistic properties of the models such as moments, moment generating function, autocovariance function and the stationary property.

Definition 1. *Let R_n be the discrete-time surplus process defined as follows*

$$R_n = u + n\pi - \sum_{i=1}^{n}\sum_{j=1}^{N_i} C_{ij}, \tag{1}$$

where u is the positive initial reserves of the business; π is the premium rate; $\{C_{ij},\, i,\, j = 1, 2, \ldots\}$ is a sequences of the claim size of claim number j^{th} in period i; $\{C_{ij},\, i,\, j = 1, 2, \ldots\}$ are assumed to be independent and identically distributed (i.i.d.) with a light-tailed distribution whose moment generating function (m.g.f.) $m_C(z)$; and N_i is the claim count in the i^{th} period. Moreover, we define $N_{(n)} = \sum_{i=1}^{n} N_i$ be the aggregate claim number for n period; $W_i = \sum_{j=1}^{N_i} C_{ij}$ be the aggregate claim size of period i; $S_n = \sum_{i=1}^{n} W_i$ be the net loss process; and $S_n - n\pi$ be the aggregate net loss process.

Several distributions have been used to model the claim counts distribution $N_i, i = 1, 2, \ldots, n$ such as Poisson distribution (Gourieroux, & Jasiak, 2004; Quddus, 2008); Negative Binomial distribution (Brijs et al., 2008); Copula model (Frees, & Wang, 2006; Zhao, & Zhou, 2012); and Integer-valued-time series for counts (Ma et al., 2015). In this paper, we consider the case where the distribution of claim counts follows the negative binomial MA(1) (NBMA(1)). The model is based on the binomial thinning operator, or simply thinning operator of γ on the integer-value Y defined as

$$\gamma \circ Y = \sum_{j=1}^{Y} \delta_j$$

where $\{\delta_j, j = 1, 2, \ldots\}$ is a sequence of i.i.d. Bernoulli random variable with mean γ and are independent from Y. Based on the binomial thinning, we then define the negative binomial moving average model as in Definition 2 below. Its probabilistic properties are also provided in Proposition 1.

Definition 2. *Let $\{N_n,\, n \in \mathbb{N}\}$ be the negative binomial MA(1) process (NBMA(1)) defined as*

$$N_n = \gamma \circ \varepsilon_{n-1} + \varepsilon_n = \sum_{j=1}^{\varepsilon_{n-1}} \delta_{n-1,j} + \varepsilon_n \quad \text{for } n = 1, 2, \ldots, \tag{2}$$

where $\{\varepsilon_i,\, i = 0, 1, \ldots\}$ is a sequence of i.i.d. negative binomial random variable with parameter (α, p) whose probability mass function defined as

$$f_Y(y) = \binom{y+\alpha-1}{y} p^y (1-p)^{\alpha} \tag{3}$$

and $\{\delta_{n-1,j}, n, j = 1, 2, \ldots\}$ is a sequences of i.i.d. Bernoulli random variables with mean γ.

Proposition 1. *Let* $\{N_n,\ n \in \mathbb{N}\}$ *defined as Definition 2, then* $\{N_n\ n \in \mathbb{N}\}$ *has the following properties.*

(a) $\{N_n,\ n \in \mathbb{N}\}$ *is a stationary process.*

(b) $G_{N_n}(z) = \left(\frac{1-p}{1-pz}\right)^{\alpha}\left(\frac{1-p}{1-p(\bar{\gamma}+\gamma z)}\right)^{\alpha}$ *for* $n = 1, 2, \ldots,$ $\bar{\gamma} = 1-\gamma$, $z < \frac{1}{p}$ *and* $1 - p(\bar{\gamma}+\gamma z) > 0$

(c) $\mathrm{E}(N_n) = \frac{\alpha p}{1-p}(1+\gamma)$ *for* $n = 1, 2, \ldots$

(d) $\mathrm{Var}(N_n) = \frac{\alpha p(\gamma^2 p+\gamma(1-p)+1)}{(1-p)^2}$ *for* $n = 1, 2, \ldots$

(e) $\mathrm{Cov}(N_n, N_{n-k}) = \begin{cases} \frac{\alpha\gamma p}{(1-p)^2} & \text{for } k = 1 \\ 0 & \text{for } k > 1 \end{cases}$

Proof. To prove (a) , consider the probability generating function (p.g.f.) of $\{N_n,\ n \in \mathbb{N}\}$. We know that $\{\varepsilon_i,\ i = 0, 1, \ldots\}$ is a sequence of i.i.d. $NB(\alpha, p)$ random variable. For $n = 1$,

$$\begin{aligned} G_{N_1}(z) &= \mathrm{E}(z^{N_1}) \\ &= \mathrm{E}\left(z^{\gamma\circ\varepsilon_0+\varepsilon_1}\right) \\ &= \mathrm{E}\left(z^{\varepsilon_1}\right)\mathrm{E}\left(z^{\gamma\circ\varepsilon_0}\right) \\ &= G_{\varepsilon_1}(z)G_{\varepsilon_0}(1-\gamma+\gamma z). \end{aligned}$$

For $n > 1$, we use the fact that $\mathrm{E}\left(z^{\varepsilon_1}\right) = \mathrm{E}\left(z^{\varepsilon_2}\right) = \cdots = \mathrm{E}\left(z^{\varepsilon_n}\right)$ and $\mathrm{E}\left(z^{\gamma\circ\varepsilon_0}\right) = \mathrm{E}\left(z^{\gamma\circ\varepsilon_1}\right) = \cdots = \mathrm{E}\left(z^{\gamma\circ\varepsilon_{n-1}}\right)$. Therefore, the p.g.f. of N_n is

$$\begin{aligned} G_{N_n}(z) &= \mathrm{E}(z^{N_n}) \\ &= \mathrm{E}\left(z^{\gamma\circ\varepsilon_{n-1}+\varepsilon_n}\right) \\ &= \mathrm{E}\left(z^{\varepsilon_n}\right)\mathrm{E}\left(z^{\gamma\circ\varepsilon_{n-1}}\right) \\ &= \mathrm{E}\left(z^{\varepsilon_1}\right)\mathrm{E}\left(z^{\gamma\circ\varepsilon_0}\right) \\ &= G_{\varepsilon_1}(z)G_{\varepsilon_0}(1-\gamma+\gamma z). \end{aligned}$$

Hence,

$$G_{N_1}(z) = G_{N_n}(z) \quad \text{for all} \quad n \in \mathbb{N}$$

That means $\{N_n,\ n \in \mathbb{N}\}$ is a stationary process.
(b) Since$\{\varepsilon_i, i = 0, 1, 2, \ldots\}$ is a negative binomial with parameter (α, p) whose distribution defined in equation (3),

$$\begin{aligned} G_{N_n}(z) &= G_{\varepsilon_1}(z)G_{\varepsilon_0}(1-\gamma+\gamma z) \\ &= \Big(\frac{1-p}{1-pz}\Big)^{\alpha}\Big(\frac{1-p}{1-p(\bar{\gamma}+\gamma z)}\Big)^{\alpha} \end{aligned}$$

for all $n = 1, 2, \ldots, |z| < \frac{1}{p}$ and $1 - p(\bar{\gamma}+\gamma z) > 0$.

(c) Sine $G_{N_n}(z) = \mathrm{E}(z^{N_n})$ for all $n \in \mathbb{N}$, we can use the m.g.f. $G_{N_n}(z)$ to find $\mathrm{E}(N_n)$ as follows

$$\begin{aligned} \mathrm{E}(N_n) &= \frac{dG_{N_n}(z)}{dz}\Big|_{z=1} \\ &= \Big(\frac{\alpha(1-p)^{\alpha}p}{(1-pz)^{\alpha+1}}\Big)\Big(\frac{1-p}{1-p(\bar{\gamma}+\gamma z)}\Big)^{\alpha}\Big|_{z=1} \\ &\quad + \Big(\frac{1-p}{1-pz}\Big)^{\alpha}\Big(\frac{\alpha(1-p)^{\alpha}\gamma p}{(1-p(\bar{\gamma}+\gamma z))^{\alpha+1}}\Big)\Big|_{z=1} \\ &= \frac{\alpha p}{1-p} + \frac{\alpha\gamma p}{1-p} \\ &= \frac{\alpha p}{1-p}(1+\gamma) \end{aligned}$$

(d) For $n = 1, 2, \ldots$, we have

$$\begin{aligned} \mathrm{E}(N_n^2) &= \frac{d^2G_{N_n}(z)}{dz^2}\Big|_{z=1} + \frac{dG_{N_n}(z)}{dz}\Big|_{z=1} \\ &= \frac{\alpha p^2(\alpha\gamma^2+2\alpha\gamma+\gamma^2+\alpha+1)}{(1-p)^2} + \frac{\alpha p}{1-p}(1+\gamma). \end{aligned}$$

Consequently,

$$\begin{aligned} \mathrm{Var}(N_n) &= \frac{\alpha p^2(\alpha\gamma^2+2\alpha\gamma+\gamma^2+\alpha+1)}{(1-p)^2} \\ &\quad + \frac{\alpha p}{1-p}(1+\gamma) - \Big(\frac{\alpha p}{1-p}(1+\gamma)\Big)^2 \\ &= \frac{\alpha p(\gamma^2 p+\gamma(1-p)+1)}{(1-p)^2}. \end{aligned}$$

(e) Since $\{\delta_{n-1,j}, n, j = 1, 2, ...\}$ is a sequences of i.i.d. Bernoulli random variables, we obtain $\mathrm{E}(\delta_{11}) = \mathrm{E}(\delta_{12}) = \cdots = \gamma$.
Then

$$\begin{aligned} \mathrm{E}(\gamma\circ\varepsilon_{n-1}) &= \mathrm{E}(\sum_{j=1}^{\varepsilon_{n-1}}\delta_{n-1,j}) \\ &= \mathrm{E}(E(\sum_{j=1}^{\varepsilon_{n-1}}\delta_{n-1,j}|\varepsilon_{n-1})) \\ &= \mathrm{E}(\varepsilon_{n-1}\,\mathrm{E}(\delta_{11})) \\ &= \mathrm{E}(\delta_{11})\,\mathrm{E}(\varepsilon_{n-1}) \\ &= \gamma\,\mathrm{E}(\varepsilon_{n-1}). \end{aligned} \tag{4}$$

Next, consider

$$\begin{aligned} \mathrm{E}((\gamma\circ\varepsilon_{n-1})\varepsilon_{n-1}) &= \mathrm{E}(\varepsilon_{n-1}\sum_{j=1}^{\varepsilon_{n-1}}\delta_{n-1,j}) \\ &= \mathrm{E}\left(\mathrm{E}(\varepsilon_{n-1}\sum_{j=1}^{\varepsilon_{n-1}}\delta_{n-1,j}|\varepsilon_{n-1})\right) \\ &= \mathrm{E}\left(\varepsilon_{n-1}\,\mathrm{E}(\sum_{j=1}^{\varepsilon_{n-1}}\delta_{i-1,j}|\varepsilon_{n-1})\right) \\ &= \mathrm{E}(\varepsilon_{n-1}^2\,\mathrm{E}(\delta_{11})) \\ &= \mathrm{E}(\delta_{11})\,\mathrm{E}(\varepsilon_{n-1}^2). \end{aligned} \tag{5}$$

For $k = 1$, we have

$$\begin{aligned} \mathrm{Cov}(N_n, N_{n-1}) &= \mathrm{Cov}(\gamma\circ\varepsilon_{n-1}+\varepsilon_n, \gamma\circ\varepsilon_{i-2}+\varepsilon_{n-1}) \\ &= \mathrm{Cov}(\gamma\circ\varepsilon_{n-1}, \varepsilon_{n-1}) \\ &= \mathrm{E}((\gamma\circ\varepsilon_{n-1})\varepsilon_{n-1}) - \mathrm{E}(\gamma\circ\varepsilon_{n-1})\,\mathrm{E}(\varepsilon_{n-1}). \end{aligned} \tag{6}$$

Substituting equations (4) and (5) into (6) then

$$\begin{aligned} \mathrm{Cov}(N_n, N_{n-1}) &= \mathrm{E}(\delta_{11})\,\mathrm{E}(\varepsilon_{n-1}^2) - \mathrm{E}(\delta_{11})\,\mathrm{E}^2(\varepsilon_{n-1}) \\ &= \mathrm{E}(\delta_{11})\,\mathrm{Var}(\varepsilon_{n-1}) \\ &= \frac{\alpha\gamma p}{(1-p)^2} \end{aligned}$$

For $k \geq 2$, we have

$$\begin{aligned} \mathrm{Cov}(N_n, N_{n-k}) &= \mathrm{Cov}(\gamma\circ\varepsilon_{n-1}+\varepsilon_n, \gamma\circ\varepsilon_{n-k-1}+\varepsilon_{n-k}) \\ &= 0 \end{aligned}$$

□

3 APPROXIMATIONS TO RUIN PROBABILITIES

In this section, we give a brief review of the time of ruin and the Lundberg adjustment coefficient approximation. We also derive the adjustment coefficient, approximate the ruin probability of the NBMA(1) risk model and discuss a special case when the distribution of claim size is an exponential distribution.

Definition 3. *Let* T *be the times of ruin, the first time that the surplus becomes negative, defined as follows*

$$T = \inf\{n|R_n \leq 0, n \in \mathbb{N}^+\}. \tag{7}$$

Then the ruin probability given the initial capital u *is given by*

$$\Psi(u) = P\{T < \infty|R_0 = u\}. \tag{8}$$

The ruin probability is difficult to calculate in general and approximations are usually required in many applications. For example, Cossette et al. (2011) proposed the approximation to ruin probability of a dependent sequence and does not require the assumption of the claim size C_{ij}. Using the asymptotic Lundberg-type result

$$\lim_{u\to\infty} -\frac{\ln(\Psi(u))}{u} = R$$

where R is the Lundberg adjustment coefficient.
Let the adjustment coefficient function $c(z)$ defined as following

$$c(z) = \lim_{n\to\infty}\frac{1}{n}c_n(z),$$

where $c_n(z)$ is the cumulate generating function of the aggregate net loss process defined by

$$c_n(z) = \ln E(e^{z(S_n-n\pi)}).$$

For $z_0 > 0$, if the adjustment coefficient function $c(z)$ for all $0 < z < z_0$ exists, then there also exists $z \in (0, z_0)$ such that $c(z) = 0$ and the positive zero-root z is the adjustment coefficient R. Then the ruin probability $\Psi(u)$ can be approximated by

$$\Psi(u) \simeq e^{-Ru}. \tag{9}$$

3.1 Adjustment coefficient function and the solution of the adjustment coefficient function

Theorem 1. *Let R_n be the discrete-time surplus process in Definition 1 where N_i be a NBMA(1) process in Definition 2. The adjustment coefficient function $c(z)$ of R_n is*

$$c(z) = \alpha \log\Big(\frac{1-p}{1-p(\bar{\gamma} m_C(z) + \gamma m_C^2(z))}\Big) - \pi z, \tag{10}$$

for $z \in \mathbb{R}^+$ such that $m_C(z) < \frac{1}{p}$, $1 - p(\bar{\gamma} + \gamma m_C(z)) > 0$, and $1 - p(\bar{\gamma} m_C(z) + \gamma m_C^2(z)) > 0$.

Proof. Consider the aggregate net loss process $c_n(z)$ defined as

$$c_n(z) = \log m_{S_n}(z) - n\pi z, \tag{11}$$

then

$$c(z) = \lim_{n\to\infty} \frac{1}{n} \log m_{S_n}(z) - \pi z. \tag{12}$$

Since $\{C_{i,j}, i, j = 1, 2, \ldots\}$ is a sequences of i.i.d., we have $m_{C_{11}}(z) = m_{C_{12}}(z) = \cdots = m_C(z)$. From $S_n = \sum_{i=1}^{n} \sum_{j=1}^{N_i} C_{ij}$,

$$\begin{aligned} m_{S_n}(z) &= \mathrm{E}(e^{zS_n}) \\ &= \mathrm{E}\big(\mathrm{E}(e^{z(\sum_{j=1}^{N_1} C_{1j} + \cdots + \sum_{j=1}^{N_n} C_{nj})} | N_1, \ldots, N_n)\big) \\ &= \mathrm{E}\Big(\prod_{j=1}^{N_1} \mathrm{E}(e^{zC_{1j}}) \prod_{j=1}^{N_2} \mathrm{E}(e^{zC_{2j}}) \cdots \prod_{j=1}^{N_n} \mathrm{E}(e^{zC_{nj}})\Big) \\ &= \mathrm{E}((m_C(z))^{N_1+N_2+\cdots+N_n}) \\ &= G_{N_{(n)}}(m_C(z)), \end{aligned} \tag{13}$$

where $G_{N_{(n)}}(z)$ is the p.g.f of $N_{(n)}$. Then

$$\begin{aligned} G_{N_{(n)}}(z) &= \mathrm{E}(z^{N_1+N_2+\cdots+N_n}) \\ &= \mathrm{E}(z^{(\gamma\circ\varepsilon_0+\varepsilon_1)+(\gamma\circ\varepsilon_1+\varepsilon_2)+\cdots+(\gamma\circ\varepsilon_{n-1}+\varepsilon_n)}) \\ &= \mathrm{E}(z^{\varepsilon_n}) \mathrm{E}(z^{\gamma\circ\varepsilon_0}) \prod_{i=1}^{n-1} \mathrm{E}(z^{\gamma\circ\varepsilon_i+\varepsilon_i}) \\ &= \mathrm{E}(z^{\varepsilon_n}) \mathrm{E}(z^{\sum_{j=1}^{\varepsilon_0} \delta_{1j}}) \prod_{i=1}^{n-1} \mathrm{E}(z^{\varepsilon_i + \sum_{j=1}^{\varepsilon_i} \delta_{i,j}}). \end{aligned} \tag{14}$$

Because the first part of equation (14) is the p.g.f of $NB(\alpha, p)$, we obtain

$$\mathrm{E}(z^{\varepsilon_n}) = \Big(\frac{1-p}{1-pz}\Big)^{\alpha} \text{ for } 0 < z < \frac{1}{p}. \tag{15}$$

Because $\{\delta_{i-1,j}, i, j = 1, 2, ...\}$ is a sequence of i.i.d. Bernoulli random variables with mean γ, $\mathrm{E}(z^{\delta_{11}}) = \mathrm{E}(z^{\delta_{12}}) = \cdots = \mathrm{E}(z^{\delta_{i,j}}) = \bar{\gamma} + \gamma z$ for all $i = 0, 1, \ldots,$ and $j = 1, 2, \ldots$

Then, the second part of equation (14) is calculated as follows

$$\begin{aligned} \mathrm{E}(z^{\sum_{j=1}^{\varepsilon_0} \delta_{1j}}) &= \mathrm{E}\big(\mathrm{E}(z^{\sum_{j=1}^{\varepsilon_0} \delta_{1j}} | \varepsilon_0)\big) \\ &= \mathrm{E}\Big(\prod_{j=1}^{\varepsilon_0} \mathrm{E}(z^{\delta_{1j}})\Big) \\ &= \mathrm{E}\big((\bar{\gamma} + \gamma z)^{\varepsilon_0}\big) \\ &= \Big(\frac{1-p}{1-p(\bar{\gamma} + \gamma z)}\Big)^{\alpha}, \end{aligned} \tag{16}$$

for $z \in \{z \in \mathbb{R}^+ | 1 - p(\bar{\gamma} + \gamma z) > 0\}$.

For the third part of equation (14), we get

$$\begin{aligned} \mathrm{E}(z^{\varepsilon_i + \sum_{j=1}^{\varepsilon_i} \delta_{i,j}}) &= \mathrm{E}\big(\mathrm{E}(z^{\varepsilon_i + \sum_{j=1}^{\varepsilon_i} \delta_{i,j}} | \varepsilon_i)\big) \\ &= \mathrm{E}\Big(z^{\varepsilon_i} \prod_{j=1}^{\varepsilon_i} \mathrm{E}(z^{\delta_{i,j}})\Big) \\ &= \mathrm{E}\big((\bar{\gamma} z + \gamma z^2)^{\varepsilon_i}\big) \\ &= \Big(\frac{1-p}{1-p(\bar{\gamma} z + \gamma z^2)}\Big)^{\alpha}, \end{aligned} \tag{17}$$

for $z \in \{z \in \mathbb{R}^+ | 1 - p(\bar{\gamma} z + \gamma z^2) > 0\}$.

Substituting equations (15), (16) and (17) into equation (14), we obtain

$$\begin{aligned} G_{N_{(n)}}(z) = &\Big(\frac{1-p}{1-pz}\Big)^{\alpha} \Big(\frac{1-p}{1-p(\bar{\gamma} + \gamma z)}\Big)^{\alpha} \\ &\times \Big(\frac{1-p}{1-p(\bar{\gamma} z + \gamma z^2)}\Big)^{\alpha(n-1)} \end{aligned} \tag{18}$$

where $z \in \{z \in \mathbb{R}^+ | z < \frac{1}{p} \cap 1 - p(\bar{\gamma} + \gamma z) > 0 \cap 1 - p(\bar{\gamma} z + \gamma z^2) > 0 \cap 1 - p(\bar{\gamma} m_C(z) + \gamma m_C^2(z)) > 0\}$.

Therefore,

$$\begin{aligned} m_{S_n}(z) = &\Big(\frac{1-p}{1-pm_C(z)}\Big)^{\alpha} \Big(\frac{1-p}{1-p(\bar{\gamma} + \gamma m_C(z))}\Big)^{\alpha} \\ &\times \Big(\frac{1-p}{1-p(\bar{\gamma} m_C(z) + \gamma m_C^2(z))}\Big)^{\alpha(n-1)}, \end{aligned} \tag{19}$$

for $z \in \mathbb{R}^+$ such that $m_C(z) < \frac{1}{p}$, $1 - p(\bar{\gamma} + \gamma m_C(z)) > 0$, and $1 - p(\bar{\gamma} m_C(z) + \gamma m_C^2(z)) > 0$.

Consequently,

$$\begin{aligned} c_n(z) = &\alpha \log\Big(\frac{1-p}{1-pm_C(z)}\Big) + \alpha \log\Big(\frac{1-p}{1-p(\bar{\gamma} + \gamma m_C(z))}\Big) \\ &+ \alpha(n-1) \log\Big(\frac{1-p}{1-p(\bar{\gamma} m_C(z) + \gamma m_C^2(z))}\Big) - n\pi z, \end{aligned}$$

for $z \in \mathbb{R}^+$ such that $m_C(z) < \frac{1}{p}$, $1 - p(\bar{\gamma} + \gamma m_C(z)) > 0$, and $1 - p(\bar{\gamma} m_C(z) + \gamma m_C^2(z)) > 0$.

Hence,

$$c(z) = \alpha \log\Big(\frac{1-p}{1-p(\bar{\gamma} m_C(z) + \gamma m_C^2(z))}\Big) - \pi z,$$

for $z \in \mathbb{R}^+$ such that $m_C(z) < \frac{1}{p}$, $1 - p(\bar{\gamma} + \gamma m_C(z)) > 0$, and $1 - p(\bar{\gamma} m_C(z) + \gamma m_C^2(z)) > 0$. □

Let

$$\pi = \big(\frac{\alpha}{1-p}\big) p(1+\gamma) E(C)(1+\rho) \tag{20}$$

for $\rho > 0$, and $D = \{z \in \mathbb{R}^+ | m_C(z) < \frac{1}{p} \cap 1 - p(\bar{\gamma} + \gamma m_C(z)) > 0 \cap 1 - p(\bar{\gamma} m_C(z) + \gamma m_C^2(z)) > 0\}$. From the assumptions of $\{C_{ij}, i, j = 1, 2, \ldots\}$ i.i.d. light-tailed distributions, we obtain that $z > 0$ and D is nonempty.

Lemma 1. *From the expression for adjustment coefficient function of $NBMA(1)$, the equation $c(z) = 0$ has the unique solution in D.*

Proof. To prove the lemma, we want to claim that

(a) $\frac{dc(z)}{dz}\Big|_{z=0} < 0$,

(b) $\frac{d^2c(z)}{dz^2} > 0$ for $z \in D$,

(c) There exists $z^* \in \overline{D}$ such that $\lim_{z_1 \to z^{*-}} c(z) = +\infty$ where $\overline{D}$ is the closure of D.

(a) Note that,

$$\frac{dc(z)}{dz} = -\pi + \frac{\alpha p(\bar{\gamma} m_C'(z) + 2\gamma m_C(z) m_C'(z))}{1 - p(\bar{\gamma} m_C(z) + \gamma m_C^2(z))}$$

then,

$$\frac{dc(z)}{dz}\Big|_{z=0} = -\pi + \Big(\frac{\alpha}{1-p}\Big) p(1+\gamma) E(C) < 0$$

(b) Since $z \in D$, $m_C'(z) > 0$, and $m_C''(z) > 0$;

$$\begin{aligned} \frac{d^2c(z)}{dz^2} = &\frac{\alpha p(\bar{\gamma} m_C''(z) + 2\gamma m_C(z) m_C''(z) + 2\gamma (m_C'(z))^2)}{1 - p(\bar{\gamma} m_C(z) + \gamma m_C^2(z))} \\ &+ \frac{\alpha p \big(\bar{\gamma} m_C'(z) + 2\gamma m_C(z) m_C'(z)\big)^2}{\big(1 - p(\bar{\gamma} m_C(z) + \gamma m_C^2(z))\big)^2} > 0. \end{aligned}$$

(c) Consider $1 - p(\bar{\gamma} m_C(z) + \gamma m_C^2(z))$ for $z \in D$.

Note that

$$1 - p(\bar{\gamma} m_C(0) + \gamma m_C^2(0)) = 1 - p > 0.$$

Because $m_C(z)$ is increasing and continuous function on $[0, \infty)$.

Then we obtain that

$$1 - p(\bar{\gamma} m_C(z) + \gamma m_C^2(z)) \tag{21}$$

is decreasing and continuous function. We also obtain

$$\lim_{z \to z^{*-}} 1 - p(\bar{\gamma} m_C(z) + \gamma m_C^2(z)) = 0,$$

and

$$1 - p(\bar{\gamma} m_C(z) + \gamma m_C^2(z)) \geq 0 \text{ for all } 0 \leq z \leq z^*.$$

Hence,

$$\lim_{z \to z^{*-}} \alpha \log\Big(\frac{1-p}{1-p(\bar{\gamma} m_C(z) + \gamma m_C^2(z))}\Big) = +\infty.$$

□

3.2 *Special case: exponential claim sizes*

In this section, we consider the cases when the claim amounts can be modeled by an exponential distribution. That is, the case when $\{C_{ij},\, i,\, j = 1,\, 2, \ldots\}$ is a sequences of i.i.d. exponential distribution with parameter β where the m.g.f. is $m_{C_{11}}(z) = \frac{1}{1-z/\beta}$. The adjustment coefficient function in equation (10) becomes

$$c(z) = \alpha \log\Big(\frac{1-p}{1-p(\frac{\bar{\gamma}\beta}{\beta - z} + \frac{\gamma\beta^2}{(\beta-z)^2})}\Big) - \pi z \tag{22}$$

where $z \in \{z \in \mathbb{R}^+ | 1 - p(\frac{\bar{\gamma}\beta}{\beta-z} + \frac{\gamma\beta^2}{(\beta-z)^2}) > 0\}$.
Let $C_{ij} \sim Exp(\beta)$ for $i, j = 1, 2, \ldots$, we obtain

$$\pi = \left(\frac{\alpha}{1-p}\right)\frac{p}{\beta}(1+\gamma)(1+\rho)$$

and we obtain $D = \{z \in \mathbb{R}^+ | 1 - p\left(\frac{\bar{\gamma}}{\beta} + \gamma\big(\frac{\beta}{\beta-z}\big)^2\right) > 0 \text{ and } z < \beta\}$.

4 APPROXIMATION TO THE VALUE-AT-RISK

In this section, we derive an approximation to the value-at-risk at the confidence level λ, $\text{VaR}_\lambda(S_n)$, for NBMA(1) process, where $S_n = \sum_{i=1}^{n}\sum_{j=1}^{N_i} C_{ij}$ be the net loss process in Definition 1 and N_i be a NBMA(1) process in Definition 2.
Then the $\text{VaR}_\lambda(S_n)$ is defined as

$$\text{VaR}_\lambda(S_n) = \inf\{k \in \mathbb{R} | F_{S_n}(k) > \lambda\} \tag{23}$$

where $F_{S_n}(k)$ is the cumulative distribution function of S_n. Refer to equation (19), the m.g.f. of S_n, it is hard to obtain the distribution of S_n. Therefore, we apply the Fast Fourier Transform (FFT) algorithm (Gray, & Pitts, 2012) to obtain an approximation of the density function of F_{S_n}. From equation (13), we know that

$$m_{S_n}(z) = G_{N_{(n)}}(m_C(z))$$

where $m_C(z)$ is the m.g.f. of C_{ij} for all $i, j = 1, 2, \ldots$
Let $\phi_C(\theta)$ be the characteristic function of C_{ij} $(i, j = 1, 2, \ldots)$. We obtain the characteristic function of S_n as

$$\begin{aligned}\phi_{S_n}(\theta) &= G_{N_{(n)}}\Big(\phi_C(\theta)\Big) \\ &= \Big(\frac{1-p}{1-p\phi_C(\theta)}\Big)^{\alpha}\Big(\frac{1-p}{1-p(\bar{\gamma} + \gamma\phi_C(\theta))}\Big)^{\alpha} \\ &\quad \times \Big(\frac{1-p}{1-p(\bar{\gamma}\phi_C(\theta) + \gamma\phi_C^2(\theta))}\Big)^{\alpha(n-1)}\end{aligned}$$

where $\theta \in \{\theta \in \mathbb{R}^+ | \phi_C(\theta) < \frac{1}{p} \cap 1 - p(\bar{\gamma} + \gamma\phi_C(\theta)) > 0 \cap 1 - p(\bar{\gamma}\phi_C(\theta) + \gamma\phi_C^2(\theta)) > 0\}$.
By applying the inverse FFT algorithm, we obtain an approximation to density of S_n and, consequently, an approximation to the $F_{S_n}(k)$. Finally, $\text{VaR}_\lambda(S_n)$ is obtained.

5 NUMERICAL EXPERIMENTS AND SIMULATIONS

In this section, we provide the example to calculate the adjustment coefficient and approximation to the ruin probability of risk model based on NBMA(1) claim counts process. We also show the calculation of value at risk of the net loss process S_n of period 10th at the confidence levels $\lambda = 0.90$ and 0.95 shown in Table 5.2.

5.1 *Calculation of adjustment coefficient of risk model based on NBMA(1)*

Let R_n be the discrete-time surplus process defined as definition 1, and $\{N_i,\, i = 1,\, 2, \ldots, n\}$ is sequences of $NBMA(1)$ claim counts process defined as definition 2. Let $\{C_{ij},\, i,\, j = 1,\, 2, \ldots\}$ is sequence of i.i.d. exponential distribution with parameter β. Then $m_{C_{i,j}}(z) = \frac{1}{1-z/\beta}$ for all $i, j = 1, 2, \ldots$ We define $u = 2$, $(\alpha, p) = (0.8, 0.2)$, $\beta = 0.5$, and $\rho = 0.4$. Table 1 shows the adjustment coefficients z^* and the upper bound of the ruin probability of R_n, $\Psi_{R_n}(u)$, by varying $\gamma = (0, 0.25, 0.5, 0.75, 1)$. The results show that the ruin probability increases when the adjustment coefficient decreases, and the ruin probability increases as a function of the dependence level γ. This seems to be reasonable because the greater value of γ, the greater the number of claim.

Table 1: The adjustment coefficients z_1^* and the upper bound $\Psi_{R_n}(u)$

γ	0	0.25	0.50	0.75	1
z^*	0.1268	0.1059	0.0953	0.0884	0.0836
$\Psi_{R_n}(u)$	0.7760	0.8091	0.8265	0.8379	0.8461

5.2 *Calculation of the value-at-risk for model NBMA(1)*

Let the time period n be 10 and divide the domain of $\{C_{i,j},\, i, j = 1, 2, \ldots\}$ to be 2^{10} parts whose length of steps are 0.04 for the FFT distribution approximation. Table 2 shows $\text{VaR}_\lambda(S_{10})$ for the confidence level $\lambda = 0.90$ and 0.95. The results show that the $\text{VaR}_\lambda(S_n)$ increases as a function of the dependence level γ.

Table 2: The value-at-risk for NBMA(1) model for the confidence level λ

γ	0	0.25	0.50	0.75	1
$\text{VaR}_{0.90}(S_{10})$	9.86	12.10	14.18	16.14	17.98
$\text{VaR}_{0.95}(S_{10})$	12.50	15.30	17.78	20.06	22.18

6 CONCLUSION

In this study, we have proposed an alternative risk model based on NBMA(1) claim counts process. The new model can deal with the overdispersed claim-counts data. The adjustment coefficient function of risk model based on NMMA(1) was provided. We also proved that the adjustment coefficient function has a unique solution. Finally, we showed some calculations of ruin probability and the value-at-risk of the net loss process.

ACKNOWLEDGEMENTS

The first author would like to thank the development and promotion of science and technology talents project (DPST) and the department of mathematics and computer science of Chulalongkorn university for the financial support.

REFERENCES

Brijs, T., Karlis, D., & Wets, G. (2008). Studying the effect of weather conditions on daily crash counts using a discrete time-series model. *Accident Analysis and Prevention*, 40(3), 1180–1190.

Byers, A. L., Allore, H., Gill, T. M., & Peduzzi, P. N. (2003) Application of negative binomial modeling for discrete outcomes: A case study in aging research, *Journal of Clinical Epidemiology*, 56(6), 559-564.

Chin, H. C., & Quddus, M. A. (2003). Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. *Accident Analysis Prevention*, 35(2), 253 – 259.

Cossette, H., Marceau, E., & Maume-Deschamps, V. (2010). Discrete time risk models based on time series for count random variables. *ASTIN Bulletin*, 40(1), 123–150.

Cossette, H., Marceau, E., & Maume-Deschamps, V. (2011). Adjustment coefficient for risk processes in some dependent contexts. *Methodology and Computing in Applied Probability*, 13(4), 695–721.

Frees, E. W., & Wang, P. (2006). Copula credibility for aggregate loss models. *Insurance: Mathematics and Economics*, 38(2), 360-373.

Gourieroux, C., & Jasiak, J. (2004). Heterogeneous INAR(1) model with application to car insurance. *Insurance Mathematics and Economics*, 34(2), 177-192.

Gray, R. J. & Pitts, S. M. (2012). The fast Fourier transform algorithm. *Risk Modelling in General Insurance* (pp. 119-124). Cambridge: Cambridge University Press.

Hu, X., Zhang, L., & Sun, W. (2018). Risk model based on the first-order integer-valued moving average process with compound Poisson distributed innovations, *Scandinavian Actuarial Journal*, 2018(5), 412-425.

Ma, D., Wang, D., & Cheng, J. (2015). Bidimensional discrete-time risk models based on bivariate claim count time series. *Journal of Inequalities and Applications*, 2015(1), 105.

Quddus, M. A. (2008). Time series count data models: An empirical application to traffic accidents. *Accident Analysis & Prevention*, 40(5), 1732-1741.

Zhao, X., & Zhou, X. (2012). Copula models for insurance claim numbers with excess zeros and time-dependence. *Insurance: Mathematics and Economics*, 50(1), 191–199.

Wilcoxon Rank Sum Resampling Test for Two Samples with Clustered Data

Prayad Sangngam[1*] and Wipawan Laoarun[2]

[1,2]Department of Statistics, Faculty of Science, Silpakorn University, NakornPathom, Thailand
*Corresponding Email: sangngam_p@su.ac.th
Email: laoarun_w@silpakorn.edu

ABSTRACT

The Wilcoxon rank sum test is commonly used to test the equality of two independent population distributions. In many practical situations, the data in each group are clustered. The Wilcoxon rank sum test was developed for testing the differences of two groups with clustered data by Rosner et al. (2003). This paper proposed the Wilcoxon rank sum test for clustered data using within-cluster resampling method. We showed that the proposed test has an asymptotic normal distribution. The efficiency of the proposed test was compared to the Rosner's test and Datta and Satten (2005) test by using simulation study. The three equal size groups are 5, 10 and 15 clusters and the cluster sizes of three are considered. The coefficient of correlation between observations in a cluster was set to be 0.3, 0.4, 0.5, 0.6 and 0.7. The correlation structure of observations in a cluster is exchangeable. The effect size equals to 0.0, 0.1, 0.2, 0.3, 0.4, 0.5 and 0.6. The results showed that the proposed test can maintain the size of the test for all situations. In some situations, the power of the proposed test is higher than that of the Rosner's test.

Keywords: clustered data; power of the test; Wilcoxon rank sum; two independent samples

1 INTRODUCTION

In many studies, the researchers are interested in testing the null hypothesis that the two independent samples have been drawn from the same population or from the populations with equal means. In parametric statistics, the two independent t-test is widely used. This test requires that the random samples are drawn from normal distribution. If the assumptions of t-test cannot be met, the Wilcoxon rank sum test (Wilcoxon, 1945) which is nonparametric procedure can be used. The common assumption of the t-test and Wilcoxon rank sum test is that all observations are independent.

For many situations, the data are the clusters of correlated observations. The cluster may be a family, a litter, a laboratory and a region or strata. Examples of clustered data are the repeated measurements of blood pressure for a single object, the socio-economic characteristic of households in a block or the body mass index of siblings. Wu et al. (1988) showed that the F-test for clustered data leads to an inflated type I error rate. This means that the probability of type I error is higher than a given significance level. In addition, the type I error rate increases as the intra-correlation increases.

In parametric approach, many researchers have considered different procedures to test the mentioned hypothesis with correlated clustered data. Most of the theoretical research for clustered data assumes a parametric model. Wu et al. (1988) adjusted the F-test statistic by using intra-correlation so that the adjusted statistic has approximately the F distribution with the same degrees of freedom as those of the F-test statistic. Rao et al. (1993) proposed a two-stage general least squared test by transforming the observations to be uncorrelated ones. The two mentioned tests depend on the unknown intra-correlation. However, Lahiri and Li (2009) proposed an alternative test which does not require the estimation of the intra-correlation.

In nonparametric approach, only a small amount of literature exists for incorporating clustered data. Rosner and Grove (1999) considered the combination of clustered data in the Mann-Whitney U test. The estimates of correlation parameters were used to correct the estimated variance of the test statistic. The test has appropriate type I error rate in balanced design with as few as 20 clusters per group. Rosner et al. (2003) introduced a large sample randomization test for the clustered data by applying the Wilcoxon rank sum test. Rosner et al. (2006) extended the signed rank test to the clustered data setting. The above nonparametric researches are constructed for one sample or two independent samples. Datta and Satten (2005) developed a nonparametric rank-sum test for clustered data using resampling one observation per cluster. When the number of clusters is small, the asymptotic normal distribution of Datta and Satten' test may be violated.

Therefore, this research will develop an alternative nonparametric test statistic for clustered data using the resampling approach. The asymptotic distribution of the proposed test will be derived. The type I error rate of the proposed test is investigated. The statistical power of the proposed test and the existing tests are compared by simulation study

2 EXISTING NONPARAMETRIC TESTS FOR CLUSTERED DATA

Let X_{ij} be the j observation in the cluster i for $1 \le i \le N$, $1 \le j \le n_i$, where n_i is the cluster size of the i-th cluster. The indicator δ_{ij} denotes the group of the samples; $\delta_{ij} = 1$ if X_{ij} belongs to the first sample and $\delta_{ij} = 0$ if X_{ij} belongs to the second sample. The data presented in the form of $(\mathbf{X}, \boldsymbol{\delta}) = \left\{ \left(X_{ij}, \delta_{ij} \right) : 1 \le j \le n_i, 1 \le i \le N \right\}$. We assume that clusters are independent while the observations within cluster are not. The hypothesis to be test is that there is no difference between the location parameters.

2.1 The RGL Test

Rosner et al. (2003) proposed the Wilcoxon rank-sum test for clustered data, namely, the RGL test. Let R_{ij} be the rank of X_{ij} based on the combined samples of all observations. The sum of rank from first sample is assigned to be statistical test. Let $\delta_{ij} = \delta_i$ for all $1 \le j \le n_i$. The RGLtest can be defined as

$$W_c = \sum_{i=1}^{N} \delta_i R_{i+} \quad ,$$

where $R_{i+} = \sum_{j=1}^{n_i} R_{ij}$ is the sum of observation ranks in the i-th cluster.

The RGL method assumes that the observations in a given cluster are exchangeable. The exact distribution of W_c is considered based on random permutation conditioning on the sum of observation ranks in the i-th cluster, R_{i+}. In order to derive the distribution of the test under null hypothesis, the clusters were partitioned by using the cluster size. Let G_{max} be the maximum of cluster sizes. The statistic W_c can be written as

$$W_c = \sum_{g=1}^{G_{max}} \sum_{i \in I_g} \delta_i R_{i+} = \sum_{g=1}^{G_{max}} W_g$$

where I_g is the set of indices of cluster size g and $W_g = \sum_{i \in I_g} \delta_i R_{i+}$.

Let N_g be the number of clusters of size g. Let m_g and n_g be the number of cluster of size g from first and second samples,

respectively. If N is small, the distribution of W_C conditioning on R_{i+} can be generated by combining all possible permutations of R_{i+} in W_g for a given cluster of size g. The total number of permutation is $\prod_{g=1}^{G_{max}} \binom{N_g}{m_g}$. If N is large, the computation is intensive. The RGL asymptotic statistic test is

$$Z = \frac{W_c - m_g\left(R_{++g}/N_g\right)}{\sqrt{\mathrm{Var}\left(W_c\right)}} \sim N(0,1) \qquad (2.1)$$

where $R_{++g} = \sum_{i\in I_g} R_{i+}$ and $\mathrm{Var}\left(W_c\right) = \frac{m_g n_g}{N_g\left(N_g-1\right)} \sum_{i\in I_g}\left(R_{i+} - \frac{R_{++,g}}{N_g}\right)^2$.

Under mild condition, the statistic test Z has an asymptotic standard normal distribution. For imbalance of sample size between two samples, Rosner et al. (2003) showed that the statistical test may results inefficiency. If some clusters from either sample have the different size from other, the permutation will be ignored as no permutation of these clusters can be made.

2.2 The DS Test

Datta and Satten (2005) proposed a statistical test in term of Monte Carlo test, namely, DS test. This test is motivated by the proposal of within cluster resampling of Hoffman et al. (2001). There are two advantages of this resampling method. Firstly, the specification of correlation structure may not be required. Secondly, this method remains valid in the presence of ignorable cluster size. The statistical test is based on sampling an observation from each cluster.

Let X_i^* be the random observation from the i-th cluster corresponding to its group of samples, δ_i^* for $i = 1,2,\ldots,N$. Under the resampling method, the pair of $\left(X_i^*, \delta_i^*\right)$ are independent and it can be used to construct the standard Wilcoxon rank-sum statistic defined as

$$W^* = \frac{1}{N+1}\sum_{i=1}^{N} \delta_i^* R_i^*$$

where R_i^* is the rank of X_i^* among all random observations $\left\{X_j^* : 1 \le j \le N\right\}$. The statistical test is given by averaging W^* over all possible sample of $\left(X_i^*, \delta_i^*\right)$ as

$$Z = \frac{S - E(S)}{\sqrt{\hat{\mathrm{Var}}(S)}} \qquad (2)$$

where $S = E\left(W^* \mid \mathbf{X}, \boldsymbol{\delta}\right)$. Datta and Satten (2005) calculated the quantities need to compute the proposed statistical test. Let $F_i(x) = \frac{1}{n_i}\sum_{k=1}^{n_i} I\left(X_{ik} \le x\right)$ be the empirical distribution function of observation from the i-th cluster and $F_i^-(x) = \frac{1}{n_i}\sum_{k=1}^{n_i} I\left(X_{ik} < x\right)$. They showed that $S = \frac{1}{N+1}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\frac{\delta_{ik}}{n_i}\left\{1 + \frac{1}{2}\sum_{j\ne i}\left[F_j\left(X_{ij}\right) + F_j^-\left(X_{ij}\right)\right]\right\}$. The expectation of the statistic S is $E(S) = \frac{1}{2}\sum_{i=1}^{N}\frac{n_{i1}}{n_i}$ where n_{i1} is the number of membership of first sample in i-th cluster. The variance of S can be estimated by $\hat{\mathrm{Var}}(S) = \sum_{i=1}^{N}\left[\hat{W}_i - E\left(W_i\right)\right]^2$, where

$$\hat{W}_i = \frac{1}{2n_i(N+1)}\sum_{k=1}^{n_i}\left[(N-1)\delta_{ik} - \sum_{j\ne i}\frac{n_{jl}}{n_j}\right]\left[\hat{F}\left(X_{ik}\right) - \hat{F}^-\left(X_{ik}\right)\right],$$

$$E\left(W_i\right) = \frac{N}{2(N+1)}\left[\frac{n_{i1}}{n_i} - \frac{1}{N}\sum_{j=1}^{N}\frac{n_{1j}}{n_j}\right] \text{ and } \hat{F} = \sum_{i=1}^{N} n_i F_i \Big/ \sum_{i=1}^{N} n_i \text{ is the pool}$$

empirical distribution function of the observation. An alternative estimator of the variance is given by

$$\hat{\mathrm{Var}}(S) = \mathrm{Var}\left(W^*\right) - E\left[\mathrm{Var}\left(W^* \mid \mathbf{X}, \mathbf{g}\right)\right].$$

The authors also showed that the statistical test has asymptotic standard normal under mild conditions. The asymptotic theory may be violated for testing contralateral data.

3 THE PROPOSED TEST

Since the RGL test ignores the permutation of a cluster when one sample shows up at a cluster size. In addition, if N is small, the asymptotic theory of the DS test may be violated. Assume that $n_i \ge 2$ for $i = 1,2,\ldots,N$. We consider the within cluster resampling method. The proposed test is constructed by drawing two observations from each cluster. Let $\left(X_{i1}^*, X_{i2}^*\right)$ be a random observations from the i-th cluster and δ_i^* be its group membership. In each cluster, the sampling is the simple random sampling without replacement. Let m and n be the numbers of cluster from the first and the second groups, respectively, where $m+n=N$. The RGL statistic for the random sample is

$$W_c = \sum_{i=1}^{N} \delta_i R_{i+}^* \quad ,$$

where $R_{i+}^* = \sum_{j=1}^{2} R_{ij}$ is the sum of observation ranks in the i-th cluster. Let $R_{++}^* = N(2N+1)$ denote the sum of ranks of all random observations. The process is replicated at a large number of times, Q. Let $W_{c,q}$ denote a resample RGL statistic for the q-th resample; $q = 1,2,\ldots,Q$ .The within-cluster resampling test is constructed as average of the Q resample based tests. The within-cluster resample statistic is given by

$$\bar{W}_c = \frac{1}{Q}\sum_{q=1}^{Q} W_{c,q} \ .$$

The within-cluster resampling statistic is the average of identically distribution but dependent estimates. The central limit theorem is not directly applicable. However, when the number of clusters N is large, under the usual regularity conditions, we can show that the statistic $\bar{W}_c$ has asymptotic normal distribution by rewriting the average as the sum of independent clusters. The proposed test is given by

$$Z = \frac{\bar{W}_c - E\left(\bar{W}_c\right)}{\sqrt{\hat{\mathrm{Var}}\left(\bar{W}_c\right)}} \sim N(0,1) \qquad (3)$$

where $E\left(\bar{W}_c\right) = m\frac{R_{++}^*}{N}$ and $\hat{\mathrm{Var}}\left(\bar{W}_c\right)$ is a consistent estimator of $\mathrm{Var}\left(\bar{W}_c\right)$. The consistent estimator of $\mathrm{Var}\left(\bar{W}_c\right)$ is

$$\hat{\mathrm{Var}}\left(\bar{W}_c\right) = \frac{1}{Q}\sum_{q=1}^{Q}\left[\frac{mn}{N(N-1)}\sum_{i\in I}\left(R_{i+,q} - \frac{R_{++}^*}{N}\right)^2\right] - \frac{(Q-1)}{Q}S_w^2,$$

where $S_W^2 = \frac{1}{Q-1}\sum_{q=1}^{Q}\left(W_{c,q} - \bar{W}_c\right)^2$.

4 SIMULATION STUDY

In this section, we study the properties of the proposed test (PT) statistic and compare to the RGL and DS tests using simulation study. We consider the situation where all elements of a cluster belong to same group which is requirement for using the proposed and the RGL tests. Let N_1 and N_2 denote the number of clusters from first and second groups respectively. When the sample sizes in each sample are equal, the power of the test is high, so we set $N_1 = N_2$. We generate data $X_{ij} = \exp\left(Y_{ij}\right) + \delta_i d$ where $\mathbf{Y}_i = \left(Y_{i1}, Y_{i2}, \ldots, Y_{in_i}\right)$ were independent multivariate normal with mean 0 and exchangeable covariance matrix $(1-\rho)\mathbf{I} + \rho\mathbf{1}$ where $\mathbf{I}$ is the identity matrix of size n_i x n_i and $\mathbf{1}$ is the n_i x n_i matrix of all elements equal to 1.The three equal size groups consisting of 5, 10 and 15 clusters with the cluster sizes of three (n_i =3)

are considered. The coefficient of correlation between observations in a cluster is set to be 0.3, 0.4, 0.5, 0.6 and 0.7. The effect size (d) equals to 0.0, 0.1, 0.2, 0.3, 0.4, 0.5 and 0.6. The significant level is set to be 0.05. For each situation the rejection rate is obtained from 1,000 replicates. The results are summarized in Table 1-3.

Table 1. The rejection rate of PT, RGL and DS methods for the number of clusters in each sample is 5 ($N_1 = N_2$ =5).

ρ	d	PT	RGL	DS
0.3	0	0.044	0.034	0.055
	0.1	**0.051**	0.044	0.059
	0.2	**0.074**	0.069	0.091
	0.3	**0.099**	**0.099**	0.128
	0.4	**0.129**	0.120	0.158
	0.5	0.188	**0.193**	0.238
	0.6	0.240	**0.248**	0.279
0.4	0	0.043	0.038	0.059
	0.1	**0.047**	**0.047**	0.061
	0.2	**0.067**	0.066	0.091
	0.3	**0.114**	0.113	0.147
	0.4	**0.120**	0.112	0.135
	0.5	0.170	**0.179**	0.220
	0.6	0.216	**0.220**	0.268
0.5	0	0.040	0.042	0.057
	0.1	0.045	**0.048**	0.058
	0.2	**0.070**	0.066	0.088
	0.3	**0.109**	0.108	0.144
	0.4	**0.109**	0.107	0.124
	0.5	0.167	**0.173**	0.214
	0.6	0.192	**0.199**	0.254
0.6	0	0.043	0.044	0.060
	0.1	0.044	**0.049**	0.063
	0.2	**0.065**	0.064	0.080
	0.3	**0.094**	0.093	0.134
	0.4	**0.099**	0.096	0.124
	0.5	0.152	**0.160**	0.190
	0.6	0.184	**0.186**	0.234
0.7	0	0.050	0.052	*0.067*
	0.1	0.043	**0.044**	0.066
	0.2	**0.060**	0.058	0.078
	0.3	0.095	**0.101**	0.134
	0.4	0.092	**0.094**	0.112
	0.5	0.146	**0.157**	0.187
	0.6	**0.180**	**0.180**	0.216

Note: The boldface indicates the higher power between PT and RGL tests**.** The italic indicates that the empirical size is out of Cochran's (1947) criterion.

Table 1 and 2 show that when the numbers of clusters in each samples equal to 5 and 10, the empirical size $(d = 0)$ of both the PT and the RGL are close to the significant level 0.05. However the empirical size of DS is far from the true size (0.05) except for $\rho = 0.3$ and $N_i = 5$. For given number of clusters in two samples equal to 5, when $d < 0.5$ and $\rho < 0.5$, the PT test can give the higher empirical power than the RGL tests. Given $N_i = 10$, when the effect size is small $(d < 0.3)$ the empirical power of the PS test is higher than that of the RGL test.

Table 2. The rejection rate of PT, RGL and DS methods for the number of clusters in each sample is 10 ($N_1 = N_2$ =10).

ρ	d	PT	RGL	DS
0.3	0	0.048	0.051	0.060
	0.1	**0.078**	0.077	0.085
	0.2	**0.095**	0.087	0.098
	0.3	0.173	**0.177**	0.188
	0.4	0.279	**0.282**	0.304
	0.5	0.367	**0.368**	0.392
	0.6	**0.461**	0.460	0.486
0.4	0	0.044	0.044	0.059
	0.1	**0.057**	0.056	0.065
	0.2	**0.102**	0.095	0.110
	0.3	0.181	**0.188**	0.202
	0.4	0.230	**0.232**	0.259
	0.5	0.326	**0.336**	0.356
	0.6	0.424	**0.413**	0.446
0.5	0	0.050	0.052	0.059
	0.1	**0.070**	0.065	0.079
	0.2	**0.083**	0.082	0.091
	0.3	0.154	**0.163**	0.173
	0.4	0.242	**0.249**	0.262
	0.5	0.309	**0.323**	0.340
	0.6	0.394	**0.392**	0.414
0.6	0	0.049	0.051	*0.061*
	0.1	**0.071**	0.066	0.073
	0.2	**0.084**	0.083	0.088
	0.3	0.148	**0.150**	0.164
	0.4	0.229	**0.234**	0.245
	0.5	0.293	**0.304**	0.323
	0.6	0.361	**0.363**	0.384
0.7	0	0.052	0.046	0.059
	0.1	**0.068**	0.067	0.074
	0.2	**0.080**	0.077	0.083
	0.3	**0.139**	0.137	0.157
	0.4	0.216	**0.220**	0.234
	0.5	0.277	**0.280**	0.303
	0.6	0.335	**0.338**	0.359

Note: The boldface indicates that the higher power between PT and RGL tests**.** The italic indicates that the empirical size is out of Cochran's (1947) criterion.

Table 3 illustrates that for the number of clusters equal to 15, the empirical sizes $(d = 0)$ of all tests are close to the significant level 0.05.

The DS test performs well in all situations. In some situation, the power of the PT test is higher than that of RGL test.

The empirical powers of all tests increase as the effect sizes and the number of clusters in the sample increase. However, the empirical powers of three tests decrease as the coefficient of correlation increases.

Table 3. The rejection rate of PT, RGL and DS methods for the number of clusters in each sample is 15 ($N_1 = N_2 = 15$).

ρ	d	PT	RGL	DS
0.3	0	0.057	0.050	0.055
	0.1	0.075	**0.078**	0.084
	0.2	0.152	**0.151**	0.161
	0.3	0.239	**0.250**	0.261
	0.4	0.414	**0.417**	0.439
	0.5	0.533	**0.533**	0.550
	0.6	**0.696**	0.694	0.707
0.4	0	0.054	0.053	0.057
	0.1	0.067	**0.072**	0.078
	0.2	0.140	**0.144**	0.152
	0.3	0.214	**0.226**	0.243
	0.4	0.371	**0.381**	0.393
	0.5	**0.494**	0.488	0.502
	0.6	0.639	**0.651**	0.661
0.5	0	0.054	0.052	0.056
	0.1	0.066	**0.071**	0.076
	0.2	0.130	**0.131**	0.140
	0.3	0.199	**0.205**	0.221
	0.4	0.345	**0.347**	0.369
	0.5	**0.458**	0.455	0.473
	0.6	0.595	**0.599**	0.612
0.6	0	0.052	0.054	0.058
	0.1	0.070	0.072	0.077
	0.2	**0.127**	**0.127**	0.135
	0.3	0.186	**0.192**	0.203
	0.4	**0.321**	0.318	0.334
	0.5	**0.427**	0.423	0.437
	0.6	**0.564**	0.556	0.572
0.7	0	0.055	0.056	0.057
	0.1	0.068	**0.070**	0.075
	0.2	**0.121**	0.118	0.125
	0.3	0.174	**0.176**	0.185
	0.4	0.300	**0.303**	0.315
	0.5	**0.388**	0.385	0.402
	0.6	**0.525**	0.524	0.539

Note : The boldface indicates that the higher power between PT and RGL tests**.**

5 DISCUSSIONS

Clustered data are usually occurred in scientific research. In this study, we proposed the Wilcoxon rank sum test for clustered data using within-cluster resampling method. We compare the efficiency of the proposed test to the RGL and DS tests by using simulation study. When the number of clusters in two samples is small, the PT test can maintain the size of the test because the PT test uses two observations per cluster. In contrast, when the numbers of clusters in two samples is small, the empirical size of the DS test is far from the significant level because the DS test uses only one observation per cluster. The estimated variances of the PT and the DS test statistics may be negative. Therefore, from this study, the proposed test may be the better alternative test for testing that there is no difference between the two location parameters for clustered data.

ACKNOWLEDGEMENTS

This work was supported by Silpakorn University Research and Development Institute, Thailand. The authors greatly appreciate**.** In addition, we would like to thank the reviewers for their comments.

REFERENCES

Cochran, W.G. (1947). Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics*, , *3*(1), 22-38.

Datta, S., & Satten, G.A. (2005). Rank-sum tests for clustered data. *Journal of the American Statistical Association*, *100*(471), 908-915.

Hoffman, E.B., Sen, P.K., & Weinberg, C.R. (2001). Within- cluster resampling. *Biometrika*, *88*(4), 1121-1134.

Lahiri, P., & Li, Y. (2009). A new alternative to the standard F test for clustered data. *Journal of Statistical Planning and Inference*, *139*(10), 3430-3441.

Rao, J.N.K., Sutradhar, B.C., & Yue, K. (1993). Generalized least squares F-test in regression analysis with two-stage cluster samples. *Journal of the American Statistical Association*, *88*(424), 1388-1391.

Rosner, B., & Grove, D. (1999). Use of the Mann–Whitney U- test for clustered data. *Statistics in medicine*, *18*(11), 1387-1400.

Rosner, B., Glynn, R. J., & Ting Lee, M. L. (2003). Incorporation of clustering effects for the Wilcoxon rank sum test: a large-sample approach. *Biometrics*, *59*(4), 1089-1098.

Rosner, B., Glynn, R.J., & Ting Lee, M.L. (2006). The Wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics*, *62*(1), 185-192.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bullitin*, , *1*(6), 80-83.

Wu, C.F.J., Holt, D., & Holmes, D.J. (1988). The effect of two-stage sampling on the F statistic. *Journal of the American Statistical Association*, *83*(401), 150-159

Bayesian Estimation for Masking Exponential Data under Noise Multiplication

Phuwanat Boonmee* and Wuttichai Srisodaphol

Department of Statistics, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand
*Corresponding Email: tpootth@hotmail.com
Email: wuttsr@kku.ac.th

ABSTRACT

This article proposes a method for statistical disclosure control in data analysis. The proposed method to control statistical disclosure involves masking exponential data with noise multiplication. In order to keep the data protected and ensure that the results of the data analysis are still useful for research, we propose such a method for the estimator under noise multiplication masking exponential data by using the Bayesian method with conjugate prior distribution and show that the noise multiplication method can yield accurate interference in our simulation.

Keywords: exponential data, noise multiplication, statistical disclosure control

1 INTRODUCTION

Potentially sensitive data ranging from tax and health records, survey and census data to educational information often serve as a basis for different areas of research. Although the use of such private information in research must be dealt with great care in order to protect the identity of the volunteers participating in the research, sometimes certain confidential information belonging to individual people or organizations is inadvertently disclosed leading to possible abuse of the information. To address this problem, masking the data before subjecting them for analysis or publication is necessary. In this regard, Statistical Disclosure Control (SDC) is considered. SDC ensures that the confidential data cannot be identified from the published data, no matter how detailed the data are. There are two principles, (1) to protect confidentiality and (2) to ensure that the results of the data analysis are still useful for the research. Some popular SDC methods that have been used for masking data include top/bottom coding, multiple imputation (MI), noise addition and noise multiplication.

Many researchers have carried work concerning the SDC methods. Little (1993) presented a model-based likelihood theory that is applicable to the masked data and the analysis of microdata files. In general, likelihood theories require information about the masking procedure, which can be viewed as a process for selecting the values that are to be masked and a mechanism for masking the selected values. This study concerned the masking methods (randomized response, subsampling of cases or variables, deletion/addition, imputation, aggregation, noise injection and simulation of artificial records) which applied when the data were corrected and supplied to the user for analysis. Kim and Winkler (2003) developed the SDC method under noise multiplication (the original data are multiplied by the noise) and logarithmic transformation. They used data from the Internal Revenue Service in 1991. In a simulation study, they compared the mean and standard deviation of the transformed data for noise multiplication and logarithmic transformation to those of the original data. This study showed that noise multiplication exhibited better data masking than logarithmic transformation. An and Little (2007) considered two alternative methods to top coding for the SDC, that is, non-parametric and parametric MI methods. From the original data, the values of the original data that are greater than a particular value (fixed) are deleted, and they imputed their values by many methods. They considered these methods for inferences about the mean of original data subject to top coding. The following eight methods were considered (before deletion (BD), top coding (TC), log-normal maximum likelihood (LNML), hot deck MI (HDMI), log-normal MIC (LNMIC), log-normal MID (LNMID), power normal MIC (PNMIC) and power normal MID (PNMID)). By considering the root mean square error (RMSE), they found that the HDMI method provided similar results to the original data and were therefore superior to other methods. Among the parametric methods, they found that the LNMID method performed the best and it performed almost as well as HDMI. Nayak et al. (2011) considered the SDC method for noise multiplication of tabular magnitude data. They derived some properties of a balanced noise method from the masked data. They transformed the original data by noise random variable, called the noise multiplied random variable. They estimated parameters based on the transformed data by sample random sampling with replacement (SRSWR). Analyzing their data, they found that the sample moments and correlations based on the original data can be recovered unbiasedly from the masked data. Klein et al. (2014) developed methodology for the SDC under a noise multiplication method for three parametric distributions: exponential, normal and lognormal distribution. They transformed the original data for these three distributions by noise random variable, called the noise multiplied random variable. The likelihood-based data analysis methods (Maximum likelihood estimation and EM algorithm) are used to estimate the unknown parameters of these distributions. They derived the E and M-step for computing the Maximum likelihood estimators of the parameters. The simulation results were only shown for the lognormal distribution. They compared the efficiency of estimators for noise multiplication method, multiple imputation method and original data by root mean square error, bias, standard deviation, mean of estimated standard deviation, coverage probability of the nominal 0.95 level of confidence interval and expected length of the confidence interval relative to the expected length of the confidence interval computed on the original data. The noise multiplication method for lognormal data provided similar result to the original data. The bias was close to zero. The coverage probability was close to the nominal coverage probability of 0.95. The estimated standard deviation was small. The noise multiplication method showed better inferences than the multiple imputation method.

In terms of SDC, noise multiplication is one of popular methods for masking data in the case of parametric distribution. In this study, we consider the data under exponential distribution, since this distribution is used for analyzing life time data. We propose a noise multiplication method to mask such data. We use the Bayesian method with conjugate prior distribution for estimating the parameter in the exponential data and show that noise multiplication method can yield accurate inference under biased and mean squared error of the estimators.

2 DATA ANALYSIS

We address the methodology for estimating the parameter under masking data. We use Bayesian method with conjugate prior distribution to estimate the parameter. The simulation study is used to compare the performance of the estimators between Maximum likelihood estimator and Bayesian estimators. The methodology is explained as follows.

2.1 Data Description

Let Y be exponential distribution random variable with parameter θ and the probability density function is given by $f(y|\theta)=\frac{1}{\theta}e^{\frac{-y}{\theta}}$ for $y>0$ and $\theta>0$, such that $y_1, y_2, ..., y_n$ denote independent and identically distributed from exponential distribution. Let R be random noise with known parameter and probability density

function is $h(r)$, such that $r_1, r_2, ..., r_n$ denote independent and identically distributed from random noise distribution. We assume that R is customized noise distribution (Klein et al., 2014) with probability density function

$$h(r) = \frac{a^{a+1}}{\Gamma(a+1)} r^{-a-2} e^{-\frac{a}{r}}$$

for $r > 0$ and $a > 1$. We write the original data (Y) by multiplying it with the customized noise random variable (R), then the noise multiplied random variable is $Z = Y \times R$. We use the theorem of distribution of product of two random variables to find the probability density function of Z that is

$$g(z) = \int f(\frac{z}{r}) h(r) r^{-1} dr .$$

Next, we will estimate the parameter θ of the distribution of random variable Y under the distribution of random variable Z by using Bayesian estimation with conjugate prior distribution.

2.2 Parameter Estimation

In the part of estimation, we use Bayesian estimation to estimate the parameter θ under the distribution of random variable Z. By the theorem of distribution of product of two random variables, the probability density function of Z is obtained by

$$\begin{aligned} g_\theta(z) &= \int_0^\infty \frac{1}{\theta} e^{-\frac{z}{r\theta}} \frac{a^{a+1}}{\Gamma(a+1)} r^{-a-2} e^{-\frac{a}{r}} r^{-1} dr \\ &= \frac{1}{\theta} \frac{a^{a+1}}{\Gamma(a+1)} \int_0^\infty e^{-\frac{z}{r\theta} - \frac{a}{r}} r^{-a-3} dr \\ &= \frac{1}{\theta} \frac{a^{a+1}}{\Gamma(a+1)} \frac{\Gamma(a+2)}{(\frac{z}{\theta}+a)^{a+2}} \int_0^\infty \frac{(\frac{z}{\theta}+a)^{a+2}}{\Gamma(a+2)} e^{-(\frac{z}{\theta}+a)\frac{1}{r}} r^{-a-3} dr \\ &= \frac{1}{\theta} \frac{a^{a+1}}{\Gamma(a+1)} \frac{\Gamma(a+2)}{(\frac{z}{\theta}+a)^{a+2}} \\ &= \frac{(a+1)a^{a+1}}{\theta(\frac{z}{\theta}+a)^{a+2}} \end{aligned}$$

for $z > 0$. The likelihood function of θ given $z_1, z_2, ..., z_n$ is given by

$$L(\theta \mid z_1, z_2, ..., z_n) = \theta^{-n} (a+1)^n a^{n(a+1)} \prod_{i=1}^n (\frac{z_i}{\theta} + a)^{-a-2} .$$

The conjugate prior distribution of θ is inverse gamma distribution, $\theta \sim IG(\alpha, \beta)$,

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{-(\alpha+1)} e^{-\frac{\beta}{\theta}}$$

where $\theta > 0$ and $\alpha, \beta > 0$. The marginal function of Z, $m(z_1, z_2, ..., z_n)$ is obtained by

$$\begin{aligned} m(z_1, z_2, ..., z_n) &= \int L(\theta \mid z_1, z_2, ..., z_n) p(\theta) d\theta \\ &= \int_0^\infty \left(\theta^{-n} (a+1)^n a^{n(a+1)} \prod_{i=1}^n (\frac{z_i}{\theta} + a)^{-a-2} \right) \left(\frac{\beta^a}{\Gamma(\alpha)} \theta^{-\alpha-1} e^{-\frac{\beta}{\theta}} \right) d\theta \\ &= \int_0^\infty \frac{\beta^a}{\Gamma(\alpha)} e^{-\frac{\beta}{\theta}} \theta^{-\alpha-n-1} (a+1)^n a^{n(a+1)} \prod_{i=1}^n (\frac{z_i}{\theta} + a)^{-a-2} d\theta \\ &= \frac{\beta^a}{\Gamma(\alpha)} (a+1)^n a^{n(a+1)} \int_0^\infty \theta^{-\alpha-n-1} e^{-\frac{\beta}{\theta}} \prod_{i=1}^n (\frac{z_i}{\theta} + a)^{-a-2} d\theta . \end{aligned}$$

Then, the posterior distribution of θ given $z_1, z_2, ..., z_n$ is obtained by

$$p(\theta \mid z_1, z_2, ..., z_n) = \frac{L(\theta \mid z_1, z_2, ..., z_n) p(\theta)}{m(z_1, z_2, ..., z_n)}$$

$$= \frac{\left[\theta^{-n} (a+1)^n a^{n(a+1)} \prod_{i=1}^n (\frac{z_i}{\theta} + a)^{-a-2} \right] \left[\frac{\beta^a}{\Gamma(\alpha)} \theta^{-\alpha-1} e^{-\frac{\beta}{\theta}} \right]}{\frac{\beta^a}{\Gamma(\alpha)} (a+1)^n a^{n(a+1)} \int_0^\infty \theta^{-\alpha-n-1} e^{-\frac{\beta}{\theta}} \prod_{i=1}^n (\frac{z_i}{\theta} + a)^{-a-2} d\theta}$$

$$= \frac{\theta^{-n-\alpha-1} e^{-\frac{\beta}{\theta}} \prod_{i=1}^n (\frac{z_i}{\theta} + a)^{-a-2}}{\int_0^\infty \theta^{-\alpha-n-1} e^{-\frac{\beta}{\theta}} \prod_{i=1}^n (\frac{z_i}{\theta} + a)^{-a-2} d\theta} .$$

The Bayes estimator of θ is the mean of posterior distribution of θ that is

$$\begin{aligned} \hat{\theta}_{Bayes} &= E(\theta \mid z_1, z_2, ..., z_n) \\ &= \int_0^\infty \theta \, p(\theta \mid z_1, z_2, ..., z_n) d\theta \\ &= \int_0^\infty \theta \frac{\theta^{-n-\alpha-1} e^{-\frac{\beta}{\theta}} \prod_{i=1}^n (\frac{z_i}{\theta} + a)^{-a-2}}{\int_0^\infty \theta^{-\alpha-n-1} e^{-\frac{\beta}{\theta}} \prod_{i=1}^n (\frac{z_i}{\theta} + a)^{-a-2} d\theta} d\theta . \end{aligned}$$

3 SIMULATION STUDY AND RESULTS

We perform a Monte Carlo simulation using R program to compute the mean square error and bias. We compare the mean square error of the Bayes estimator using conjugate prior distribution with the mean square error of the Maximum likelihood estimator (Klein et al., 2014). We use R program in a simulation study and use 5000 iterations of simulation. Y follows the exponential distribution with parameter θ, R follows the customized noise distribution with parameter a, prior distribution of θ is inverse gamma with parameter α, β. We denote parameters in 8 cases as follows,

Case 1: θ = 1, a = 1.1, α = 3, β = 4,
Case 2: θ = 1, a = 1.3, α = 3, β = 4,
Case 3: θ = 10, a = 1.1, α = 4, β = 30,
Case 4: θ = 10, a = 1.3, α = 4, β = 30,
Case 5: θ = 50, a = 1.1, α = 2, β = 50,
Case 6: θ = 50, a = 1.3, α = 2, β = 50,
Case 7: θ = 100, a = 1.1, α = 1.5, β = 50,
Case 8: θ = 100, a = 1.3, α = 1.5, β = 50.

The mean square error of Bayes estimator ($\hat{\theta}_{Bayes}$) and Maximum likelihood estimator ($\hat{\theta}_{MLE}$) can be computed respectively as

$$MSE(\hat{\theta}_{Bayes}) = \frac{1}{5000} \sum_{i=1}^{5000} (\hat{\theta}_{Bayes_i} - \theta)^2$$

and

$$MSE(\hat{\theta}_{MLE}) = \frac{1}{5000} \sum_{i=1}^{5000} (\hat{\theta}_{MLE_i} - \theta)^2 .$$

Also, we compute the bias of the Bayes estimator and the Maximum likelihood estimator by

$$bias(\hat{\theta}_{Bayes}) = \frac{1}{5000} \sum_{i=1}^{5000} (\hat{\theta}_{Bayes_i} - \theta)$$

and

$$bias(\hat{\theta}_{MLE}) = \frac{1}{5000} \sum_{i=1}^{5000} (\hat{\theta}_{MLE_i} - \theta) .$$

For estimation of parameter θ in masking exponential data, Bayes estimator ($\hat{\theta}_{Bayes}$) has closed result in terms of mean square error more than MLE estimator ($\hat{\theta}_{MLE}$) in all case. The bias is close to zero in both methods. While Bayes estimator ($\hat{\theta}_{Bayes}$) has under estimation in some cases, MLE estimator ($\hat{\theta}_{MLE}$) has over estimation in all of the

cases. Bayes estimator ($\hat{\theta}_{Bayes}$) has point estimator close to θ and has lower standard deviations than MLE estimator ($\hat{\theta}_{MLE}$) in all cases. In the appendix part, the results of our simulation study are shown in Table 1 Table 2 and Table 3 with sample size n = 30, 50, 100, respectively.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to Prof. Bimal K. Sinha from Department of Mathematics and Statistics, University of Maryland Baltimore County, MD, USA for his instructive guidance, constructive comments and suggestions. The authors gratefully acknowledge financial support from the Department of Statistics, Faculty of Science, Khon Kaen University.

REFERENCES

An, D., & Little, R.J. (2007). Multiple imputation: an alternative to top coding for statistical disclosure control. *Journal of Royal Statistical Society: Series A (Statistics in Society), 170(4)*, 923-40.

Glen, A.G., Leemis, L.M., & Drew, J.H. (2004). Computing the distribution of the product of two continuous random variable. *Computational Statistics & Data Analysis*, *44(3)*, 451-464.

Kim, J.J., & Winkler, W.E. (2003). Multiplicative noise for masking continuous data. (Research Report Series, 2003-01). Washington D.C.: Statistical Research Division, U.S. Census Bureau of the Census.

Klein, M., Mathew, T., & Sinha, B.K. (2014). Likelihood based inference under noise multiplication. *Thailand Statistician*, *12(1)*, 1-23.

Little, R.J. (1993). Statistical analysis of masked data, *Journal of Official Statistics Stockholm*, *9*, 407-426.

Nayak, T.K., Sinha, B.K., & Zayatz, L. (2011). Statistical properties of multiplicative noise masking for confidentiality protection, *Journal of Official Statistics*, *27(3)*, 527-544.

APPENDIX

Table 1: The simulation result of Bayes estimator and Maximum Likelihood estimator for the exponential data with parameter θ based on sample size n = 30

θ	a,α,β	$MSE(\hat{\theta}_{Bayes})$	$MSE(\hat{\theta}_{MLE})$	$bias(\hat{\theta}_{Bayes})$	$bias(\hat{\theta}_{MLE})$	$\hat{\theta}_{Bayes}$	$\hat{\theta}_{MLE}$	$std(\hat{\theta}_{Bayes})$	$std(\hat{\theta}_{MLE})$
1	(1.1,3,4)	0.06164	0.07441	0.00373	0.02843	1.00373	1.02843	0.24828	0.27133
	(1.3,3,4)	0.05666	0.06769	0.00288	0.02558	1.00288	1.02558	0.23804	0.25893
10	(1.1,4,30)	4.70626	7.14340	0.05609	0.29020	10.05609	10.29020	2.16888	2.65718
	(1.3,4,30)	4.48466	6.66772	0.01636	0.22621	10.01636	10.22621	2.11785	2.57252
50	(1.1,2,50)	148.68370	179.58170	0.18824	1.42236	50.18824	51.42236	12.19336	13.32644
	(1.3,2,50)	141.80400	169.23450	0.10762	1.23935	50.10762	51.23935	11.90885	12.95114
100	(1.1,1.5,50)	640.97900	727.09050	0.20340	2.71660	100.20340	102.71660	25.31928	26.83011
	(1.3,1.5,50)	591.64390	664.61000	-0.12453	2.17561	99.87547	102.17560	24.32585	25.69064

Remark: $std(\hat{\theta}_{Bayes})$ and $std(\hat{\theta}_{MLE})$ are standard deviations of Bayes estimator and Maximum Likelihood estimator.

Table 2: The simulation result of Bayes estimator and Maximum Likelihood estimator for the exponential data with parameter θ based on sample size n = 50

θ	a,α,β	$MSE(\hat{\theta}_{Bayes})$	$MSE(\hat{\theta}_{MLE})$	$bias(\hat{\theta}_{Bayes})$	$bias(\hat{\theta}_{MLE})$	$\hat{\theta}_{Bayes}$	$\hat{\theta}_{MLE}$	$std(\hat{\theta}_{Bayes})$	$std(\hat{\theta}_{MLE})$
1	(1.1,3,4)	0.03825	0.04277	-0.00129	0.01349	0.99871	1.01349	0.19559	0.20638
	(1.3,3,4)	0.03472	0.03859	-0.00165	0.01197	0.99835	1.01197	0.18634	0.19609
10	(1.1,4,30)	3.13555	4.05944	0.01686	0.15861	10.01686	10.15861	1.77085	2.00875
	(1.3,4,30)	3.06204	3.91416	0.00118	0.13045	10.00118	10.13045	1.75004	1.97432
50	(1.1,2,50)	92.59991	103.43440	-0.15006	0.58430	49.84994	50.58430	9.62268	10.15449
	(1.3,2,50)	90.00221	100.13780	0.02690	0.71341	50.02690	50.71341	9.48786	9.98242
100	(1.1,1.5,50)	404.21750	436.79650	0.43666	1.96347	100.43670	101.96350	20.10243	20.80932
	(1.3,1.5,50)	343.64530	370.01620	0.41980	1.82684	100.41980	101.82680	18.53477	19.15078

Table 3: The simulation result of Bayes estimator and Maximum Likelihood estimator for the exponential data with parameter θ based on sample size n = 100

θ	a,α,β	$MSE(\hat{\theta}_{Bayes})$	$MSE(\hat{\theta}_{MLE})$	$bias(\hat{\theta}_{Bayes})$	$bias(\hat{\theta}_{MLE})$	$\hat{\theta}_{Bayes}$	$\hat{\theta}_{MLE}$	$std(\hat{\theta}_{Bayes})$	$std(\hat{\theta}_{MLE})$
1	(1.1,3,4)	0.01885	0.01997	0.00106	0.00863	1.00106	1.00863	0.13730	0.14107
	(1.3,3,4)	0.01825	0.01926	0.00041	0.00736	1.00041	1.00737	0.13511	0.13861
10	(1.1,4,30)	1.81710	2.07237	-0.00264	0.06998	9.99736	10.06998	1.34813	1.43801
	(1.3,4,30)	1.69005	1.91940	0.02322	0.09176	10.02322	10.09176	1.29994	1.38252
50	(1.1,2,50)	47.28493	50.04824	-0.00133	0.37538	49.99867	50.37538	6.87709	7.06522
	(1.3,2,50)	45.51049	48.01519	0.01029	0.35739	50.01029	50.35739	6.74681	6.92077
100	(1.1,1.5,50)	199.40770	207.26580	0.14374	0.90731	100.14370	100.90730	14.12186	14.36955
	(1.3,1.5,50)	0.01885	0.01997	0.00106	0.00863	1.00106	1.00863	0.13730	0.14107

Adaptive Neuro Fuzzy Inference System (ANFIS) for Predicting the Hourly Temperature in Pattani, Thailand

Tri Wijayanti Septiarini[1*] and Salang Musikasuwan[2]

[1]Department of Mathematics and Computer Science, Faculty of Science and Technology,
Prince of Songkla University, Pattani, Thailand
*Corresponding Email: triwijayantiseptiarini@gmail.com
[2]Department of Mathematics and Computer Science, Faculty of Science and Technology,
Prince of Songkla University, Pattani, Thailand
Email: salang.m@psu.ac.th

ABSTRACT

The objectives of this study were (i) to construct adaptive neuro fuzzy inference system (ANFIS) model for predicting hourly temperature in Pattani, Thailand, and (ii) to compare the predicting performance with statistical methods by using root mean square error (RMSE) and mean square error (MSE) as forecasting evaluation tool. The observation data used in this study were from automatic weather station (AWS) in Pattani, Thailand, collected during January, 18th 2014 to January, 18th 2015 (in total 8,640 data series). In ANFIS method, the model combined the learning capabilities of a neural network and reasoning capabilities of fuzzy logic in order to increase predicting ability. The statistical methods used in this study were ARIMA (Autoregressive Moving Average) and Exponential Smoothing. The data partitions consisted of 80%-20%, 70%-30%, 75%-25%, 65%-35%, and 60%-40% of training and testing data, respectively. The differences partitions of data set were investigated. The results showed that ANFIS had the smallest evaluation value for all data partitions. However, the performance of forecasting method cannot be guaranteed from either classical or modern forecasting method.

Keywords: fuzzy; neural network; predicting; statistical methods

1 INTRODUCTION

Temperature is a part of the environmental system. Temperature can provide effects in any field such as chemical reaction, human health, human activity, etc. According to Svec and Stevenson (2007), the enhanced awareness of temperature risk has enhanced the need for effective weather hedging and risk management programs, driving the demand for weather derivatives. Schemes for hourly temperature forecasting have been mainly developed in the context of short- or long-term load forecasting and power utility management.

Nowadays, forecasting activities have an important part in our daily life. Every day the temperature forecast informs us that how the temperature will be for the next days. We can prevent deep damage by forecasting, for instance, the coming of storms or typhoons. Eynard et al. (2011) stated that the complex variations of temperature and the plenty of historical data suggested by the computational intelligence data-based techniques would be appropriate models to predict temperature. The forecasting with 100% accuracy may be impossible, but it can be deal by decreasing the forecasting errors or increasing the speed of the forecasting process.

Consider with solving temperature problems, many researchers have constructed various methods or models. Chen and Hwang (2000) have proposed fuzzy time series model for forecasting temperature in Taipei. Lee et al. (2007) have constructed fuzzy logical relationship and genetic algorithm for forecasting temperature in Taipei, Taiwan. In 2007, Svec and Stevenson applied weather derivatives for modeling and predicting the temperature in Australia. In the later year, Lee et al. (2008) have proposed high-order fuzzy logical relationship and genetic simulated annealing techniques for forecasting temperature in Taipei, Taiwan. Wang and Chen (2009) used automatic clustering techniques and two-factor high-order fuzzy time series to predict the temperature in Taiwan. Dupuis (2011) presented forecasting temperature in US cities in order to price temperature derivatives on CME. In the same year, Eynard et al. (2011) have succeeded to construct wavelet-based multi-resolution analysis and artificial neural network for predicting temperature.

However, there were competitions among researchers to predict temperature. But, it is quite hard to find the study of predicting the temperature in Pattani, Thailand. This study presented adaptive neuro fuzzy inference system (ANFIS) to predict the hourly temperature in Pattani, Thailand. It performance was evaluated and compared with statistical methods by using root mean square error (RMSE) and mean square error (MSE). According to Pahlavani and Delavar (2014), the integrating neural network and fuzzy systems can be fused with the learning capability of neural network in expression function of fuzzy inference system. Thus, ANFIS can optimize the performance of fuzzy model by tuning the parameter in membership function.

2 ADAPTIVE NEURO FUZZY INFERENCE SYSTEM (ANFIS)

The pioneer of ANFIS is J.S. Roger Jang in 1992. ANFIS is a new kind of neural network which is a combination of fuzzy logic and neural network (Wang, 2015). ANFIS constructs a fuzzy inference system (FIS) which the membership function parameters are optimized by using a neural network. According to Tarno et al. (2013), ANFIS provides four types of membership function which were identified in fuzzy inference system, i.e. triangular membership function, trapezoidal membership function, Generalized Bell membership function, and Gaussian membership function. ANFIS provide a tool for the fuzzy model to learn the data set, in order to tune the membership function parameters that best allow the associated fuzzy inference system to track the given input/output. The learning methods which are used in ANFIS are either a backpropagation algorithm or the combination with least squares which is a hybrid algorithm.

2.1 ANFIS Architecture

Figure 1 shows the typical architecture of ANFIS. For instance, ANFIS has two inputs (x,y) and one output (z).

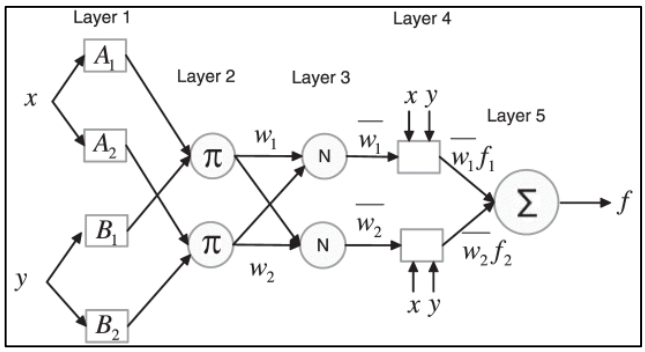


Figure 1: The architecture of Adaptive Neuro Fuzzy Inference System (Adopted from Wang and Ning, 2015).

Layer 1 (Fuzzy layer)
Each node in the fuzzy layer is the degree of membership function $\left(\mu_{A_i}(x)\right)$ from input.
Layer 2 (Product Layer)
Every node in this layer is a circle node which multiples from the incoming signal. Fuzzy operator AND is applied in this step in order to get the product out. For examples,

$$w_i = \mu_{A_i}(x) \times \mu_{B_i}(x), i = 1,2.$$

The firing strength of $i-th$ rule is represented by a node in the second layer.
Layer 3 (Normalized Layer)
Each node in this layer is circle node. In the third layer, it calculates normalized firing strengths which is computed by the ratio of the firing strength of $i-th$ rule to the sum of all firing strength rules as follow

$$\overline{w_i} = \frac{w_i}{w_1 + w_2}, i = 1,2.$$

The normalized firing strength is the output of this step.
Layer 4 (De-fuzzy Layer)
Each node in this layer is a square node or called adaptive node. The output of this step can be calculated as this function

$$\overline{w_i}\, f_i = \overline{w_i}(p_i x + q_i y + r_i),$$

where
$\overline{w_i}$ is defined as a normalized firing strength of layer 3,
$\{p_i, q_i, r_i\}$ is defined as the parameters set (consequent parameters).
Layer 5 (Total Output Layer)
In the fifth layer, the single node calculates the overall output as the summation all of the incoming signals from the previous layer.

$$overall\ output = \sum_i \overline{w_i}\, f_i = \frac{\sum_i w_i f_i}{\sum_i w_i}.$$

2.2 Learning Algorithm of ANFIS

The learning process in ANFIS is the changing parameters set of membership function in order to get the optimal parameters and solution (Tarno *et al.*, 2013). There are two kinds of ANFIS learning algorithm which are backpropagation and hybrid algorithm. In this study, the hybrid algorithm was applied. A hybrid algorithm is a combination of backpropagation and least square method. According to Jang (1993), in the hybrid algorithm, premise and consequent parameters pass the network backward and forward, respectively. In a forward way, least square method will identifies the consequent parameters when the input passed into layer 4. Another way, backward step, gradient descent will identifies the premise parameters.

3 METHODS

According to Tarno et al. (2013), there are three main steps to construct ANFIS which are preprocessing data, establishing the rules, and evaluating the performance. The implementation ANFIS for forecasting temperature in Pattani, Thailand can be described as following:

1. Data collection
   The data were obtained from Automatic Weather Station (AWS) in Pattani, Thailand, collected during January, 18th 2014 to January, 18th 2015 (in total 8,640 data series).
2. Preprocessing data
   After fixed missing data, the data were separated into two groups which are training and testing. According to Dobbin and Simon (2011), the optimal proportion of training dataset is within range 40%-80%. And the best proportion for splitting datasets commonly used 1/2 training or 2/3 training. Thus, the proportions of this study were 60%:40%, 65%:35%, 70%:30%, 75%:25%, and 80%:20% of training and testing datasets, respectively. The various proportions were utilized to analyze the impact for the forecasting results. The fixed data is represented in Figure. 2.

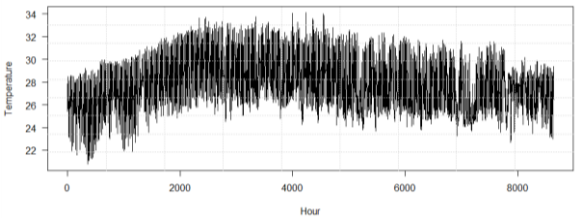


Figure 2: Time series plot of the hourly Pattani temperature.

3. Input selection
   In this study, two inputs were considered to get output by considering examining process.
4. Determining the membership function
   After examining the process, the ANFIS model was constructed which has seven of generalized bell functions (gbellmf) with linear outputs. The membership function for input 1 and 2 were shown in Figure 3 and 4, respectively.

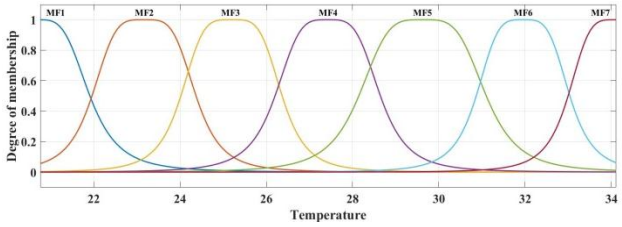


Figure 3: The membership function for input 1.

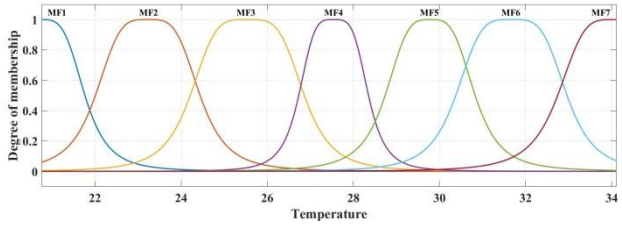


Figure 4: The membership function for input 2.

5. Generating fuzzy rules
   From the system, 49 of the fuzzy rules were generated. The number of fuzzy set output equals to the number of rules.
6. Determining the learning algorithm
   The hybrid method was selected to learning the model. The architecture of ANFIS model in this study can be seen in Figure 5. According to the architecture of ANFIS, the model is consist of 7 membership function in each input to get 1 output in the of process.

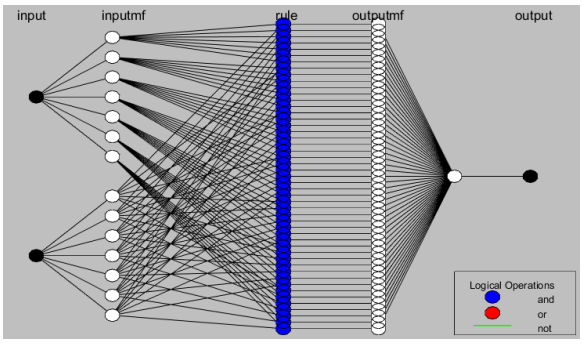


Figure 5: The ANFIS architecture.

7. Tuning the parameters of the fuzzy inference system
   The parameters of fuzzy inference system were tuned in this step. Fuzzy inference system was used for reasoning rule to get the fuzzy output. It has been training until 100 epochs in order to obtain the optimal solution. The training error plots for all data proportions are shown in Figure 6-10. The 60% training data, training error is going down from 0.5592 to 0.5576. For 65% training data, training error is decreasing from 0.572 to 0.565. The

training error also decreased from 0.575 to 0.5705 for 70% training data. The training error of 75% training data decreased from 0.575 to 0.5715. And the last proportion, 80% training data, the training error decreased from 0.5685 to 0.5646. For all proportion had been treated such as the training error decreasing along the epoch time step.

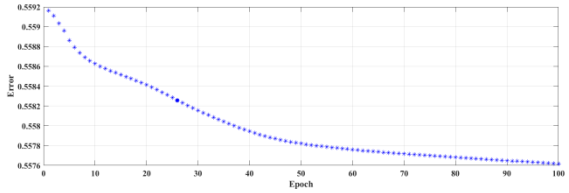


Figure 6: The training error of ANFIS model for 60% training data.

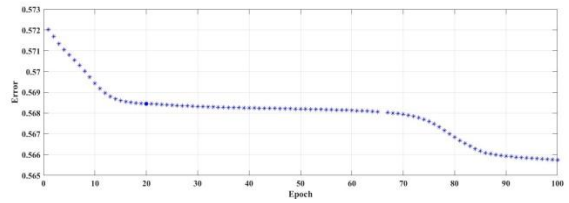


Figure 7: The training error of ANFIS model for 65% training data.

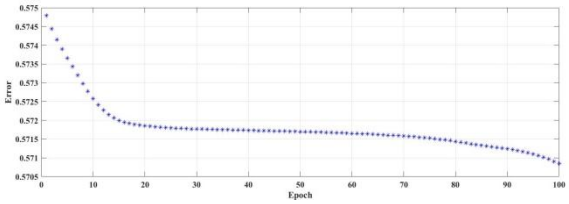


Figure 8: The training error of ANFIS model for 70% training data.

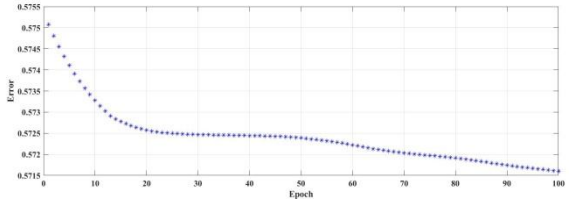


Figure 9: The training error of ANFIS model for 75% training data.

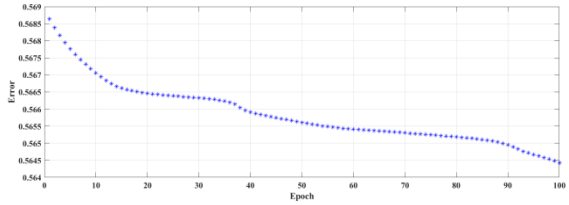


Figure 10: The training error of ANFIS model for 80% training data.

8. Forecasting and evaluating the performance

After achieving the significant model, the predicted value was determined from training and check data. Then the model was evaluated by using root mean square error (RMSE) and mean square error (MSE).

$$RMSE = \sqrt{\sum_{t=1}^{s} \frac{\left(actual(t) - predict(t)\right)^2}{s}},$$

$$MSE = \sum_{t=1}^{s} \frac{\left(actual(t) - predict(t)\right)^2}{s},$$

where

s is defined as the number of predicted data,

t is defined as the time step (hourly).

Once, the ANFIS model has been constructed, the autoregressive integrated moving average (ARIMA) and exponential smoothing methods were implemented to construct the models and compared the forecasting performance with ANFIS.

4 RESULTS AND DISCUSSIONS

In this study, the ANFIS model performance was compared with ARIMA and exponential smoothing. Table 1 shows the MSE values for all models with testing data. And Table 2 shows the RMSE values for all models with testing data. According the evaluation value, the ANFIS model had the smallest evaluation value for all proportion. And the smallest error was in 75%:25%, training and testing data. The evaluation value can give evidence that ANFIS is an adequate predictive model.

Table 1 The MSE values of the testing data.

Data Proportion	ANFIS	ES	ARIMA
60%:40%	1.024449	5.270776	0.511381
65%:35%	0.333375	5.054555	0.492895
70%:30%	0.322702	4.064305	0.465337
75%:25%	0.310621	3.538892	0.439578
80%:20%	0.332759	3.342112	0.434597

Table 2 The RMSE values of the testing data.

Data Proportion	ANFIS	ES	ARIMA
60%:40%	1.012151	2.295817	0.715109
65%:35%	0.577386	2.248234	0.702065
70%:30%	0.568068	2.016012	0.682156
75%:25%	0.557334	1.881194	0.663006
80%:20%	0.576853	1.828144	0.659240

Figures 11-15 show the time series plot of the forecasting results for ANFIS, ARIMA, and simple exponential smoothing forecasted for all proportion data. From the time series plot, it can be seen that the difference is quite small for ANFIS and ARIMA.

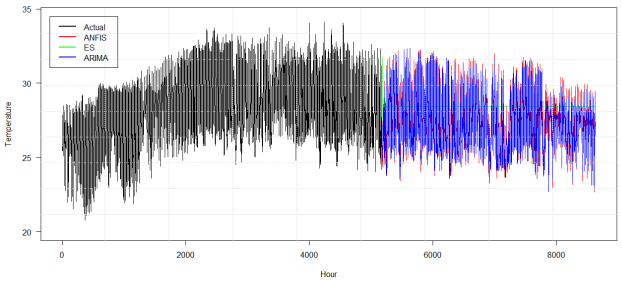


Figure 11: The forecasting result of all models for 60% training data.

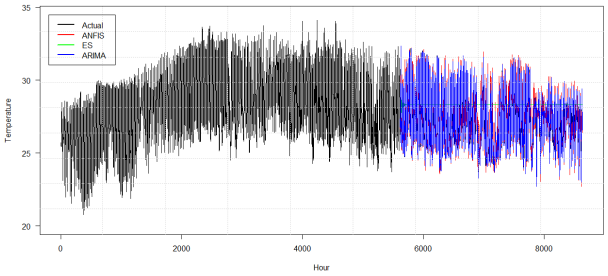


Figure 12: The forecasting result of all models for 65% training data.

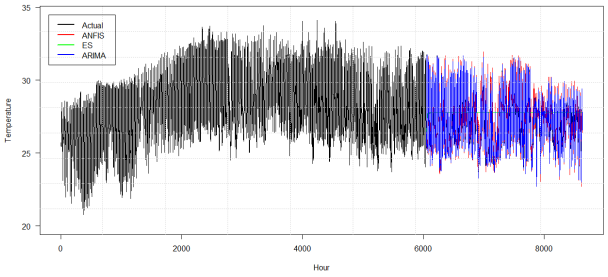


Figure 13: The forecasting result of all models for 70% training data.

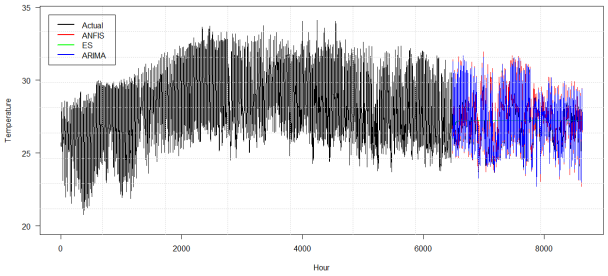


Figure 14: The forecasting result of all models for 75% training data.

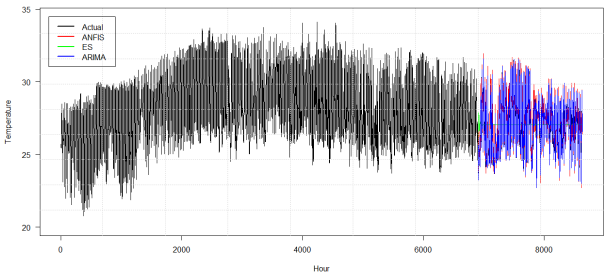


Figure 15: The forecasting result of all models for 80% training data.

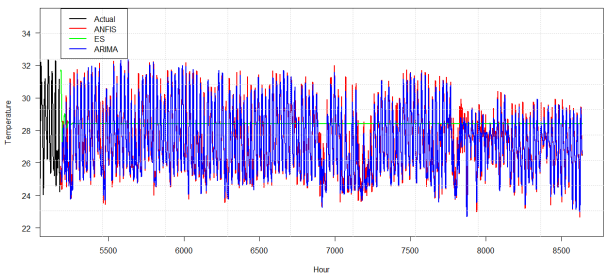


Figure 16: The plot of the checking data of all models for 60% training data.

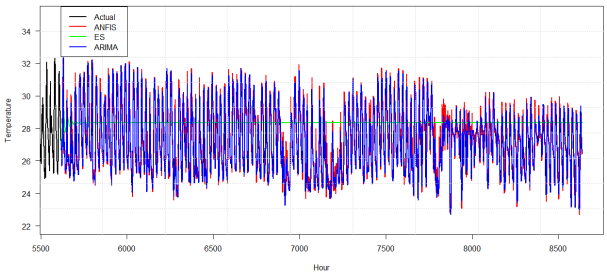


Figure 17: The plot of the checking data of all models for 65% training data.

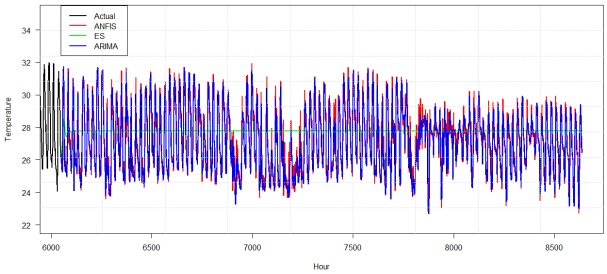


Figure 18: The plot of the checking data of all models for 70% training data.

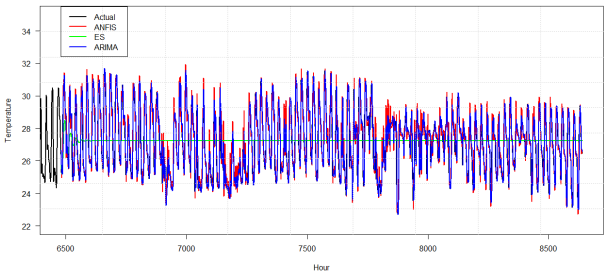


Figure 19: The plot of the checking data of all models for 75% training data.

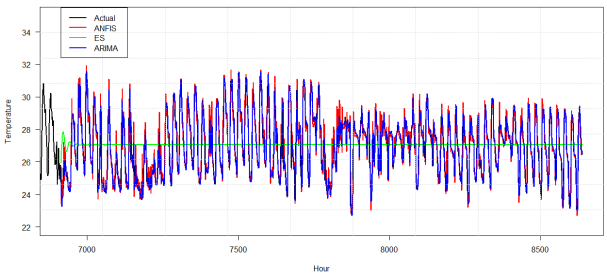


Figure 20: The plot of the checking data of all models for 80% training data.

In order to make clearly, the checking forecasting result has been plotted for all proportion as shown in Figure 16-20. Yet, the performance of simple exponential smoothing does not as good as other methods, since the forecasting result is almost similar in each time step. ARIMA is a classical forecasting method which can gain pretty good results. The performance of ANFIS, which is a computational artificial intelligence data-based technique, is performed the best in this study.

5 CONCLUSIONS

It can be concluded that the 75% training data is the best proportion for constructing ANFIS model for forecasting the hourly temperature in Pattani, Thailand. Yet, each time series data has unique characteristic which need to analyze. The accuracy of ANFIS model is

influenced by many factors such as the number of the membership function, the type of membership function, the selection of input, the number of iteration, and the type of output. According to the result of this study, the ANFIS model was constructed with seven of the generalized bell function (gbell) membership function. The optimal ANFIS model was compared with ARIMA and exponential smoothing. The result stated that ANFIS had the smaller error in 75% training data than statistical methods which were 0.557334 for RMSE and 0.310621 for MSE. It means that ANFIS is able to predict more accurately than another methods.

ACKNOWLEDGEMENTS

This work was supported by SAT-ASEAN Scholarship for International Student of Faculty of Science and Technology. This work is also (partially) supported by the Centre of Excellence in Mathematics, Commission on Higher Education, Thailand. Finally, the authors would like to thank the reviewers for their helpful suggestions.

REFERENCES

Chen, S. M., & Hwang, J.R. (2000). Temperature prediction using fuzzy time series. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *30*(2), 263-275.

Dobbin, K.K., & Simon, R.M. (2011). Optimally splitting cases for training and testing high dimensional classifiers. *BMC medical genomics*, *4*(1), 31.

Dupuis, D.J. (2011). Forecasting temperature to price CME temperature derivatives. *International Journal of Forecasting*, *27*(2), 602-618.

Eynard, J., Grieu, S., & Polit, M. (2011). Wavelet-based multi-resolution analysis and artificial neural networks for forecasting temperature and thermal power consumption. *Engineering Applications of Artificial Intelligence*, *24*(3), 501-516.

Jang, J.S. (1993). ANFIS: adaptive-network-based fuzzy inference system. *IEEE transactions on systems, man, and cybernetics*, *23*(3), 665-685.

Lee, L.W., Wang, L.H., & Chen, S.M. (2007). Temperature prediction and TAIFEX forecasting based on fuzzy logical relationships and genetic algorithms. *Expert Systems with Applications*, *33*(3), 539-550.

Lee, L.W., Wang, L.H., & Chen, S. M. (2008). Temperature prediction and TAIFEX forecasting based on high-order fuzzy logical relationships and genetic simulated annealing techniques. *Expert Systems with Applications*, *34*(1), 328-336.

Pahlavani, P., & Delavar, M.R. (2014). Multi-criteria route planning based on a driver's preferences in multi-criteria route selection. *Transportation research part C: emerging technologies*, *40*, 14-35.

Svec, J., & Stevenson, M. (2007). Modelling and forecasting temperature based weather derivatives. *Global Finance Journal*, *18*(2), 185-204.

Tarno, Subanar, Rosadi, D., & Suhartono. (2013). Analysis of financial time series data using adaptive neuro-fuzzy inference system (ANFIS). *IJCSI International Journal of Computer Science Issues*, (10), 491-496.

Wang, N.Y., & Chen, S.M. (2009). Temperature prediction and TAIFEX forecasting based on automatic clustering techniques and two-factors high-order fuzzy time series. *Expert Systems with Applications*, *36*(2), 2143-2154.

Wang, J.S., & Ning, C.X. (2015). ANFIS Based time series prediction method of bank cash flow optimized by adaptive population activity PSO algorithm. *Information*, *6*(3), 300-313.

A Comparison of Parameter Estimation with Penalized Regression Analysis on High-Dimensional Data

Benjamas Rungsaranon* and Autcha Araveeporn

King Mongkut's Institute of Technology Ladkrabang /Department of Statistics/Faculty of Science, Bangkok, Thailand
*Corresponding Email: benjamasrung99@gmail.com
Email: kaautcha@hotmail.com

ABSTRACT

The objective of this research is to compare the parameter estimation of multiple linear regression models which are consisted of dependent variable and independent variables. Normally the number of independent variables is less than the number of sample sizes, so the ordinary least squares give a unique solution. But the number of independent variables are larger than number of sample sizes, this data is called as high-dimensional data. The traditional regression analysis has a problem in case of high-dimensional data. To overcome this problem, penalized regression analysis concerns to solve high-dimensional data. In this case, we focus to estimate the parameter of the ridge regression, lasso and elastic net method which called penalized regression analysis. Ridge regression is to choose the unknown ridge parameter by cross-validation, so ridge estimator is evaluated by adding ridge parameter on the penalty term. Lasso (least absolute shrinkage and selection operator) is added the penalty term as the scaled sum of the absolute value of the coefficients. The elastic net is mixed between ridge regression and lasso on the penalty term. The criterion of comparison is the average mean square errors. This study examines the residual distribution from a normal distribution, contaminated normal distribution, and Weibull distribution. The number of independent variables is focused on 11, 12, 13, 14 and 15 when the sample sizes are specified by 10. Another case, the number of independent variable is focused on 16, 17, 18, 19 and 20 when the sample sizes are specified by 15. The data are obtained through simulation using a Monte Carlo technique with 1,000 replications for each case. The results are found that elastic net is satisfied when the residuals are simulated from normal and Weibull distributions in all cases. When the residuals are contaminated normal distribution, the elastic net lasso outperforms in most cases, except that lasso gives the best results with independent variable = 12, 15, sample sizes = 10 and independent variable = 17, 20, sample sizes = 15 on large contaminated data.

Keywords: elastic net; lasso; ridge regression

1. INTRODUCTION

Regression analysis is a set of statistical technique for estimating the relationship model among dependent variable and independent variable. Regression analysis consists of simple and multiple regression analysis. The simple regression is focused on the relationship between a dependent variable and one independent variable, but the multiple regression is also used a dependent variable and two or more independent variables. The performance of regression analysis depends on the assumption such as linear relationship, normality homoscedasticity, homoscedasticity, and no multicollinearity.

Regression analysis is widely used for estimating linear regression model and forecasting future values. Many techniques for carrying out linear regression model have been developed several methods. Familiar methods such as ordinary least squares (OLS) method, that shown a type of linear least squares method for estimating unknown parameters in linear regression model. The OLS is chosen to create the linear regression model when the independent variable is less than the sample sizes. When the independent variable is larger than the sample size, it is called high-dimensional data, thus the OLS could not estimate parameter in this case. However penalized regression analysis has been proposed for estimating parameter on linear regression model based on high-dimensional data. Using high-dimensional with penalized regression, a multiple regression model is fitted by using bone mineral density as dependent variable and genome-wide DNA-methylations as independent variables. (Lien et al., 2018)

Ridge regression, lasso, and elastic net are known in a class of penalized regression analysis. The ridge regression was proposed by Hoerl and Kennard (1970) that was a popular estimation of regression with multicollinearity among independent variables. When multicollinearity occurs, OLS is unbiased estimator, and variance is high value. By adding a constant value of bias to the regression estimates, ridge regression reduced the standard errors.

Lasso (least absolute shrinkage and selection operator) was introduced in order to improve the prediction accuracy and interpretability of regression models by altering the model fitting process to select only an independent variable for using in the final model rather than using all of them. It was developed by using the penalty for both fitting and the penalization of the coefficients that studied by the statistician, Tibshirani (1996). However, when the independent variable is highly correlated, the lasso estimator performs unsatisfactorily.

Zou and Hastie (2005) proposed the elastic net to improve performance of lasso in multicollinearity. The elastic net method can also be viewed as a penalized least squares method where the penalty term is a convex combination of the ridge penalty and lasso penalty. It is particularly useful when there are much more independent variable than the sample sizes.

In this research, we focus on the multiple regression analysis in terms of high-dimensional data based on penalized regression analysis. This concept is to estimate parameter of regression coefficient that let the objective function to be minimized under conditions of penalty term. The penalty term has shown in different forms that made the different coefficient estimators. Finally we interest three methods of penalized regression analysis such as ridge regression, lasso, and elastic net methods. The criterion to select the best method is minimum average mean square errors (MSE).

2. PENALIZED REGRESSION ANALYSIS

Penalized regression analysis is considered in form of multiple linear regression model depended on dependent variable (y) and independent variable (X). The multiple linear regression models can be written in matrix form as

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon},$$

where

$\underset{\sim}{y}$ is a vector of dependent variables $n \times 1$,

X is a matrix of independent variables $n \times (p+1)$,

$\underset{\sim}{\beta}$ is a vector of multiple regression coefficient $(p+1) \times 1$,

$\underset{\sim}{\varepsilon}$ is a vector of error $n \times 1$ and $\underset{\sim}{\varepsilon} \sim N(0, \sigma^2 I)$,

n is sample sizes,

p is number of independent variables.

Normally the coefficient of multiple linear regression models can be estimated by the ordinal least squares (OLS) method from

$$\hat{\underset{\sim}{\beta}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j X_{ij})^2 = (X^T X)^{-1} X^T \underset{\sim}{y}.$$

For high-dimensional data, when the number of independent variables is larger than sample sizes $(p > n)$, the matrix $X^T X$ is a singular matrix, thus it can't be calculated inverse matrix. It follows that we can't estimate the estimator from OLS method.

All of the above mentioned estimation procedures are from the OLS of view. It is also possible to consider the penalized regression analysis. This method can be viewed as a number of independent variable more sample size depend on the penalty function.

2.1 Ridge Regression Method

Hoerl and Kennard (1970) proposed that potential instability of the OLS estimator could be improved by adding a small constant value (λ) to the diagonal matrix $X^T X$ before taking its inverse. The result is the ridge regression estimator as

$$\hat{\underset{\sim}{\beta}}_R = (X^T X + \lambda I_{p+1})^{-1} X^T \underset{\sim}{y}.$$

Ridge regression places a particular of the constraint on the parameter $(\hat{\underset{\sim}{\beta}}_R)$ is chosen to minimize the penalized sum of squares :

$$\hat{\underset{\sim}{\beta}}_R = \underset{\beta}{\operatorname{argmin}} \left[\sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j X_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right], \lambda > 0.$$

Where the penalty term is $\lambda \sum_{j=1}^{p} \beta_j^2$, the tuning parameter serves λ to control the amount of shrinkage. It can be used to obtain an estimated parameter with smaller mean square error. The cross-validation method is chosen to find the smallest tuning parameter (λ) that yields the generalized cross-validation prediction error (Boonstra et al., 2015).

2.2 Lasso Method

Lasso method is the most widely used method for choosing which independent variables is to select as the stepwise selection, which only improves prediction accuracy in certain cases, such as when only a few independent variables have a strong relationship with the dependent variable. Also, at the time, ridge regression was the most popular technique for improving prediction accuracy. Ridge regression improves prediction error by shrinking large regression coefficients in order to reduce overfitting, but it does not perform independent selection.

Lasso is able to achieve both of these goals by forcing the sum of the absolute value of the regression coefficients to be less than a fixed value, which forces certain coefficients to be set to zero, effectively choosing a simpler model that does not include those coefficients (Tibshirani, 2011). This idea is similar to ridge regression, in which the sum of the squares of the coefficients is forced to be less than a fixed value. The Lasso estimator ($\hat{\underset{\sim}{\beta}}_L$) proposed by Tibshirani (1996) following:

$$\hat{\underset{\sim}{\beta}}_L = \underset{\beta}{\operatorname{argmin}} \left[\sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j X_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right], \lambda > 0 .$$

Where the penalty function is $\lambda \sum_{j=1}^{p} |\beta_j|$, the tuning parameter λ determines the $\hat{\underset{\sim}{\beta}}_L$ that shrunk towards 0. The cross-validation method is used to try out different values of λ, while the other data are used to assess the predictive performance of models of the different complexities. Lasso estimators for all values of λ can be computed through a modification of the LARS algorithm (Efron et al., 2004).

2.3 Elastic Net Method

Zou and Hastie (2005) proposed a new regularization technique that called the elastic net. Elastic net method is to combine between ridge regression and Lasso method that it does automatic variable selection and continuous shrinkage, and it can select groups of correlated variables. Elastic net estimator ($\hat{\underset{\sim}{\beta}}_E$) is estimated as the following:

$$\hat{\underset{\sim}{\beta}}_E = \underset{\beta}{\operatorname{argmin}} \left[\sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j X_{ij})^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right]$$

$, 0 < \lambda_1 + \lambda_2 < 1.$

Elastic net penalty is a convex combination of the ridge and lasso penalty. When $\lambda_1 = 0$, the elastic net becomes simple ridge regression. The tuning parameter λ_1 and λ_2 are selected by cross-validation (Hastie et al., 2009).

3. SCOPE OF RESEARCH

The scope of this research is considered as follows:

3.1 Generate residual (ε) from multiple linear regression models from 3 distributions.

3.1.1 The normal distribution is generated from mean (μ) and standard deviation (σ) in 2 sets such as $N(\mu, \sigma^2) = N(0,1)$, and $N(0,9)$. The normal distribution function is

$$f(\varepsilon) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\varepsilon-\mu}{\sigma}\right)^2}, -\infty < \varepsilon < \infty.$$

3.1.2 The contaminated normal distribution is generated from the normal distribution $N(\mu, \sigma_1^2) = N(0,1)$, and percent of contaminated data as 10% $(p' = 0.1)$ on normal distribution $N(\mu, \sigma_2^2) = N(0,25)$ and $N(0,100)$. The contaminated normal distribution function is

$$f(\varepsilon) = (1-p')N(\mu, \sigma_1^2) + p' N(\mu, \sigma_2^2).$$

3.1.3 The Weibull distribution is generated from scale factor (α) 1 shape parameter (β) 3 and 5 or called $W(\alpha, \beta) = W(1,3)$ and $W(1,5)$. The Weibull distribution is given by

$$f(\varepsilon) = \frac{\beta}{\alpha}\left(\frac{\varepsilon}{\alpha}\right)^{\beta-1} e^{-\left(\frac{\varepsilon}{\alpha}\right)^{\beta}}, -\infty < \varepsilon < \infty.$$

3.2 Generate independent variables (X) from normal distribution with $N(\mu, \sigma^2) = N(0,2)$. The number of independent variables (p) are focused on 11, 12, 13, 14 and 15 when sample sizes (n) is specified by 10. Another group of sample sizes (n) is defined by 15, and the number of independent variables (p) is defined by 16, 17, 18, 19 and 20.

3.3 Define regression coefficient (β) as one for all situations.

3.4 The dependent variable $(\underset{\sim}{y})$ is obtained from $\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$.

3.5 The R program is used to generate data at 1,000 replications for each case.

3.6 The criterion is considered the mean square errors between simulate data $(\underset{\sim}{y})$ and estimated data $(\hat{\underset{\sim}{y}})$ that can be computed by

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}.$$

4. RESULTS

The results of this research are presented the average mean square errors based on ridge regression, lasso, and elastic net. Table 1-3 are shown the various residuals such as normal, contaminated normal, Weibull distribution.

TABLE 1 : The average mean square error on normal distribution $(N(\mu,\sigma^2))$, sample sizes (n), the number of independent variables (p) of ridge regression, lasso and elastic net methods.

Residuals	n	p	Ridge	Lasso	Elastic Net
$N(0,1)$	10	11	27.6467	6.2450	**4.8227**
		12	30.0424	8.5207	**6.4655**
		13	31.8397	8.9704	**7.2823**
		14	33.7792	11.7642	**9.6232**
		15	36.3571	14.5881	**11.3298**
	15	16	39.8654	7.1738	**4.8367**
		17	40.8375	9.8699	**6.6379**
		18	43.4204	10.7431	**8.5590**
		19	45.0963	13.6072	**10.5857**
		20	47.8418	15.4750	**11.4757**
$N(0,9)$	10	11	33.2892	11.4343	**9.8586**
		12	36.2911	13.0946	**11.2592**
		13	38.1075	14.4731	**12.6249**
		14	39.8210	17.9240	**14.3015**
		15	41.8102	18.8864	**16.3756**
	15	16	45.7509	13.0236	**10.1875**
		17	46.8766	15.7449	**12.5993**
		18	49.3107	16.5633	**14.1860**
		19	51.1913	18.9570	**15.8870**
		20	53.7133	21.0638	**16.5765**

From Table 1, the elastic is presented the minimum average mean square errors for all sample size when the residuals are simulated from normal distribution. When the independent variable is large, the average mean square error is large too.

TABLE 2 : The average mean square error on contaminated normal distribution with percent and variance of contaminated data , sample sizes (n), the number of independent variables (p) of ridge regression, lasso and elastic net methods.

Residuals	n	p	Ridge	Lasso	Elastic net
10%, $\sigma_2^2 = 25$	10	11	77.6205	49.4937	**49.4778**
		12	81.5141	54.3089	**52.3294**
		13	76.5887	48.6484	**47.1712**
		14	83.6092	56.9226	**54.3684**
		15	80.8305	54.3198	**51.7821**
	15	16	87.6211	56.1982	**51.4032**
		17	89.2121	56.0319	**52.2743**
		18	93.6419	62.4168	**57.9068**
		19	94.8645	58.6434	**55.9436**
		20	97.6708	61.0101	**59.8755**
10% , $\sigma_2^2 = 100$	10	11	842.1847	739.7596	**739.7215**
		12	898.0900	**800.8394**	807.3138
		13	778.0641	699.6539	**693.9046**
		14	852.4576	790.1886	**779.9636**
		15	794.4773	**711.0069**	711.9731
	15	16	852.2075	792.8478	**779.8263**
		17	855.6464	**791.9745**	801.3105
		18	872.5684	811.1464	**808.3667**
		19	895.6993	819.2871	**796.7144**
		20	905.4148	**859.8209**	863.7218

From Table 2, the elastic is presented the minimum average mean square errors for all sample size when the variance of contaminated data is 25. When the variance of contaminated data is large, some case of lasso show the minimum average mean square errors.

TABLE 3: The average mean square error on Weibull distribution $(W(\alpha,\beta))$, sample sizes (n) ,the number of independent variables (p) of ridge regression, lasso and elastic net methods.

Residuals	n	p	Ridge	Lasso	Elastic net
$W(1,3)$	10	11	34.1418	11.8084	**10.0682**
		12	35.9092	13.5087	**11.6946**
		13	37.4478	15.2590	**13.0802**
		14	40.1458	17.5545	**15.5832**
		15	42.6393	19.4498	**17.0847**
	15	16	45.8025	11.9634	**9.9336**
		17	47.9186	16.9710	**12.6808**
		18	49.3024	16.9777	**13.2299**
		19	50.7277	18.9437	**15.7200**
		20	53.2725	22.1116	**17.3166**
$W(1,5)$	10	11	46.5962	21.1480	**19.7157**
		12	48.1250	24.3577	**21.3607**
		13	49.6316	24.5045	**22.0646**
		14	52.5075	27.8925	**25.8245**
		15	55.2659	29.3421	**27.1035**
	15	16	58.1672	23.4177	**21.5652**
		17	60.4853	28.7203	**24.3836**
		18	61.8641	30.2465	**25.5338**
		19	62.6749	30.6176	**26.8236**
		20	65.7522	33.3725	**29.4614**

From Table 3, the elastic is presented the minimum average mean square errors for all sample size when the residuals are simulated from Weibull distribution. When the independent variable is large, the average mean square errors are large as normal distribution.

5. CONCLUSIONS

In this research, we focused on the multiple regression model when residuals are generate on normal, contaminated normal, and Weibull distribution. The ridge regression, lasso, and elastic net are used to estimate parameter on multiple regression models via high-

dimension data. The elastic net method is a good performance when the residuals are generated on normal and Weibull distribution in all cases. Therefore, the lasso method is a good fit when the residuals are played on contaminated normal distribution in some case especially large outlier data. The results are shown that the lasso and elastic net are superior over ridge regression methods. It is expected because the lasso and elastic net consisted of tuning parameter on the penalty function which controlled the interpolating function. On future work, we may focus on Bayesian lasso and Bayesian elastic net to use an expanded hierarchy with conjugate prior for the multiple regression models.

REFERENCES

Boonstra, P.S., Mukherjee, B., & Taylar, J.M. (2015). A small-sample choice of the tuning parameter in ridge regression. *Statistics Sinica*, 25(3), 1185.

Efron, B., Hastie, T., Johnson, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of statistics*, 32(2), 407-499.

Hoerl, A.E., & Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55- 67.

Lien, T.G., Borgan, Ø., Reppe, S., Gautvik, K., & Glad, I.K. (2018). Integrated analysis of DNA-methylation and gene expression using high-dimensional penalized regression: a cohort study on bone mineral density in postmenopausal women. *BMC medical genomics*, 11(1), 24.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1). 267-288.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,*73*(3), 273-282.

Hastie, T., Tibshirani, R., & Fried, J.H. (2009). The elements of statistical learning: data mining inference and prediction. New York: Springer.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.

Payment Data: Stylized Facts and Private Consumption Indicators[a]

Godchagon Panyamanotham

Statistics and Data Management Department, Bangkok, Thailand
GodchagP@bot.or.th

ABSTRACT

Payment data are able to reflect consumers' spending and go hand-in-hand with economic activities. In the paper, stylized facts are analyzed and presented, also the relationship between payment data and private consumption over time is assessed. Evidently, it is found that card usage has shown a decline in trend, for both value and number of transactions while electronic banking transactions are on the rise. On payment methods, cash transactions are still in preference, followed by cards and electronic banking payments respectively. Regarding the construction of private consumption indicator by using payment data, indicators derived from payment variables are coincident with the private consumption expenditure, with some degree of robustness (correlation coefficient greater than 0.5). Such findings suggest that payment data can be used as an alternative indicator for not only tracking private consumption, but for monitoring changes in payment behavior of spenders in general.

Keywords: payment data; private consumption; spending

1 INTRODUCTION

Payment systems facilitate economic activities, by linking trading of goods and services with settlement process. Recently, a number of countries have begun utilizing payment data for contemplating private consumption and predicting economic growth since the data can reflect consumers' spending as well as being timely available. Galbraith and Tkacz (2015) studied the use of payment data to forecast the Gross Domestic Product (GDP) of Canadian economy by exploiting debit & credit card transactions and cheque clearing data where such data mirroring private consumption representing the main composition of the GDP. Following a similar study by Verbaan et al. (2017), debit card payment data was adopted to forecast and nowcast consumption growth in the Netherlands. Findings from two studies are relatively similar in the sense that by encapsulating debit card data in the model, it helps improving predictive power by reducing root mean squared error (RMSE). Apart from scholastic researches, the Bank Indonesia (BI) has constructed indicators for tracking private consumption through payment data, it's also being used for assessing developments of non-cash transactions in different parts of the country. The study revealed the majority of Indonesian still prefer cash transactions, despite supports from the government to promote non-cash payments since 2004 (Peranginangin, 2017).

In Thailand, payment data prepared by the Bank of Thailand, has been primarily used for monitoring payment systems stability. Rungsun et al. (2007) examined the relationship between payment data and economic and financial transactions. What can be concluded from the study is interbank cheque and credit card data can be used to track economic activities. Contradictorily, there was no relationship between electronic funds transfer data and state of the economy. With the benefit of hindsight, such findings were likely to be resulted by effects of data constraint during the observation period. Unlike these days, consumers have increasingly accessed financial services through electronic devices or online channels. Besides, those electronic payment data has been compiled regularly (monthly basis) since 2005 and being handled for economic analysis purposes.

The paper is divided into 3 parts. The first section reveals stylized facts of card usage and electronic banking data. The second part covers correlation testing, explanatory variables – payment data and obtained results – the correlation coefficient with private consumption. The final portion discusses conclusion and recommendation for further study.

2 PAYMENT DATA AND STYLIZED FACTS

The Bank of Thailand (BOT) has collected payment data since 2005. By law and regulations, financial institutions and electronic payment service providers are obligated to report transactional data in aggregate format to the BOT, the reported payment data are basically compiled as parts of payment systems statistics and published on the BOT website. In the study, 3 payment datasets comprising as follows are analyzed:

(1) Credit Card Summary (CCS)
(2) Card Usage Summary (CUS)
(3) Electronic Banking Summary (EBS)

Regarding the payment datasets, details of payment dimensions such as transaction type, card type and payment instrument or channel are available. To apply payment data for private consumption tracking, understanding definition and context of payments are prerequisites. The stylized facts derived from each dataset will be individually presented in subsequent sections.

2.1 Credit Card Summary (CCS)

Credit Card Summary (CCS), the dataset captures credit card transactions composed of domestic and oversea card spending including cash advances in terms of value and volume (number of transactions). Over the time span, it portraits a rising trend but at diminishing rate due to the BOT had enforced tightening policy on credit card approval (Figure 1) for the sake of curbing soaring household debts. Later, during 2016-17, value of total spending per transaction had declined for the reason of changing consumers' behavior. Shoppers have rapidly shifted their purchasing pattern from onsite shopping to online transactions. It is obvious that more and more transactions had taken place online yet in less value per purchase. This explains the falling off in value per transaction (Figure 2). When mining credit card statistics in depth, the data unearths that majority of domestic spending are from high-income cardholder (Figure 3), nevertheless, those income range information was submitted for card application purpose and it has not been updated as time went by.

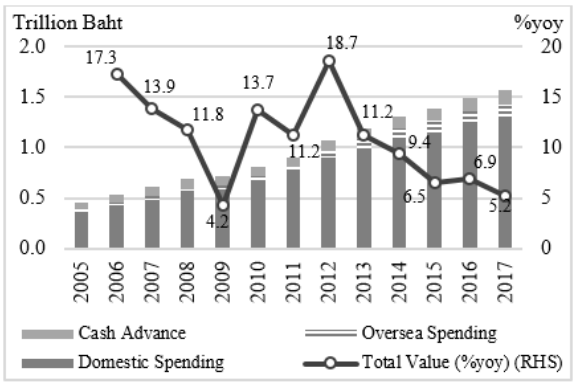


Figure 1: Value of credit card spending and growth (%yoy) classified by transaction type

[a] Some contents of this paper were published in Panyamanotham (2018).

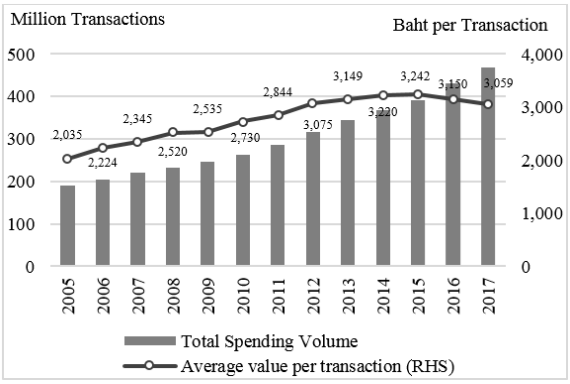


Figure 2: Volume of credit card spending and value per transaction

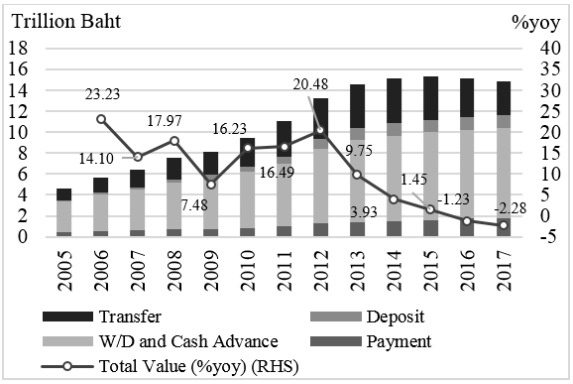


Figure 4: Value of card usage classified by transaction type

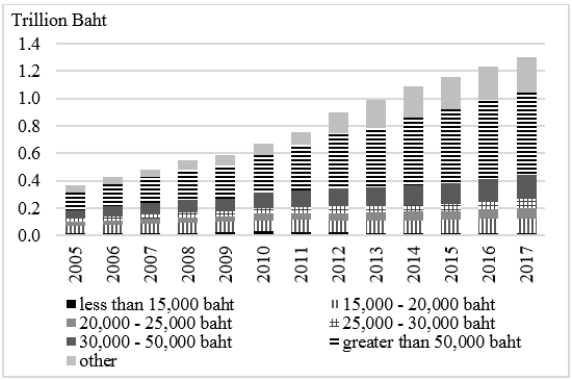


Figure 3: Domestic spending by credit card classified by cardholder income range

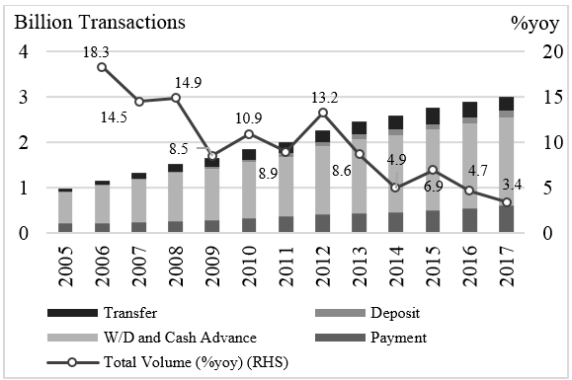


Figure 5: Volume of card usage classified by transaction type

2.2 Card Usage Summary (CUS)

Apart from credit card data, the CUS has other cards information contained. Basically, the CUS embraces debit, credit, ATM and other type of card usage, i.e., personal loan cards. From the perspective of data report, unlike the CCS where a bulk of data providers are banks and card services companies, CUS has additional data contributors – the Specialized Financial Institutions (SFIs). The CUS, thus, is a comprehensive dataset for card usage transactions.

By comparing cards classified by card type, in terms of usage, the stylized fact reveals that in 2017 the combination of ATM and debit cards usage in value account for 88% of total usage, followed by credit and other cards with shares of 11% and 1% respectively. Regarding the channel, card usage is made through ATM/ADM/CDM, especially for cash withdrawals and advances transactions. Owing to this, it is fair to conclude that people are still in favor of cash transactions.

Year after year, card usage has shown a declining trend in value and number of transactions, remarkably transactions related to "funds transfer" (Figure 4, 5), caused by consumers' adoption of online funds transfer. It is observed that such technology has lent some supports to enhance growth in both value and volume of transactions via "mobile phone" in particular. For greater details, further discussion is in the next part.

2.3 Electronic Banking Service Summary (EBS)

Electronic Banking Summary (EBS) is a dataset which gathering all transactions done through electronic or online channel (e-Banking), however, it does not include card usage transactions. Factually, major transaction purposes cover funds transfer (46%), payment transaction (35%) and payroll (13%). Empirically, most transactions were conducted via Internet Banking[1], followed by Direct Credit[2] and Mobile Banking[3]. (Figure 6)

Regarding the prevailing mobile banking technology, e-Banking transactions done through this channel has increased substantially in the past three years from 2015 to 2017 (Figure 7) and become the most popular channel in 2017, with a share of 59.7%. Factors supporting the success are 1) an introduction of PromptPay[4] Scheme in mid of 2016 by the BOT that offering quick, convenient and secure payments with minimal or no fees, 2) an increase in rate of people's smart phone possession where mobile phones have become parts of their lives and 3) consumers' confidence in online security related to financial transactions.

[1] Internet Banking is an electronic banking transaction by customers through computer network, excluding all card usage transactions through online channel.

[2] Direct Credit is an electronic transfer system which enables a payer transfers funds directly into bank account of a payee. It is commonly used by employer to make a transfer of periodic compensation to employees' account.

[3] Mobile banking is a service provided by Financial Institutions which allow customers to make financial transactions on a mobile device (cell phone, tablet, etc.).

[4] PromptPay, initiated by the Bank of Thailand in collaboration with the Government, is a service that enables one to receive and transfer funds, using Citizen ID or mobile phone number instead of a bank account number, via electronic channels – namely internet banking, mobile banking and ATMs.

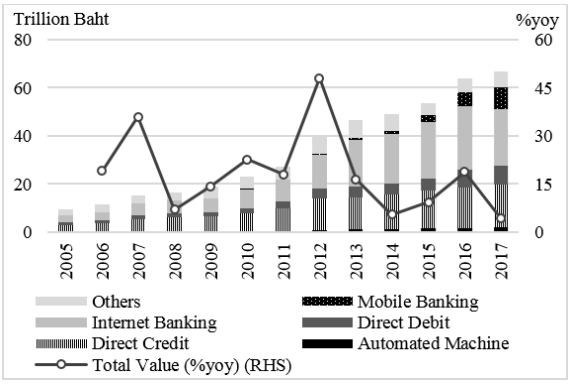


Figure 6: Value of e-Banking transaction classified by service type[5]

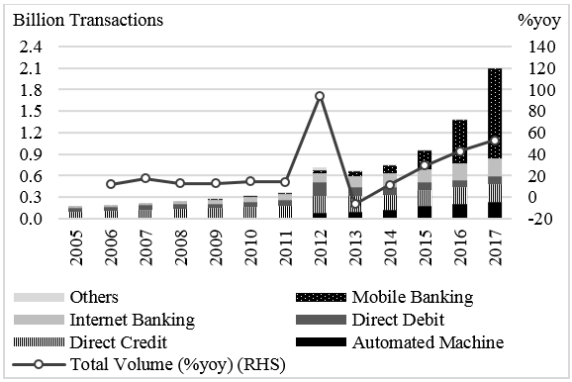


Figure 7: Volume of e-Banking transaction classified by service type[5]

Stylized facts derived from three payment datasets can be concluded as 1) card usage has declined in volume and value, while electronic banking transactions are boosted by changing consumers' behavior since 2015, shifting from card-based funds transfer to e-Banking based transactions and 2) Evidently, cash transactions are preferred mean of payments for daily spending relative to credit card and electronic banking (Figure 8).

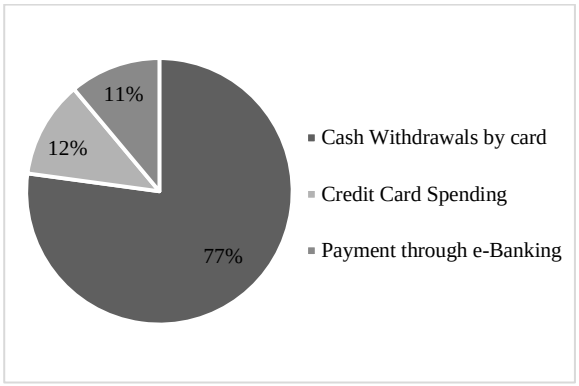


Figure 8: The share of cash withdrawals by card, credit card spending and e-Banking payment transaction in 2017

3 PAYMENT DATA AND PRIVATE CONSUMPTION

Private Consumption Index (PCI)[6], a monthly tracking indicator, is constructed by the Bank of Thailand (BOT) for monitoring quarterly Private Consumption Expenditure (PCE). In this section, the focus is on employing payment data as supplementary indicators for tracking the PCE by justifying certain payment data that can be applied to ensue state of the economy. It follows that those selective data are then paired with private consumption variables, in form of, the rate of change (%yoy). By measuring value of Cross-Correlation Coefficient of the pairs, it demonstrates explanatory capability or reflectivity of selected variables proceeding hand-in-hand with private consumption. In the paper, if values of Cross-Correlation Coefficient is greater than 0.5, it implicates those payment data variables can serve as good indicators.

3.1 Methodology

Cross-Correlation Coefficient (r) is a measure of similarity of two series as a function of the displacement of one relative to the other. The value lies between -1 and +1 indicating the degree of linear dependence between two series. If the value approaches -1, implies an inverse linear relationship. Conversely, if the value inclines to +1, indicates a positive linear relationship. The Cross-Correlation Coefficient formula is:

$$r_{X_i Y_j} = \frac{N \sum X_i Y_j - (\sum X_i)(\sum Y_j)}{\sqrt{[N \sum X_i^2 - (\sum X_i)^2][N \sum Y_j^2 - (\sum Y_j)^2]}}$$

where $r_{X_i Y_j}$ is covariance of X_i and Y_j divided by product of standard deviations

X_i is payment data variables (i = 1, 2, 3, 4)

Y_j is private consumption expenditure (PCE) (j = 1, 2, 3, 4)

N is number of observations

Cross-Correlation Coefficient can also indicate characteristic of payment data variables that could be leading, lagging or coincident indicators in relation to the PCE. There are 3 possibilities as follows:

Case 1: If Cross-Correlation Coefficient is highest at period t + i (i > 0), regarding this, payment data variables are leading indicators for the PCE – described as payment transactions taken place before private consumption (occurred ahead of an i period).

Case 2: If Cross-Correlation Coefficient is highest at period t, regarding this, payment data variables are coincident indicators for the PCE – described as payment transactions taken place at the same time as private consumption occurred.

Case 3: If Cross-Correlation Coefficient is highest at period t - i (i > 0), regarding this, payment data variables are lagging indicators for the PCE – described as payment transactions taken place after private consumption (occurred after an i period).

3.2 Variables

In this paper, payment data variables (X_i) are explanatory variables. Data available quarterly from 2005 to 2017 in aggregate value. To eliminate inflationary effects over time, those variable are technically deflated by the Consumer Price Index (CPI). Descriptions of variables are shown below.

(1) Credit Card Spending (X_1) is drawn from the CCS. The data covers all domestic spending made by credit cards. The variable is anticipated to well mirror the consumption since the purpose of credit card purchase is mainly for goods and services consumption.

(2) Credit and Debit Card Usage at the POS- point of sales (X_2) is drawn from the CUS. The data covers domestic purchase of goods and services by credit and debit card through the EFTPOS[7]. The variable is anticipated to reflect consumption via card usage where the "Credits" accounts for 80% of aggregate spending, nonetheless, it shows a diminishing trend, whereas the "Debits" have continually gained more share in payments.

(3) Cash Withdrawals and Advances by card (X_3) is drawn from the CUS. The data covers transactions including cash withdrawals and advances for all card types through ATM/ADM/CDM machines. One obstacle to be noted, it is not feasible to identify transaction purpose accurately, for instance, if one has withdrawn cash from the ATM, it is not possible to justify whether such transaction is done for liquidity purpose (cash holding), spending or cash transfer to third party / repayments.

(4) Payments through e-Banking (X_4) (excluding card usage) is drawn from the EBS. The data covers goods and services payments

[5] By e-Banking service type, others are referred to 3 service types such as Telephone Banking, Office/PC Banking and Other.

[6] Private Consumption Index (PCI) is a monthly indicator constructed from various economic variables, for instance, Sales of goods and services, Sales of car and motorcycle, Import of textiles, Household electricity consumption, Sales of benzene, gasohol and diesel and Tourists' expenditure.

[7] EFTPOS = Electronic Funds Transfer at Point of Sales

made by residents via electronic banking services. Channels are confined as follows: (1) Internet Banking, (2) Mobile Banking, (3) Telephone Banking, (4) Office & PC Banking and (5) ATM. The variable is anticipated to exhibit growing momentum, and expected to have increasing role in comparison to cash and card payments in near future.

For "Explained Variable", the quarterly Private Consumption Expenditure (PCE) (Y_j), compiled by the National Economic and Social Development Board (NESDB), is adopted. Since the PCE is compiled in real term and the variables are also deflated, so the measurement of coefficient will not be distorted by price effects. The model logically includes total PCE and its components (value in real term) such as the PCE (Y_1), Non-Durable PCE (Y_2), Semi-Durable PCE (Y_3) and Service PCE (Y_4) but excludes the Durable PCE since it accounts for 55% of vehicle purchases – where majority of the purchases financed by loans and paid by cheque or cash.

3.3 Result

From empirical findings, a terse conclusion can be explained as the defined payment variables are coincident with the PCE and all of its components – in growth term, as evidenced that the Cross-Correlation Coefficient has reached its maximum at period t, connoting payment transactions correspondingly echo present consumption. Results from the Table 1 can be succinctly described as:

1) Credit Card Spending (X_1) tends to correlate with the PCE (Y_1) and its components Y_2, Y_3 and Y_4, with correlation greater than 0.5 ;
2) Credit Card and Debit Card Spending at POS (X_2) exclusively gravitates towards the Semi-Durable PCE (Y_3) ;
3) Cash Withdrawals and Advances by card (X_3) and Payment through e-Banking (X_4) move well with the PCE (Y_1) but not its components.

Table 1: Correlation between payment data variables and PCE

DATASET	CCS	CUS		EBS
Variables	Credit card spending (X_1)	Credit card and debit card spending at the point of sales (X_2)	Cash withdrawals and advances by card (X_3)	Payment through e-Banking (X_4)
PCE (Y_1)	0.59	0.45	0.62	0.71
PCE (Non-Durable) (Y_2)	0.55	0.31	0.33	0.45
PCE (Semi-Durable) (Y_3)	0.56	0.52	0.37	0.36
PCE (Service) (Y_4)	0.57	0.46	0.48	0.43

Note: Values in Table 1 referred to correlation coefficient between X_i and Y_j.

4 CONCLUSION

Stylized facts derived from 3 payment datasets (CCS, CUS, and EBS) reveal that people still prefer cash transactions, confirmed by the records of card usage for cash withdrawals and advances through automated machines - ATM/ADM/CDM. On one hand, card transactions have clearly presented a downward trend, on the other hand, electronic banking services, to be precise "funds transfer", are prospering. Regarding the adoption of payment data variables as supplementary source to follow-up private consumption, it can be concluded that (1) Domestic Credit Card Spending, (2) Cash Withdrawals and Advances by cards via machines and (3) Payments made by residents through e-Banking services: some of payment data variables exhibit robust relationship (highly correlated) with private consumption that is pro-cyclical.

One worth point to be mentioned is under the current payment data structure, transaction purpose, in other words, spending purpose cannot be identified, i.e., purpose of cash withdrawals or funds transfer transactions. As payment technology has advanced unstoppably, and in order to adopt new payment data to monitor economic conditions, changing consumers' behavior in payment patterns should be taken into account. With regards to the first Payment Systems Act launched in 2017, yet effective from April 2018 onwards, a new set of payment data report is designed to reduce previous data constraints as well as being able to cope with fast changing payments ecosystem, for instance, the reports cover modern payment methods, business type of merchant and location of services. Thus, the newly acquired data will enable the BOT to monitor, oversee and regulate payment systems stability more effectively. In addition, these data will be useful for research works and economic analytics in a more profound manner.

On top of things, central banks around the globe have expressed their interests to exploit new data combined with existing database to track economic activities. This will generate unusual perceptions for data users, for example, Spain, Italy and the U.S. have embraced searching words from Google Trends application to catch economic pulses. Presumably, those words could reflect consumers' interest of something at each point of time. As for Thailand, the BOT has made experiments by using micro-data of funds transfer – classified by transaction purpose "Payroll" to follow footprints of private consumption based on the ground of disposable income. Furthermore, the BOT has explored the possibility to work on large volume data, with high granularity (Big Data), which is timely available, for policy researches. This kind of data would have brought advantages for policymakers to be able to monitor and project the economy with better foresight. Unlike in the old days, decisions were solely based on macro-economic data obtained from central data administration unit with 1-3 month lag time. Due to such constraints, it unintentionally limited the efficacy of policies. From data usage aspect, to attain the best of both worlds outcome in time to come, an integrated use of existing and new formatted data would yield a synergistic result for policymakers by enhancing analytical capability based on new data dimension which helps deepening the understanding of state of economy in a well-rounded manner. By conceiving different features of data that were once imperceptible will exalt the likelihood of policy decisions to achieve favorable consequences in proactive and reactive ways.

ACKNOWLEDGEMENTS

The author is grateful to Napat Phongluangtham and Jaruphan Wanitthanankun for valuable suggestions, and to Teerapap Pangsapa, Paphatsorn Sawaengsuksant and Kiattikhun Samritpiam for their assistance during the study. In addition, the author also wishes to thank Krit Chalermdumrichai for comments and editing completely.

REFERENCES

Allan, C. (2013, July). Examining the Usefulness of the Electronic Card Transactions Data as an Indicator for the New Zealand Economy: Some Preliminary Evidence, Paper Presented at the 54th Annual New Zealand Association of Economists Conference of Moto Economic and Public Policy Research, New Zealand.

Ardizzi, G., Emiliozzi, S., Marcucci, J. and Monteforte, L. (2018, March). News and Consumer Card Payment, Paper Presented at Banca d'Italia Workshop Harnessing Big Data & Machine Learning for Central Banks, Italy.

Galbraith, J. W., & Tkacz, G. (2015). Nowcasting GDP with electronic payments data (No. 10). European Central Bank.

Gil, M., Pérez, J. J., Urtasun, A., & Sánchez, A. J. (2017) Nowcasting private consumption: traditional indicators, uncertainty measures, and the role of internet search query data.

Hataiseree, R., Nakornthab, D., & Boonsiri, J. (2007). Payment System Data and Economic and Financial Activities: Some Empirical Evidences from Thailand. BOT Working Papers, (Also available via the Internet: https://www.bot.or.th/english/paymentsystems/publication/psresearchpaper/workingpaper/workingpaper_2007_01.pdf).

Kholodilin, K. A., Podstawski, M., & Siliverstovs, B. (2010). Do Google searches help in nowcasting private consumption? A real-time evidence for the US.

Nakamura, K., Kawata, H., Tanaka, M., & Uemae, L. (2016). The Consumption Activity Index. BOJ Reports & Research Papers.

Panyamanotham, G. (2018). Payment Data: Stylized Facts and Private Consumption Indicators. (Also available via the Internet: https://www.bot.or.th/Thai/Statistics/Articles/Doc_Lib_statistics Horizon/Stylized%20facts.pdf).

Peranginangin, F. (2017). Payment System Statistics to Support Policy Formulation in Indonesia.

Verbaan, R., Bolt, W., & van der Cruijsen, C. (2017). Using debit card payments data for nowcasting Dutch household consumption. *De Nederlandsche Bank Working Paper*. *571*

The Use of Information Criteria for Selecting Number of Knots in Natural Cubic Spline Volatility Estimation

Jetsada Laipaporn* and Phattrawan Tongkumchum

Prince of Songkla University, Pattani, Thailand
*Corresponding Email: laipaporn.j@gmail.com
Email: phattrawan@gmail.com

ABSTRACT

The low-frequency volatility has been used as the indicator of the change in the financial market stability due to the macroeconomic situations**.** Previous studies estimated this volatility by applying spline function to the series of financial assets**'** returns and using an information criterion to specify the optimum number of knots**.** However, some studies especially in case applying natural cubic spline function to estimate the low-frequency volatility mostly selected the number of knots subjectively**.** Therefore, this study tried to compare the performance of two widely used information criteria, Akaike**'**s Information Criterion **(**AIC**)** and Bayesian Information Criterion **(**BIC**)**, for selecting the number of knots of natural cubic spline volatility model**.** The results of the Monte Carlo simulation found that BIC selected the less number of knot and under-parameterized the model while the empirical results shown that AIC was likely to use too many number of knots and over-parameterized the model**.**

Keywords: Volatility; Natural Cubic Spline Function; Number of knots; Information criteria

1 INTRODUCTION

Financial volatility is one of key indicators of market stability. The higher volatility indicates that the index of the stock market has a wider range of changing and potential losses are higher. Therefore, some investors may not be able to withstand that risk and decide to delay their investment. During low volatility, the stock market index also changes but slightly. The market is more stable and the value of possible losses is lower. So investment banks can reduce their reserve requirement due to reduced risk. According to the influences of volatility in the stock market on investment, most investors consider volatility to be an important information for their decision-making.

Financial volatility cannot be measured directly like weight or height. Thus, there are so many approaches to estimate volatility through it proxy, return series. Those approaches were different to each other due to the objective of each study (Poon, 2005). Among those studies, they found that spline function is a suitable function for modeling the low-frequency volatility (Farida *et al.*, 2018; Laipaporn & Tongkumchum, 2017; Engle & Rangel, 2008). Those study assumed that the financial volatility was hypothetically divided into two parts. The low-frequency volatility was defined as the slow-moving part of financial volatility, indicating the long-run change of market stability (Engle & Rangel, 2008). Additionally, it governed the cyclical moving of the financial volatility (Awalludin & Saelim, 2016) and related to the change of the macroeconomic factors (Engle & Rangel, 2008). Though, it differed to the other part, the high-frequency volatility that indicated the change of the index according to the most recent information.

To apply spline function for estimating volatility, the previous studies have shown that using the too many number of knots might provide overfitted model (Engel & Rangel, 2008). Therefore, it needed to define an appropriated number of knots that provided the most explainable volatility model. Engel and Rangel (2008) used Bayesian Information Criterion (BIC) for selecting the number of knots of the exponential quadratic spline function in low-frequency volatility estimation, differed to Liu et al. (2015) which used Akaike's Information Criterion (AIC) with the same function. Some studies, such as Farida et al. (2018), Laipaporn and Tongkumchum (2017) and Awalludin and Saelim (2016), subjectively chose the suitable number of knots of the natural cubic spline functions for their volatility model.

This study tried to apply the information criterion for selecting the number of knots of the natural cubic spline volatility model followed Laipaporn and Tongkumchum (2017). The performances of two widely used information criteria, AIC and BIC, were assessed by the Monte Carlo simulation to identify which information criteria was suitable for specifying the number of knots in natural cubic spline volatility model. Furthermore, the empirical results of using these two information criteria for selecting the natural cubic spline volatility model among various number of knots of two stock market index, Stock Exchange of Thailand index(SET) and Strait Time index (STI), during 1997-2017 was also presented.

This paper is organized as follows. Section 2 describes data and methodology used in this study. Section 3 and section 4 informs the Monte Carlo simulation result and the empirical results of the natural cubic spline volatility of two stock market index respectively. The last section is a conclusion.

2 METHODS

This study might divide into 2 parts. The first is Monte Carlo simulation. This simulation was conducted to compare the performance between two information criteria in selecting number of knots in natural cubic spline volatility model given the true volatility was previously identified. The information criteria which provided the volatility model that better fitted to the given volatility was indicated as the preferred criteria. Another part of this study was applying the information criteria for specifying the natural cubic spline volatility model of each stock market index. The details of data and methodology were described as follows.

2.1 Data

2.1.1 Simulated returns series

This study assumed that the daily returns series ($R_t^{i,j}$) had zero mean. Each series which contained 5,000 daily returns indexed by trading day (t), was simulated as the multiplicative combination of two components, the known volatility and the noise series, which parameterized by the following formula.

$$R_t^{i,j} = \sigma_t^j \varepsilon_t^i \tag{1}$$

According to Engel and Rangel (2008), the low-frequency volatility was the unconditional volatility, which was not constant but gradually changed by the time. Consequently, three kinds of pre-specified volatility (σ_t^j), which included the low fluctuated volatility, the moderate fluctuated volatility and the high fluctuated volatility, were assumed as an additive combination of a single sinusoidal function of time, trading day (t), followed Saejiang *et al.* (2001). j identified the kind of these volatilities. Besides that, the 100 series of random noise ε_t^i, all were assumed having fat-tailed distribution, were generated by transforming the white noise, z_t^i, followed Huber (1964) as this formula.

$$\varepsilon_t^i = \begin{cases} c + a(z_t^i - c) \\ z_t^i \\ -c + a(z_t^i + c) \end{cases} \tag{2}$$

The constant values c and a were 1.25 and 2.5, respectively. Totally there were 300 simulated returns series in this simulation. These three

pre-specified volatilities and three absolute returns series from 300 simulated returns series with respect to each pre-specified volatility are shown in Figure 1.

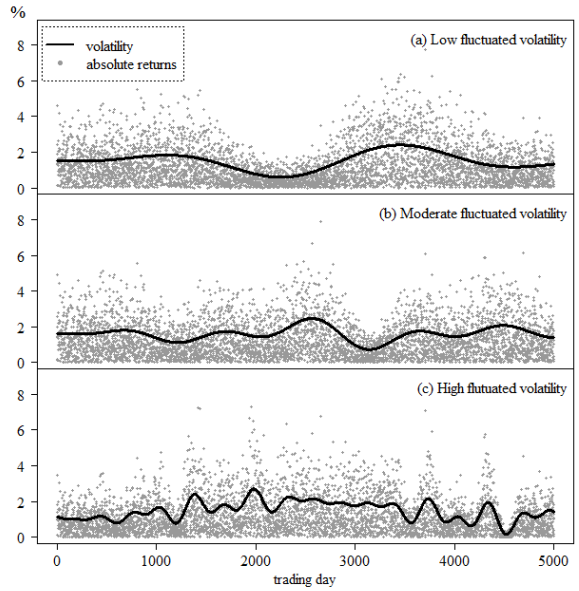


Figure 1: three kinds of pre-specified volatility with the example of the absolute returns series of each volatility

2.1.2 *Returns series of two stock markets index*

The index of two stock markets, SET and STI, during 1997-2017 were obtained from yahoo finance website. The daily returns of each stock market index were calculated as following equation.

$$R_t = \log \frac{I_t}{I_{t-1}} \tag{3}$$

Where R_t is the log return and I_t is the daily stock index at time t. Time series plots of daily stock index and their corresponding daily absolute returns are shown in Figure 2.

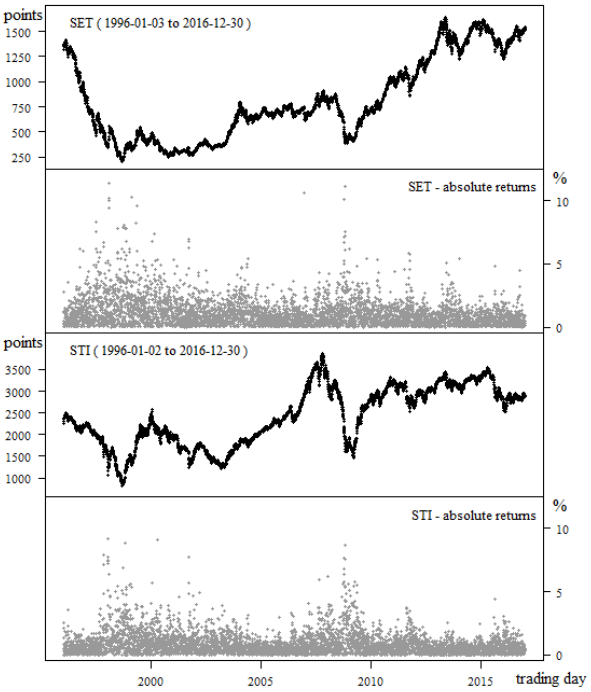


Figure 2: Time series plots of the Stock Exchange of Thailand index and the Strait Time index with their daily absolute returns

2.2 *Natural cubic spline volatility model*

Spline function has been an attractive and flexible non-parametric method for curve estimation (Silverman, 1985). This function has been used for approximating the shape of curvilinear function without the necessity of pre-specifying the mathematical form of the function (Suits et al., 1978). The natural cubic spline function was a spline function that was linear in the distant past and future and practically fitted to the dependent variable for extracting the variation pattern of that series (Wahba, 1975). In financial context, the natural cubic spline function has been widely used as an interpolation technique to estimate yield curve of the financial assets (Hastie et al., 2009; Greene, 2002; Engle & Russell, 1998).

Laipaporn and Tongkumchum (2017) used natural cubic spline function for modeling volatility. This model supposed that the returns series had two multiplicative components The first component was the conditional volatility (S_t) which was modeled by the natural cubic spline function with equi-spaced knots and the second component was white noises (Z_t). Thus, the returns and volatility models were parameterized as follows.

$$R_t = S_t Z_t \tag{4}$$

$$S_t = a + bt + \sum_{k=1}^{p} c_k \, (t - t_k)_+^3 \tag{5}$$

where t denoted time which $t_1 < t_2 < \ldots < t_p$ were specified knots and an additive term, $(t - x)_+$ was $t - x$ for $t > x$ and zero otherwise. Since this spline function was linear outside the boundary knots, t_1 and t_p, the coefficients of quadratic and cubic were 0 for $t < t_1$ and $t > t_p$. To satisfy these constraints, the cubic spline functions in equation 5 became

$$S_t = a + bt + \sum_{k=1}^{p} c_k \left[(t - t_k)_+^3 - \frac{t_p - t_k}{t_p - t_{p-1}} \left(t - t_{p-1}\right)_+^3 + \frac{t_{p-1} - t_k}{t_p - t_{p-1}} \left(t - t_p\right)_+^3 \right] \tag{6}$$

The parameters of this function were estimated by maximizing the log likelihood function with respect to the returns series. The log likelihood function (L) was defined as follows,

$$L = \sum_{t=1}^{n} \left[-\log(s_t) - \frac{R_t^2}{2S_t^2} \right] \tag{7}$$

where S_t was natural cubic spline volatility following the equation 6 and R_t was the return on day t.

2.3 *The set of the number of equi-spaced knots*

Number of knots effected the estimated natural cubic spline volatility. Increasing number of equi-spaced knots provided the volatility model that was more fitted to the returns series. Successively, the estimated volatility with respect to that model became more varied. However, in order to prevent overfitting, the volatility model needed an optimal number of knots. To obtain that number, first was setting the set of possible number and second was electing the appropriate number by the selection criteria.

The natural cubic spline function required two boundary knots. Thus, the first member in the set of possible number was 3. With 3 knots, the whole returns series were separated into 4 parts. To increase the knots from 3 knots to the next one, the additional knots were put in the middle of each separated parts. So the next number in the set was 3+4 = 7. By repeating this procedure, the members in the set of number of knot consequently included 3, 7, 15, 31, 63 and so on.

2.4 *The number of knots selection criteria*

Increasing number of knots was increasing more parameters in the natural cubic spline volatility model. It made the volatility model more sensitive to the changes of daily returns. Generally, root mean squared error (RMSE) was used to indicate the goodness of fit of the model. It shown how well the estimates was fitted to the observed data. Therefore, this study employed RMSE for selecting number of knots that made the natural cubic spline volatility model most fitted to pre-specified volatility not to the simulated returns and the number of knots

that gave the least RMSE was chosen as the preferred number. RMSE was calculated as follows.

$$RMSE = \sqrt{\frac{\sum_t^n (\sigma_t - S_t)^2}{n}} \quad (8)$$

where σ_t was the pre-specified volatility and S_t was the estimated one.

In case of modeling unknown volatility, it could not apply RMSE to indicated the optimum number of knots like the case of simulation. The other criteria that have been used for specifying spline model in previous studies were AIC and BIC (Engel & Rangel, 2008; Liu et al., 2015). Both AIC and BIC were broadly used for specifying parsimonious model from the set of candidate models (Burnham and Anderson, 2002). These two criteria were formulated as follows.

$$AIC = -L + 2P \quad (9)$$

$$BIC = -L + P\log(n) \quad (10)$$

where L is maximum likelihood value. P is number of parameters and n is number of observations (Hastie et al., 2009; Venables & Ripley, 2002). The first term of two criteria was likelihood value which indicated the goodness of fit of the model whereas the second term was penalized term with respect to the number of parameters in the model. The weight of penalized term of AIC was constant but the weight of BIC was not constant. It was higher when number of observation increased (Burnham & Anderson, 2002).

RMSE as well as AIC and BIC were calculated after each estimation in Monte Carlo simulation, while only AIC and BIC were calculated in empirical study. The lowest value of RMSE AIC and BIC indicated the best approximate model according to each criterion.

3 MONTE CARLO SIMULATION RESULTS

Since each returns series in this simulation included 5,000 daily returns, subsequently the maximum number in the set of number of knots was 127. This number was obtained by trying to add more knots until it did not provide the lower value of RMSE, AIC and BIC. So the number of knots in the set included 3, 7, 15, 31, 63 and 127 knots.

All 300 returns series were used for estimating natural cubic spline volatility six times with six different number of knots, so each returns series had six candidate models and six values of RMSE, AIC and BIC. Figure 3 is the boxplots that summarized the values of RMSE, AIC and BIC obtained by this simulation. These three statistics were grouped by the kind of pre-specified volatility of the simulated returns series.

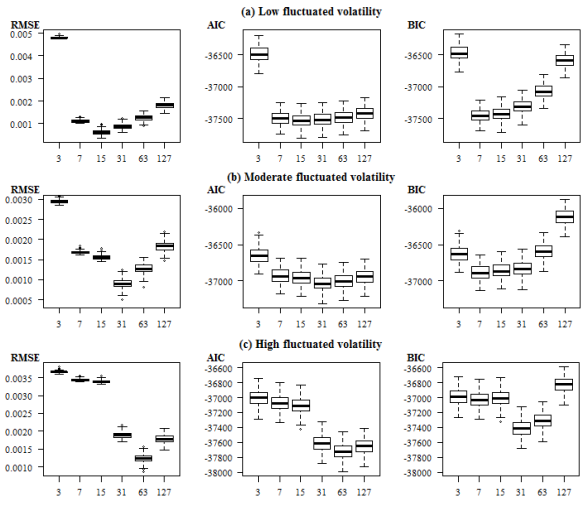


Figure 3: boxplot summarized the values of AIC, BIC and RMSE of the natural cubic spline volatility models with respect to the pre-specified volatility and the number of knots

RMSE values precisely shown that the simulated returns series generated from the same pre-specified volatility used the same number of knots for the appropriate natural cubic spline volatility model. The optimum number of knots for low, moderate and high fluctuated volatility possibly were 15, 31 and 63, respectively. This evidence shown that the more fluctuate volatility needed the more knots for natural cubic spline volatility model.

Comparing to AIC and BIC, the results shown that the optimum number of knots indicated by AIC were more likely to the number indicated by RMSE, differed to BIC which provide the smaller number than the other two criteria.

After identifying the number of knots that provided the least values of each criteria, RMSE and AIC specified the same number of knots for all simulated returns series which were 15, 31 and 63 for low, moderate and high fluctuated volatility, respectively. While BIC gave several different number. Most of number indicated by BIC were less than the number provided by the other two criteria.

These results implied that the number of knots indicated by BIC provided the estimated volatility that was underfitted to the pre-specified volatility when compared to RMSE, whereas AIC specified the well fitted model. Details of comparison shown in Table 1.

Table 1: The number of knots the provided the least criterion's value classified by kinds of pre-specified volatility and knot selection criteria (number of simulated returns series shown in the brackets)

Criteria	Low fluctuated volatility	Moderate fluctuated volatility	High fluctuated volatility
RMSE	15 (100)	31 (100)	63 (100)
AIC	15 (100)	31 (100)	63 (100)
BIC	7 (90)	7 (96)	31 (100)
	15 (10)	15 (2)	
		31 (2)	

4 RESULTS

The number of daily returns of SET and STI were 5135 and 5252, respectively. Likewise, the maximum number in the set of number of knots was 127 and there were six candidate volatility models for both series. Only AIC and BIC were calculated and compared. Their values were shown in Figure 4.

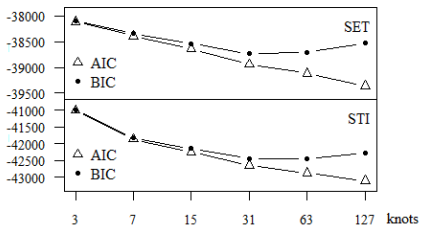


Figure 4: the values of AIC and BIC of the natural cubic spline volatility models of SET and STI with respect to the number of knots

BIC values were likely to lower when applying the more number of knots to the model. The natural cubic spline volatility model with 31 knots provided the lowest value of BIC for both series. However, BIC values became greater for the volatility model with the number of knots greater than 31. These BIC values still behaved like the values in Monte Carlo simulation. They could specify the appropriate model from the set of candidate models. Contrast to AIC, their values were smaller as increasing the number of knots. They differed to the AIC values in simulation which had a lowest value as a point for identifying the appropriate model from the set of candidate models.

Figure 5 and Figure 6 shown the low-frequency volatilities with their corresponding absolute returns series for both SET and STI series. These volatilities estimated by the natural cubic spline volatility model with 31 knots indicated by BIC, 127 knots indicated by AIC and 63 knots which subjectively selected.

The same as simulation results, BIC was likely to under-parameterize the model and the volatilities estimated by the model specified by BIC poorly traced the daily returns variation. Meanwhile, the number of knots indicated by AIC seem to provide too fluctuated estimated volatility. The model specified by AIC became over-parameterization.

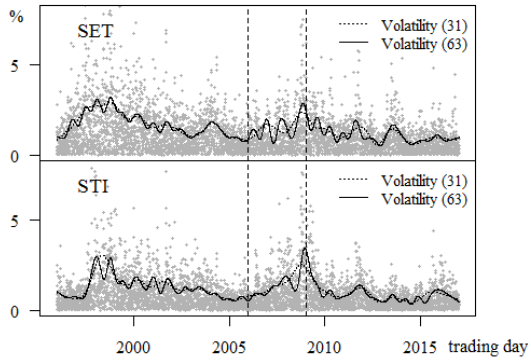


Figure 5: the low-frequency volatilities of SET and STI in case of applying 31 and 63 knots with the natural cubic spline volatility model and their corresponding absolute returns series

The estimated volatility by the model with 63 knots was higher fluctuated than the volatility estimated by the model with 31 knots but it was better to explain the variation of daily returns than another one especially in the period during 2006-2008 as shown in Figure 5.

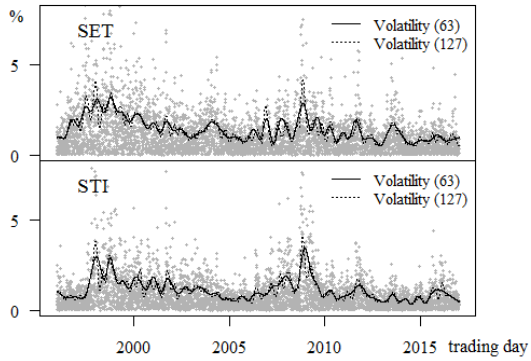


Figure 6: the low-frequency volatilities of SET and STI in case of applying 63 and 127 knots with the natural cubic spline volatility model and their corresponding absolute returns series

The estimated volatility by the model with 63 knots was less fluctuated than the volatility estimated by the model with 127 knots but their capability to trace the changes of return variation were not significantly different as seen in Figure 6. Therefore, the natural cubic spline volatility models with 63 knots were selected from the set of candidate models as the parsimonious model for estimating low-frequency volatility for both SET and STI.

5 CONCLUSIONS

The results of the Monte Carlo simulation precisely shown that BIC under-parameterized the model. This criterion specified the natural cubic spline volatility model with the number of knot that provided under estimated volatility. The empirical results also shown that the low-frequency volatilities of SET and STI estimated by the model which specified by BIC were less capability to trace the variation of their daily returns.

AIC seemly performed well in Monte Carlo simulation. The volatility models for each simulated returns series selected by AIC from the set of candidate models were fitted well to the pre-specified volatility. But empirical results shown that AIC provided too fluctuated low-frequency volatility and over-parameterized volatility model.

This study concluded that using BIC as number of knots selection criteria made the natural cubic spline volatility model under-parameterized. However, it needed more consideration for using AIC as a criterion for specifying the number of knots of natural cubic spline volatility model.

ACKNOWLEDGEMENTS

The authors would like to thank Emeritus Professor Don McNeil for his kindly suggestions.

REFERENCES

Awalludin, S.A. & Saelim, R. 2016. Modelling the volatility and assessing the performance of the model. The 20th International Annual Symposium on Computational Science and Engineering, Faculty of Science, Kasetsart University, Bangkok, July 27-29, 2016.

Engle, R.F. & Rangel, J. G. 2008. The Spline-GARCH model for low-frequency volatility and its global macroeconomic causes. *The Review of Financial Studies*, 21(3), 1187-1222.

Engle, R.F. and Russell, J. R. 1998. Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica*, 66(5), 1127-1162.

Farida, Makaje N., Tongkumchum P., Phon-On A. & Laipaporn J. 2018. Natural cubic spline model for estimating volatility, *International Journal on Advanced Science, Engineering and Information Technology*, 8(4). DOI:10.18517/ijaseit.8.4.3107

Greene, W.H. 2002. *Econometric Analysis*. Prentice Hall, U.S.A, pp. 199-200.

Hastie, T., Tibshirani, R. & Friedman, J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, New York, pp.141-148.

Huber, P.J. 1964. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1), 73-101.

Laipaporn, J. and Tongkumchum, P. 2017. Maximum likelihood estimation of non-stationary variance. 2nd ISI Regional Statistics Conference, 20-24 March 2017. Bali, Indonesia.

Liu, S., Tang, T., McKensie, A.M. & Liu, Y. 2015. Low-frequency volatility in China's gold futures market and its macroeconomic determinants. *Mathematical Problems in Engineering*. Article ID 646239. Doi:10.1155/2015/646239

Poon, S.H. 2005. *A Practical Guide to Forecasting Financial Market Volatility*. John Wiley and Sons, U.K., pp. 1-17.

Saejiang, S., Pornwiriyamongkol, W. & McNeil, D. 2001. Time series analysis of banking share returns in Thailand. *Songklanakarin Journal of Science and Technology*, 23(3), 443-448.

Silverman, B.W. 1985. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47(1), 1-52.

Suits, D.B., Mason, A. & Chan, L. 1978. Spline function fitted by standard regression methods. *The Review of Economics and Statistics*, 60(1), 132-139.

Venables, W.N. & Ripley, B.D. 2002. *Modern Applied Statistics with S. Springer-Verlag New York*, pp. 414-418.

Wahba, G. 1975. Smoothing noisy data with spline function. *Numerische Mathematik*, 24(5), 383-393

การเปรียบเทียบวิธีการคำนวณดัชนีความสามารถของกระบวนการสำหรับการแจกแจงเลขชี้กำลังและไวบูล

กฤษณะ ลาน้ำเที่ยง[1*] และ รัตนา เลิศสุวรรณศรี[2]

[1]สาขาวิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยแม่โจ้ จังหวัดเชียงใหม่
*อีเมลผู้ประสานงาน: k.lanumteang@mju.ac.th
[2]สาขาวิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ ศูนย์รังสิต จังหวัดปทุมธานี
อีเมล: rattana@mathstat.sci.tu.ac.th

บทคัดย่อ

การวิเคราะห์ความสามารถของกระบวนการคือการประเมินความผันแปรของกระบวนการผลิต โดยการวิเคราะห์ความผันแปรของกระบวนการเทียบกับพิกัดข้อกำหนดเฉพาะของผลิตภัณฑ์ ซึ่งมีข้อสมมติเบื้องต้นว่าข้อมูลที่ศึกษาต้องมีการแจกแจงปรกติ ในกรณีที่ข้อมูลที่ศึกษาไม่ได้มีการแจกแจงปรกติโดยส่วนใหญ่เราจะแปลงข้อมูลเพื่อให้ข้อมูลมีการแจกแจงปรกติ ในกรณีที่ไม่แปลงข้อมูลโปรแกรมสำเร็จรูปทางสถิติ เช่น MINITAB จะใช้วิธีเปอร์เซ็นไทล์ และวิธีของ MINITAB ในการคำนวณดัชนีความสามารถของกระบวนการ ดังนั้นผู้ศึกษาจึงสนใจเปรียบเทียบประสิทธิภาพวิธีการคำนวณดัชนีความสามารถของกระบวนการเมื่อข้อมูลไม่ได้มีการแจกแจงปรกติและไม่แปลงข้อมูลโดยพิจารณาข้อมูลของกระบวนการผลิตที่มีการแจกแจงเลขชี้กำลังและไวบูล ซึ่งพิจารณาการคำนวณดัชนีความสามารถของกระบวนการด้วยวิธีเปอร์เซ็นไทล์และวิธีของMINITAB เปรียบเทียบกับวิธีคลาสสิกที่มีข้อสมมติเบื้องต้นของการแจกแจงปรกติ ทำการจำลองสถานการณ์บนโปรแกรม R โดยการทดลองซ้ำ 10,000 รอบ เกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพ คือ ความเอนเอียง ความแปรปรวน และความคลาดเคลื่อนกำลังสองเฉลี่ย ผลการศึกษาพบว่า ในภาพรวมของสถานการณ์ที่ศึกษาวิธีเปอร์เซ็นไทล์ให้ค่าความเอนเอียงและ ความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำกว่าวิธีอื่น และผู้ศึกษาได้นำเสนอการคำนวณดัชนีความสามารถของกระบวน การทั้ง 3 วิธี สำหรับกระบวนการผลิตน้ำดื่มบรรจุขวดของบริษัทกรณีศึกษาแห่งหนึ่ง

คำสำคัญ: การวิเคราะห์ความสามารถของกระบวนการ ขีดจำกัดข้อกำหนด ขีดจำกัดกระบวนการตามธรรมชาติ ดัชนีความสามารถของกระบวนการ

Abstract

Process capability analysis is to measure how well a process performs and to quantify process variability. It also can be used to estimate the natural capability of the process corresponding to specifications of the products. An important assumption underlying the process capability is that their usual interpretation is based on a normal distribution of process output. If the underlying distribution is non-normal, then we often deal with the data transformation to normal distribution. On the other hand, some statistical package such as MINITAB provides alternative methods including percentile approach and MINITAB approach. Therefore, this study aims to compare performance of the methods calculating the process capability index based on non-normal distribution and without transformation. We considered the process characteristics data based on exponential and Weibull distribution. The percentile approach and MINITAB approach which provided on MINITAB for non-normal distribution were compared to the classical approach. The simulation technique was applied by running 10,000 times in each scenarios. Bias, variance and mean square error of estimation were the criteria of efficiency comparisons. Overall, the percentile approach shows the good performance in terms of the lowest bias and mean square error. An application of the process capability analysis of three considered methods was done for a case study of the bottle production of drinking water.

Keywords: process capability analysis; specification limits; natural process limit; process capability indices

1 บทนำ

ในกระบวนการผลิตใด ๆ ส่วนประกอบสำคัญที่เป็นปัจจัยหลักที่ทำให้เกิดผลผลิตที่ดีก็คือ คน เครื่องจักร วัตถุดิบ และกระบวนการทำงาน ถ้าองค์ประกอบต่าง ๆ นี้ไม่มีความบกพร่อง สินค้าหรือผลิตภัณฑ์ที่ผลิตได้ก็จะอยู่ในระดับมาตรฐานมีคุณภาพน่าเชื่อถือและทำให้ผู้บริโภคมีความพึงพอใจในตัวสินค้าและผลิตภัณฑ์ แต่ในความเป็นจริงในกระบวนการผลิตมักจะเกิดความผันแปรอยู่เสมอ ตั้งแต่ คน เครื่องจักร และวัตถุดิบ ซึ่งความผันแปรเหล่านี้จะทำให้คุณภาพของผลิตภัณฑ์ที่ผลิตได้นั้นไม่คงที่เกิดการเปลี่ยนแปลงไปตามความผันแปรดังกล่าว อาจจะส่งผลให้ผลิตภัณฑ์นั้นใช้ไม่ได้หรือไม่สามารถยอมรับได้หากมีบางส่วนที่บกพร่องผิดปรกติเกินขอบเขตเฉพาะที่กำหนด ดังนั้นเพื่อให้ผลิตภัณฑ์ที่บกพร่องอยู่ในเกณฑ์ที่พอยอมรับได้ไม่ถูกปฏิเสธ จึงจำเป็นที่จะต้องมีการควบคุมคุณภาพสินค้าและผลิตภัณฑ์โดยการควบคุมการผันแปรที่เกิดขึ้นในกระบวนการ ซึ่งหนึ่งในเทคนิคที่ใช้เป็นแนวทางในการควบคุมความผันแปรข้างต้นคือ การควบคุมคุณภาพเชิงสถิติ (Statistical Quality Control; SQC) โดยการประยุกต์ใช้ระเบียบวิธีการทางสถิติ เก็บรวบรวมข้อมูล วิเคราะห์และเปรียบเทียบแสดงข้อมูลที่ได้ เพื่อแก้ไขปัญหาต่างๆในระบบการผลิต โดยเราจะเก็บตัวอย่างสินค้ามาตรวจสอบ เพื่อวัดดูค่าคุณสมบัติต่าง ๆ แล้วนำข้อมูลที่ได้มาคำนวณและประเมินตามระเบียบวิธีทางสถิติ หลังจากนั้นนำมาเปรียบเทียบกับข้อกำหนดหรือมาตรฐานเพื่อที่จะพิจารณาว่าสินค้านั้นมีคุณภาพดีหรือไม่ (Howell, 1952) ในการควบคุมคุณภาพเชิงสถิติอาจจะดำเนินการโดยใช้เครื่องมือต่าง ๆ เช่น ใบตรวจสอบ แผนภูมิพาเรโต ฮิสโทแกรม และแผนภูมิควบคุม

นอกจากการพิจารณาและควบคุมกระบวนการผลิตว่าอยู่ภายใต้สภาวะการควบคุมจริงหรือไม่ ซึ่งหากพบว่ากระบวนการไม่อยู่ภายใต้สภาวะการควบคุมก็จำเป็นต้องกลับไปแก้ไขกระบวนการก่อน แต่ถ้าหากพบว่ากระบวนการอยู่ภายใต้การควบคุมจริงแล้ว เราอาจจะศึกษาและวิเคราะห์ความสามารถของกระบวนการต่อไป ในกรณีที่ตัวแปรบ่งชี้คุณภาพเป็นตัวแปรเชิงปริมาณ เราสามารถวิเคราะห์ความสามารถของกระบวนการ (Process Capability Analysis) โดยการประเมินความผันแปรของกระบวนการ (อาจอยู่ในรูปของฟังก์ชันความน่าจะเป็นที่จะระบุทั้งรูปทรงค่ากลางและประมาณการกระจายของการแจกแจง) และวิเคราะห์ความผันแปรนี้กับพิกัดข้อกำหนดเฉพาะของผลิตภัณฑ์ตลอดจนพิจารณาแหล่งความผันแปรต่างๆ เพื่อหาทางลดความผันแปรที่ศึกษาต่อไป (Plopanichcharean, 2001) ซึ่งแนวทางในการประเมินความสามารถของกระบวนการอาจจะพิจารณาจากดัชนีวัดความสามารถของกระบวนการ (Process Capability Index; C_p)

$$C_p = \frac{\text{ความคลาดเคลื่อนอนุโลมที่ยอมให้เกิด}}{\text{ความสามารถของกระบวนการ}} \quad (1)$$

จากสมการที่ (1) เมื่อแทนความคลาดเคลื่อนอนุโลมที่ยอมให้เกิดด้วยผลต่างของพิกัดข้อกำหนดเฉพาะด้านบน (Upper Specification Limit; USL) และพิกัดข้อกำหนดเฉพาะด้านล่าง (Lower Specification Limit; LSL) และแทนความสามารถของกระบวนการด้วยความคลาดเคลื่อนอนุโลมโดยธรรมชาติ (Natural Tolerance Limit; NTL) คือ 6σ เมื่อ σ คือ ส่วนเบี่ยงเบนมาตรฐานของกระบวนการ จะได้ว่า

$$C_p = \frac{USL - LSL}{6\sigma} \quad (2)$$

สำหรับการประเมินดัชนีความสามารถของกระบวนการเมื่อกำหนดพิกัดข้อกำหนดเฉพาะแบบด้านเดียวนั้น จะพิจารณาการกระจายของกระบวนการเพียงครึ่งหนึ่ง (คือ 3σ) กับระยะห่างระหว่างค่าเฉลี่ยของกระบวนการและพิกัดความคลาดเคลื่อนอนุโลม ดังสมการที่ (3) และ (4)

$$C_{pu} = \frac{USL - \mu}{3\sigma} \quad (3)$$

$$C_{pl} = \frac{\mu - LSL}{3\sigma} \quad (4)$$

ในการคำนวณดัชนีความสามารถของกระบวนการตามสมการที่ (2)–(4) นั้น นอกจากจะมีข้อสมมติว่ากระบวนการผลิตอยู่ภายใต้การควบคุมแล้ว ยังมีข้อสมมติว่าตัวแปรที่ศึกษาต้องมีการแจกแจงปรกติ อนึ่งในการกำหนดสัญลักษณ์สำหรับสมการที่ (1)–(4) ในตำราหรือบทความต่าง ๆ อาจจะมีความแตกต่างกันไป เช่น Montgomery (2001) จะเรียกดัชนีนี้ว่า "Process Capability Ratio (PCR)"

แนวทางในการแก้ไขปัญหาเมื่อตัวแปรที่ศึกษาไม่ได้มีการแจกแจงปรกติ (Non-normal Distribution) ตามข้อสมมติเบื้องต้น อาจใช้ระเบียบวิธีการแปลงข้อมูลให้มีการแจกแจงปรกติด้วยวิธีการแปลงของ Box-CoxJohnson หรือวิธีการแปลงอื่น ๆ เช่น Watthanacheewakul (2017) ได้ศึกษาวิธีการแปลงของ Manly Yeo-Johnson และ Nelson สำหรับข้อมูลที่มีการแจกแจงไวบูลเพื่อที่จะคำนวณดัชนีความสามารถของกระบวนการตามสมการที่ (2) ซึ่งผลการศึกษานี้พบว่าการแปลงข้อมูลด้วยวิธีข้างต้นสามารถแปลงข้อมูลไวบูลเป็นการแจกแจงปรกติได้และให้ผลการคำนวณดัชนีความสามารถของกระบวนการที่ไม่แตกต่างกัน อนึ่งโปรแกรมสำเร็จรูปทางสถิติ เช่น MINITAB ก็มีฟังก์ชันสำหรับการแปลงข้อมูลเพื่อการคำนวณดัชนีความสามารถของกระบวนการเมื่อข้อมูลไม่ได้มีการแจกแจงปรกติ (Solution center, 2007) อย่างไรก็ตามแนวคิดในการแปลงข้อมูลต้องอาศัยความรู้พื้นฐานทางสถิติอย่างลึกซึ้งและในบางกรณีข้อมูลที่ศึกษาเมื่อแปลงแล้วอาจจะไม่ได้มีการแจกแจงปรกติตามข้อสมมติเบื้องต้นที่ต้องการ หนึ่งในทางเลือกในการแก้ไขปัญหาดังกล่าวคือการใช้วิธีการประมาณความสามารถของกระบวนการ (การกระจายของกระบวนการ) และการประมาณพารามิเตอร์แสดงตำแหน่งด้วยค่าเปอร์เซ็นไทล์ (Percentile Approach) และอีกหนึ่งทางเลือกคือวิธีการแปลงกลับความน่าจะเป็นสะสมจากค่าสังเกตในตัวอย่างสู่การแจกแจงปรกติมาตรฐาน ซึ่งทั้งสองวิธีนี้ง่ายและสะดวกในการคำนวณ โดยโปรแกรมสำเร็จรูป MINITAB ก็มีฟังก์ชันในการคำนวณดัชนีความสามารถของกระบวนทั้งสองวิธีนี้เมื่อข้อมูลไม่ได้มีการแจกแจงปรกติด้วย (Solution center, 2007) ดังนั้นเพื่อเป็นแนวทางในการเลือกใช้วิธีการคำนวณดัชนีเพื่อประเมินความสามารถของกระบวนการ ผู้ศึกษาจึงทำการศึกษาเปรียบเทียบประสิทธิภาพของวิธีการคำนวณดัชนีความสามารถของกระบวนการ เมื่อ

ข้อมูลไม่ได้มีการแจกแจงปรกติโดยพิจารณากรณีศึกษาตัวแปรบ่งชี้คุณภาพที่มีการแจกแจงเลขชี้กำลังและการแจกแจงไวบูล ซึ่งเป็นข้อมูลที่สำคัญในงานด้านการควบคุมภาพกระบวนการโดยเฉพาะการวิเคราะห์ความเชื่อถือได้และอายุการใช้งานของผลิตภัณฑ์

2 วิธีดำเนินการศึกษา

จากวัตถุประสงค์ของการศึกษาที่ต้องการเปรียบเทียบวิธีการคำนวณดัชนีความสามารถของกระบวนการเมื่อข้อมูลมีการแจกแจงเลขชี้กำลังและไวบูลผู้ศึกษาใช้วิธีการจำลองแบบด้วยเทคนิคมอนติคาร์โลโดยใช้โปรแกรม R ในแต่ละสถานการณ์ที่ศึกษาทำซ้ำ 10,000 รอบ และเพื่อให้ผู้อ่านเข้าใจถึงแนวคิดของการวิเคราะห์ความสามารถของกระบวนการและวิธีการคำนวณดัชนีความสามารถของกระบวนการ ผู้ศึกษาได้สรุปเนื้อหาและวิธีการคำนวณดัชนีความสามารถของกระบวนการต่าง ๆ ที่ใช้ในการเปรียบเทียบประสิทธิภาพในการศึกษาครั้งนี้ ตามรายละเอียดดังต่อไปนี้

2.1 ดัชนีความสามารถของกระบวนการ

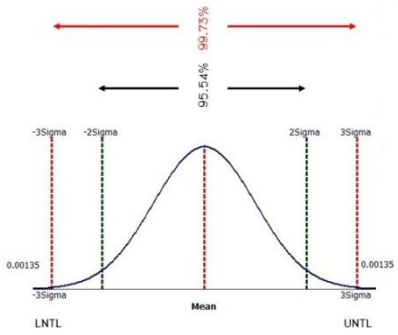
ความสามารถของกระบวนการ (Process Capability) คือความผันแปรโดยธรรมชาติของผลิตภัณฑ์ที่เกิดขึ้นจากกระบวนการที่ศึกษาซึ่งเป็นผลมาจากการวัดงานที่ได้รับการผลิตจากกระบวนการที่ศึกษาและความสามารถโดยธรรมชาติความสม่ำเสมอของผลิตภัณฑ์ที่ผลิตได้จากกระบวนการที่อยู่ภายใต้สภาวะควบคุมดังนั้นการวิเคราะห์ความสามารถของกระบวนการ (Process Capability Analysis; PCA) คือการประเมินความผันแปรของกระบวนการและวิเคราะห์ความผันแปรนี้กับพิกัดข้อกำหนดเฉพาะของผลิตภัณฑ์ตลอดจนพิจารณาแหล่งความผันแปรต่างๆ เพื่อหาทางลดความผันแปรที่ศึกษาต่อไป (Plopanichcharean, 2001) ซึ่งในกระบวนการผลิตใด ๆ สำหรับตัวแปรบ่งชี้คุณภาพเชิงปริมาณมักจะมีการกำหนดพิกัดข้อกำหนดเฉพาะด้านล่าง (Lower Specification Limit; LSL) และพิกัดข้อกำหนดเฉพาะด้านบน (Upper Specification Limit; USL) โดยจะถือว่าผลิตภัณฑ์ที่ผลิตได้จากกระบวนการผลิตตรงตามข้อกำหนดถ้ามีค่าอยู่ระหว่าง LSL และ USL

นอกจากกระบวนผลิตใด ๆ ที่มีการกำหนดพิกัดข้อกำหนดเฉพาะในกระบวนการผลิตย่อมมีความผันแปรหรือความคลาดเคลื่อนตามธรรมชาติที่เกิดขึ้น หรืออาจจะเรียกว่า "ขอบเขตเพื่อการยินยอมให้เกิดความคลาดเคลื่อนสำหรับกระบวนการ" (Sukchareonpong, 2000) ซึ่งขีดจำกัดกระบวนการตามธรรมชาติ โดยทั่วไปคำนวณจากค่าพารามิเตอร์ของประชากรหรือจากตัวอย่างที่มีขนาดใหญ่ เช่น สำหรับตัวแปรสุ่ม X ที่มีการแจกแจงปรกติซึ่งมีค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานเท่ากับ μ และ σ ตามลำดับขีดจำกัดกระบวนการตามธรรมชาติมีค่าห่างจากค่าเฉลี่ยของประชากร $\pm 3\sigma$ กล่าวคือ เมื่อกำหนดให้ UNTL คือ ขีดจำกัดเกณฑ์ความคลาดเคลื่อนตามธรรมชาติบน (Upper Natural Tolerance Limit) และ LNTL คือ ขีดจำกัดเกณฑ์ความคลาดเคลื่อนตามธรรมชาติล่าง (Lower Natural Tolerance Limit) จะได้ว่า

$$P(\text{LNTL} < X < \text{UNTL}) = 0.9973 \quad (5)$$

หรือ $$P(X < \text{LNTL}) = P(X > \text{UNTL}) = 0.00135 \quad (6)$$

ซึ่งแสดงดังรูปที่1



รูปที่ 1 ลักษณะของความผันแปรจากสาเหตุธรรมชาติ

โดยทั่วไปในการประเมินความสามารถของกระบวนการอาจจะพิจารณาจากดัชนีวัดความสามารถของกระบวนการ (Process Capability Index; C_p) ซึ่งพิจารณาจากอัตราส่วนความคลาดเคลื่อนอนุโลมที่ยอมให้เกิด (คือ USL–LSL) และความสามารถของกระบวนการด้วยความคลาดเคลื่อนอนุโลมโดยธรรมชาติ (คือ 6σ) ดังที่แสดงในสมการที่ (1)–(2) สำหรับพิกัดกำหนดเฉพาะแบบสองทาง ซึ่งจะถือว่ากระบวนการผลิตมีความสามารถในระดับดีถ้าค่า C_p ไม่ต่ำกว่า 1.33 หรือสัดส่วนของเสียที่พบไม่เกิน 64 ส่วนในล้านส่วน (Part Per Million; ppm) และการประเมินจากดัชนีสำหรับพิกัดข้อกำหนดเฉพาะด้านเดียวดังสมการที่ (3)–(4) จะถือว่ากระบวนการผลิตที่ทำอยู่มีความสามารถในระดับดีถ้าค่า C_p ไม่ต่ำกว่า 1.25 ซึ่งในการคำนวณดัชนีความสามารถของกระบวนการจากตัวอย่างขนาด n ตามสมการที่ (2)–(4) เรามีความจำเป็นต้องประมาณค่าพารามิเตอร์ต่าง ๆ ของกระบวนการดังรายละเอียดที่จะกล่าวถึงในหัวข้อ 2.1.1–2.1.3

2.1.1 Classical Approach หรือ Normality Approach

ภายใต้ข้อสมมติของการแจกแจงปรกติ เมื่อสุ่มตัวอย่างขนาด n จากประชากรดังกล่าว ด้วยวิธีการอนุมานทางสถิติสามารถประมาณค่าเฉลี่ยประชากร μ ด้วยค่าเฉลี่ยตัวอย่าง $\bar{X}$ และประมาณส่วนเบี่ยงเบนมาตรฐานของประชากร σ ด้วยส่วนเบี่ยงเบนมาตรฐานของตัวอย่าง s นอกจากนี้ในกระบวนการควบคุมคุณภาพเชิงสถิติเราอาจจะใช้แผนภูมิควบคุมในการพิจารณาว่ากระบวนการผลิตอยู่ภายใต้การควบคุมหรือไม่ เช่น ในกรณีที่ใช้แผนภูมิควบคุมค่าเฉลี่ยและพิสัยเราสามารถประมาณค่า σ ด้วย $\frac{\bar{R}}{d_2}$ เมื่อ $\bar{R}$ คือค่าเฉลี่ยของพิสัยจากแต่ละกลุ่มตัวอย่างย่อย และ d_2 ได้จากตารางตัวประกอบสำหรับแผนภูมิควบคุมซึ่งขึ้นอยู่กับขนาดตัวอย่างย่อย (Montgomery, 2001) สำหรับในการจำลองสถานการณ์ของการศึกษาครั้งนี้ ผู้ศึกษาจะใช้ค่าเฉลี่ยตัวอย่าง $\bar{X}$ และส่วนเบี่ยงเบนมาตรฐานของตัวอย่าง s ในการประมาณค่าเฉลี่ยของประชากร μ และส่วนเบี่ยงเบนมาตรฐานของประชากร σ ตามลำดับ ซึ่งเมื่อแทนค่าในสมการที่ (2)–(4) จะได้ดัชนีความสามารถของกระบวนการดังนี้

$$C_{np} = \frac{USL - LSL}{6s} \quad (7)$$

$$C_{npu} = \frac{USL - \bar{x}}{3s} \quad (8)$$

$$C_{npl} = \frac{\bar{x} - LSL}{3s} \quad (9)$$

จากสมการที่ (7) จะพบว่าการคำนวณดัชนีความสามารถของกระบวนการสำหรับพิกัดกำหนดเฉพาะแบบสองทางไม่ได้มีการพิจารณาค่าพารามิเตอร์แสดงตำแหน่งของกระบวนการหรือค่าเฉลี่ยของกระบวนการ ซึ่งการคำนวณดังกล่าวอยู่ภายใต้ข้อสมมติของการแจกแจงปรกติ โดยสามารถกล่าวได้ว่าค่าเฉลี่ยของกระบวนการอยู่กึ่งกลางระหว่าง USL และ LSL แต่ในสถานการณ์จำลองแบบของการศึกษาครั้งนี้ ทำการศึกษาภายใต้การแจกแจงเลขชี้กำลังและไวบูล ดังนั้นผู้ศึกษาจะพิจารณาดัชนีความสามารถของกระบวนการเมื่อกำหนดพิกัดเฉพาะแบบสองทางจากค่าต่ำสุดระหว่าง C_{npl} และ C_{npu} ซึ่งแสดงดังสมการที่ (10)

$$C_{npk} = \min\{C_{npl}, C_{npu}\} \quad (10)$$

2.1.2 Percentile Approach

จากหัวข้อที่ผ่านมา เราจะพบว่าการคำนวณดัชนีความสามารถของกระบวนดังกล่าวพัฒนาภายข้อสมมติของการแจกแจงปรกติ ในกรณีที่ตัวแปรบ่งชี้ลักษณะคุณภาพของกระบวนการไม่ได้มีการแจกแจงปรกติ ถ้าเราไม่แปลงข้อมูลให้มีการแจกแจงปรกติแล้วคำนวณดัชนีความสามารถของกระบวนการตามวิธีในหัวข้อ 2.1.1 อาจจะดำเนินการได้โดยการประมาณการกระจายของกระบวนการ (ส่วนเบี่ยงเบนมาตรฐานของตัวแปรที่ศึกษา) จากผลต่างของตำแหน่งในเปอร์เซ็นไทล์จากข้อมูลในตัวอย่างที่สัมพันธ์กับการประมาณส่วนเบี่ยงเบนมาตรฐานของกระบวนการ โดยการพิจารณาค่าเปอร์เซ็นไทล์ที่ 0.135, 50 และ 99.865 จากค่าสังเกตของตัวอย่างขนาด n ซึ่งพบว่าค่าผลต่างของเปอร์เซ็นไทล์ที่ 99.865 และ 0.135 นั้นมีค่าโดยประมาณเท่ากับ $6\hat{\sigma}$ ภายใต้การแจกแจงปรกติ (Montgomery, 2001) วิธีการคำนวณดัชนีความสามารถของกระบวนการโดยการพิจารณาค่าเปอร์เซ็นไทล์นี้เป็น default ในโปรแกรม MINITAB กรณีที่กระบวนการไม่ได้มีการแจกแจงปรกติ และองค์การระหว่างประเทศว่าด้วยการมาตรฐาน (International Organization for Standardization; ISO) ก็แนะนำให้ใช้วิธีการนี้ในการคำนวณดัชนีความสามารถของกระบวนการ(ซึ่งอ้างอิงใน help ของโปรแกรมสำเร็จรูป MINITAB สำหรับ Process Capability (Non-normal Distribution)) สำหรับบทความนี้จะเรียกวิธีการคำนวณดัชนีความสามารถของกระบวนการนี้ว่า "วิธีเปอร์เซ็นไทล์" ซึ่งสามารถสรุปขั้นตอนการคำนวณได้ดังต่อไปนี้

1) จากค่าสังเกตของตัวอย่างขนาด n พิจารณาค่าเปอร์เซ็นไทล์ที่ 0.135, 50 และ 99.865 โดยกำหนดให้มีค่าเท่ากับ $X_{0.00135}$, $X_{0.50}$ และ $X_{0.99865}$ ตามลำดับ

2) เมื่อกำหนด USL แทนพิกัดข้อกำหนดเฉพาะด้านบนและ LSL แทนพิกัดข้อกำหนดเฉพาะด้านล่าง จะได้ดัชนีความสามารถของกระบวนการสำหรับข้อกำหนดเฉพาะแบบสองทาง คือ

$$C_{pp} = \frac{USL - LSL}{X_{0.99865} - X_{0.00135}} \quad (11)$$

และสำหรับดัชนีความสามารถของกระบวนการเมื่อกำหนดพิกัดข้อกำหนดเฉพาะแบบด้านเดียวเฉพาะขอบเขตบน และดัชนีความสามารถของกระบวนการเมื่อกำหนดพิกัดข้อกำหนดเฉพาะแบบด้านเดียวเฉพาะขอบเขตล่าง สามารถคำนวณจากสมการที่ (12) และ (13) ตามลำดับ

$$C_{ppu} = \frac{USL - X_{0.50}}{X_{0.99865} - X_{0.50}} \quad (12)$$

$$C_{ppl} = \frac{X_{0.50} - LSL}{X_{0.50} - X_{0.00135}} \quad (13)$$

จะเห็นว่าวิธีเปอร์เซ็นไทล์นี้ไม่ข้อสมมติเบื้องต้นเกี่ยวกับการแจกแจงของตัวแปรที่ศึกษา

2.1.3 MINITAB Approach

ในโปรแกรมสำเร็จรูป MINITAB นอกจากจะนำเสนอวิธีการคำนวณดัชนีความสามารถของกระบวนการเมื่อข้อมูลไม่ได้มีการแจกแจงปรกติด้วยวิธีเปอร์เซ็นไทล์ ดังที่ได้กล่าวในหัวข้อ 2.1.2 นั้น MINITAB ได้เสนอทางเลือกในการคำนวณดัชนีความสามารถของกระบวนการโดยใช้แนวคิดการแปลงกลับสัดส่วนของตัวอย่างที่มีค่าสูงกว่า USL และสัดส่วนของตัวอย่างที่มีค่าต่ำกว่า LSL ไปสู่การแจกแจงปรกติมาตรฐาน โดยใช้แนวคิดของฟังก์ชันการแจกแจงสะสม (Cumulative Distribution Function; cdf) แล้วใช้ค่าจากการแจกแจงปรกติมาตรฐานไปคำนวณดัชนีความสามารถของกระบวนการต่อไป ในบทความนี้จะเรียกวิธีนี้ว่า "วิธีของ MINITAB" ซึ่งสามารถสรุปวิธีการคำนวณตามขั้นตอนต่อไปนี้

1) จากค่าสังเกตของตัวอย่างขนาด n คำนวณสัดส่วนของตัวอย่างที่มีค่าสูงกว่า USL กำหนดให้เท่ากับ P_u และคำนวณสัดส่วนของตัวอย่างที่มีค่าต่ำกว่า LSL กำหนดให้เท่ากับ P_l

2) แปลงกลับค่า P_u และ P_l ไปสู่การแจกแจงปรกติมาตรฐานโดยใช้แนวคิดฟังก์ชันการแจกแจงสะสม (cdf) กล่าวคือ คำนวณค่า z_u และ z_l ที่สอดคล้องกับเงื่อนไข $P(Z > z_u) = P_u$ และ $P(Z < z_l) = P_l$ ตามลำดับ

3) จะได้ดัชนีความสามารถของกระบวนการสำหรับข้อกำหนดเฉพาะแบบสองทาง คือ

$$C_{mp} = \frac{z_u - z_l}{6} \quad (14)$$

สำหรับดัชนีความสามารถของกระบวนการเมื่อกำหนดพิกัดข้อกำหนดเฉพาะแบบด้านเดียวเฉพาะขอบเขตบน และดัชนีความสามารถของกระบวนการเมื่อกำหนดพิกัดข้อกำหนดเฉพาะแบบด้านเดียวเฉพาะขอบเขตล่าง จะคำนวณจากสมการที่ (15) และ (16) ตามลำดับ

$$C_{mpu} = \frac{z_u}{3} \quad (15)$$

$$C_{mpl} = \frac{-z_l}{3} \quad (16)$$

4) และเนื่องจากข้อมูลที่ศึกษาไม่ได้มีการแจกแจงปรกติ ค่าเฉลี่ยของกระบวนการอาจจะไม่ได้อยู่กึ่งกลางระหว่างค่า ULS และ LSL ดังนั้นเราอาจจะพิจารณาดัชนีความสามารถของกระบวนการเมื่อกำหนดพิกัดเฉพาะแบบสองทางจาก

$$C_{mpk} = \min\{C_{mpl}, C_{mpu}\} \quad (17)$$

จะเห็นได้ว่าวิธีการของ MINITAB ง่ายและสะดวกในการคำนวณดัชนีความสามารถของกระบวนการโดยที่ไม่ต้องมีข้อสมมติเบื้องต้นเกี่ยวกับการแจกแจงของตัวแปรที่ศึกษา อย่างไรก็ตามในทางปฏิบัติอาจจะพบปัญหาถ้าจากตัวอย่างขนาด n ไม่พบหน่วยตัวอย่างที่มีค่าต่ำกว่าพิกัดกำหนดเฉพาะขอบเขตล่างและไม่พบหน่วยตัวอย่างที่มีค่าสูงกว่าพิกัดกำหนดเฉพาะขอบเขตบนเลย ซึ่งจะทำให้ได้ค่า P_u และ P_l เท่ากับ 0 และเราไม่สามารถคำนวณค่า z_u และ z_l ได้ แนวทางในการแก้ไขปัญหาสำหรับกรณีการจำลองสถานการณ์ที่ศึกษา เมื่อเกิดกรณีดังกล่าวผู้ศึกษาจะกำหนดให้ $z_u = 4$ และ $z_l = -4$ ซึ่งจะทำให้ได้ดัชนีความสามารถของกระบวนการเท่ากับ 1.33 ตามเกณฑ์ขั้นต่ำที่แสดงว่ากระบวนการมีความสามารถในระดับดี

2.2 การกำหนดสถานการณ์ที่ศึกษา

ผู้ศึกษากำหนดสถานการณ์ที่ศึกษาทั้งหมด 400 สถานการณ์ ดังต่อไปนี้

2.2.1กำหนดขนาดตัวอย่าง n 5 ระดับ คือ 30, 50, 100, 200 และ 500

2.2.2 กำหนดพารามิเตอร์สำหรับการแจกแจงเลขชี้กำลัง 4 รูปแบบ คือ มี scale parameter β ซึ่ง $\beta \in$ {1, 2, 5, 10} และกำหนดพารามิเตอร์สำหรับการแจกแจงไวบูล 6 รูปแบบ คือ มี shape parameter α โดยที่ $\alpha \in$ {1.5,2} และ scale parameter β ซึ่ง $\beta \in$ {1, 2, 5}

2.2.3 กำหนดดัชนีความสามารถของกระบวนการสำหรับพิกัดข้อกำหนดเฉพาะแบบสอง เท่ากับ 1.00 และ 1.33 และพิกัดข้อกำหนดเฉพาะแบบทางเดียวทั้งด้านบนและล่างเท่ากับ 1.00

ซึ่งแต่ละสถานการณ์ตาม 2.2.1–2.2.3 ทำการจำลองแบบซ้ำ 10,000 รอบ โดยใช้โปรแกรม R

2.3 เกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพ

สำหรับการศึกษาครั้งนี้ ดัชนีความสามารถของกระบวนการจะเสมือนพารามิเตอร์ที่เราต้องการประมาณค่า ดังนั้นเกณฑ์ที่ใช้ในการเปรียบประสิทธิภาพของแต่ละวิธีจากการจำลองสถานการณ์ต่าง ๆ ผู้ศึกษาจึงพิจารณาจากความเอนเอียง (Bias) ความแปรปรวน (Variance; Var) และความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error; MSE) ของการประมาณค่า ซึ่งจากการจำลองแบบ 10,000 รอบ เมื่อกำหนดให้ $\hat{C}_p$ แทนค่าดัชนีความสามารถของกระบวนการแต่ละวิธีที่ใช้ในการประมาณค่า C_p ดังนั้นจะได้ว่า

ความเอนเอียงของ $\hat{C}_p$

$$Bias = \frac{1}{10{,}000}\sum_{i=1}^{10{,}000}\hat{C}_{pi} - C_p$$

ความแปรปรวนของ $\hat{C}_p$

$$Var = \frac{1}{10{,}000}\sum_{i=1}^{10{,}000}(\hat{C}_{pi} - \bar{\hat{C}}_p)^2$$

เมื่อ $\bar{\hat{C}}_p = \frac{1}{10{,}000}\sum_{i=1}^{10{,}000}\hat{C}_{pi}$

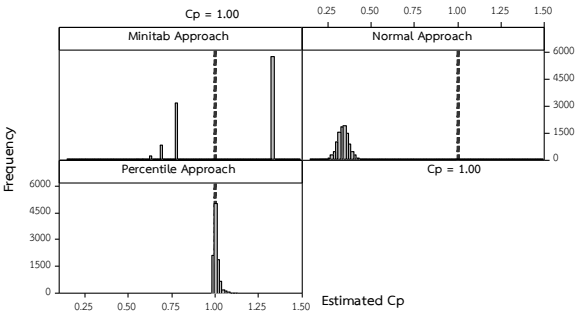
ความคลาดเคลื่อนกำลังสองเฉลี่ยของ $\hat{C}_p$

$$MSE = \frac{1}{10{,}000}\sum_{i=1}^{10{,}000}(\hat{C}_{pi} - C_p)^2$$

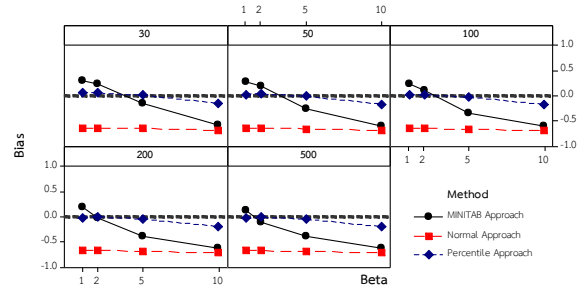
3 สรุปผลการศึกษา

3.1 ผลการจำลองสถานการณ์

สำหรับการแจกแจงเลขชี้กำลัง เมื่อพิจารณาความเอนเอียงของ $\hat{C}_p$ พบว่า วิธีคลาสสิกให้ค่าความเอนเอียงที่สูงกว่าวิธีเปอร์เซ็นไทล์และวิธีของ MINITAB และค่าดัชนีความสามารถของกระบวนการที่คำนวณด้วยวิธีคลาสสิกมีค่าต่ำกว่าค่าที่แท้จริงในทุกสถานการณ์ที่ศึกษา โดยภาพรวมพบว่าวิธีเปอร์เซ็นไทล์ให้ค่าความเอนเอียงที่ต่ำที่สุด ซึ่งวิธีเปอร์เซ็นไทล์และวิธีของ MINITAB จะให้ค่าประมาณดัชนีความสามารถของกระบวนการที่สูงกว่าค่าที่แท้จริงในกรณีที่ Scale Parameter มีค่าน้อยและตัวอย่างขนาดเล็ก แต่เมื่อ Scale Parameter และตัวอย่างขนาดมีค่าเพิ่มขึ้นทั้งสองวิธีก็จะให้ค่าประมาณดัชนีความสามารถของกระบวนการที่ต่ำกว่าค่าที่แท้จริงเช่นเดียวกับวิธีคลาสสิก เมื่อพิจารณาการแจกแจงของ $\hat{C}_p$ จากการจำลองสถานการณ์ พบว่า $\hat{C}_p$ ของวิธีคลาสสิกมีลักษณะการแจกแจงใกล้เคียงปรกติ รายละเอียดดังรูปที่ 2-4

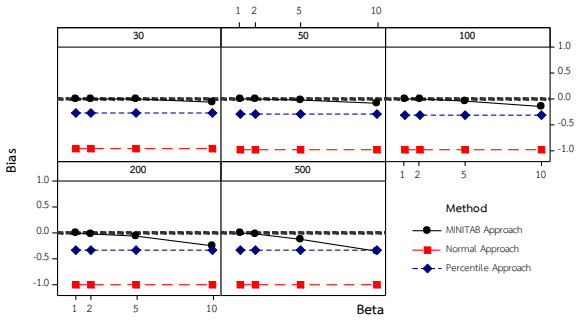


รูปที่ 2 การแจกแจงของ $\hat{C}_p$ เมื่อ C_p = 1.00 กรณี n=100, Exp(2)



Panel variable: Sample Size

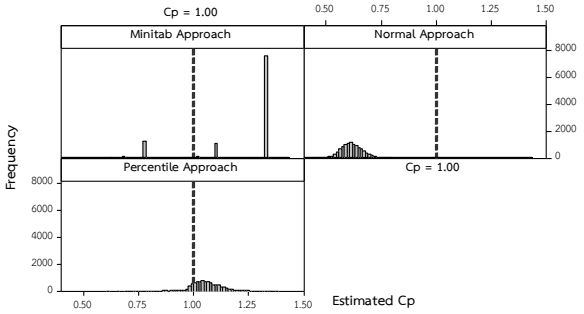
รูปที่ 3 ความเอนเอียงของ $\hat{C}_p$ เมื่อ C_p = 1.00 กรณีการแจกแจงเลขชี้กำลัง; Exp(β)



Panel variable: Sample Size

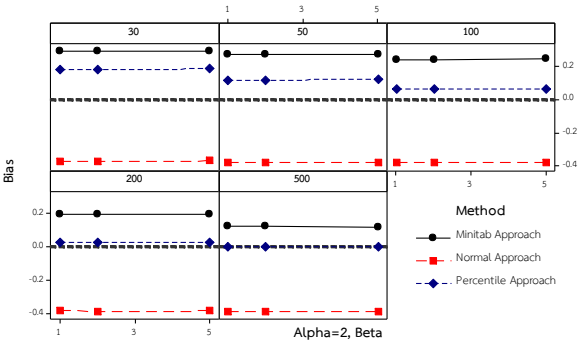
รูปที่ 4 ความเอนเอียงของ $\hat{C}_p$ เมื่อ C_p = 1.33 กรณีการแจกแจงเลขชี้กำลัง; Exp(β)

เมื่อพิจารณาความเอนเอียงของ $\hat{C}_p$ สำหรับสถานการณ์ที่ศึกษากรณีการแจกแจงไวบูล พบว่า วิธีคลาสสิกให้ค่าความเอนเอียงที่สูงกว่าวิธีเปอร์เซ็นไทล์และวิธีของ MINITAB และค่าดัชนีความสามารถของกระบวนการที่คำนวณด้วยวิธีคลาสสิกมีค่าต่ำกว่าค่าที่แท้จริงในทุกสถานการณ์ที่ศึกษา ซึ่งให้ผลที่คล้ายคลึงกับกรณีการแจกแจงเลขชี้กำลังโดยภาพรวมเมื่อกำหนดค่า $C_p = 1.00$ พบว่าวิธีเปอร์เซ็นไทล์ให้ค่าความเอนเอียงที่ต่ำที่สุด ซึ่งวิธีเปอร์เซ็นไทล์และวิธีของ MINITAB จะให้ค่าประมาณดัชนีความสามารถของกระบวนการที่สูงกว่าค่าที่แท้จริงในเกือบทุกสถานการณ์ที่ศึกษา แต่สำหรับกรณีที่กำหนดค่า $C_p = 1.33$ พบว่าวิธีของ MINITAB จะให้ค่าความเอนเอียงที่ต่ำกว่าวิธีอื่น ๆและในภาพรวมของสถานการณ์การแจกแจงไวบูลที่ศึกษาทั้งหมด ทุกวิธีจะให้ค่าประมาณดัชนีความสามารถของกระบวนการที่ต่ำกว่าค่าที่แท้จริงเมื่อพิจารณาการแจกแจงของ $\hat{C}_p$ จากการจำลองสถานการณ์ พบว่า $\hat{C}_p$ ของวิธีคลาสสิกและวิธีเปอร์เซ็นไทล์มีลักษณะการแจกแจงใกล้เคียงปรกติ รายละเอียดดังรูปที่ 5-7



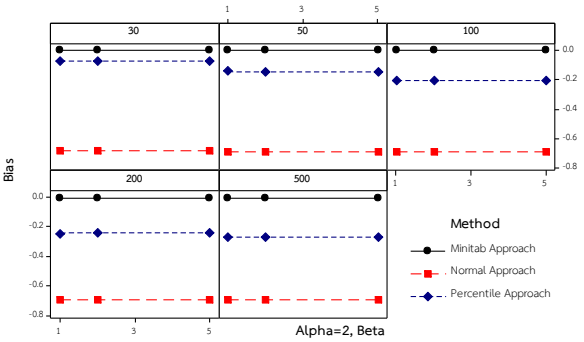
Panel variable: Method

รูปที่ 5 การแจกแจงของ $\hat{C}_p$ เมื่อ C_p = 1.00 กรณี n=100, Wei(2,2)



Panel variable: Sample Size

รูปที่ 6 ความเอนเอียงของ $\hat{C}_p$ เมื่อ C_p = 1.00 กรณีการแจกแจงไวบูล; Wei(2,β)



Panel variable: Sample Size

รูปที่ 7 ความเอนเอียงของ $\hat{C}_p$ เมื่อ C_p = 1.33 กรณีการแจกแจงไวบูล; Wei(2,β)

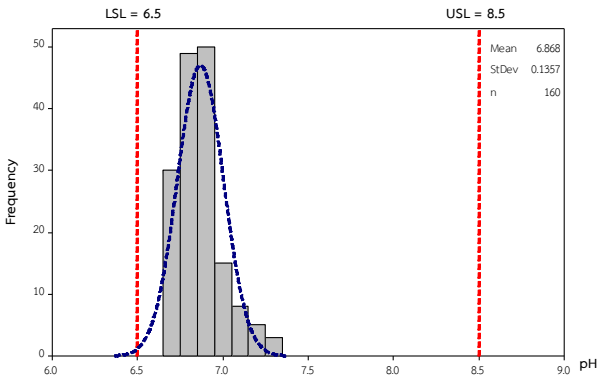
เมื่อพิจารณาความแปรปรวนของดัชนีความสามารถของกระบวนการสำหรับการแจกแจงเลขชี้กำลัง เมื่อกำหนด C_p=1.00 โดยภาพรวมวิธีคลาสสิกจะให้ค่าความแปรปรวนที่ต่ำกว่าวิธีอื่น ๆ แต่เมื่อ C_p=1.33 วิธีเปอร์เซ็นไทล์จะให้ค่าความแปรปรวนที่ต่ำที่สุด สำหรับทุกสถานการณ์ที่ศึกษาของการแจกแจงไวบูลเมื่อ C_p=1.00 วิธีคลาสสิตแสดงค่าความแปรปรวนที่ต่ำที่สุด แต่ในกรณีที่ C_p=1.33 วิธี MINITAB จะให้ค่าความแปรปรวนที่ต่ำที่สุด รายละเอียดดังตารางที่ ผ.1-ผ.2

เมื่อพิจารณาความคลาดเคลื่อนกำลังสองเฉลี่ยของดัชนีความสามารถของกระบวนการสำหรับการแจกแจงเลขชี้กำลัง เมื่อกำหนด C_p=1.00 โดยภาพรวมวิธีวิธีเปอร์เซ็นไทล์จะให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยที่ต่ำที่สุด รองลงมาคือวิธีของ MINITAB แต่เมื่อ C_p=1.33 วิธีของ MINITAB จะมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุด ในส่วนของการจำลองสถานการณ์การแจกแจงไวบูล เมื่อC_p=1.00 โดยภาพรวมวิธีวิธีเปอร์เซ็นไทล์จะให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยที่ต่ำที่สุด และสำหรับ C_p=1.33 วิธีของ MINITAB จะให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุด รายละเอียดดังตารางที่ ผ.1-ผ.2

3.2 ผลการวิเคราะห์ข้อมูลกรณีศึกษา

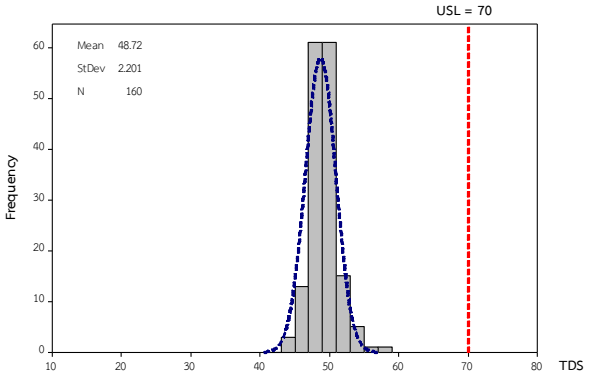
ผู้ศึกษาใช้ข้อมูลจากกรณีศึกษาเรื่อง "การลดของเสียในกระบวนการผลิตน้ำดื่มบรรจุขวด กรณีศึกษาบริษัทน้ำใสใจจริง จำกัด" (Jaroenkul and Lanumteang, 2018) เพื่อแสดงการประยุกต์ใช้วิธีการประเมินดัชนีความสามารถของกระบวนการทั้ง 3 วิธี ซึ่งพิจารณาค่าความเป็นกรดเบส (pH) และค่าของแข็งที่ละลายในน้ำ (TDS) ทางบริษัทกรณีศึกษามีข้อกำหนดเฉพาะของค่า pH ของน้ำดื่มทุกขวดต้องมีค่าระหว่าง 6.5-8.5 และข้อกำหนดของ TDS ต้องไม่เกิน 70 มิลลิกรัม/ลิตร ซึ่งข้อมูลที่ศึกษาเก็บรวบรวมจากการกระบวนการผลิตระหว่างวันที่ 8 ถึง 19 มกราคม 2561 รวม 10 วันทำการชักตัวอย่างทุก ๆ หนึ่งชั่วโมง ครั้งละ 2 ขวด ตรวจวัดลักษณะ pH และ TDS หนึ่งขวด ส่วนอีกหนึ่งขวดที่เหลือจะเก็บไว้เป็นตัวอย่างอ้างอิง ซึ่งมีหน่วยตัวอย่างรวมทั้งหมด 160 ขวด

เมื่อพิจารณาค่า pH ของตัวอย่างพบว่า มีค่าเฉลี่ยเท่ากับ 6.868 ซึ่งไม่อยู่กึ่งกลางระหว่างข้อกำหนดเฉพาะด้านล่างและบน (6.5-8.5) และค่อนไปทางข้อกำหนดเฉพาะขอบเขตล่าง ส่วนเบี่ยงเบนมาตรฐาน 0.136 ซึ่งแสดงว่า pH ของน้ำดื่มบรรจุขวดมีการกระจายน้อยไม่พบหน่วยตัวอย่างที่มีค่า pH ต่ำกว่าข้อกำหนดเฉพาะขอบเขตล่างและสูงเกินกว่าที่กำหนด การแจกแจงของ pH ในกระบวนการผลิตนี้มีลักษณะเบ้ขวา รายละเอียดดังรูปที่8 การวิเคราะห์ความสามารถของกระบวนการด้วยดัชนีทั้ง 3 วิธี พบว่ามีค่า C_{np} = 0.90, C_{pp} = 2.00 และ C_{zp} = 1.33 ตามลำดับ ซึ่งเมื่อพิจารณาการคำนวณโดยวิธี Normal Approach จะถือว่ากระบวนการนี้ไม่มีความสามารถ แต่ถ้าพิจารณาการคำนวณโดยวิธี Percentile Approach และ MINITAB Approach จะถือว่ากระบวนการผลิตนี้มีความสามารถในระดับดี



รูปที่ 8 การแจกแจง pH ของตัวอย่างน้ำดื่มบรรจุขวด

เมื่อพิจารณาค่า TDS ของตัวอย่างพบว่า มีค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน เท่ากับ 48.719 และ 2.201 มิลลิกรัม/ลิตร ตามลำดับ ไม่พบหน่วยตัวอย่างที่มีค่า TDS สูงกว่าข้อกำหนดเฉพาะขอบเขตบนที่กำหนด (70 มิลลิกรัม/ลิตร) การแจกแจงของ TDS ในกระบวนการผลิตนี้มีลักษณะใกล้เคียงปรกติ รายละเอียดดังรูปที่9 การวิเคราะห์ความสามารถของกระบวนการด้วยดัชนีทั้ง 3 วิธี พบว่ามีค่า C_{np} = 3.22, C_{pp} = 2.33 และ C_{zp} = 1.33 ตามลำดับซึ่งจากผลการคำนวณดัชนีความสามารถจากทั้ง 3 วิธี จะถือว่ากระบวนการผลิตนี้มีความสามารถในระดับดี



รูปที่ 9 การแจกแจง TDS ของตัวอย่างน้ำดื่มบรรจุขวด

4 อภิปรายผลการศึกษาและข้อเสนอแนะ

ผู้ศึกษาทำการเปรียบเทียบประสิทธิภาพการคำนวณดัชนีความสามารถของกระบวนการเมื่อข้อมูลไม่ได้มีการแจกแจงปรกติและกรณีที่ไม่แปลงข้อมูล โดยพิจารณากรณีศึกษา การแจกแจงเลขชี้กำลังและไวบูล โดยภาพรวมของทุกสถานการณ์ที่ศึกษาพบว่า วิธีเปอร์เซ็นไทล์มีประสิทธิภาพดีที่สุด โดยให้ค่าความเอนเอียงและค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำว่าวิธีอื่น สำหรับทุกสถานการณ์ที่ศึกษาวิธีคลาสสิกหรือวิธีภายใต้ข้อสมมติการแจกแจงปรกติจะให้ค่าประมาณดัชนีความสามารถของกระบวนการที่ต่ำกว่าค่าที่แท้จริง

ในบางสถานการณ์ที่กำหนดให้ C_p = 1.33 วิธีของ MINITAB จะให้ค่าความเอนเอียง ความแปรปรวนและค่าความความคลาดเคลื่อนกำลัง

สองเฉลี่ยต่ำกว่าวิธีอื่น ซึ่งอาจจะเป็นผลเนื่องจากในสถานการณ์นี้ระดับความสามารถของกระบวนการอยู่ในระดับดี ส่งผลให้ข้อมูลตัวอย่างจากการจำลองสถานการณ์ไม่มีหน่วยตัวอย่างที่มีลักษณะที่สูงหรือต่ำกว่าข้อกำหนดพิกัดเฉพาะด้านบนและด้านล่าง ซึ่งเป็นปัญหาของวิธี MINITAB ที่ผู้ศึกษาได้กล่าวถึงในหัวข้อ 2.1.3 โดยในกรณีนี้ผู้ศึกษาได้กำหนดให้ค่าดัชนีความสามารถของกระบวนการที่คำนวณโดยวิธี MINITAB เท่ากับค่าที่แท้จริง 1.33 จึงทำให้ค่าความเอนเอียง ความแปรปรวนและความคลาดเคลื่อนกำลังสองเฉลี่ยมีค่าต่ำ

จากข้อมูลในกรณีศึกษา เมื่อพิจารณาการแจกแจงของ TDS ของน้ำดื่มบรรจุขวดซึ่งพบว่ามีการแจกแจงใกล้เคียงปรกติ ผลการคำนวณดัชนีความสามารถของกระบวนการโดยวิธีคลาสสิกให้ค่าที่สูงกว่าวิธีอื่น ๆ แต่กรณีค่าความเป็นกรดเบส pH ที่ข้อมูลมีลักษณะเบ้ขวา วิธีคลาสสิกแสดงค่าดัชนีความสามารถของกระบวนการที่ต่ำกว่าวิธีเปอร์เซ็นไทล์และวิธีของ MINITAB ซึ่งสอดคล้องกับผลของการจำลองสถานการณ์

ในการศึกษาครั้งนี้ผู้ศึกษาเลือกพิจารณาวิธีการคำนวณดัชนีความสามารถของกระบวนการที่ง่ายและสะดวกในทางปฏิบัติ โดยเฉพาะวิธีเปอร์เซ็นไทล์ที่ไม่มีข้อสมมติของการแจกแจงข้อมูล ในการศึกษาครั้งต่อไปอาจจะพิจารณาเปรียบเทียบวิธีการดังกล่าวกับการคำนวณที่มีขั้นตอนที่ซับซ้อนขึ้น เช่น วิธีการแปลงข้อมูลตามการศึกษาของ Watthanacheewakul (2017) หรือการวิเคราะห์โครงข่ายประสาทเทียมตามการศึกษาของ Abbasi (2009)

เอกสารอ้างอิง

Abbasi, B. (2009). A neural network applied to estimated process capability of non-normal process. *Expert Systems with Application*, 36(2), 3039-3100.

Howell, J. M. (1952). Statistical quality control. *Mathematics Magazine*, 25(3), 155-157.

Jaroenkul, K. and Lanumteang, K. (2018). Defect reduction in production of bottled drinking water: A Case Study of the Namsai Jailing Co., Ltd. *In Proceeding of the National Undergraduate Conference on Statistics 2018,* 98-114. ChiangMai Thailand. (in Thai).

Montgomery, D. C. (2001). *Introduction to Statistical Quality Control.* John Wiley & Sons (New York).

Plopanichcharean, K. (2001). *Process Capability Analysis.*Technology Promotion Association (THAILAND-JAPAN), (in Thai).

Solution Centre. (2007). Quality Process Improvement, Minitab 15 Statistical Software. (in Thai).

Sukchareonpong, P. (2000). Engineering Quality Control. Se-Education. (in Thai).

Watthanacheewakul, L. (2017). Calculating the process capability ratio for weibull data. *Naresuan University Journal: Science and Technology (NUJST),* 25(1), 44-56.

ภาคผนวก

ตารางที่ ผ.1 ความแปรปรวนและความคลาดเคลื่อนกำลังสองเฉลี่ยของ $\hat{C}_p$ สำหรับข้อกำหนดพิกัดเฉพาะสองทางกรณีการแจกแจงเลขชี้กำลัง; Exp(β)

β	n	C_p= 1.00						C_p= 1.33					
		Variance of $\hat{C}_p$			MSE of $\hat{C}_p$			Variance of $\hat{C}_p$			MSE of $\hat{C}_p$		
		Normal	Percentile	MINITAB	Normal	Percentile	MINITAB	Normal	Percentile	MINITAB	Normal	Percentile	MINITAB
1	30	*0.0031	0.0047	0.0247	0.4240	*0.0068	0.1083	0.0032	0.0033	*0.0008	0.9584	0.0791	*0.0008
1	50	*0.0019	0.0027	0.0311	0.4312	*0.0031	0.1039	0.0019	0.0010	*0.0008	0.9711	0.0895	*0.0008
1	100	*0.0010	0.0026	0.0374	0.4376	*0.0026	0.0929	0.0010	*0.0003	0.0011	0.9823	0.0984	*0.0011
1	200	*0.0005	0.0035	0.0398	0.4419	*0.0037	0.0761	0.0005	*0.0001	0.0018	0.9866	0.1030	*0.0018
1	500	*0.0002	0.0031	0.0356	0.4438	*0.0036	0.0502	0.0002	*0.0000	0.0023	0.9908	0.1058	*0.0023
2	30	*0.0032	0.0033	0.0688	0.4249	*0.0054	0.1184	0.0032	0.0032	*0.0019	0.9617	0.0793	*0.0019
2	50	0.0019	*0.0010	0.0780	0.4324	*0.0016	0.1108	0.0019	*0.0011	0.0022	0.9707	0.0897	*0.0022
β	n	C_p= 1.00						C_p= 1.33					
		Variance of $\hat{C}_p$			MSE of $\hat{C}_p$			Variance of $\hat{C}_p$			MSE of $\hat{C}_p$		
2	100	0.0010	*0.0002	0.0834	0.4399	*0.0003	0.0915	0.0010	*0.0002	0.0039	0.9840	0.0984	*0.0039
2	200	0.0005	*0.0001	0.0624	0.4426	*0.0001	0.0626	0.0005	*0.0001	0.0057	0.9878	0.1030	*0.0058
2	500	0.0002	*0.0000	0.0181	0.4454	*0.0000	0.0301	0.0002	*0.0000	0.0087	0.9907	0.1059	*0.0091
5	30	*0.0029	0.0031	0.1436	0.4388	*0.0031	0.1698	*0.0031	0.0033	0.0119	0.9605	0.0796	*0.0121
5	50	0.0019	*0.0010	0.0882	0.4463	*0.0014	0.1571	0.0019	*0.0010	0.0162	0.9725	0.0903	*0.0167
5	100	0.0010	*0.0003	0.0229	0.4524	*0.0014	0.1475	0.0010	*0.0002	0.0231	0.9835	0.0990	*0.0248
5	200	0.0005	*0.0001	0.0043	0.4562	*0.0017	0.1468	0.0005	*0.0001	0.0288	0.9871	0.1037	*0.0332

5	500	0.0002	*0.0000	0.0014	0.4584	*0.0020	0.1495	0.0002	*0.0000	0.0340	0.9916	0.1065	*0.0499
10	30	*0.0024	0.0052	0.0249	0.4879	*0.0307	0.3714	*0.0031	0.0032	0.0448	0.9614	0.0817	*0.0489
10	50	*0.0015	0.0024	0.0080	0.4943	*0.0327	0.3756	0.0019	*0.0010	0.0530	0.9744	0.0922	*0.0612
10	100	*0.0008	0.0010	0.0031	0.5001	*0.0349	0.3789	0.0010	*0.0002	0.0661	0.9836	0.1014	*0.0902
10	200	*0.0004	*0.0004	0.0015	0.5031	*0.0359	0.3804	0.0005	*0.0001	0.0638	0.9901	*0.1060	0.1201
10	500	*0.0002	*0.0002	0.0006	0.5053	*0.0368	0.3822	0.0002	*0.0000	0.0334	0.9932	*0.1088	0.1518

หมายเหตุ: * แสดงค่าความแปรปรวนและความคลาดเคลื่อนกำลังสองเฉลี่ยของ $\hat{C}_p$ ต่ำที่สุดในสถานการณ์ที่ศึกษา

ตารางที่ ผ.2 ความแปรปรวนและความคลาดเคลื่อนกำลังสองเฉลี่ยของ $\hat{C}_p$ สำหรับข้อกำหนดพิกัดเฉพาะสองทางกรณีการแจกแจงไวบูล; Wei(α,β)

α	β	n	C_p= 1.00						C_p= 1.33					
			Variance of $\hat{C}_p$			MSE of $\hat{C}_p$			Variance of $\hat{C}_p$			MSE of $\hat{C}_p$		
			Normal	Percentile	MINITAB	Normal	Percentile	MINITAB	Normal	Percentile	MINITAB	Normal	Percentile	MINITAB
1.5	1	30	*0.0049	0.0126	0.0251	0.2560	*0.0258	0.1085	0.0050	0.0133	*0.0005	0.6869	0.0465	*0.0005
1.5	1	50	*0.0027	0.0063	0.0305	0.2600	*0.0113	0.1036	0.0029	0.0054	*0.0009	0.6922	0.0586	*0.0009
1.5	1	100	*0.0014	0.0036	0.0363	0.2630	*0.0046	0.0926	0.0014	0.0018	*0.0009	0.6989	0.0736	*0.0009
1.5	1	200	*0.0007	0.0029	0.0396	0.2645	*0.0030	0.0767	0.0007	*0.0006	0.0017	0.7021	0.0844	*0.0018
1.5	1	500	*0.0003	0.0022	0.0353	0.2653	*0.0023	0.0505	0.0003	*0.0002	0.0021	0.7042	0.0926	*0.0021
1.5	2	30	*0.0049	0.0128	0.0251	0.2562	*0.0262	0.1088	0.0048	0.0131	*0.0008	0.6867	0.0470	*0.0008
1.5	2	50	*0.0027	0.0062	0.0320	0.2597	*0.0111	0.1036	0.0029	0.0054	*0.0006	0.6925	0.0581	*0.0006
1.5	2	100	*0.0014	0.0038	0.0371	0.2629	*0.0047	0.0926	0.0014	0.0018	*0.0010	0.6986	0.0735	*0.0010
1.5	2	200	*0.0007	0.0029	0.0406	0.2644	*0.0029	0.0757	0.0007	*0.0006	0.0015	0.7012	0.0843	*0.0015
1.5	2	500	*0.0003	0.0022	0.0356	0.2654	*0.0023	0.0501	0.0003	*0.0002	0.0022	0.7036	0.0924	*0.0023
1.5	5	30	*0.0049	0.0129	0.0246	0.2554	*0.0263	0.1086	0.0051	0.0134	*0.0007	0.6838	0.0464	*0.0007
1.5	5	50	*0.0028	0.0063	0.0296	0.2602	*0.0114	0.1042	0.0029	0.0053	*0.0009	0.6923	0.0583	*0.0009
1.5	5	100	*0.0014	0.0036	0.0377	0.2628	*0.0046	0.0931	0.0014	0.0017	*0.0012	0.6986	0.0736	*0.0012
1.5	5	200	*0.0007	0.0030	0.0404	0.2649	*0.0030	0.0759	0.0007	*0.0006	0.0016	0.7020	0.0845	*0.0016
1.5	5	500	*0.0003	0.0022	0.0355	0.2654	*0.0024	0.0502	0.0003	*0.0002	0.0026	0.7038	0.0925	*0.0026
2.0	1	30	*0.0070	0.0227	0.0244	0.1481	*0.0551	0.1090	0.0075	0.0284	*0.0006	0.4739	0.0343	*0.0006
2.0	1	50	*0.0041	0.0122	0.0306	0.1498	*0.0256	0.1040	0.0044	0.0138	*0.0008	0.4769	0.0345	*0.0008
2.0	1	100	*0.0020	0.0059	0.0378	0.1498	*0.0095	0.0934	0.0020	0.0051	*0.0013	0.4809	0.0475	*0.0013
2.0	1	200	*0.0010	0.0036	0.0401	0.1502	*0.0042	0.0772	*0.0010	0.0020	0.0017	0.4828	0.0610	*0.0017
2.0	1	500	*0.0004	0.0023	0.0357	0.1507	*0.0023	0.0507	*0.0004	0.0007	0.0021	0.4842	0.0730	*0.0022
2.0	2	30	*0.0073	0.0227	0.0241	0.1475	*0.0545	0.1087	0.0077	0.0279	*0.0007	0.4705	0.0331	*0.0007
2.0	2	50	*0.0042	0.0119	0.0300	0.1493	*0.0252	0.1038	0.0044	0.0129	*0.0008	0.4776	0.0345	*0.0008
2.0	2	100	*0.0019	0.0062	0.0367	0.1499	*0.0096	0.0925	0.0020	0.0051	*0.0010	0.4814	0.0476	*0.0010
2.0	2	200	*0.0010	0.0038	0.0395	0.1510	*0.0044	0.0768	*0.0010	0.0021	0.0017	0.4825	0.0607	*0.0017
2.0	2	500	*0.0004	0.0024	0.0354	0.1509	*0.0024	0.0505	*0.0004	0.0007	0.0022	0.4842	0.0731	*0.0022
2.0	5	30	*0.0075	0.0238	0.0247	0.1470	*0.0572	0.1087	0.0076	0.0280	*0.0006	0.4725	0.0340	*0.0006
2.0	5	50	*0.0041	0.0123	0.0304	0.1488	*0.0263	0.1042	0.0044	0.0132	*0.0011	0.4760	0.0340	*0.0011
2.0	5	100	*0.0020	0.0059	0.0358	0.1500	*0.0096	0.0939	0.0021	0.0050	*0.0008	0.4817	0.0477	*0.0008
2.0	5	200	*0.0010	0.0037	0.0397	0.1507	*0.0043	0.0770	*0.0010	0.0021	0.0015	0.4832	0.0606	*0.0015
2.0	5	500	*0.0004	0.0023	0.0359	0.1512	*0.0023	0.0499	*0.0004	0.0008	0.0024	0.4844	0.0730	*0.0024

หมายเหตุ: * แสดงค่าความแปรปรวนและความคลาดเคลื่อนกำลังสองเฉลี่ยของ $\hat{C}_p$ ต่ำที่สุดในสถานการณ์ที่ศึกษา

การเปรียบเทียบการทดสอบค่าเฉลี่ยหรือค่ากลาง 2 ประชากรของการแจกแจงปรกติปลอมปนเมื่อความแปรปรวนไม่เท่ากัน

อัชฌา อระวีพร[1*] ศุภวิชญ์ ลีลาพีระพันธ์[2] ชุติมา โพชรา[3] สุพัฒน์ เทียนรุ่งโรจน์[4] และ อนุวัตร นามบุญศรี[5]

[1]ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง กทม.
*อีเมลผู้ประสานงาน: kaautcha@hotmail.com
[2,3,4,5]ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง กทม.
อีเมล: supavit_mound@hotmail.com

บทคัดย่อ

การวิจัยนี้เป็นการวิจัยเชิงจำลองมีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบค่าเฉลี่ยหรือค่ากลาง 2 ประชากร โดยใช้การทดสอบที การทดสอบซี การทดสอบแมนท์-วิทนี ยู วิธีบูตสแตรป และการทดสอบแวน-เดอร์ วาเดนท์ โดยศึกษาจากข้อมูลที่สุ่มมาจากประชากรที่มีการแจกแจงปรกติปลอมปนทั้ง 2 ประชากรเมื่อความแปรปรวนทั้งสองกลุ่มไม่เท่ากัน กำหนดขนาดตัวอย่าง (n_1, n_2) เท่ากับ (20,20) (50,50) (20,25) และ (50,70) ในการประมาณค่าความผิดพลาดแบบที่ 1 โดยกำหนดค่าเฉลี่ยของประชากรทั้งสองกลุ่ม (μ_1, μ_2) เท่ากับ (0,0) และประมาณค่ากำลังการทดสอบ กำหนดค่าเฉลี่ยของประชากรทั้งสองกลุ่ม (μ_1, μ_2) เท่ากับ (0,2) ที่ระดับนัยสำคัญ 2 ระดับคือ 0.01 และ 0.05 โดยใช้โปรแกรมอาร์ในการจำลองและวิเคราะห์ข้อมูล ทำการจำลองข้อมูลซ้ำ 1,000 รอบในแต่ละสถานการณ์ จากผลการวิจัย พบว่า สถิติไม่อิงพารามิเตอร์ คือ การทดสอบแวน-เดอร์ วาเดนท์ ส่วนใหญ่มีกำลังการทดสอบสูงที่สุด รองลงมาคือวิธีบูตสแตรป และแมนท์-วิทนี ยู

คำสำคัญ: การทดสอบที การทดสอบแมนท์-วิทนี ยู การทดสอบแวน-เดอร์ วาเดนท์ วิธีบูตสแตรป

ABSTRACT

This research is a simulating research that aimed to study and to compare for testing of two independent population means or medians by using t-test, Z-test, Mann – Whitney U Test, bootstrap method, and Van der Waerden test. The data is generated from two populations based on contaminated normal distribution and unequal variance. The sample sizes (n_1, n_2) are set as (20,20), (50,50), (20,25), and (50,70). For estimating type I error, the mean of two populations (μ_1, μ_2) are defined equal by (0,0). The mean of two populations (μ_1, μ_2) are defined unequal by (0,2) for estimating power of a test. The significant levels are focused on two levels at 0.01 and 0.05. R program is used for simulation and data analysis with 1,000 times for each situation. The results can be found that nonparametric statistics exhibits the highest power of a test such as Van der Waerden test, bootstrap method, and Mann-Whithney U test, respectively.

Keywords: t-test; Mann – Whitney U test; Van der Waerden test; Bootstap method

1. บทนำ

ในการสมมติฐานเพื่อเปรียบเทียบความแตกต่างของค่าเฉลี่ยหรือค่ากลางของประชากร 2 กลุ่มที่เป็นอิสระ โดยการทดสอบที่นิยมใช้ในการวิเคราะห์ข้อมูล ได้แก่ สถิติอิงพารามิเตอร์ เช่น การทดสอบซี (Z-test) การทดสอบที (t-test) โดยพิจารณาจากข้อมูลว่าเป็นการแจกแจงปรกติหรือไม่ และค่าความแปรปรวนของประชากรทั้ง 2 กลุ่มว่าเท่ากันหรือไม่เท่ากัน และเมื่อข้อมูลไม่เป็นไปตามข้อตกลงเบื้องต้นของสถิติอิงพารามิเตอร์จึงสามารถเลือกใช้สถิติไม่อิงพารามิเตอร์ เช่น การทดสอบแมนท์-วิทนี ยู (Man-Whitney U test) การทดสอบแวน-เดอร์ วาเดนท์ (Van der waerden) วิธีบูตสแตรป (Bootstrap method)

การแจกแจงปรกติปลอมปน (Contaminated Normal Distribution) เป็นการแจกแจงผสมระหว่างการแจกแจงปรกติ 2 การแจก

แจง โดยการแจกแจงหนึ่งมีข้อมูลเป็นรูปแบบของการแจกแจงปรกติที่ค่าเฉลี่ยและความแปรปรวนปกติ และอีกกลุ่มหนึ่งจะถูกสุ่มมากจากข้อมูลที่มีค่าเฉลี่ยปกติแต่มีค่าความแปรปรวนสูงมากโดยถ่วงน้ำหนักด้วยค่าความน่าจะเป็น ซึ่งเรียกว่าค่าสัดส่วนการปลอมปน (Contamination proportion) ซึ่งค่าที่ได้จะแสดงค่านอกเกณฑ์ (outlier) โดยมักเกิดเมื่อพนักงานขาดความชำนาญ หรือมีการเปลี่ยนแปลงวัตถุดิบ วัสดุ หรืออุปกรณ์ ทำให้ดูเสมือนว่ามีประชากรเล็ก ๆ เกิดขึ้น (Bakker, 2014)

จากการศึกษางานวิจัยที่เกี่ยวข้องพบว่า Songthong (2013) ศึกษาประสิทธิภาพการทดสอบอิงพารามิเตอร์และการทดสอบไม่อิงพารามิเตอร์ พบว่า การทดสอบที และ การทดสอบแวน-เดอร์ วาเดนท์ สามารถควบคุมความผิดพลาดแบบที่ 1 ได้และมีกำลังทดสอบสูงสุดเมื่อตัวอย่างทั้ง 2 มีขนาดเล็ก สำหรับการทดสอบแมนท์ – วิทนี ยู สามารถควบคุมความผิดพลาดแบบที่ 1 ได้และมีกำลังทดสอบสูงสุด เมื่อตัวอย่างทั้ง 2 กลุ่มมีขนาดกลางและใหญ่ Songthong (2014) ศึกษาความแกร่งและกำลังการทดสอบของการทดสอบอิงพารามิเตอร์และการทดสอบไม่อิงพารามิเตอร์ การศึกษาพบว่า เมื่อประชากรมีการแจกแจงปรกติ การทดสอบที และ การทดสอบแมนท์ – วิทนี ยู มีความเหมาะสมที่จะสามารถควบคุมความผิดพลาดแบบที่ 1 ได้ และมีกำลังการทดสอบสูงสุด

จากที่ได้กล่าวมาข้างต้นนี้ ผู้วิจัยจึงสนใจที่จะศึกษาการเปรียบเทียบการทดสอบค่าเฉลี่ยหรือค่ากลางของ ประชากรที่เป็นอิสระ 2 กัน เมื่อใช้การทดสอบที (t-test) การทดสอบซี (Z-test) การทดสอบแมนท์-วิทนี ยู (Mann-Whitney U test) วิธีบูตสแตรป (Bootstrap method) และการทดสอบแวน-เดอร์ วาเดนท์ (Van der Waerden Test) เพื่อหาว่าการทดสอบใดมีประสิทธิภาพสูงสุดสำหรับข้อมูลที่มีการแจกแจงการแจกแจงปรกติปลอมปนทั้ง 2 กลุ่มประชากร โดยใช้โปรแกรมอาร์ (R) ในการจำลองและวิเคราะห์ข้อมูล

2. วิธีการวิจัย

ในการวิจัยครั้งนี้เป็นการวิจัยเชิงทดลอง เพื่อศึกษาค่าความผิดพลาดแบบที่ 1 และกำลังการทดสอบของการทดสอบของประชากร 2 กลุ่มในกรณีค่าเฉลี่ยเท่ากันและความแปรปรวนไม่เท่ากัน กรณีที่ค่าเฉลี่ยไม่เท่ากันและความแปรปรวนไม่เท่ากัน

2.1 ขอบเขตการวิจัย

ในการวิจัยครั้งนี้กำหนดสถานการณ์ในการศึกษาเปรียบเทียบดังนี้

2.1.1 กำหนดให้ประชากรทั้ง 2 ประชากรเป็นอิสระกัน

2.1.2 ศึกษาในกรณีที่ตัวอย่างขนาดเท่ากันและไม่เท่ากัน ดังนี้

ขนาดตัวอย่าง (n_1, n_2)	
เท่ากัน $(n_1 = n_2)$	ไม่เท่ากัน $(n_1 \neq n_2)$
(20,20) , (50,50)	(20,25) , (50,70)

2.1.3 ศึกษาในกรณีที่ค่าความแปรปรวนทั้งสองกลุ่มไม่เท่ากัน โดยเกณฑ์ที่ใช้กำหนดค่าความแตกต่างของค่าความแปรปรวน เรียกว่า ค่านอนเซนทรัลลิตี้พารามิเตอร์ ϕ (Noncentrality Parameter) โดยคำนวณจาก

$$\phi = \sqrt{\frac{\sum_{i=1}^{2} \frac{(\sigma_i^2 - \bar{\sigma}^2)^2}{2}}{\sigma_1^2}}$$

เมื่อ σ_1^2 เป็น ค่าความแปรปรวนของประชากรที่มีค่าต่ำสุด

σ_i^2 เป็น ค่าความแปรปรวนของประชากรที่ i โดย $i = 1, 2$

$\bar{\sigma}^2$ เป็น ค่าเฉลี่ยความแปรปรวนของประชากรทั้ง กลุ่ม 2

มีรายละเอียดดังตารางต่อไปนี้

ระดับความแตกต่างของความแปรปรวน	ความแปรปรวนแต่ละประชากร $\sigma_1^2 : \sigma_2^2$	ϕ
มีความแตกต่างกันน้อย $(0 < \phi < 1.5)$	4: 6.2	0.55
มีความแตกต่างกันปานกลาง $(1.5 \leq \phi < 3)$	4 : 13.48	2.37

ข้อมูลของทั้งสองประชากรสุ่มมาจากการแจกแจงปรกติปลอมปน (Contaminated Normal Distribution) ด้วยค่าเฉลี่ย (μ) และค่าความแปรปรวน (σ^2) ที่มีสัดส่วนการปลอมปน (p) เท่ากับ 0.1 ค่านอกกลุ่มจากการแจกแจงปรกติที่มีค่าสเกลแฟคเตอร์ (c) เท่ากับ 5 และ 10 โดยมีฟังก์ชันความหนาแน่นความน่าจะเป็นดังนี้

$$f(x) = (1-p)N(\mu_1, \sigma^2) + (p)N(\mu_2, c^2\sigma^2)$$

สามารถสรุปเป็นตารางได้ดังนี้

กรณีค่าเฉลี่ยเท่ากันและระดับความแปรปรวนแตกต่างกันเล็กน้อย

ประชากรกลุ่มที่ 1	ประชากรกลุ่มที่ 2
$(0.9)N(0,4) +$ $(0.1)N(0,(5^2 \times 4))$	$(0.9)N(0,6.2) +$ $(0.1)N(0,(5^2 \times 6.2))$
$(0.9)N(0,4) +$ $(0.1)N(0,(10^2 \times 4))$	$(0.9)N(0,6.2) +$ $(0.1)N(0,(10^2 \times 6.2))$

กรณีค่าเฉลี่ยเท่ากันและระดับความแปรปรวนแตกต่างกันปานกลาง

ประชากรกลุ่มที่ 1	ประชากรกลุ่มที่ 2
$(0.9)N(0,4)+$ $(0.1)N(0,(5^2\times 4))$	$(0.9)N(0,13.48)+$ $(0.1)N(0,(5^2\times 13.48))$
$(0.9)N(0,4)+$ $(0.1)N(0,(10^2\times 4))$	$(0.9)N(0,13.18)+$ $(0.1)N(0,(10^2\times 13.48))$

กรณีค่าเฉลี่ยไม่เท่ากันและระดับความแปรปรวนแตกต่างกันเล็กน้อย

ประชากรกลุ่มที่ 1	ประชากรกลุ่มที่ 2
$(0.9)N(0,4)+$ $(0.1)N(0,(5^2\times 4))$	$(0.9)N(2,6.2)+$ $(0.1)N(2,(5^2\times 6.2))$
$(0.9)N(0,4)+$ $(0.1)N(0,(10^2\times 4))$	$(0.9)N(2,6.2)+$ $(0.1)N(2,(10^2\times 6.2))$

กรณีค่าเฉลี่ยไม่เท่ากันและระดับความแปรปรวนแตกต่างกันปานกลาง

ประชากรกลุ่มที่ 1	ประชากรกลุ่มที่ 2
$(0.9)N(0,4)+$ $(0.1)N(0,(5^2\times 4))$	$(0.9)N(2,13.48)+$ $(0.1)N(2,(5^2\times 13.48))$
$(0.9)N(0,4)+$ $(0.1)N(0,(10^2\times 4))$	$(0.9)N(2,13.18)+$ $(0.1)N(2,(10^2\times 13.48))$

2.1.4 ในทุกขนาดตัวอย่างและทุกกรณี กำหนดระดับนัยสำคัญ 2 ระดับ คือ 0.01 และ 0.05

2.1.5 ใช้โปรแกรม *R* เวอร์ชัน 3.4.3 ในการจำลองและวิเคราะห์ข้อมูล

2.2 สถิติทดสอบ

ในการทดสอบสมมติฐานเพื่อเปรียบเทียบค่าเฉลี่ยหรือค่ากลางระหว่างประชากร 2 กลุ่มที่เป็นอิสระกันสถิติอิงพารามิเตอร์ที่ใช้ทดสอบสำหรับสมมติฐานดังกล่าวในการศึกษาครั้งนี้มี 2 การทดสอบ คือ การทดสอบที และการทดสอบซี มีสมมติฐานในการทดสอบดังนี้

สมมติฐานว่าง $H_0:\ \mu_1=\mu_2$

สมมติฐานทางเลือก $H_1:\ \mu_1\neq\mu_2$

โดยที่ μ_1 และ μ_2 แทน ค่าเฉลี่ยของประชากรกลุ่มที่ 1 และ 2 ตามลำดับ

*2.2.1 การทดสอบที (t-test) (Kuh*aratanachai, 2013)

เป็นการทดสอบอิงพารามิเตอร์ที่ใช้ทดสอบความแตกต่างของค่าเฉลี่ยระหว่างประชากร 2 กลุ่มที่เป็นอิสระกัน จะสามารถใช้การทดสอบนี้ได้ก็ต่อเมื่อข้อมูลเป็นไปตามข้อกำหนดเบื้องต้น คือ ประชากร 2 กลุ่มอิสระกัน ประชากร 2 กลุ่มมีการแจกแจงปรกติ ตัวอย่างทั้ง 2 กลุ่มมีขนาดเล็ก เมื่อไม่ทราบค่าความแปรปรวนทั้งสองกลุ่ม (σ_1^2,σ_2^2) แต่ทราบว่าทราบว่าความแปรปรวนทั้งสองกลุ่มไม่เท่ากัน $(\sigma_1^2\neq\sigma_2^2)$

สถิติทดสอบ $t_{cal} = \dfrac{\bar{x}_1-\bar{x}_2}{\sqrt{\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}}}$

โดยมีองศาอิสระ $\upsilon = \dfrac{\left(\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{S_1^2}{n_1}\right)^2}{n_1-1}+\dfrac{\left(\dfrac{S_2^2}{n_2}\right)^2}{n_2-1}}$

เมื่อ

$\bar{x}_1,\bar{x}_2$ คือ ค่าเฉลี่ยของตัวอย่างของกลุ่มตัวอย่างที่ 1 และ 2

S_1^2,S_2^2 คือ ค่าความแปรปรวนของตัวอย่างของกลุ่มที่ 1 และ 2

n_1,n_2 คือ จำนวนตัวอย่างของกลุ่มที่ 1 และ 2

อาณาเขตวิกฤต : $t_{cal}>t_{\frac{\alpha}{2};\upsilon}$ หรือ $t_{cal}<-t_{\frac{\alpha}{2};\upsilon}$

2.2 2.การทดสอบซี (Z-test) (Kuharatanachai, 2013)

เป็นการทดสอบอิงพารามิเตอร์ที่ใช้ทดสอบความแตกต่างของค่าเฉลี่ยระหว่างประชากร กลุ่มที่เป็นอิสระกัน จะสามารถใช้การทดสอบ 2 กลุ่ม 2 ซีได้ก็ต่อเมื่อข้อมูลเป็นไปตามข้อกำหนดเบื้องต้น คือ ประชากร อิสระต่อกัน ประชากร2 กลุ่มมีการแจกแจงปรกติ ตัวอย่างทั้ง กลุ่มมี 2 ขนาดใหญ่ โดยใช้ทฤษฎีบทขีดจำกัดส่วนกลาง (Central Limit Theorem)

สถิติทดสอบ $Z_{cal} \approx \dfrac{\bar{x}_1-\bar{x}_2}{\sqrt{\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}}}$

อาณาเขตวิกฤต : $Z_{cal}>Z_{\frac{\alpha}{2}}$ หรือ $Z_{cal}<-Z_{\frac{\alpha}{2}}$

สถิติไม่อิงพารามิเตอร์ที่ใช้ทดสอบสำหรับสมมติฐานดังกล่าวในการศึกษาครั้งนี้มีการทดสอบแมนท์-วิทนี ยู วิธีบูตสแตรป และการทดสอบแวน-เดอร์ วาเดนท์ มีสมมติฐานในการทดสอบดังนี้

สมมติฐานว่าง $H_0:\ M_1=M_2$

สมมติฐานทางเลือก $H_1:\ M_1\neq M_2$

โดยที่ M_1 และ M_2 แทน ค่ามัธยฐานของประชากรกลุ่มที่ 1 และ 2

2.2.3 การทดสอบแมนท์-วิทนี ยู (Mann-Whitney U test)

(Janthasorn, 1998)

การทดสอบแมนท์-วิทนี ยู มักนิยมใช้เพื่อหลีกเลี่ยงกรณีที่ข้อมูลไม่เป็นไปตามข้อกำหนดของสถิติอิงพารามิเตอร์เช่น เมื่อข้อมูลมีมาตราต่ำกว่าแบบอันตรภาค และข้อมูลไม่เป็นการแจกแจงปรกติ

สถิติทดสอบ $Z_{cal} \approx \frac{T - n_1 n_2 / 2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}}$

เมื่อขนาดตัวอย่างใหญ่ $(n_1, n_2 \geq 20)$ (Janthasorn, 1998)

กำหนดให้ $T = S - \frac{n_1(n_1 - 1)}{2}$

และ S ผลรวมลำดับของตัวอย่าง n_1 และ n_2

อาณาเขตวิกฤต : $Z_{cal} > Z_{\frac{\alpha}{2}}$ หรือ $Z_{cal} < -Z_{\frac{\alpha}{2}}$

2.2.4 วิธีบูตสแตรป (Bootstrap Method)

(Boonpen et al., 2015)

เป็นการกำหนดตัวอย่างโดยสุ่มแบบคืนที่จำนวน n ตัวจากตัวอย่างสุ่ม $X_1, X_2, \ldots, X_n$ ที่มีการแจกแจงปรกติปลอมปน ซึ่งจะได้ตัวอย่างสุ่มชุดใหม่ คือ $X_1^*, X_2^*, \ldots, X_n^*$ ซึ่งเรียกว่าตัวอย่างบูตสแตรป และคำนวณค่า $\bar{X}_j^*$ ของตัวอย่างบูตสแตรป เมื่อ j = 1,..., B โดยที่ B เป็นจำนวนการทำซ้ำ

การคำนวณค่าวิธีบูตสแตรป

กำหนดให้ $\hat{D}_j = \bar{X}^*_{j(1)} - \bar{X}^*_{j(2)}$

$$\hat{\eta}_B = \frac{1}{B}\sum_{j=1}^{B} \hat{D}_j$$

$$SE_{\hat{\eta}_B} = \sqrt{\frac{1}{B-1}\sum_{j=1}^{B}(\hat{D}_j - \hat{\eta}_B)^2}$$

โดย B คือ จำนวนครั้งการทำซ้ำแบบบูตสแตรป

$\hat{\eta}_B$ คือ ค่าประมาณผลต่างของค่าประมาณชุดใหม่ของประชากรแต่ละกลุ่ม

$\hat{D}_j$ คือ ผลต่างของค่าประมาณชุดใหม่ของประชากรแต่ละกลุ่ม

$\bar{X}^*_{j(1)}, \bar{X}^*_{j(2)}$ คือ ค่าประมาณชุดใหม่ของประชากรกลุ่มที่ 1 และ 2 ที่ได้จากการสุ่มแบบแทนที่ $= \sum_{i=1}^{n} \frac{X_i^*}{n}$; i คือ ตัวอย่างบูตสแตรปชุดที่ i

จากทฤษฎีบทขีดจำกัดส่วนกลาง (Central Limit Theorem)

สถิติทดสอบ $Z_{cal} \approx \frac{\hat{\eta}_B}{SE_{\hat{\eta}_B}}$

อาณาเขตวิกฤต : $Z_{cal} > Z_{\frac{\alpha}{2}}$ หรือ $Z_{cal} < -Z_{\frac{\alpha}{2}}$

2.2.5 การทดสอบแวน-เดอร์ วาเดนท์ (Van der Waerden)

(Songthong, 2014)

การทดสอบแวน-เดอร์ วาเดนท์ เป็นตัวสถิติ ทดสอบที่ถูกค้นพบโดยนักคณิตศาสตร์ชาวดัตถ์ ชื่อว่า Bartel Leendert van der Waerden ในปี 1952 ใช้ทดสอบความแตกต่างระหว่างค่ากลาง 2 กลุ่ม

สถิติทดสอบ $T_1 = \frac{1}{S^2}\sum_{i=1}^{k} n_i \bar{A}_i^2$

เมื่อ $A_{ij} = \phi^{-1}\left[\frac{R(X_{ij})}{N+1}\right]$ คือ คะแนนมาตรฐานตัวที่ j กลุ่มที่ i เมื่อ $N = n_1 + n_2$

$\bar{A}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} A_{ij}$, $i = 1, 2, \ldots, k$ คือ ค่าเฉลี่ยของคะแนนมาตรฐานกลุ่มที่ i

$S^2 = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i} A_{ij}^2$ คือ ค่าความแปรปรวนของคะแนนมาตรฐาน

อาณาเขตวิกฤต : $T_1 > \chi^2_{\alpha,1}$ โดย $\chi^2_{\alpha,1}$ คือค่าวิกฤตที่เปิดได้จากตารางแจกแจงไคกำลังสอง

2.3 ขั้นตอนการดำเนินงาน

2.3.1 จำลองข้อมูลและสุ่มตัวอย่างที่ใช้ในการวิจัยด้วยโปรแกรม R กำหนดค่าเฉลี่ยหรือค่ากลางเท่ากันและค่าความแปรปรวนไม่เท่ากัน

2.3.2 ทำการทดสอบเปรียบเทียบค่าเฉลี่ยหรือค่ากลางทั้ง 5 การทดสอบ และบันทึกจำนวนครั้งที่ปฏิเสธสมมติฐานว่าง ทำซ้ำจนครบ 1,000 ครั้ง

2.3.3 หาค่าประมาณความผิดพลาดแบบที่ 1 โดยนำจำนวนครั้งที่ปฏิเสธสมมติฐานหลักหารด้วย 1,000

2.3.4 เปรียบเทียบค่าประมาณความผิดพลาดแบบที่ 1 ของแต่ละการทดสอบกับเกณฑ์ของ Bradlay โดยการทดสอบจะสามารถควบคุมค่าประมาณความผิดพลาดแบบที่ 1 ได้ ก็ต่อเมื่อ α มีค่าอยู่ในช่วง (0.5α ,0.15α) ในการศึกษาครั้งนี้กำหนดระดับนัยสำคัญ (α) 2 ระดับคือ 0.01 และ 0.05 ดังนั้นช่วงค่าประมาณความผิดพลาดแบบที่ 1 คือ [0.005,0.015] และ [0.025,0.075] ตามลำดับ ถ้าค่าประมาณความผิดพลาดแบบที่ 1 จากการทดลองอยู่ในช่วงที่กำหนด จะสรุปได้ว่าการทดสอบทั้งหมดนั้นสามารถควบคุมความผิดพลาดแบบที่ 1 ได้

2.3.5 จำลองข้อมูลและสุ่มตัวอย่างที่ใช้ในการวิจัยด้วยโปรแกรม R กำหนดค่าเฉลี่ยหรือค่ากลางไม่เท่ากันและค่าความแปรปรวนไม่เท่ากัน ทำการทดสอบเปรียบเทียบค่าเฉลี่ยหรือค่ากลางทั้ง 5 การ

ทดสอบ เพื่อหาค่าประมาณกำลังการทดสอบ โดยนับจำนวนครั้งที่ปฏิเสธสมมติฐานหลักหารด้วย 1,000

2.3.6 เปรียบเทียบค่าประมาณกำลังการทดสอบที่สามารถควบคุมค่าประมาณความผิดพลาดแบบที่ 1 ได้ โดยการทดสอบที่มีค่าประมาณกำลังการทดสอบที่สูงที่สุด จะถือว่าเป็นการทดสอบที่มีประสิทธิภาพดีที่สุด เมื่อเทียบกับการการทดสอบของการศึกษาครั้งนี้

3. ผลการวิจัย

ผลการวิจัยสรุปได้ 2 ส่วนใหญ่ๆ คือ ค่าประมาณความผิดพลาดแบบที่ 1 ในตารางที่ 1และ 2 และสถิติทดสอบที่ให้ค่าประมาณกำลังการทดสอบในตารางที่ 3 และ 4 โดยกำหนดให้

T หมายถึง การทดสอบที

Z หมายถึง การทดสอบซี

MWU หมายถึง การทดสอบแมน์-วิทนี ยู

BT หมายถึง วิธีบูตสแตรป

VW หมายถึง การทดสอบแวน-เดอร์ วาเดนท์

3.1 ค่าประมาณความผิดพลาดแบบที่ 1

ประชากรที่มีการแจกแจงปรกติปลอมปนที่มีสัดส่วนการปลอมปน (p) เท่ากับ 0.1 เมื่อความแปรปรวนไม่เท่ากัน ที่ระดับนัยสำคัญ 0.01 โดยพิจารณาความสามารถในการควบคุมค่าประมาณความผิดพลาดแบบที่ 1 ตามเกณฑ์ของ Bradley สรุปผลได้ดังตารางที่ 1 และ 2

ตารางที่ 1: ค่าประมาณความผิดพลาดแบบที่ 1 ที่ระดับนัยสำคัญ 0.01

ความแปรปรวน $(\sigma_1^2:\sigma_2^2)$	c	การทดสอบ	ขนาดตัวอย่าง (n_1,n_2)			
			เท่ากัน		ไม่เท่ากัน	
			(20,20)	(50,50)	(20,25)	(50,70)
4:6.2	5	T	0.002	0.001	0.002	0.002
		Z	0.007*	0.011*	0.009*	0.01*
		MWU	0.003	0.004	0.002	0.002
		BT	0.007*	0.013*	0.013*	0.009*
		VW	0.007*	0.013*	0.009*	0.005*
	10	T	0.001	0.001	0.001	0.004
		Z	0.004	0.007*	0.004	0.005*
		MWU	0.008*	0.006*	0.005*	0.008*
		BT	0.008*	0.007*	0.003	0.006*
		VW	0.013*	0.009*	0.008*	0.011*
ความแปรปรวน $(\sigma_1^2:\sigma_2^2)$	c	การทดสอบ	ขนาดตัวอย่าง (n_1,n_2)			
			เท่ากัน		ไม่เท่ากัน	
			(20,20)	(50,50)	(20,25)	(50,70)
4:13.48	5	T	0.001	0.003	0.001	0.002
		Z	0.009*	0.01*	0.009*	0.007*
		MWU	0.004	0.001	0.006*	0.002
		BT	0.052	0.01*	0.011*	0.007*
		VW	0	0.002	0.007*	0.002
	10	T	0	0.003	0	0.002
		Z	0.001	0.008*	0.002	0.005*
		MWU	0.007*	0.008*	0.005*	0.001
		BT	0.004	0.008*	0.003	0.006*
		VW	0.011*	0.01*	0.007*	0.008*

หมายเหตุ * หมายถึง สามารถการควบคุมความผิดพลาดแบบที่ 1 ได้ ตามเกณฑ์ของ Bradley

จากตารางที่ 1 จะได้สถิติทดสอบที่สามารถในการควบคุมค่าประมาณความผิดพลาดแบบที่ 1 โดยแบ่งเป็น สถิติอิงพารามิเตอร์ คือ การทดสอบซี สำหรับสถิติไม่อิงพารามิเตอร์ คือ การทดสอบแวน-เดอร์ วาเดนท์ สูงสุด รองลงมาคือ วิธีบูตสแตรปและการทดสอบแมนท์-วิทนี ยู

ตารางที่ 2 : ค่าประมาณความผิดพลาดแบบที่ 1 ที่ระดับนัยสำคัญ 0.05

ความแปรปรวน $(\sigma_1^2:\sigma_2^2)$	c	การทดสอบ	ขนาดตัวอย่าง (n_1,n_2)			
			เท่ากัน		ไม่เท่ากัน	
			(20,20)	(50,50)	(20,25)	(50,70)
4:6.2	5	T	0.01	0.014	0.013	0.021
		Z	0.041*	0.046*	0.046*	0.045*
		MWU	0.021	0.023	0.018	0.021
		BT	0.05*	0.059*	0.051*	0.044*
		VW	0.049*	0.044*	0.049*	0.044*
	10	T	0.008	0.016	0.008	0.017
		Z	0.032*	0.043*	0.037*	0.048*
		MWU	0.017	0.017	0.017	0.023
		BT	0.04*	0.044*	0.041*	0.049*
		VW	0.044*	0.037*	0.037*	0.041
4:13.48	5	T	0.014	0.019	0.011	0.025*
		Z	0.04*	0.044*	0.047*	0.049*
		MWU	0.025*	0.029*	0.024	0.027*
		BT	0.053*	0.048*	0.052*	0.051*
		VW	0.04*	0.052*	0.037*	0.044*
	10	T	0.006	0.016	0.011	0.01
		Z	0.037*	0.036*	0.039*	0.036*
		MWU	0.023	0.037*	0.02	0.015
		BT	0.04*	0.039*	0.041*	0.04*
		VW	0.035*	0.052*	0.047*	0.034*

หมายเหตุ * หมายถึง สามารถควบคุมความผิดพลาดแบบที่ 1 ได้ ตามเกณฑ์ของ Bradley

จากตารางที่ 2 จะได้สถิติทดสอบที่สามารถในการควบคุมค่าประมาณความผิดพลาดแบบที่ 1 โดยแบ่งเป็น สถิติอิงพารามิเตอร์ คือ การทดสอบซี สำหรับสถิติไม่อิงพารามิเตอร์ คือ การทดสอบแวน-เดอร์ วาเดนท์ วิธีบูตสแตรป

3.2 ค่าประมาณกำลังการทดสอบ

ประชากรที่มีการแจกแจงปรกติปลอมปนที่มีสัดส่วนการปลอมปน (p) เท่ากับ 0.1 เมื่อความแปรปรวนไม่เท่ากัน ที่ระดับนัยสำคัญ 0.01 และ 0.05 ตัวสถิติที่ให้ค่าประมาณกำลังการทดสอบสูงสุดของแต่ละการทดสอบจะแสดงในกรณีที่การทดสอบสามารถควบคุมค่าประมาณความผิดพลาดแบบที่ 1 ได้เท่านั้น

ตารางที่ 3 : ค่าประมาณกำลังการทดสอบ ที่ระดับนัยสำคัญ 0.01

ความแปรปรวน $(\sigma_1^2 : \sigma_2^2)$	c	การทดสอบ	ขนาดตัวอย่าง (n_1, n_2)			
			เท่ากัน		ไม่เท่ากัน	
			(20,20)	(50,50)	(20,25)	(50,70)
4:6.2	5	T	-	-	-	-
		Z	0.217	0.482	0.259	0.565
		MWU	-	-	-	-
		BT	0.229	0.493	0.273	0.574
		VW	0.298*	0.787*	0.347*	0.835*
	10	T	-	-	-	-
		Z	-	0.162	-	0..184
		MWU	0.253	0.378	0.295	0.812*
		BT	0.132	0.165	-	0.0189
		VW	0.285*	0.719*	0.309*	0.792
4:13.48	5	T	-	-	-	-
		Z	0.136	0.299	0.148	0.337
		MWU	-	-	0.136	-
		BT	0.149*	0.31*	0.164*	0.342*
		VW	0.146	-	0.155	-
	10	T	-	-	-	-
		Z	-	0.114	-	0.089
		MWU	0.138	0.446*	0.154	-
		BT	-	0.121	-	0.091
		VW	0.148*	0.424	0.164*	0.473*

หมายเหตุ - หมายถึงไม่พิจารณากำลังการทดสอบเนื่องจากไม่สามารถควบคุมค่าประมาณความผิดพลาดแบบที่ 1

* หมายถึง ค่าประมาณกำลังการทดสอบสูงที่สุดในสถานการณ์นั้น

จากตารางที่ 3 การทดสอบที่มีค่าประมาณกำลังทดสอบสูงที่สุดกรณีข้อมูลสุ่มจากประชากรที่มีการแจกแจงปรกติปลอมปนที่ระดับนัยสำคัญ 0.01 ส่วนใหญ่ คือ การทดสอบแวน-เดอร์ วาเดนท์ รองลงมาคือ วิธีบูตสแตรป

ตารางที่ 4 : ค่าประมาณกำลังการทดสอบ ที่ระดับนัยสำคัญ 0.05

ความแปรปรวน $(\sigma_1^2 : \sigma_2^2)$	c	การทดสอบ	ขนาดตัวอย่าง (n_1, n_2)			
			เท่ากัน		ไม่เท่ากัน	
			(20,20)	(50,50)	(20,25)	(50,70)
4:6.2	5	T	-	-	-	-
		Z	0.443	0.692	0.44	0.771
		MWU	-	-	-	-
		BT	0.457	0.695	0.448	0.776
		VW	0.563*	0.992*	0.609*	0.958*
	10	T	-	-	-	-
		Z	0.025	0.338	0.249	0.393
		MWU	-	-	-	-
		BT	0.267	0.341	0.268	0.399
		VW	0.535*	0.886*	0.557*	0.94*
4:13.48	5	T	-	-	-	0.463
		Z	0.297	0.507	0.322	0.573
		MWU	0.334	0.695	-	0.785
		BT	0.306	0.52	0.329	0.575
		VW	0.388*	0.721*	0.406*	0.787*
	10	T	-	-	-	-
		Z	0.179	0.238	0.191	0.258
		MWU	-	0.642	-	-
		BT	0.19	0.245	0.201	0.263
		VW	0.359*	0.661*	0.353*	0.748*

หมายเหตุ - หมายถึงไม่พิจารณากำลังการทดสอบเนื่องจากไม่สามารถควบคุมค่าประมาณความผิดพลาดแบบที่ 1

* หมายถึง ค่าประมาณกำลังการทดสอบสูงที่สุดในสถานการณ์นั้น

จากตารางที่ 4 การทดสอบที่มีค่าประมาณกำลังการทดสอบสูงที่สุดกรณีข้อมูลสุ่มจากประชากรที่มีการแจกแจงปรกติปลอมปนที่ระดับนัยสำคัญ 0.05 การทดสอบแวน-เดอร์ วาเดนท์ให้ค่าประมาณกำลังการทดสอบสูงสุดในทุกกรณี

4. สรุปผลการวิจัย

จากผลการวิเคราะห์ความสามารถในการควบคุมค่าประมาณความผิดพลาดแบบที่ 1 และค่าประมาณกำลังการทดสอบ ของการทดสอบ 5 ตัว ได้แก่ การทดสอบที การทดสอบซี การทดสอบแมนท์-วิทนี ยู วิธีบูตสแตรป และการทดสอบแวน-เดอร์ วาเดนท์ จากการจำลองข้อมูลตามลักษณะต่างๆ ที่ได้กำหนดไว้ในขอบเขตการวิจัย ข้อมูลสุ่มมาจากประชากรที่มีการแจกแจงปรกติปลอมปน ตัวสถิติที่มีค่าประมาณกำลังการทดสอบสูงที่สุด คือสถิติไม่อิงพารามิเตอร์ ได้แก่ การทดสอบแวน-เดอร์ วาเดนท์ มากที่สุดเกือบทุกกรณี รองมาคือวิธีบูตสแตรป และการทดสอบแมนท์-วิทนี ยู ซึ่งผลการวิจัยสอดคล้องกับงานวิจัยของ Bakker (2014) ซึ่งการทดสอบทีมีประสิทธิภาพต่ำที่สุดเมื่อเทียบกับการทดสอบ แมนท์-วิทนี ยู

ดั้งนั้นในกรณีที่ต้องการทดสอบสมมติฐานของค่าเฉลี่ยและค่ากลางของ 2 ประชากร ถ้าทราบว่าข้อมูลมีค่านอกกลุ่ม ควรใช้สถิติไม่อิงพารามิเตอร์ เช่น การทดสอบแวน-เดอร์ วาเดนท์ ในการทดสอบสมมติฐานจะทำให้การทดสอบมีความถูกต้องมากขึ้น

เอกสารอ้างอิง

Bakker, M. (2014). Outlier removal, sum scores, and the inflation of the type i error rate in independent sample t test: The power of alternatives and recommendations. *Psychol Methods*, (19)3, 409-427.

Boonpen, S., Chomtee, B., & Hitunwong, A. (2015). A comparison of intervals estimation methods for scale parameter of the two-parameter weibull distribution, *Thai Science and Technology Journal, 23(4)*, 579-587 (in Thai).

Janthasorn, U. (1998). Nonparametrics. Department of statistics, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang (in Thai).

Kuharatanachai, C. (2013). Introduction to Statistics, Department of statistics, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang (in Thai).

Songthong, M. (2013). A study of efficiency of parametric and nonparametric statistics in testing of central difference between two populations. *KKU Science Journal*, 41(1), 226-238 (in Thai).

Songthong, M. (2014). Robustness and power of the test of parametric and nonparametric statistics in testing of central difference between two populations for Likert-type data 5 point, *Thai Science and Technology Journal*, 22(5), 605-609 (in Thai).

Joint Monitoring Mean and Variability Using Adaptive Kalman Filter

Pairoj Khawsithiwong

Department of Statistics, Faculty of Science, Silpakorn University, Nakorn Phatom, Thailand
Corresponding Email: khawsithiwong_p@su.ac.th

ABSTRACT

For detecting changes in the mean vector and the covariance matrix of an autocorrelated process simultaneously, the multivariate control chart based on the adaptive Kalman filter is proposed. Since the adaptive Kalman filter provides the precise mean estimates. Therefore, such the mean estimates and the corresponding residuals can be applied to monitor the process means and the process variability, respectively. To construct the proposed control chart, a control function is obtained to combine two terms of effects of process mean and variability changes, i.e. respectively, one is the distance between the current mean estimates and the in-control means and the other one is the Gaussian quantile function of the chi-square cumulative probability of the standardized distance of the residuals. A performance study shows that, in most cases, the average run length of the proposed control chart is smaller than the combined charting scheme of MEWMA and MEWMC control charts and the combined charting scheme of MEWMA and EWMAD control charts.

Keywords: single variable control chart; adaptive filter; autocorrelated process

1 INTRODUCTION

An effectiveness of a traditional control chart for detecting parameter changes in a multivariate autocorrelated process is absolutely influenced by the autocorrelation. Such the disadvantage should be alleviate or eliminated on any ways or taken into account. One approach aims to fit an exact process time series model correctly to obtain one step ahead forecast errors or residuals which are independent. Those residuals can be applied to entire traditional control charts directly. The other approach is to modify control limits of the traditional control charts by taking an effect of an autocorrelation into account when observations are demanded. See Woodall and Montgomery (2014) for more details.

In practice, both the process mean and variability could be affected by special causes, simultaneously. Additionally, when a process is out of control, the estimation of means and variances impacts on each other as shown in Figure 1, i.e. when the process mean changes (Figure 1(a)), the EWMV control chart using to monitor a change in variance is affected significantly as in Figure 1(b). For inexperience practitioners, it is rational to combine the information of process mean and dispersion changes in one function. The single variable control chart is plausibly constructed to monitor both changes in a process at a same time. In present, many single variable control charts have been proposed for both cases of univariate and multivariate. See Cheng and Thaga (2006), McCracken and Chakraborti (2013).

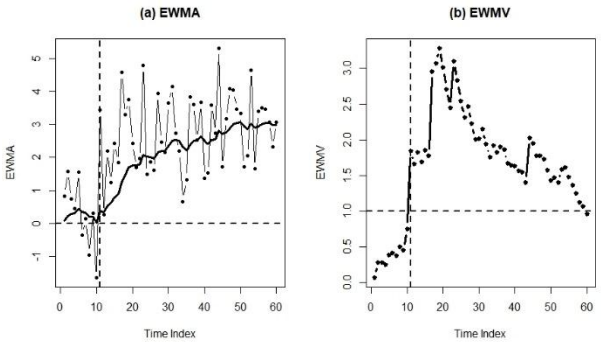


Figure 1: An effect of process mean change to variance

In this work, the single variable control chart is developed to monitor both process means and variability simultaneously in a multivariate autocorrelated process. Mean estimates from the adaptive Kalman filter are obtained to approximate current process means precisely. This can imply that process mean changes should be less affected to variance estimates. Thus, the proposed control chart could perform favorably in such those situations. A simulation study is conducted with an average run length (ARL) as a preferred criterion.

2 ADAPTIVE KALMAN FILTER

Khawsithiwong et al. (2011) developed the adaptive Kalman filter, called AMGLF, to deal with a problem of linear estimation and prediction for a linear discrete-time stochastic process which possesses properties of observability and reachability. This process is represented by the state space model consisting of the state equation and the measurement equation, respectively, given by

$$\underline{X}_{t+1} = A_t\underline{X}_t + \underline{W}_t \tag{1}$$

$$\underline{Y}_t = C_t\underline{X}_t + \underline{V}_t \tag{2}$$

where $\underline{Y}_t$ is a $k \times 1$ vector of measurements, $\underline{X}_t$ is an $r \times 1$ state vector, A_t is an $r \times r$ transition matrix and C_t is a $k \times r$ measurement matrix. A $r \times 1$ system noise term $\underline{W}_t$ is assumed to be distributed as an zero mean multivariate generalized Laplace (MGL) random vector with a scale parameter Σ_W and a shape parameter λ_W , denoted by $\underline{W}_t \sim MGL_r(\underline{0},\ \Sigma_W,\ \lambda_W)$, and a $k \times 1$ measurement noise term $\underline{V}_t$ is a Gaussian random vector denoted by $\underline{V}_t \sim MGL_k(\underline{0},\ \Sigma_V,\ 2)$. In addition, the scale parameter matrices Σ_W and Σ_V are assumed to be known positive definite matrices.

The MGL distribution was introduced by Ernst (1998) as a class of symmetric multivariate probability models depending on the value of a shape parameter. The shape parameter λ distinguishes between members of the family such as the multivariate Laplace $(\lambda = 1)$, the multivariate normal $(\lambda = 2)$, and the multivariate uniform $(\lambda \to \infty)$ distribution. Suppose a $k \times 1$ random vector $\underline{Z}$ has an MGL distribution denoted by $\underline{Z} \sim MGL_k\left(\underline{\mu},\ \Sigma,\ \lambda\right)$ with the joint density function defined as

$$f(\underline{Z}) = \frac{\lambda\Gamma\left(\frac{k}{2}\right)}{2\pi^{\frac{k}{2}}\Gamma\left(\frac{k}{\lambda}\right)}|\Sigma|^{-\frac{1}{2}}exp\left\{-\left[\left(\underline{Z}-\underline{\mu}\right)'\Sigma^{-1}\left(\underline{Z}-\underline{\mu}\right)\right]^{\frac{1}{2}}\right\} \tag{3}$$

where $\Gamma(\cdot)$ denotes a gamma function and $|\cdot|$ is an absolute value. The mean vector and the covariance matrix of $\underline{Z}$ are respectively given by

$$E(\underline{Z}) = \underline{\mu} \tag{4}$$

And

$$COV(\underline{Z}) = \frac{\Gamma\left(\frac{k+2}{\lambda}\right)}{k\Gamma\left(\frac{k}{\lambda}\right)}\Sigma \tag{5}$$

By means of the least squares approach, an unbiased minimum variance state estimate and its covariance matrix can be obtained

$$\underline{\hat{X}}_{t|t} = \underline{\hat{X}}_{t|t-1} + K_t\left(\underline{Y}_t - C_t\underline{\hat{X}}_{t|t-1}\right) \quad (6)$$

and $$P_{t|t} = (I - K_tC_t)P_{t|t-1} \quad (7)$$

where $$\underline{\hat{X}}_{t|t-1} = A_{t-1}\underline{\hat{X}}_{t-1|t-1} \quad (8)$$

$$P_{t|t-1} = A_{t-1}P_{t-1|t-1}A'_{t-1} + \frac{\Gamma\left(\frac{r+2}{\lambda_W}\right)}{r\Gamma\left(\frac{r}{\lambda_W}\right)}\Sigma_W \quad (9)$$

and $$K_t = P_{t|t-1}C'_t\left(C_tP_{t|t-1}C'_t + \frac{1}{2}\Sigma_V\right)^{-1} \quad (10)$$

To implement the AMGLF filter, the shape parameter of the system noise term λ_W in (9) should be estimated at each time point. A pseudo state estimation error is defined as $\underline{Z}_t = \underline{\hat{X}}_{t|t} - \underline{\hat{X}}_{t|t-1} = K_t\left(\underline{Y}_t - C_t\underline{\hat{X}}_{t|t-1}\right)$ which is assumed to be an MGL distributed random vector with a scale parameter matrix Σ_Z and a time-varying shape parameter $\lambda_{Z(t)}$, denoted by $\underline{Z}_t \sim MGL_r(\underline{0},\ \Sigma_Z,\ \lambda_{Z(t)})$. The approximate value of the time-varying shape parameter $\hat{\lambda}_{Z(t)}$ is obtained by maximizing a likelihood function of the pseudo state estimation error with the scale parameter matrix $\Sigma_Z = 2\delta_t^2K_t\left(C_tP_{t|t-1}C'_t + \frac{1}{2}\Sigma_V\right)K'_t$ where $\delta_t = \frac{\hat{\lambda}_{Z(t-1)}}{2}$ is a time-varying adaptive factor. Then the one-step ahead state estimate $\underline{\hat{X}}_{t+1|t}$ and the state forecast error covariance matrix $P_{t+1|t}$ are given

$$\underline{\hat{X}}_{t+1|t} = A_t\underline{\hat{X}}_{t|t} \quad (11)$$

And $$P_{t+1|t} = A_tP_{t|t}A'_t + \frac{\Gamma\left(\frac{r+2}{\hat{\lambda}_{Z(t)}}\right)}{r\Gamma\left(\frac{r}{\hat{\lambda}_{Z(t)}}\right)}\Sigma_W \quad (12)$$

Figure 2 shows a performance of the AMGLF filter when a process mean shifted. It indicates that the mean estimates of the AMGLF filter can describe the process accurately. Also, the effect of mean shift does not affect to the residuals significantly.

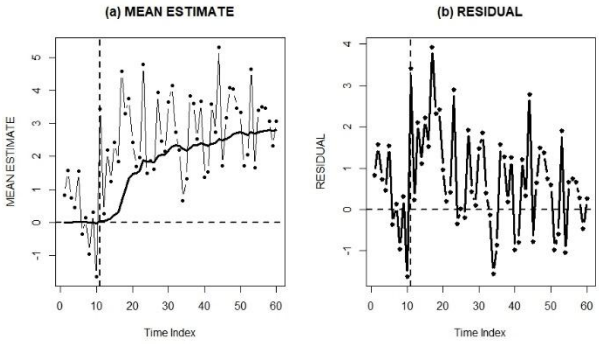


Figure 2: Mean estimates and residuals of the AMGLF filter

3 PROPOSED CONTROL CHART

To construct the proposed single variable control chart based on AMGLF filter, named SAMGL control chart, the residual $\underline{e}_t = \underline{Y}_t - \underline{\overline{Y}}_t$ is obtained to detect changes in process variability where the mean estimate $\underline{\overline{Y}}_t = C_t\underline{\hat{X}}_{t|t}$. Suppose a process is in control, the residual $\underline{e}_t$ is distributed as Gaussian random vector, i.e. $\underline{e}_t \sim N_k\left(\underline{0},\ 2\left(C_tP_{t|t}C'_t + \frac{1}{2}\Sigma_V\right)\right)$. The standardized distance of $\underline{e}_t$ is defined

$$d_t = \frac{1}{2}\left(\underline{Y}_t - \underline{\overline{Y}}_t\right)'\left(C_tP_{t|t}C'_t + \frac{1}{2}\Sigma_V\right)^{-1}\left(\underline{Y}_t - \underline{\overline{Y}}_t\right) \quad (13)$$

where the distance d_t is distributed as a chi square random variable with k degrees of freedom, denoted by $d_t \sim \chi^2_k$. For detecting an effect of process mean shifts, a quantity b_t is obtained

$$b_t = \frac{1}{2}\left(\underline{\overline{Y}}_t - \underline{\mu}_0\right)'\left(C_tP_{t|t}C'_t + \frac{1}{2}\Sigma_V\right)^{-1}\left(\underline{\overline{Y}}_t - \underline{\mu}_0\right) \quad (14)$$

where $\underline{\mu}_0$ is the mean vector of the in-control process. A quantity b_t trends to zero when a process is in control.

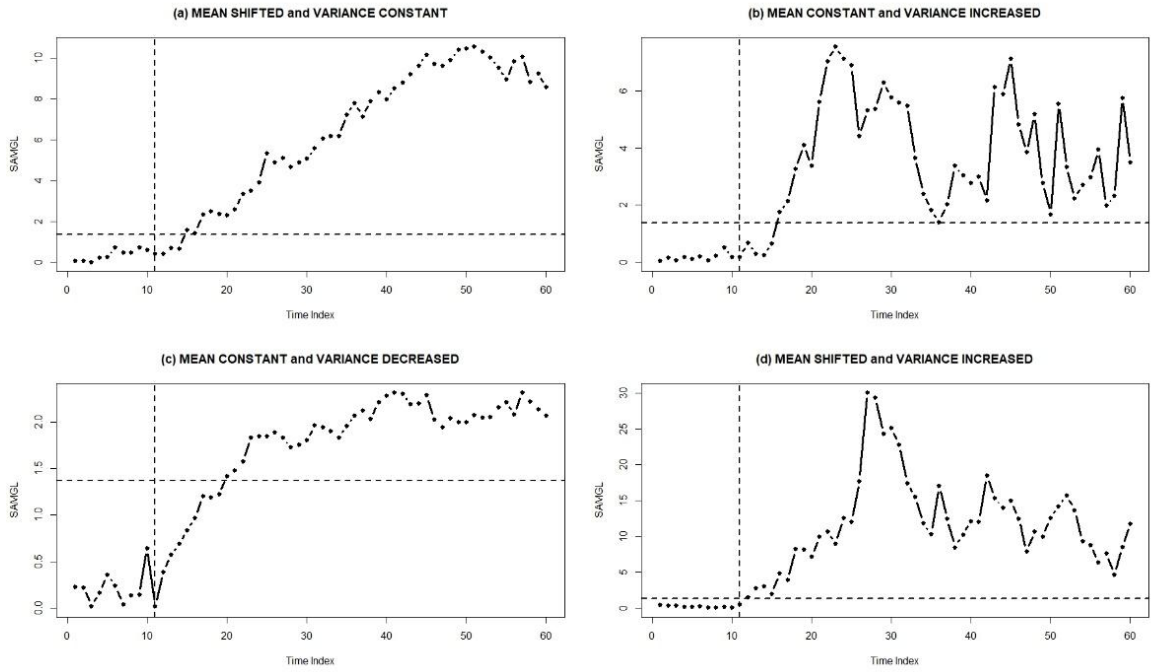


Figure 3: Performance of the SAMGL control chart when process means and dispersion changed; (a) process means shifted, (b) process variance increased, (c) process variance decreased, and (d) process means shifted and process variance increased

Table 1: ARL values of the SAMGL chart, (MEWMA, MEWMC) scheme, and (MEWMA, EWMAD) scheme

		δ_v						
	δ_m	0.2	0.6	0.8	1	1.2	1.4	2
SAMGL	0	210.79	425.89	443.26	**249.75**	**82.13**	**25.73**	3.17
	0.25	93.21	120.71	91.81	**51.68**	**24.45**	**11.60**	2.75
	0.50	13.40	12.97	**10.80**	**7.95**	**5.63**	**4.16**	**2.08**
	0.75	**3.53**	**3.44**	**3.24**	**2.94**	**2.62**	**2.32**	**1.67**
	1	**2.12**	**2.09**	**2.02**	**1.93**	**1.82**	**1.71**	**1.44**
	2	**1.18**	**1.18**	**1.18**	**1.18**	**1.17**	**1.17**	1.14
	3	1.06	1.06	1.06	**1.06**	**1.06**	**1.06**	1.06
(MEWMA, MEWMC)	0	101.05	426.62	398.09	250.50	98.54	28.29	**2.78**
Based on Kalman filter	0.25	87.05	114.21	91.44	65.60	39.71	17.70	**2.61**
	0.50	12.28	11.89	11.21	10.13	8.73	6.61	2.18
	0.75	3.89	3.93	3.94	3.87	3.71	3.27	1.78
	1	2.62	2.63	2.62	2.56	2.40	2.17	1.49
	2	1.21	1.19	1.18	1.17	1.16	1.14	**1.10**
	3	**1.02**	**1.02**	**1.02**	1.02	1.03	1.03	**1.03**
(MEWMA, EWMAD)	0	**70.74**	**131.30**	**225.84**	247.52	135.80	42.55	3.55
Based on Kalman filter	0.25	**59.12**	**82.05**	**82.29**	66.68	45.68	22.92	3.26
	0.50	**12.00**	**11.73**	11.17	10.26	9.12	7.38	2.58
	0.75	3.91	3.94	3.95	3.91	3.83	3.50	2.01
	1	2.63	2.64	2.63	2.60	2.49	2.31	1.63
	2	1.31	1.29	1.26	1.24	1.21	1.19	1.13
	3	1.03	1.03	1.03	1.03	1.03	1.03	1.04
(MEWMA, MEWMC)	0	93.01	649.35	591.02	249.75	73.25	20.92	2.43
Based on AMGLF filter	0.25	128.01	237.87	123.64	55.27	25.07	10.85	2.22
	0.50	19.58	15.62	11.22	7.82	5.55	4.07	1.80
	0.75	3.55	3.39	3.16	2.87	2.58	2.28	1.51
	1	2.08	2.06	2.00	1.93	1.81	1.68	1.32
	2	1.17	1.16	1.15	1.14	1.13	1.12	1.08
	3	1.02	1.02	1.02	1.02	1.02	1.03	1.03
(MEWMA, EWMAD)	0	74.05	146.28	273.82	250.00	89.70	26.68	2.89
Based on AMGLF filter	0.25	77.76	123.58	105.35	55.49	25.73	12.25	2.57
	0.50	18.35	15.14	11.14	7.87	5.65	4.26	1.98
	0.75	3.55	3.39	3.16	2.88	2.62	2.34	1.60
	1	2.08	2.06	2.01	1.94	1.84	1.73	1.38
	2	1.21	1.20	1.19	1.17	1.16	1.14	1.10
	3	1.03	1.03	1.03	1.03	1.03	1.03	1.03

To enhance the sensitivity of the SAMGL chart, the exponentially weighted moving average statistic is required

$$s_t = \alpha q_t + (1-\alpha)s_{t-1}, \quad 0 < \alpha < 1 \tag{15}$$

where α is a smoothing parameter and q_t is a Gaussian quantile function of chi-square cumulative probability of the distance d_t defined as

$$q_t = \Phi^{-1}\big(H_k(d_t)\big) \tag{16}$$

where $\Phi^{-1}(\cdot)$ is an inverse standard Gaussian distribution function and $H_k(\cdot)$ is a chi-square distribution function with k degrees of freedom. This realizes that the quantile q_t has a standard normal distribution. Further, s_t has also a normal distribution with zero mean and the variance $\sigma_s^2 = \frac{\alpha}{2-\alpha}$.

The control function consisting of both effects of process mean and variability changes is given

$$SAMGL_t = |s_t| + \eta b_t, \quad \eta = \tfrac{2-\alpha}{\alpha} \tag{17}$$

where $|s_t|$ is distributed as a half normal random variable with mean $E(|s_t|) \le \sqrt{\frac{2}{\pi}}$ and the variance $V(|s_t|) \le \left(1-\frac{2}{\pi}\right)\left(\frac{\alpha}{2-\alpha}\right)$. When a process is in control, the expected value and the variance of $SAMGL_t$ are $E(SAMGL_t) = E(|s_t|)$ and $V(SAMGL_t) = V(|s_t|)$, respectively.

Eventually, the SAMGL control chart is established with the asymptotic control limit as

$$UCL = E(SAMGL_t) + \gamma\sqrt{V(SAMGL_t)} = \sqrt{\frac{2}{\pi}} + \gamma\sqrt{\left(1-\frac{2}{\pi}\right)\left(\frac{\alpha}{2-\alpha}\right)} \tag{18}$$

$$cl = E(SAMGL_t) = \sqrt{\frac{2}{\pi}} \tag{19}$$

where the constant γ is evaluated to satisfy a preferred specific in-control ARL. UCL and cl are named respectively the upper control limit and the central line. Figure 3 shows a performance of the SAMGL control chart for various situations of process changes. This indicates that the SAMGL chart can detect all changes in process means and process dispersion. The SAMGL chart seem to be well-mannered when process means shift and/or a process dispersion increases as shown in Figure 3(a), 3(b), 3(d). Figure 3(c) shows that the power of the SAMGL chart is seemingly deteriorated when a process dispersion decreases.

4 PERFORMANCE STUDY

Consider the 5-variate autocorrelated process represented by (1) and (2) where matrices A_t and C_t are the identity matrices, I. The system noise term $\underline{W_t}$ and the measurement noise term $\underline{V_t}$ are independently distributed as Gaussian random vector with zero mean vectors and covariance matrices Σ_W and Σ_V, respectively. The structure of both system and measurement noise covariance matrices is based on $\Sigma = \sigma^2(1-\rho)I + \sigma^2\rho J$ where J is a matrix of one and ρ is a correlation coefficient of the noise term. All values of the covariance structure are set to the system noise variance $\sigma_W^2 = 0.0001$ and the measurement noise variance $\sigma_V^2 = 1$ as well as correlation coefficients of the system and measurement noise terms $\rho_W = \rho_V = 0.4$.

In the investigation, process changes are set to only the first variable at time 11 of the time series of length 60. A mean shift is generated by means of the first variate system noise term $W_{1,t=11}$ with mean $\delta_m\sigma_V$ where a magnitude of mean shift $\delta_m = 0,\ 0.25,\ 0.50,\ 0.75,\ 1,\ 2,$ and 3. Also, a variance change is set to the first variate measurement noise term $V_{1,t=11}$ with variance $\delta_v\sigma_V$ where a magnitude of variance change $\delta_v = 0.2,\ 0.6,\ 0.8,\ 1,\ 1.2,\ 1.4,$ and 2.

A Monte Carlo simulation consisting of 5,000 iterations is conducted to compare a performance of the SAMGL chart with those of combined charting schemes based on the traditional Kalman filter and the AMGLF filter, i.e. one scheme is a combination of the multivariate exponentially weighted moving average (MEWMA) chart (Lowry et al., 1992) and the multivariate exponentially weighted moving covariance matrix (MEWMC) chart (Hawkins and Maboudou-Tchao, 2008) and the other scheme is a combination of the MEWMA chart and the exponentially weighted moving average of squared distance (EWMAD) chart (Khawsithiwong & Yatawara, 2007). To accomplish the exponentially weighted moving average charting approach, the smoothing parameter set to be 0.2. Furthermore, the ARL is used in a comparison as a preferred criterion with the in-control ARL, $\text{ARL}_0 = 250$.

Results in Table 1 show that when the process variability is in control, ARL values of the SAMGL chart are smallest for all magnitudes of process mean shifts. In case of the in-control process mean, the SAMGL chart also obtains the least ARLs as the process variance increases. However, when the process variance decreases, the (MEWMA, EWMAD) charting scheme performs well with the minimal ARLs.

In situations of simultaneous process changes in both the process mean and the process variance, ARLs of the SAMGL chart are smallest when the process variance grows up. In contrast, when a process variance decreases, the (MEWMA, EWMAD) charting scheme performs well in case of small magnitudes of the process mean shift. However, when the process mean get larger, the SAMGL chart is able to detect changes faster than others.

When the AMGLF filter is applied to the combined charting schemes. Performances of both the combined charting schemes become similar to those of the SAMGL chart. However, the (MWMA, EWMAD) charting scheme obviously performs well as a decrease in the process variance. In addition, a masking effect of the process variance decrease causes the (MEWMA, EWMAD) charting scheme difficult to detect the process mean shift (Gan, 1995). Finally, the adaptive AMGLF filter can improve an effectiveness of the control charts since the process means can be estimated precisely.

5 SUMMARY

The SAMGL single variable control chart is developed to jointly monitor both process means and variability of a multivariate autocorrelated process. The AMGLF filter is demanded to estimate process means adaptively. Further, the control function is established to obtain both effects of process changes. Eventually, the SAMGL control chart is constructed and its performance is investigated through the simulation experiment. It results that, in most cases, a performance of the SAMGL chart is superior to the combined charting schemes based on Kalman filter excluding a situation of the process variance decrease and a situation of the process variance decrease and the small process mean shift. Additionally, the adaptive AMGLF filter can enhance a capability of the control charts. Ultimately, the SAMGL chart is a simplified way in making a decision whether a process is out of control.

ACKNOWLEDGEMENTS

The anonymous reviewers are acknowledged for their critical suggestions and helpful revisions.

REFERENCES

Cheng, S.W., & Thaga, K. (2006). Single variables control charts: an overview. *Quality and Reliability Engineering International*, 22(7), 811-820.

Ernst, M.D. (1998). A multivariate generalized Laplace distribution. *Computational Statistics*, 13(2), 227-232.

Gan, F.F. (1995). Joint monitoring of process mean and variance using exponentially weighted moving average control charts. *Technometrics*, 37(4), 446-453.

Hawkind, D.M. & Maboudou-Tchao, E.M. (2008). Multivariate exponentially weighted moving covariance matrix. *Technometrics*, 50(2), 155-166.

Khawsithiwong, P. & Yatawara, N. (2007). Monitoring process variability with individual measurements following elliptically contoured distributions. *Communications in Statistics—Simulation and Computation*, 36(3), 699-718.

Khawsithiwong, P., Yatawara, N., & Pongsapukdee, V. (2011). Adaptive Kalman filtering with multivariate generalized Laplace system noise. *Communications in Statistics—Simulation and Computation*, 40(9), 1278-1290.

Lowry, C.A., Woodall, W.H., Champ, C.W., & Rigdon, S.E. (1992). A multivariate exponentially weighted moving average control chart. *Technometrics*, 34(1), 46-53.

McCracken, A.K. & Chakraborti, S. (2013). Control charts for joint monitoring of mean and variance: an overview. *Quality Technology and Quantitative Management*, 10(1), 17-36.

Woodall, W.H. & Montgomery, D.C. (2014). Some current directions in the theory and application of statistical process monitoring. *Journal of Quality Technology*, 46(1), 78-94.

Interval Estimation for the Standard Deviation of the Lognormal Distribution

Prapassiri Moongprachachon, Prachya Thongtasee, Jintana Jitthaisong and Patarawan Sangnawakij*

Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Thailand
*Corresponding Email: patarawan.s@gmail.com

ABSTRACT

The objective of this study was to develop the interval estimator for the standard deviation of the lognormal distribution, based on the unbiased estimator. We conducted the efficiency of the proposed estimator with an existing interval estimator derived based on a biased estimator, using the coverage probability, average length, and relative bias via simulations. The results showed that the coverage probability of the proposed estimator was greater than that of the compared estimator. The average lengths of the two confidence intervals were identical. However, the relative bias of the proposed interval estimator was smaller. The methods were applied to a real example on the air pollution from a part of Bangkok.

Keywords: confidence interval; simulation study; Stirling's approximation; unbiased estimator

1 INTRODUCTION

The lognormal distribution is a probability model widely used in diverse disciplines, including industrial engineering, medical research, social sciences, and natural sciences. In statistical inference, the construction of confidence interval for parameter, such as mean and standard deviation, is important. There have been numerous studies introduced confidence intervals for parameter in the lognormal distribution, see for example, Verrill (2003), Harvey and Merwe (2012), and Niwitpong (2013). In the recent, Tang and Yeh (2016) presented approximate confidence interval for the lognormal standard deviation by transformation the lognormal variable to the normal. Based on the theory, if the random variable X is lognormally distributed, then Y=log(X) has a normal distribution. However, their estimator constructed based on a biased estimator. We therefore develop the interval estimator for the standard deviation of the lognormal distribution, using the unbiased estimator. The performance of our confidence interval and that of Tang and Yeh (2016) was evaluated using simulations. A real data was used to conclude the findings.

2 METHODS

Suppose that $X = (X_1, X_2, \dots, X_n)$ be a random variable from a lognormal distribution. The probability model of X is

$$f_X(X; \mu, \sigma^2) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(\log(x)-\mu)^2}{2\sigma^2}\right\},$$

where $x > 0$, $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$. The mean and variance of X are $E(X) = \exp\{\mu + \sigma^2/2\}$ and $\mathrm{Var}(X) = \exp\{2\mu + \sigma^2\} \cdot (\exp\{\sigma^2\} - 1)$, respectively.

According to Tang and Yeh (2016), logarithm of $\mathrm{Var}(X)$ can be estimated by $\log[\mathrm{Var}(X)] \cong 2\mu + \sigma^2 + \log(\sigma^2)$. A point estimator is given as $2\bar{Y} + S^2 + \log(S^2)$, where $\bar{Y}$ and S^2 are denoted as the sample mean and variance of a lognormal variable Y, respectively. Therefore, an approximate $100(1-\alpha)\%$ confidence interval for $\log[\mathrm{Var}(X)]$ can be derived by

$$(2\bar{Y} + S^2 + \log(S^2)) \pm z_{\frac{\alpha}{2}} \sqrt{\mathrm{Var}\left(2\bar{Y} + S^2 + \log(S^2)\right)}.$$

Using Stirling's approximation, $\mathrm{Var}\left(2\bar{Y} + S^2 + \log(S^2)\right)$ is estimated as $\widehat{\mathrm{Var}}(2\bar{Y} + S^2 + \log(S^2)) \cong \frac{4S^2}{n} + \frac{2+6S^2}{n-1} + \frac{2S^4-2S^2}{n+1}$. Thus, the lower and upper limits for the standard deviation, $SD(X)$, are given by

$$CI_{TY} = (LB, UB),$$

where

$$LB = \sqrt{\exp\left(A - z_{\frac{\alpha}{2}}\sqrt{\frac{4S^2}{n} + \frac{2+6S^2}{n-1} + \frac{2S^4-2S^2}{n+1}}\right)},$$

$$UB = \sqrt{\exp\left(A + z_{\frac{\alpha}{2}}\sqrt{\frac{4S^2}{n} + \frac{2+6S^2}{n-1} + \frac{2S^4-2S^2}{n+1}}\right)},$$

and $A = 2\bar{Y} + S^2 + \log(S^2)$.

From the previous method, it can be seen that $2\bar{Y} + S^2 + \log(S^2)$ is the biased estimator for $2\mu + \sigma^2 + \log(\sigma^2)$. As it is well-known that a good quality of estimator is unbiased. *In this study*, the point estimator satisfied this property is applied. We consider the mean of $2\bar{Y} + S^2 + \log(S^2)$ which can be approximated by

$$E(2\bar{Y} + S^2 + \log(S^2)) \cong 2\mu + \sigma^2 + \log(\sigma^2) - \frac{1}{n-1}.$$

This follows that

$$E\left(2\bar{Y} + S^2 + \log(S^2) + \frac{1}{n-1}\right) \cong 2\mu + \sigma^2 + \log(\sigma^2),$$

so the unbiased estimator of $2\mu + \sigma^2 + \log(\sigma^2)$ is

$$2\bar{Y} + S^2 + \log(S^2) + \frac{1}{n-1}.$$

Furthermore, the estimated variance of $2\bar{Y} + S^2 + \log(S^2) + \frac{1}{n-1}$ is given as $\widehat{\mathrm{Var}}\left(2\bar{Y} + S^2 + \log(S^2) + \frac{1}{n-1}\right) \cong \frac{4S^2}{n} + \frac{2+6S^2}{n-1} + \frac{2S^4-2S^2}{n+1}$. Thus, the new $100(1-\alpha)\%$ confidence interval for $SD(X)$ is given by

$$CI_{PR} = (LB_{PR}, UB_{PR}),$$

where

$$LB_{PR} = \sqrt{\exp\left(B - z_{\frac{\alpha}{2}}\sqrt{\frac{4S^2}{n} + \frac{2+6S^2}{n-1} + \frac{2S^4-2S^2}{n+1}}\right)},$$

$$UB_{PR} = \sqrt{\exp\left(B + z_{\frac{\alpha}{2}}\sqrt{\frac{4S^2}{n} + \frac{2+6S^2}{n-1} + \frac{2S^4-2S^2}{n+1}}\right)},$$

$B = 2\bar{Y} + S^2 + \log(S^2) + \frac{1}{n-1}$

and $z_{\frac{\alpha}{2}}$ is the $(\alpha/2)$th percentile of the standard normal distribution.

3 SIMULATION RESULTS

In this section, the performance of the confidence interval was investigated using simulations by R package. The data were generated from lognormal distribution with parameter $\mu = -\sigma^2/2$ and variance $\sigma^2 =$ 0.01, 0.05, 0.10, 0.25, 0.50, 0.80, 0.95, and 1.00. Sample size (n) was set as 10, 30, 100, and 300. The confidence level $(1-\alpha)$ was given by 0.95. The number of replications was set as 10000 for each simulation constellation. Then the performance of the proposed estimator and the existing estimator was computed in terms of coverage probability, expected length, and relative bias.

The major findings of the simulation studies can be summarized as follows. As in Figure 1, the coverage probability of CI_{PR} was greater than that of CI_{TY} in all cases in the study, especially when $n \leq 30$. The coverage probabilities of these confidence intervals decreased, if σ^2 was increased. These results are as same as the results reported in Tang and Yeh (2016). The expected length of CI_{PR} was slightly greater that of CI_{TY}. However, when $n \geq 30$, the expected lengths of CI_{PR} and CI_{TY} did not differ. These are shown in Figure 2. Furthermore, the expected lengths decreased, if n was increased. In Figure 3, it can be concluded that the novel confidence interval CI_{PR} had relative bias smaller than CI_{TY}. From the results, we note that CI_{PR} performed well in terms of coverage probability level and relative bias, and its estimate was close to the true standard deviation.

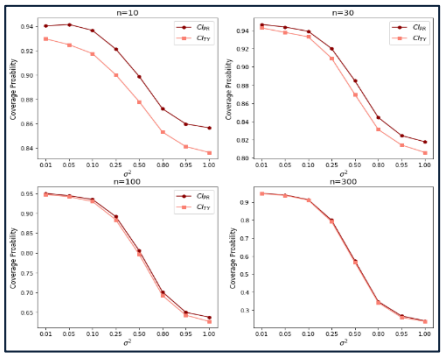


FIGURE 1: COVERAGE PROBABILITY OF THE 95% CONFIDENCE INTERVALS

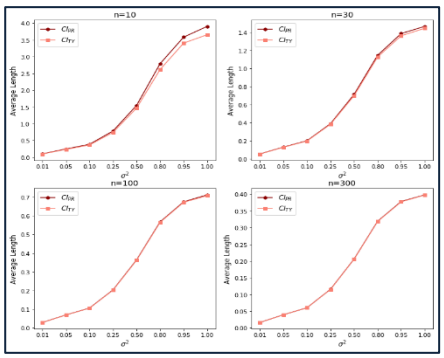


FIGURE 2: AVERAGE LENGTH OF THE 95% CONFIDENCE INTERVALS

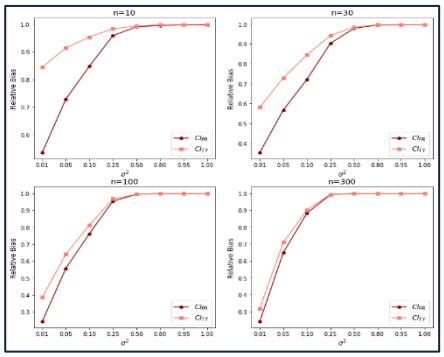


FIGURE 3: RELATIVE BIAS OF THE 95% CONFIDENCE INTERVALS

4 APPLICATIONS

The data on the air pollution were obtained from Pollution Control Department of a station in Lardprao, Wang Thonglang, Bangkok, between 6 March to 5 April 2018. The distributions of the data are presented in Figure 4 and Figure 5. From 31 sample sizes of particulate matter 2.5 (PM2.5), the mean was 33.00 μg/m^3 and standard deviation was 13.22 μg/m^3. The 95% CI_{TY} was (9.38, 17.91) with the length of interval 8.53 μg/m^3. For our method, CI_{PR} was (9.54, 18.22) with the length 8.68 μg/m^3. Therefore, the confidence intervals based on this example match the simulation results.

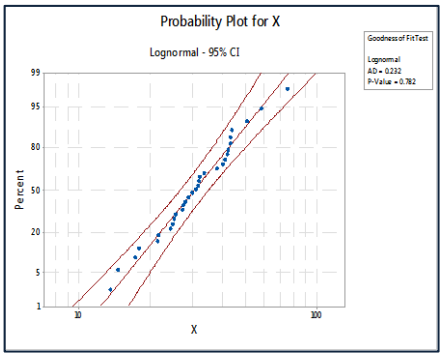


FIGURE 4: PROBABILITY PLOT OF THE REAL DATASET (X)

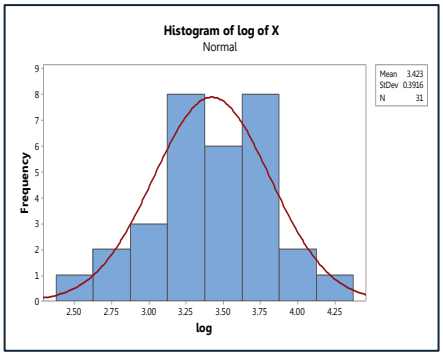


FIGURE 5: PLOT OF LOGARITHM OF X

5 CONCLUSIONS

The approximate confidence interval for the standard deviation of the lognormal distribution introduced in this work was derived based on the related unbiased estimator. The results showed that the coverage probability of our method was greater than the target probability level, if the true variance was less than 0.10. In general, it was also greater than the coverage probability of the confidence interval derived from the biased estimator. The relative bias of our estimator was also better. Clearly, the novel confidence interval had a good performance. We therefore recommend this estimator to estimate the standard deviation of the lognormal distribution.

ACKNOWLEDGEMENTS

The authors would like to thank Faculty of Science and Technology, Thammasat University, Thailand for the financial support for this work.

REFERENCES

Harvey, J., & Merwe, A.J. (2012). Bayesian confidence intervals for means and variances of lognormal and bivariate lognormal distributions. *Journal of Statistical Planning and Inference, 142*(6), 1294-1309.

Niwitpong, S. (2013). Confidence intervals for coefficient of variation of lognormal distribution with restricted parameter space. *Applied Mathematical Sciences, 7*(77), 3805-3810.

Tang, S., & Yeh, A.B. (2016). Approximate confidence intervals for the log-normal standard deviation. *Quality and Reliability Engineering International, 32*, 715-725.

Verrill, S. (2003). Confidence bounds for normal and lognormal distribution coefficients of variation. *Res. Pap. FPL-RP-609. Madison, WI: US Department of Agriculture, Forest Service, Forest Products Laboratory: 13 pages, 609.*

Ridge, Lasso, and Elastic Net Regressions Where the Predictors Show Degrees of Multicollinearity

Kanyalin Jiratchayut*

Faculty of Science and Arts, Burapha University, Chanthaburi Campus, Chanthaburi 22170, Thailand
*Corresponding Email: kanyalinn@gmail.com

ABSTRACT

The objective of this research was to study parameter estimation accuracy of ridge, lasso, and elastic net regressions where the predictors show different degrees of multicollinearity and simulation datasets are in three scenarios: sparse, dense, and grouped predictors. For sparse situation, the simulation studies show that the lasso performs best when correlations among the predictors are low and moderate, while the elastic net performs best when the predictors are highly correlated. For dense and grouped situations, the elastic net has parameter estimation performance better than ridge and lasso methods do.

Keywords: elastic net; lasso; multicollinearity; ridge

1 INTRODUCTION

Regression technique is widely used for machine learning and data science. Regression analysis is a technique used for prediction and analyzing the relationship between dependent and independent (predictor) variables. Ridge (Hoerl & Kennard, 1970a, 1970b), lasso (Tibshirani, 1996), and elastic net (Zou & Hastie, 2005) are penalized regression techniques used for analyzing data where multicollinearity problem among predictor variables exists. Consider a linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,, \tag{1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of response variable, $\mathbf{X}$ is an $n \times p$ matrix of predictor variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameter of regression coefficients, $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of random errors, p is the number of predictors, and n is the number of observations. The errors are assumed to be independent identically normally distributed random variable with mean 0 and finite variance σ^2.

Ridge regression was proposed by Hoerl and Kennard (1970a, 1970b). Assuming that the response is centered and the predictors are standardized, ridge estimator is defined by

$$\widehat{\boldsymbol{\beta}}_{ridge} = \arg\min_{\boldsymbol{\beta}}\left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \theta \sum_{j=1}^{p} \beta_j^2\right), \tag{2}$$

where $\theta \geq 0$.

The least absolute shrinkage and selection operator or lasso technique (Tibshirani, 1996) is a penalized least squares procedure which can do both continuous shrinkage and variable selection. Assuming that the response is centered and the predictors are standardized, the lasso estimator is given by

$$\widehat{\boldsymbol{\beta}}_{lasso} = \arg\min_{\boldsymbol{\beta}}\left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|\right), \tag{3}$$

where the penalty parameter $\lambda \geq 0$.

Zou and Hastie (2005) proposed the elastic net to solve the regression problem in microarray genes expression data. The elastic net simultaneously does automatic variable selection and continuous shrinkage, it can select groups of correlated variables and overcomes the difficulty of $p > n$. The elastic net is based on a combination of the ridge (L_2) and the lasso (L_1) penalties. Assuming that the response is centered and the predictors are standardized, the naïve elastic net estimator is defined as follows:

$$\widehat{\boldsymbol{\beta}}_{\text{elastic net}} = \arg\min_{\boldsymbol{\beta}}[\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1], \tag{4}$$

where the penalty parameters $\lambda_1 \geq 0$, $\lambda_2 \geq 0$ and $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$ where $\alpha \in (0,1)$.

The objective of this research was to study parameter estimation accuracy of ridge, lasso, and elastic net regressions where the predictors show different degrees of multicollinearity and simulation datasets are in three scenarios: sparse, dense, and grouped predictors. In this research, we limited our attention to full rank model. This article is organized as follows. Section 2 describes simulation data and decision criteria. Results and discussion are provided in Section 3.

2 MATERIALS AND METHODS

2.1 Simulation Data

The simulation datasets are generated with sample size 100 observations from the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,, \tag{5}$$

where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I_n})$. The correlation matrix is given by $corr(x_i, x_j) = \rho^{|i-j|}$ where ρ = 0.1, 0.5, and 0.7. The simulation datasets are in three scenarios: sparse, dense, and grouped predictors.

(i) For sparse situation, the true regression coefficients are $\boldsymbol{\beta}$ = (3, 2, 1.5, 0, 0, 0, 0, 0), and $\sigma = 3$.
(ii) Scenario 2 is the same as the first one, except that $\beta_1 = \beta_2 = \cdots = \beta_8 = 0.85$.
(iii) With p = 20, the true coefficient vector is given by $\boldsymbol{\beta}$ = (0, 0, 0, 0, 0, 2, 2, 2, 2, 2, 0, 0, 0, 0, 0, 2, 2, 2, 2, 2), and $\sigma = 9$.

The 100 datasets are simulated under each scenarios. For ridge regression, the value of θ is chosen as suggested by Hoerl et al. (1975) (as cited in Draper & Smith, 1998, p.390). The estimated value of θ is

$$\hat{\theta} = ps^2/\left[\widehat{\boldsymbol{\beta}}_{LS}\right]^T\left[\widehat{\boldsymbol{\beta}}_{LS}\right], \tag{6}$$

where p is the number of parameters in the model (not counting the intercept term), s^2 is the residual mean square in the analysis of variance table obtained from the standard least squared fit of the model and $\widehat{\boldsymbol{\beta}}_{LS}$ is the least squared estimator. For the lasso, the penalty parameter λ is tuning by 10-fold cross validation method. The naïve elastic net approach is fitted with $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$ = 0.5 and the penalty parameters λ_1 and λ_2 are selected by 10-fold cross validation method.

2.2 Decision Criteria

For each method, the parameter estimation accuracy is measured by the mean square error $MSE(\widehat{\boldsymbol{\beta}}) = E\left[(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^T(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\right]$.

3 RESULTS AND DISCUSSIONS

Table 1 – 3 show the parameter estimation performance of ridge, lasso, and elastic net under each scenarios. The average of $MSE(\widehat{\boldsymbol{\beta}})$ is computed based on 100 datasets. The numbers in parenthesis are the corresponding standard errors of $MSE(\widehat{\boldsymbol{\beta}})$ estimated using the bootstrap with B = 500 resampling from the 100 $MSE(\widehat{\boldsymbol{\beta}})$'s.

When correlations between the predictors are low and moderate, the lasso has parameter estimation performance better than the others do

in sparse situation, while the elastic net performs best for dense and grouped situations. For dense situation where correlation between the predictors are low, the lasso is slightly different from the elastic net. When the predictors are highly correlated, the elastic net has parameter estimation performance better than the lasso and ridge regressions.

The lasso and elastic net are penalized regression methods which perform both parameter estimation and variable selection. They do variable selection by shrinking some coefficients toward zero. This causes the lasso and elastic net have parameter estimation performance better than ridge regression. The elastic net method has the ability to perform group selection – highly correlated predictors tend to be in or out of the model together, so the elastic net performs best when the predictors are in dense and grouped situations, or the situation where the predictors are highly correlated.

Another penalized regression method such as LqCP (Mao & Ye, 2017) is developed to solve linear regression when the multicollinearity problem exists. One of future work is to study the performance of LqCP where the datasets are in different situations.

Table 1: Sparse situation

ρ	ridge	lasso	elastic net
0.1	0.1000 (0.0060)	0.0781 (0.0058)	0.0848 (0.0053)
0.5	0.1752 (0.0101)	0.1083 (0.0079)	0.1154 (0.0075)
0.7	0.2460 (0.0160)	0.1345 (0.0134)	0.1243 (0.0096)

Table 2: Dense situation

ρ	ridge	lasso	elastic net
0.1	0.1080 (0.0050)	0.1019 (0.0053)	0.0980 (0.0046)
0.5	0.1532 (0.0077)	0.1463 (0.0068)	0.1135 (0.0052)
0.7	0.2513 (0.0175)	0.2245 (0.0128)	0.1289 (0.0095)

Table 3: Grouped situation

ρ	ridge	Lasso	elastic net
0.1	1.1118 (0.0362)	0.8239 (0.0295)	0.6771 (0.0217)
0.5	1.6196 (0.0664)	0.8558 (0.0379)	0.5884 (0.0253)
0.7	2.7725 (0.1164)	1.1601 (0.0532)	0.5616 (0.0334)

ACKNOWLEDGEMENTS

The author would like to thank the referees for their valuable suggestions.

REFERENCES

Draper, N.R., & Smith, H. (1998). *Applied regression analysis* (3rd edition). New York: John Wiley & Sons.

Hoerl, A.E., & Kennard, R.W. (1970). Ridge regression: Applications to nonorthogonal problems. *Technometrics, 12*(1), 69-82.

Hoerl, A.E., & Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67.

Hoerl, A.E., Kannard, R.W., & Baldwin, K. F. (1975). Ridge regression: some simulations. *Communications in Statistics-Theory and Methods, 4*(2), 105-123.

Mao, N., & Ye, W. (2017). Group variable selection via a combination of Lq norm and correlation-based penalty. *Advances in Pure Mathematics, 7*(01), 51-65.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological), 58*, 267-288.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B (Statistical Methodology), 67*(2), 301-320.

ช่วงความเชื่อมั่นสำหรับผลต่างระหว่างส่วนกลับของค่าเฉลี่ยประชากรปรกติ เมื่อทราบค่าสัมประสิทธิ์การแปรผันและปริภูมิพารามิเตอร์มีขอบเขต

กุลนิดา เหมะรักษ์[1] และ วราฤทธิ์ พานิชกิจโกศลกุล[2*]

[1,2]สาขาวิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ ศูนย์รังสิต
คลองหลวง ปทุมธานี 12120
อีเมล: kulnida.hem@gmail.com
*อีเมลผู้ประสานงาน: wararit@mathstat.sci.tu.ac.th

บทคัดย่อ

ในงานวิจัยนี้มีวัตถุประสงค์เพื่อเสนอช่วงความเชื่อมั่นสำหรับผลต่างระหว่างส่วนกลับของค่าเฉลี่ยประชากรปรกติ เมื่อทราบค่าสัมประสิทธิ์การแปรผันและปริภูมิพารามิเตอร์มีขอบเขต โดยช่วงความเชื่อมั่นที่ศึกษามี 3 ช่วงความเชื่อมั่น คือ ช่วงความเชื่อมั่นนัยทั่วไป (Generalized Confidence Interval: GCI) ช่วงความเชื่อมั่นโดยใช้วิธี Method of Variance Estimates Recovery (MOVER) และช่วงความเชื่อมั่นแบบประมาณโดยวิธี MOVER (Approximation Confidence Interval with Method of Variance Estimates Recovery: AMOVER) การเปรียบเทียบประสิทธิภาพของช่วงความเชื่อมั่นจะศึกษาภายใต้กรณีที่ปริภูมิพารามิเตอร์ไม่มีขอบเขตและมีขอบเขต สำหรับการหาช่วงความเชื่อมั่นภายใต้ปริภูมิพารามิเตอร์มีขอบเขตจะใช้หลักการหาช่วงความเชื่อมั่นร่วม (Intersection) ระหว่างปริภูมิพารามิเตอร์กับช่วงความเชื่อมั่นทั้ง 3 วิธี การวิจัยครั้งนี้ใช้วิธีการจำลองด้วยเทคนิคมอนติคาร์โลโดยการเปรียบเทียบค่าความน่าจะเป็นคุ้มรวมและค่าความยาวโดยเฉลี่ยของช่วงความเชื่อมั่น ผลการศึกษาพบว่า การเปรียบเทียบภายใต้สองกรณีดังกล่าวให้ผลการศึกษาที่เหมือนกัน กล่าวคือ เมื่อพิจารณาที่ค่าความน่าจะเป็นคุ้มรวม ส่วนใหญ่ช่วงโดยวิธี MOVER และ AMOVER ให้ค่าความน่าจะเป็นคุ้มรวมไม่น้อยกว่าสัมประสิทธิความเชื่อมั่นที่กำหนด และเมื่อพิจารณาที่ค่าความยาวโดยเฉลี่ยของช่วงโดยวิธี MOVER และ AMOVER ให้ความยาวช่วงที่สั้นที่สุด
คำสำคัญ: การประมาณค่าแบบช่วง ค่าวัดแนวโน้มเข้าสู่ส่วนกลาง ค่าวัดการกระจาย

Abstract

The objective of this research is to propose new confidence intervals for the difference between reciprocal of normal means with known coefficients of variation and restricted parameter spaces. The confidence intervals which interested are generalized confidence interval (GCI), confidence interval based on the method of variance estimates recovery (MOVER) and approximate confidence interval based on the method of variance estimates recovery (AMOVER). In addition, this research presents confidence intervals based on a restricted parameter spaces; that is, the confidence intervals are derived from intersection between parameter space and the confidence intervals (GCI, MOVER confidence interval and AMOVER confidence interval). Monte Carlo simulation technique is applied in this research by considering the coverage probability and expected length to compare the efficiency of the confidence intervals. The results show that the restricted confidence intervals which are derived from the MOVER and the AMOVER have more coverage probability than the 95% confidence level that is determined and have the smallest expected length
Keywords: Interval estimation, Central tendency, Dispersion measurement

1 บทนำ

โดยทั่วไปการอนุมานเชิงสถิติ (Statistical Inference) สามารถแบ่งได้เป็น 2 ลักษณะดังนี้ การประมาณค่าพารามิเตอร์ (Estimation) และการทดสอบสมมติฐาน (Test of Hypothesis) โดยในงานวิจัยนี้สนใจกรณีของการประมาณค่าพารามิเตอร์แบบช่วง มีงานวิจัยที่สนใจเกี่ยวกับการหาค่าส่วนกลับของค่าเฉลี่ยอย่างกว้างขวาง เช่น ในงานวิจัยด้านฟิสิกส์นิวเคลียร์ (Nuclear Physics) ของ Lamanna (1981) ได้ทำการศึกษาเกี่ยวกับโมเมนตัมของอนุภาค (Particle Momentum) $p=1/\mu$ โดยที่ μ คือเส้นทางการเคลื่อนที่ (Track) ของอนุภาค

หลังจากนั้นได้มีการสนใจศึกษาเกี่ยวกับการประมาณค่าแบบช่วงสำหรับส่วนกลับของค่าเฉลี่ยประชากรปรกติ โดย Wongkhao et al.

(2013) ซึ่งได้ศึกษางานวิจัยเกี่ยวกับการประมาณค่าแบบช่วงสำหรับส่วนกลับของค่าเฉลี่ยประชากรปรกติ เมื่อทราบค่าสัมประสิทธิ์การแปรผัน (Coefficient of Variation) โดยได้นำเสนอทฤษฎีและบทแทรกของค่าคาดหวัง (Expected Value) และค่าความแปรปรวน (Variance) ของส่วนกลับของค่าเฉลี่ยมาใช้ในการสร้างช่วงความเชื่อมั่นของส่วนกลับของค่าเฉลี่ย หลังจากนั้น Panichkitkosolkul (2017) ได้ปรับช่วงความเชื่อมั่นสำหรับส่วนกลับของค่าเฉลี่ยประชากรปรกติ เมื่อทราบค่าสัมประสิทธิ์การแปรผัน (Coefficient of Variation) เพื่อสร้างช่วงความเชื่อมั่นของส่วนกลับของค่าเฉลี่ยที่ประชากรปรกติมีประสิทธิภาพมากยิ่งขึ้น ต่อมาการประมาณค่าแบบช่วงสำหรับผลต่างระหว่างส่วนกลับของค่าเฉลี่ยประชากรปรกติ เมื่อทราบค่าสัมประสิทธิ์การแปรผัน (Coefficient of Variation) ได้รับความสนใจมากขึ้น เห็นได้จากในงานวิจัย Wongkhao (2014) สนใจศึกษาเกี่ยวกับการประมาณค่าแบบช่วงสำหรับผลต่างระหว่างส่วนกลับของค่าเฉลี่ย โดยเสนอช่วงความเชื่อมั่นสำหรับผลต่างระหว่างส่วนกลับของค่าเฉลี่ยประชากรปรกติ โดยใช้วิธีช่วงความเชื่อมั่นนัยทั่วไป (Generalized Confidence Interval: GCI) และวิธี Method of Variance Estimates Recovery (MOVER)

เนื่องจากงานวิจัยในหลากหลายแขนงที่สนใจศึกษานั้น ไม่ว่าจะเป็นในด้านฟิสิกส์นิวเคลียร์ ด้านวิทยาศาสตร์ชีวภาพ ด้านการเกษตร และด้านเศรษฐศาสตร์ โดยส่วนใหญ่จะเป็นกรณีที่ปริภูมิพารามิเตอร์มีขอบเขตทั้งสิ้น เช่น ค่าขอบเขตความดันเลือดของผู้ป่วย และค่าขอบเขตน้ำหนักของคน ซึ่งในงานวิจัย Mandelkern (2002) ได้กล่าวว่า วิธีการประมาณค่าช่วงความเชื่อมั่นแบบเนย์แมน (Neyman Procedure) ไม่เพียงพอที่จะใช้ในการประมาณค่าช่วงความเชื่อมั่นกรณีที่ปริภูมิพารามิเตอร์มีขอบเขต เพื่อแก้ปัญหาดังกล่าว Wang (2008) จึงได้นำเสนอการหาช่วงความเชื่อมั่นของพารามิเตอร์โดยที่ปริภูมิพารามิเตอร์มีขอบเขต ซึ่งสามารถทำได้โดยการหาช่วงความเชื่อมั่นร่วม (Intersection) ระหว่างปริภูมิพารามิเตอร์กับช่วงความเชื่อมั่นของพารามิเตอร์ที่หามาได้ด้วยวิธีการต่าง ๆ ทางสถิติ

ในการศึกษาทางฟิสิกส์นิวเคลียร์ดังกล่าวได้มีการขยายผลการศึกษาออกเป็นสองกลุ่ม กล่าวคือสนใจเพิ่มเติมเกี่ยวกับผลต่างระหว่างโมเมนตัม $1/\mu_x - 1/\mu_y$ โดยที่ μ_x และ μ_y คือเส้นทางการเคลื่อนที่ (Track) ของอนุภาค x และอนุภาค y ตามลำดับ และจากที่กล่าวข้างต้นในงานวิจัยทางฟิสิกส์นิวเคลียร์จะเป็นกรณีที่ปริภูมิพารามิเตอร์มีขอบเขตทั้งสิ้น ดังนั้นในงานวิจัยนี้จึงสนใจศึกษาการประมาณค่าช่วงความเชื่อมั่นสำหรับผลต่างระหว่างส่วนกลับของค่าเฉลี่ยประชากรปรกติ เมื่อทราบค่าสัมประสิทธิ์การแปรผัน และสนใจศึกษากรณีที่ประชากรทั้งสองกลุ่มนั้นเป็นอิสระต่อกัน โดยช่วงความเชื่อมั่นที่ศึกษามี 3 ช่วงความเชื่อมั่น คือ ช่วงความเชื่อมั่นนัยทั่วไป (Generalized Confidence Interval: GCI) ช่วงความเชื่อมั่นโดยใช้วิธี Method of Variance Estimates Recovery (MOVER) และช่วงความเชื่อมั่นแบบประมาณโดยวิธี MOVER (Approximation Confidence Interval with Method of Variance Estimates Recovery: AMOVER) การเปรียบเทียบประสิทธิภาพของช่วงความเชื่อมั่นดังกล่าวจะศึกษาภายใต้กรณีที่ปริภูมิพารามิเตอร์ไม่มีขอบเขตและมีขอบเขต สำหรับการหาช่วงความเชื่อมั่นภายใต้ปริภูมิพารามิเตอร์มีขอบเขตจะใช้หลักการหาช่วงความเชื่อมั่นร่วม (Intersection) ระหว่างปริภูมิพารามิเตอร์กับช่วงความเชื่อมั่นทั้ง 3 วิธีดังกล่าว การวิจัยครั้งนี้ใช้วิธีการจำลองด้วยเทคนิคมอนติคาร์โลโดยเกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพของช่วงความเชื่อมั่น คือค่าความน่าจะเป็นคุ้มรวมของช่วงความเชื่อมั่น (Coverage Probability: CP) และค่าความยาวโดยเฉลี่ยของช่วงความเชื่อมั่น (Expected Length: EL)

2 วิธีดำเนินการวิจัย

2.1 แผนการวิจัย

ในงานวิจัยได้จำลองข้อมูลจากโปรแกรม R โดยมีการกำหนดค่าต่าง ๆ ไว้ดังนี้

1. กำหนดขนาดตัวอย่างกลุ่มที่ 1 และ 2 เท่ากับ 50 และ 50 ตามลำดับ

2. ระดับความเชื่อมั่นที่กำหนดในงานวิจัยนี้ คือ 95%

3. กำหนดให้ข้อมูลมีการแจกแจงปรกติที่มีค่าเฉลี่ยประชากรสองกลุ่ม $\left(\mu_x = 5, \mu_y = 10\right)$ และกำหนดค่าสัมประสิทธิ์การแปรผันของประชากรทั้งสองกลุ่ม $\left(\tau_x, \tau_y\right)$ เท่ากับ ร้อยละ 5 ร้อยละ 10 ร้อยละ 15 ร้อยละ 30 และร้อยละ 50

4. กำหนดให้ขอบเขตของปริภูมิพารามิเตอร์ คือ $\mu_x \in (4.5, 5.5)$ และ $\mu_y \in (9, 11)$

5. ทำการจำลองทั้งหมด $10{,}000$ ครั้ง ในแต่ละสถานการณ์ที่กำหนด

2.2 ช่วงความเชื่อมั่นในงานวิจัย

2.2.1 ช่วงความเชื่อมั่นสำหรับผลต่างระหว่างส่วนกลับของค่าเฉลี่ยประชากร กรณีปริภูมิพารามิเตอร์ไม่มีขอบเขต

กำหนดให้ $X \sim N\left(\mu_x, \sigma_x^2\right)$ และ $Y \sim N\left(\mu_y, \sigma_y^2\right)$ จะได้ $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$, $S_x^2 = (n-1)^{-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ และ $\bar{Y} = m^{-1}\sum_{i=1}^{m} Y_i$, $S_y^2 = (m-1)^{-1}\sum_{i=1}^{m}\left(Y_i - \bar{Y}\right)^2$ เมื่อ n, m คือขนาดตัวอย่างของกลุ่ม X และ Y ตามลำดับ

1) ช่วงความเชื่อมั่นนัยทั่วไป (Generalized Confidence Interval: GCI)

Wongkhao (2014) ใช้การประมาณแบบช่วงโดยใช้ปริมาณหมุนซึ่งในงานวิจัยนี้ใช้ปริมาณหมุนนัยทั่วไป (Generalized Pivotal) คือ

$$M_\eta = M\left(X, Y, x, y, \eta, \sigma_x^2, \sigma_y^2\right) = \frac{\tau_x}{\sqrt{(n-1)S_x^2/U}} - \frac{\tau_y}{\sqrt{(m-1)S_y^2/V}}$$

โดยที่ $U = (n-1)S_x^2/\sigma_x^2 \sim \chi_{n-1}^2$ และ $V = (m-1)S_y^2/\sigma_y^2 \sim \chi_{m-1}^2$ จะได้ช่วงความเชื่อมั่นนัยทั่วไป $100(1-\alpha)\%$ สำหรับ $1/\mu_x - 1/\mu_y$ ดังนี้

$$CI_{GCI} = \left[M_{\eta(\alpha/2)}, M_{\eta(1-\alpha/2)}\right] \quad (1)$$

โดยที่ $M_{\eta(\alpha/2)}$ และ $M_{\eta(1-\alpha/2)}$ คือ ควอนไทล์ตำแหน่งที่ $\alpha/2$ และ $1-\alpha/2$ ของปริมาณหมุนทั้งหมด $5{,}000$ รอบ

(2) ช่วงความเชื่อมั่นโดยวิธี MOVER (Method of Variance Estimates Recovery)

Wongkhao (2014) ได้ใช้วิธี MOVER ในการสร้างช่วงความเชื่อมั่น $100(1-\alpha)\%$ สำหรับฟังก์ชันของค่าเฉลี่ยประชากรปรกติ $\theta_1-\theta_2$ ซึ่งในที่นี้คือ $1/\mu_x-1/\mu_y$ ดังนั้น ช่วงความเชื่อมั่น $100(1-\alpha)\%$ สำหรับ $\hat{\theta}_1-\hat{\theta}_2$ คือ

$$(L',U')=\left(\left(\hat{\theta}_1-\hat{\theta}_2\right)-\sqrt{\left(\hat{\theta}_1-l_1\right)^2+\left(u_2-\hat{\theta}_2\right)^2},\ \left(\hat{\theta}_1-\hat{\theta}_2\right)+\sqrt{\left(u_1-\hat{\theta}_1\right)^2+\left(\hat{\theta}_2-l_2\right)^2}\right) \quad (2)$$

จาก Wongkhao et al. (2013) จะได้ช่วงความเชื่อมั่นสำหรับส่วนกลับของค่าเฉลี่ยประชากรปรกติ $\theta_1=1/\mu_x$ และ $\theta_2=1/\mu_y$ คือช่วง

$$\left(l_1^*,u_1^*\right)=\left[\frac{\hat{\theta}_1}{w_x}-z_{1-\alpha/2}\sqrt{\frac{\tau_x^2}{n\bar{X}^2}},\frac{\hat{\theta}_1}{w_x}+z_{1-\alpha/2}\sqrt{\frac{\tau_x^2}{n\bar{X}^2}}\right]$$

โดยที่ $w_x=1+\sum_{k=1}^{\infty}\left(\frac{(2k)!}{2^k k!}\right)\left(\frac{\tau_x^2}{n}\right)^k$

และช่วง

$$\left(l_2^*,u_2^*\right)=\left[\frac{\hat{\theta}_2}{w_y}-z_{1-\alpha/2}\frac{\tau_y^2}{m\bar{Y}^2},\frac{\hat{\theta}_2}{w_y}+z_{1-\alpha/2}\sqrt{\frac{\tau_y^2}{m\bar{Y}^2}}\right]$$

โดยที่ $w_y=1+\sum_{k=1}^{\infty}\left(\frac{(2k)!}{2^k k!}\right)\left(\frac{\tau_y^2}{m}\right)^k$ ตามลำดับ

แทนค่า $\hat{\theta}_1=\frac{1}{\bar{X}}$, $\hat{\theta}_2=\frac{1}{\bar{Y}}$, $l_1=l_1^*$, $l_2=l_2^*$, $u_1=u_1^*$ และ $u_2=u_2^*$ ลงในช่วงความเชื่อมั่นในสมการที่ (2) ดังนั้นจะได้ช่วงความเชื่อมั่นสำหรับ $1/\mu_x-1/\mu_y$ คือ

$$CI_{MV}=\left(\left(\frac{1}{\bar{X}}-\frac{1}{\bar{Y}}\right)-\sqrt{\left(\frac{1}{\bar{X}}-l_1^*\right)^2+\left(u_2^*-\frac{1}{\bar{Y}}\right)^2},\ \left(\frac{1}{\bar{X}}-\frac{1}{\bar{Y}}\right)+\sqrt{\left(u_1^*-\frac{1}{\bar{X}}\right)^2+\left(\frac{1}{\bar{Y}}-l_2^*\right)^2}\right) \quad (3)$$

(3) ช่วงความเชื่อมั่นแบบประมาณโดยวิธี MOVER (Approximation Confidence Interval with Method of Variance Estimates Recovery: AMOVER)

ในงานวิจัยนี้ได้สนใจนำทฤษฎีบทของ Panichkitkosolkul (2017) มาประยุกต์ใช้กับวิธี MOVER โดยเรียกช่วงที่สร้างขึ้นใหม่นี้ว่า ช่วงความเชื่อมั่นแบบประมาณโดยวิธี MOVER (Approximation Confidence Interval with Method of Variance Estimates Recovery: AMOVER) ซึ่งช่วงดังกล่าวสร้างได้ดังนี้

จากงานวิจัยของ Panichkitkosolkul (2017) จะได้ช่วงความเชื่อมั่นสำหรับส่วนกลับของค่าเฉลี่ยประชากรปรกติ $\theta_3=1/\mu_x$ และ $\theta_4=1/\mu_y$ คือช่วง $\left(l_3^*,u_3^*\right)=\left[\frac{\hat{\theta}_3}{v_x}-z_{1-\alpha/2}\sqrt{\frac{\tau_x^2}{n\bar{X}^2}},\frac{\hat{\theta}_3}{v_x}+z_{1-\alpha/2}\sqrt{\frac{\tau_x^2}{n\bar{X}^2}}\right]$, $v_x=1+\frac{\tau_x^2}{n}$

และช่วง

$$\left(l_4^*,u_4^*\right)=\left[\frac{\hat{\theta}_4}{v_y}-z_{1-\alpha/2}\sqrt{\frac{\tau_y^2}{m\bar{Y}^2}},\frac{\hat{\theta}_4}{v_y}+z_{1-\alpha/2}\sqrt{\frac{\tau_y^2}{m\bar{Y}^2}}\right],\ v_y=1+\frac{\tau_y^2}{m}$$

ตามลำดับ

แทนค่า $\hat{\theta}_3=\frac{1}{\bar{X}}$, $\hat{\theta}_4=\frac{1}{\bar{Y}}$, $l_3=l_3^*$, $l_4=l_4^*$, $u_3=u_3^*$ และ $u_4=u_4^*$ ลงในช่วงความเชื่อมั่นในสมการที่ (2) ดังนั้นจะได้ช่วงความเชื่อมั่นสำหรับ $1/\mu_x-1/\mu_y$ คือ

$$CI_{AMV}=\left(\left(\frac{1}{\bar{X}}-\frac{1}{\bar{Y}}\right)-\sqrt{\left(\frac{1}{\bar{X}}-l_3^*\right)^2+\left(u_4^*-\frac{1}{\bar{Y}}\right)^2},\ \left(\frac{1}{\bar{X}}-\frac{1}{\bar{Y}}\right)+\sqrt{\left(u_3^*-\frac{1}{\bar{X}}\right)^2+\left(\frac{1}{\bar{Y}}-l_4^*\right)^2}\right) \quad (4)$$

2.2.2 ช่วงความเชื่อมั่นสำหรับผลต่างระหว่างส่วนกลับของค่าเฉลี่ยประชากร กรณีปริภูมิพารามิเตอร์มีขอบเขต

ในงานวิจัยนี้ได้นำแนวคิดกรณีปริภูมิพารามิเตอร์มีขอบเขตจากงานวิจัยของ Wang (2008) มาปรับใช้กับการหาช่วงความเชื่อมั่นสำหรับผลต่างระหว่างส่วนกลับของค่าเฉลี่ยประชากร จะได้ช่วงความเชื่อมั่นสำหรับผลต่างระหว่างส่วนกลับของค่าเฉลี่ยประชากร กรณีปริภูมิพารามิเตอร์มีขอบเขต คือ

$$CI_{\eta}\in\left[\max\left(b^{-1}-c^{-1},L_{\eta}\right),\min\left(a^{-1}-d^{-1},U_{\eta}\right)\right] \quad (5)$$

โดยที่ $\eta=\mu_x^{-1}-\mu_y^{-1}$ คือพารามิเตอร์ที่สนใจศึกษา

L_{η},U_{η} คือขีดจำกัดล่างและบนของช่วงความเชื่อมั่น η และ

$b^{-1}-c^{-1}$, $a^{-1}-d^{-1}$ คือขอบล่างและบนของปริภูมิพารามิเตอร์ $\mu_x^{-1}-\mu_y^{-1}$ เมื่อ $\mu_x\in[a,b]$ และ $\mu_y\in[c,d]$

1) ช่วงความเชื่อมั่นนัยทั่วไป (Generalized Confidence Interval: GCI)

จากช่วงความเชื่อมั่น $100(1-\alpha)\%$ ของ $\frac{1}{\mu_x}-\frac{1}{\mu_y}$ ในสมการที่ (1) สำหรับกรณีปริภูมิพารามิเตอร์มีขอบเขตจะได้

$$CI_{BGCI}\in\left[\max\left(b^{-1}-c^{-1},L_{GCI}\right),\min\left(a^{-1}-d^{-1},U_{GCI}\right)\right] \quad (6)$$

โดยที่ $L_{GCI}=M_{\eta(\alpha/2)},U_{GCI}=M_{\eta(1-\alpha/2)}$

2) ช่วงความเชื่อมั่นโดยวิธี MOVER (Method of Variance Estimates Recovery)

จากช่วงความเชื่อมั่น $100(1-\alpha)\%$ ของ $\frac{1}{\mu_x}-\frac{1}{\mu_y}$ ในสมการที่ (3) สำหรับกรณีปริภูมิพารามิเตอร์มีขอบเขตจะได้

$$CI_{BMV}\in\left[\max\left(b^{-1}-c^{-1},L_{MV}\right),\min\left(a^{-1}-d^{-1},U_{MV}\right)\right] \quad (7)$$

โดยที่ $L_{MV}=\left(\frac{1}{\bar{X}}-\frac{1}{\bar{Y}}\right)-\sqrt{\left(\frac{1}{\bar{X}}-l_1^*\right)^2+\left(u_2^*-\frac{1}{\bar{Y}}\right)^2}$

และ $U_{MV}=\left(\frac{1}{\bar{X}}-\frac{1}{\bar{Y}}\right)+\sqrt{\left(u_1^*-\frac{1}{\bar{X}}\right)^2+\left(\frac{1}{\bar{Y}}-l_2^*\right)^2}$

3) ช่วงความเชื่อมั่นแบบประมาณโดยวิธี MOVER (Approximation Confidence Interval with Method of Variance Estimates Recovery: AMOVER)

จากช่วงความเชื่อมั่น $100(1-\alpha)\%$ ของ $\frac{1}{\mu_x}-\frac{1}{\mu_y}$ ในสมการที่ (4) สำหรับกรณีปริภูมิพารามิเตอร์มีขอบเขตจะได้

$$CI_{BAMV}\in\left[\max\left(b^{-1}-c^{-1},L_{NMV}\right),\min\left(a^{-1}-d^{-1},U_{NMV}\right)\right] \quad (8)$$

โดยที่ $L_{AMV} = \left(\frac{1}{\bar{X}} - \frac{1}{\bar{Y}}\right) - \sqrt{\left(\frac{1}{\bar{X}} - l_3^*\right)^2 + \left(u_4^* - \frac{1}{\bar{Y}}\right)^2}$

และ $U_{AMV} = \left(\frac{1}{\bar{X}} - \frac{1}{\bar{Y}}\right) + \sqrt{\left(u_3^* - \frac{1}{\bar{X}}\right)^2 + \left(\frac{1}{\bar{Y}} - l_4^*\right)^2}$

2.3 เกณฑ์ในการเปรียบเทียบช่วงความเชื่อมั่นสำหรับผลต่างระหว่างส่วนกลับของค่าเฉลี่ย

2.3.1 ค่าความน่าจะเป็นคุ้มรวมของช่วงความเชื่อมั่น

ในการคำนวณค่าความน่าจะเป็นคุ้มรวมของจำนวนครั้งที่ช่วงความเชื่อมั่นครอบคลุมค่าพารามิเตอร์ $\left(1/\mu_x - 1/\mu_y\right)$ จากจำนวนครั้งในการทำซ้ำทั้งหมด $10,000$ ครั้ง โดยสามารถคำนวณได้จากสูตร $CP = \sum_{i=1}^{10,000} I_i\left(1/\mu_x - 1/\mu_y\right) \Big/ 10,000$

จากงานวิจัย Rohde CA. (2014) เมื่อกำหนดระดับนัยสำคัญที่ 0.05 จากการทดสอบสมมติฐาน $H_0 : CP \geq 0.95$ จะได้ว่า สัมประสิทธิ์ความเชื่อมั่นไม่น้อยกว่า 95% เมื่อค่าความน่าจะเป็นคุ้มรวมมีค่าไม่น้อยกว่า 0.9464 ที่ระดับนัยสำคัญ 0.05 แสดงว่าช่วงความเชื่อมั่นนั้นมีประสิทธิภาพ

2.3.2 ค่าความยาวโดยเฉลี่ยของช่วงความเชื่อมั่น

ในงานวิจัยนี้มีการทำซ้ำทั้งหมด $10,000$ ครั้ง ค่าความยาวโดยเฉลี่ยของช่างความเชื่อมั่นสามารถคำนวณได้โดยการหาผลรวมของความยาวของช่วงความเชื่อมั่นของจำนวนครั้งที่ช่วงความเชื่อมั่นครอบคลุมค่าพารามิเตอร์ $\left(1/\mu_x - 1/\mu_y\right)$ หารด้วยจำนวนครั้งของการทำซ้ำทั้งหมด ซึ่งเขียนเป็นสูตรได้ดังนี้ $EL = \sum_{i=1}^{10,000} Len_i\left(1/\mu_x - 1/\mu_y\right) \Big/ 10,000$ ดังนั้นช่วงความเชื่อมั่นที่ได้จะมีประสิทธิภาพ ก็ต่อเมื่อ ค่าความยาวโดยเฉลี่ยของช่วงความเชื่อมั่นมีค่าน้อย

2.4 ขั้นตอนในการวิจัย

1. สร้างตัวแปรสุ่มจากการแจกแจงปรกติ โดยจำลองข้อมูลในโปรแกรม R จากนั้นกำหนดขนาดตัวอย่างและค่าพารามิเตอร์ต่าง ๆ กันตามแต่ละกรณี
2. นำตัวแปรสุ่มที่ได้ในแต่ละกรณี มาคำนวณค่าเฉลี่ยและค่าความแปรปรวนของตัวอย่าง
3. คำนวณช่วงความเชื่อมั่นสำหรับผลต่างระหว่างส่วนกลับของค่าเฉลี่ยจากวิธีการประมาณช่วงความเชื่อมั่นสำหรับผลต่างระหว่างส่วนกลับของค่าเฉลี่ยประชากรทั้ง 6 วิธี ได้แก่ ช่วงแบบ $CI_{GCI}, CI_{MV}, CI_{AMV}, CI_{BGCI}, CI_{BMV}$ และ CI_{BAMV}
4. ทำการตรวจสอบว่าช่วงความเชื่อมั่นสำหรับผลต่างระหว่างส่วนกลับของค่าเฉลี่ยประชากรที่คำนวณได้ในแต่ละวิธีครอบผลต่างระหว่างคลุมค่าส่วนกลับของค่าเฉลี่ยที่กำหนดไว้หรือไม่
5. คำนวณค่าความน่าจะเป็นคุ้มรวมและค่าความยาวโดยเฉลี่ยของช่วงความเชื่อมั่นที่คำนวณได้ในแต่ละวิธี และทำการเปรียบเทียบ
6. สรุปผลการวิจัย

2.5 ผลการวิจัย

ในส่วนของการแสดงผลการวิจัยนี้ จะแสดงผลในรูปแบบตาราง โดยกำหนดสัญลักษณะต่างๆ ดังนี้

CP_{GCI}, CP_{MV}, CP_{AMV}, CP_{BGCI}, CP_{BMV} และ CP_{BAMV} หมายถึงค่าความน่าจะเป็นคุ้มรวม (Coverage Probability) ของช่วงความเชื่อมั่น CI_{GCI}, CI_{MV}, CI_{AMV}, CI_{BGCI}, CI_{BMV} และ CI_{BAMV} ตามลำดับ

EL_{GCI}, EL_{MV}, EL_{AMV}, EL_{BGCI}, EL_{BMV} และ EL_{BAMV} หมายถึง ค่าความยาวโดยเฉลี่ย (Expected Length) ของช่วงความเชื่อมั่น CI_{GCI}, CI_{MV}, CI_{AMV}, CI_{BGCI}, CI_{BMV} และ CI_{BAMV} ตามลำดับ

<u>ขีดเส้นใต้หนึ่งเส้น</u> หมายถึง ค่าความน่าจะเป็นคุ้มรวมที่ไม่น้อยกว่าค่าสัมประสิทธิ์ความเชื่อมั่นที่กำหนด

<u>ขีดเส้นใต้สองเส้น</u> หมายถึง ค่าเฉลี่ยของความยาวช่วงที่สั้นที่สุดโดยที่มีค่าความน่าจะเป็นคุ้มรวมไม่น้อยกว่าค่าสัมประสิทธิ์ความเชื่อมั่นที่กำหนด

" – " หมายถึง กรณีที่ค่าความน่าจะเป็นคุ้มรวมให้ค่าน้อยกว่าค่าสัมประสิทธิ์ความเชื่อมั่นที่กำหนด ซึ่งจะไม่นำมาพิจารณาในตารางค่าความยาวช่วงโดยเฉลี่ย

$\tau_x, \tau_y = 0.05$ และ 0.10 หมายถึง ค่าสัมประสิทธิ์การแปรผันที่มีค่าน้อย

$\tau_x, \tau_y = 0.15$ และ 0.30 หมายถึง ค่าสัมประสิทธิ์การแปรผันที่มีค่าปานกลาง

$\tau_x, \tau_y = 0.50$ หมายถึง ค่าสัมประสิทธิ์การแปรผันที่มีค่ามาก

จากตารางที่ 1 พบว่า เมื่อพิจารณาค่าสัมประสิทธิ์การแปรผัน ทุกช่วง CI_{GCI}, CI_{MV}, CI_{AMV}, CI_{BGCI}, CI_{BMV} และ CI_{BAMV} ส่วนใหญ่ให้ค่า CP_{GCI}, CP_{MV}, CP_{AMV}, CP_{BGCI}, CP_{BMV} และ CP_{BAMV} ไม่น้อยกว่าค่าสัมประสิทธิ์ความเชื่อมั่นที่กำหนด และในกรณีที่ค่าสัมประสิทธิ์การแปรผันตัวใดตัวหนึ่งมีค่ามาก ช่วงความเชื่อมั่นจะให้ค่าความน่าจะเป็นคุ้มรวมน้อยกว่าสัมประสิทธิ์ความเชื่อมั่นที่กำหนดเพียงเล็กน้อย

จากตารางที่ 2 พบว่า เมื่อพิจารณาค่าความยาวโดยเฉลี่ยของช่วงความเชื่อมั่น โดยที่ช่วงดังกล่าวให้ค่าความน่าจะเป็นคุ้มรวมไม่น้อยกว่าสัมประสิทธิ์ความเชื่อมั่นที่กำหนด จะได้ว่า ส่วนใหญ่ทุกกรณีของ $\left(\tau_x, \tau_y\right)$ ช่วง CI_{MV} CI_{AMV} CI_{BMV} และ CI_{BAMV} ให้ค่า EL_{MV} EL_{AMV} EL_{BMV} และ EL_{BAMV} ที่น้อยและใกล้เคียงกันมาก นอกจากนี้เมื่อค่าสัมประสิทธิ์การแปรผันตัวตัวหนึ่งมีค่าเพิ่มขึ้นจะทำให้ค่าความยาวช่วงโดยเฉลี่ยมีค่าเพิ่มขึ้น

ตารางที่ 1: ค่าความน่าจะเป็นคุ้มรวม (Coverage Probability) ของช่วงความเชื่อมั่นสำหรับผลต่างระหว่างส่วนกลับของค่าเฉลี่ยในกรณีที่ปริภูมิพารามิเตอร์ไม่มีขอบเขตและมีขอบเขต

τ_x	τ_y	ปริภูมิพารามิเตอร์ไม่มีขอบเขต			ปริภูมิพารามิเตอร์มีขอบเขต		
		CP_{GCI}	CP_{MV}	CP_{AMV}	CP_{BGCI}	CP_{BMV}	CP_{BAMV}
0.05	0.05	0.9487	0.9516	0.9516	0.9487	0.9516	0.9516
	0.10	0.9514	0.9514	0.9514	0.9514	0.9514	0.9514
	0.15	0.9476	0.9479	0.9479	0.9476	0.9479	0.9479
	0.30	0.9477	0.9465	0.9466	0.9477	0.9465	0.9466
	0.50	0.9519	0.9474	0.9474	0.9519	0.9474	0.9474
0.10	0.05	0.9505	0.9489	0.9489	0.9505	0.9489	0.9489
	0.10	0.9510	0.9508	0.9508	0.9510	0.9508	0.9508
	0.15	0.9493	0.9481	0.9481	0.9493	0.9481	0.9481
	0.30	0.9455	0.9481	0.9481	0.9455	0.9481	0.9481
	0.50	0.9513	0.9502	0.9499	0.9513	0.9502	0.9499
0.15	0.05	0.9465	0.9547	0.9547	0.9465	0.9547	0.9547
	0.10	0.9475	0.9518	0.9518	0.9475	0.9518	0.9518
	0.15	0.9485	0.9547	0.9547	0.9485	0.9547	0.9547
	0.30	0.9448	0.9506	0.9506	0.9448	0.9506	0.9506
	0.50	0.9455	0.9481	0.9481	0.9455	0.9481	0.9481

ตารางที่ 1 (ต่อ)

τ_x	τ_y	ปริภูมิพารามิเตอร์ไม่มีขอบเขต			ปริภูมิพารามิเตอร์มีขอบเขต		
		CP_{GCI}	CP_{MV}	CP_{AMV}	CP_{BGCI}	CP_{BMV}	CP_{BAMV}
0.30	0.05	0.9513	0.9488	0.9489	0.9513	0.9488	0.9489
	0.10	0.9488	0.9499	0.9499	0.9488	0.9499	0.9499
	0.15	0.9500	0.9491	0.9491	0.9500	0.9491	0.9491
	0.30	0.9476	0.9473	0.9473	0.9476	0.9473	0.9473
	0.50	0.9502	0.9506	0.9506	0.9502	0.9506	0.9506
0.50	0.05	0.9523	0.9486	0.9487	0.9523	0.9486	0.9487
	0.10	0.9512	0.9485	0.9486	0.9512	0.9485	0.9486
	0.15	0.9469	0.9469	0.9470	0.9469	0.9469	0.9470
	0.30	0.9497	0.9528	0.9529	0.9497	0.9528	0.9529
	0.50	0.9524	0.9502	0.9503	0.9524	0.9502	0.9503

ตารางที่ 2: ค่าความยาวโดยเฉลี่ย (Expected Length) ของช่วงความเชื่อมั่นสำหรับผลต่างระหว่างส่วนกลับของค่าเฉลี่ยในกรณีที่ปริภูมิพารามิเตอร์ไม่มีขอบเขตและมีขอบเขต

τ_x	τ_y	ปริภูมิพารามิเตอร์ไม่มีขอบเขต			ปริภูมิพารามิเตอร์มีขอบเขต		
		EL_{GCI}	EL_{MV}	EL_{AMV}	EL_{BGCI}	EL_{BMV}	EL_{BAMV}
0.05	0.05	0.0899	0.0062	0.0062	0.0530	0.0062	0.0062
	0.10	0.0899	0.0078	0.0078	0.0531	0.0078	0.0078
	0.15	0.0900	0.0100	0.0100	0.0531	0.0100	0.0100
	0.30	0.0898	0.0176	0.0176	0.0532	0.0176	0.0176
	0.50	0.0899	0.0284	0.0284	0.0533	0.0281	0.0281
0.10	0.05	0.0899	0.0114	0.0114	0.0532	0.0114	0.0114
	0.10	0.0898	0.0124	0.0124	0.0530	0.0124	0.0124
	0.15	0.0899	0.0139	0.0139	0.0532	0.0139	0.0139

	0.30	-	0.0200	0.0200	-	0.0200	0.0200
	0.50	0.0899	0.0300	0.0300	0.0531	0.0297	0.0297
	0.05	0.0897	0.0169	0.0169	0.0530	0.0169	0.0169
	0.10	0.0897	0.0175	0.0175	0.0532	0.0175	0.0175
0.15	0.15	0.0898	0.0186	0.0186	0.0531	0.0186	0.0186
	0.30	-	0.0236	0.0236	-	0.0235	0.0235
	0.50	-	0.0325	0.0325	-	0.0319	0.0319
	0.05	0.0898	0.0334	0.0334	0.0532	0.0330	0.0330
	0.10	0.0898	0.0338	0.0338	0.0530	0.0333	0.0333
0.30	0.15	0.0898	0.0343	0.0343	0.0530	0.0338	0.0338
	0.30	0.0898	0.0373	0.0373	0.0531	0.0362	0.0362
	0.50	0.0899	0.0435	0.0435	0.0532	0.0405	0.0405
	0.05	0.0898	0.0557	0.0557	0.0533	0.0466	0.0466
	0.10	0.0899	0.0560	0.0560	0.0532	0.0466	0.0466
0.50	0.15	0.0897	0.0563	0.0563	0.0531	0.0468	0.0468
	0.30	0.0899	0.0582	0.0582	0.0532	0.0474	0.0474
	0.50	0.0898	0.0623	0.0623	0.0532	0.0487	0.0487

2.6 อภิปรายและสรุปผลการวิจัย

เมื่อพิจารณาค่าความน่าจะเป็นคุ้มรวมของช่วงความเชื่อมั่น พบว่า ค่าความน่าจะเป็นคุ้มรวมในกรณีปริภูมิพารามิเตอร์ไม่มีขอบเขตและมีขอบเขตให้ค่าเท่ากันในแต่ละวิธี ดังนี้

- ช่วง CI_{GCI} และช่วง CI_{BGCI} จะให้ค่า CP_{BGCI} และ CP_{GCI} เท่ากันสำหรับทุกกรณีของ τ_x และ τ_y
- ช่วง CI_{MV} และช่วง CI_{BMV} จะให้ค่า CP_{MV} และ CP_{BMV} เท่ากันสำหรับทุกกรณีของ τ_x และ τ_y
- ช่วง CI_{AMV} และช่วง CI_{BAMV} จะให้ค่า CP_{AMV} และ CP_{BAMV} เท่ากันสำหรับทุกกรณีของ τ_x และ τ_y

เมื่อพิจารณาค่าความยาวโดยเฉลี่ยของช่วงความเชื่อมั่น พบว่า ค่าความยาวโดยเฉลี่ยของช่วงในกรณีปริภูมิพารามิเตอร์มีขอบเขตให้ค่าน้อยกว่าค่าความยาวโดยเฉลี่ยของช่วงในกรณีปริภูมิพารามิเตอร์ไม่มีขอบเขตในแต่ละวิธี ดังนี้

- ช่วง CI_{GCI} และช่วง CI_{BGCI} จะให้ค่า EL_{BGCI} น้อยกว่าค่า EL_{GCI} สำหรับทุกกรณีของ τ_x และ τ_y
- ช่วง CI_{MV} และช่วง CI_{BMV} จะให้ค่า EL_{BMV} น้อยกว่าค่า EL_{MV} สำหรับทุกกรณีของ τ_x และ τ_y
- ช่วง CI_{AMV} และช่วง CI_{BAMV} จะให้ค่า EL_{BAMV} น้อยกว่าค่า EL_{AMV} สำหรับทุกกรณีของ τ_x และ τ_y

เมื่อพิจารณาตารางที่ 1 และตารางที่ 2 ควบคู่กัน พบว่า แม้ว่าช่วง CI_{MV} CI_{AMV} CI_{BMV} และ CI_{BAMV} จะให้ค่าความน่าจะเป็นคุ้มรวมที่เท่ากัน แต่ค่าความยาวโดยเฉลี่ยของช่วงความเชื่อมั่นดังกล่าวในกรณีปริภูมิพารามิเตอร์มีขอบเขตจะให้ค่าที่น้อยกว่าในทุกกรณีของ τ_x และ τ_y เนื่องมาจากการทราบสารสนเทศเกี่ยวกับพารามิเตอร์มากขึ้น กล่าวคือทราบปริภูมิพารามิเตอร์ จะทำให้ความยาวเฉลี่ยของช่วงความเชื่อมั่นสั้นลง

จากตารางที่ 3 พบว่า ในกรณีปริภูมิพารามิเตอร์ไม่มีขอบเขตและมีขอบเขต ช่วง CI_{MV} CI_{AMV} CI_{BMV} และ CI_{BAMV} เป็นช่วงที่ให้ค่าความน่าจะเป็นคุ้มรวมที่ไม่น้อยกว่าค่าสัมประสิทธิ์ความเชื่อมั่นในทุกค่าสัมประสิทธิ์การแปรผันคิดเป็นร้อยละ 100 นั่นคือ ช่วงโดยวิธี MOVER และ AMOVER มีแนวโน้มที่จะครอบคลุมค่าจริงของพารามิเตอร์มากที่สุด

จากตารางที่ 4 พบว่า ในกรณีปริภูมิพารามิเตอร์ไม่มีขอบเขตและมีขอบเขต ช่วง CI_{MV} CI_{AMV} CI_{BMV} และ CI_{BAMV} เป็นช่วงที่ให้ค่าความยาวโดยเฉลี่ยของช่วงที่ค่อนข้างน้อยมาก และเมื่อพิจารณาช่วงที่ให้ค่าความยาวโดยเฉลี่ยของช่วงสั้นที่สุด โดยร้อยละ 100 จะเป็นช่วงที่มาจากวิธี MOVER และ AMOVER กรณีปริภูมิพารามิเตอร์มีขอบเขตทั้งสิ้น

ตารางที่ 3: ร้อยละของจำนวนกรณีมีประสิทธิภาพ เมื่อพิจารณาค่าความน่าจะเป็นคุ้มรวม

μ_x	μ_y	ร้อยละ					
		ปริภูมิพารามิเตอร์ไม่มีขอบเขต			ปริภูมิพารามิเตอร์มีขอบเขต		
		CI_{GCI}	CI_{MV}	CI_{AMV}	CI_{BGCI}	CI_{BMV}	CI_{BAMV}
5	10	88	100	100	88	100	100

ตารางที่ 4: ร้อยละของจำนวนกรณีมีประสิทธิภาพ เมื่อพิจารณาค่าความยาวช่วงโดยเฉลี่ย

μ_x	μ_y	ร้อยละ					
		ปริภูมิพารามิเตอร์ไม่มีขอบเขต			ปริภูมิพารามิเตอร์มีขอบเขต		
		CI_{GCI}	CI_{MV}	CI_{AMV}	CI_{BGCI}	CI_{BMV}	CI_{BAMV}
5	10	0	44	44	0	100	100

ข้อเสนอแนะ

ในงานวิจัยครั้งนี้ มีข้อเสนอแนะดังนี้

1. ในงานวิจัยนี้สนใจเพียงการหาค่าประมาณแบบช่วงสำหรับผลต่างระหว่างส่วนกลับของค่าเฉลี่ยประชากรปรกติ เมื่อทราบค่าสัมประสิทธิการแปรผัน และปริภูมิพารามิเตอร์มีขอบเขต ที่ระดับนัยสำคัญ 0.05 เท่านั้น จึงเป็นเรื่องที่น่าสนใจที่จะพิจารณาที่ระดับนัยสำคัญ 0.10 และ 0.01 เพื่อช่วงความเชื่อมั่นที่ได้อาจมีประสิทธิภาพมากยิ่งขึ้น

2. เนื่องจากในงานวิจัยนี้ใช้การประมาณค่าพารามิเตอร์ในการอนุมานทางสถิติ จึงเป็นเรื่องที่น่าสนใจที่จะพิจารณาการอนุมานทางสถิติโดยใช้การทดสอบสมมติฐาน เพื่อช่วงความเชื่อมั่นที่ได้อาจมีประสิทธิภาพมากยิ่งขึ้น

กิตติกรรมประกาศ

งานวิจัยนี้สำเร็จลุล่วงไปได้ด้วยดี เนื่องจากความช่วยเหลือของคณาจารย์และเจ้าหน้าที่สาขาวิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ รวมทั้งครอบครัวที่คอยให้การสนับสนุนในทุกๆเรื่องตลอดการทำงานวิจัย สุดท้ายนี้เพื่อนๆทุกคนที่คอยให้กำลังใจจนกระทั่งงานวิจัยเล่มนี้สำเร็จลุล่วงไปได้ด้วยดี

เอกสารอ้างอิง

Lamanna, E., Romano, G., & Sgarbi, C. (1981). Curvature measurements in nuclear emulsions. *Nuclear Instruments and Methods*, 187, 387-391.

Mandelkern, M. (2002). Setting confidence intervals for bounded parameters. *Statistical Science*, 17(2), 149-172.

Panichkitkosolkul, W. (2017). Confidence intervals for the reciprocal of a normal mean with a known coefficient of variation and restricted parameter space. *Pakistan Journal of Statistics and Operation Research*, 13(2), 449-461.

Rohde, C.A. (2014). Introductory Statistical Inference with the Function. *1st ed. London*, Springer.

Wang, H. (2008). Confidence intervals for the mean of a normal distribution with restricted parameter space. *Journal of Statistical Computation and Simulation*, 78(9), 829-841.

Wongkhao, A., Niwitpong, S., & Niwitpong, S. (2013). Confidence interval for the inverse of a normal mean with a known coefficient of variation. *International Journal of Mathematical, Computational, Statistical, Natural and Physical Engineering*, 7(9), 1408-1411.

Wongkhao, A. (2014). Confidence interval for Parameters of Normal. *Ph.D. Thesis, King Mongkut's University of Technology North Bangkok*.

Conflict and Natural Disaster Research Methodology: A Case Study of Aceh, Indonesia

Alisa Hasamoh[1], Stewart Lockie[2], Theresa Petray[3]

[1]Social Development Department, Humanities and Social Science, Prince of Songkla University, Pattani Campus, Thailand
*Corresponding Email: alisa.h@psu.ac.th
[2]The Cairns Institute, James Cook University, Australia
Email: stewart.lockie@jcu.edu.au
[3]College of Arts, Society and Education, James Cook University, Australia
Email: theresa.petray@jcu.edu.au

ABSTRACT

This paper describes and reflects on a research methodology used to study the inter-related social impacts of violent conflict and natural disasters in Aceh in Indonesia. The paper follows three stages in the development and implementation of the methodology. First, it presents the researcher's own experience of both natural disasters and living in a violent conflict zone and how this experience shaped the conceptualization of the research. Second, the paper discusses the importance of building relationships of trust between the researcher and interviewees and how this was approached during the collection of data. Third, the paper examines the question of how researchers' background and techniques of data analysis can be brought together and applied to reflexive, comparative studies. As a result of the existing research experience, the new research methodology and methods of comparative studies in natural disaster on top of violent conflict has emerged.

Keywords: natural disaster; conflict research; methodology; Aceh

Hysteretic Vector Autoregressive Model with Modified t-distribution Errors

Hong Than-Thi[1*], Cathy W.S. Chen[1] and Mike K.P. So[2]

[1]Department of Statistics, Feng Chia University, Taiwan
[*]Corresponding Email: tthong@mail.fcu.edu.tw
Email: chenws@mail.fcu.edu.tw
[2]Department of Information Systems, Business Statistics and Operations Management, Hong Kong University of Science and Technology, Hong Kong, China
Email: immkpso@ust.hk

ABSTRACT

This study proposes a hysteretic vector autoregressive (HVAR) model that provides a new way to understand a nonlinear multivariate model in which the regime switch may be delayed when the hysteresis variable lies in a hysteresis zone. We employ an adapted multivariate Student-t distribution in the (HVAR) model to allow for a higher degree of flexibility in the degrees of freedom for each variable. This study creates this adapted multivariate Student-t distribution from modifying the scale mixture of a normal representation of the multivariate Student-t distribution. We make use of this model to test for a causal relationship between any two target time series. Using posterior odds ratios, we overcome the limitations of the classical approach to multiple testing. Both simulated and real examples help illustrate the suggested methods herein. We apply the proposed HVAR model to investigate the causal relationship between the quarterly GDP growth rates of the U.S. and U.K. and check the lagged dependence among daily PM2.5 levels.

Keywords: hysteresis; nonlinear Granger causality; scale mixture of normal distributions; posterior odds ratio

A Comparison of Linear Regression Models for Heteroscedastic and Non-Normal Data

Raksmey Thinh, Klairung Samart* and Naratip Jansakul

Department of Mathematics and Statistics, Prince of Songkla University Songkhla 90110, Thailand
Email: 5910220104@email.psu.ac.th
*Corresponding email: klairung.s@psu.ac.th
Email: naratip.j@psu.ac.th

ABSTRACT

In common practices, heteroscedasticity and non-normality are frequently encountered when fitting linear regression models. Several methods have been proposed to handle these problems. In this research, we compared four different variance estimation methods: ordinary least squares (OLS), transform both sides (TBS), power of the mean function (POM) and exponential variance function (VEXP), dealing with three different forms of the non-constant variances under four symmetric distributions. In order to study the performance of the four methods in estimating the studied model parameters, a simulation study with various sample sizes; 20, 50, 100, and 200, was conducted. To determine the models with the best fit, bias, mean squared error (MSE) and coverage probability of the nominal 95% confidence interval were obtained. The simulation results suggest that when the true variance is extremely heteroscedastic, TBS is the best method to estimate the linear regression coefficients, while VEXP and POM methods are appealing choices in practices which is reasonably accurate with slightly bias. Moreover, we noted that OLS is acceptable with small bias when the true variance is slightly heteroscedastic.

Keywords: heteroscedasticity; symmetric distribution; variance function; simulation

Weighted D–Optimal Response Surface Designs in the Presence of Block Effects

Peang-or Yeesa[1*], John J. Borkowski[2] and Patchanok Srisuradetchai[1]

[1]Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathum Thani 12121, Thailand
*Corresponding email: peang.or2406@gmail.com
[2]Department of Mathematical Sciences, Montana State University, Bozeman, MT, 59717-2400, USA
Email: jobo@math.montana.edu
Email: patchanok@mathstat.sci.tu.ac.th

ABSTRACT

The purpose of this paper is to help the researcher in finding robust response surface designs. As we cannot ignore uncertainty of the possible reduced models prior to the data collection, the researcher could consider using experimental designs that are robust with respect to a set of potential models obtained from weak heredity (WH). Also blocking effects are incorporated into all possible models in this study. The geometric mean of D-optimalities was proposed as the weighted D-optimality criterion (D_w) to generate a robust design, and the genetic algorithm (GA) and exchange algorithm (EA) were developed to optimize the weighted D-optimality criterion. From 7 to 21 design points with 2 or 3 design variables, robust designs in hypercube were constructed with an appropriate number of blocks. Our scheme for weighting the criteria is to give more weight to a model with larger number of parameters. The resulting designs obtained from GA and EA were compared and the results showed that the GA algorithm is superior to the EA.

Keywords: D-optimality criteria; weak heredity; genetic algorithm

Indonesian Electricity Load Forecasting Using Singular Spectrum Analysis

Subanar[1*], Winita Sulandari[1,2] and Muhammad Hisyam Lee[3]

[1]Department of Mathematics, Universitas Gadjah Mada, Yogyakarta, Indonesia
*Corresponding email: subanar@ugm.ac.id
[2]Study Program of Statistics, Universitas Sebelas Maret, Surakarta, Indonesia
Email: winita@mipa.uns.ac.id
[3]Department of Mathematical Sciences, Universiti Teknologi Malaysia, Skudai, Johor, Malaysia
Email: mhl@utm.my

ABSTRACT

Electricity plays a key role in human life. This study presents several methods to forecast Indonesian electricity load demand and compares the performance of the methods. The hourly load series of Java-Bali for period 1 October to 1 December 2016 shows multiple seasonal patterns, there are different patterns between hour to hour. Singular Spectrum Analysis (SSA) is chosen because of its capability in decomposing the series into two separable components, a combination of cyclist and seasonal series and noise (irregular) components. In the beginning, the forecast values are obtained by SSA-LRF and afterward the irregular component is modeled by the fuzzy and neural network (NN). The forecast values obtained from SSA-LRF are then compared with the forecast values obtained from the combining methods, i.e. SSA-LRF-Fuzzy and SSA-LRF-NN. Based on RMSE and MAPE, the SSA-LRF-NN is the most appropriate method to predict the future values of electricity load series for next 24 hours.

Keywords: electricity; SSA; fuzzy; neural network

A Cross Sectional Assessment of Knowledge, Attitude and Practice toward Smoking among University Students in Malaysia

Busaban Chirtkiatsakul[1,2], Rohana Jani[1*]

[1]Department of Applied Statistics, Faculty of Economics and Administration, University of Malaya, Malaysia
*Corresponding email: rohanaj@um.edu.my
[2]Prince of Songkla University, Thailand
Email: busaban.c@psu.ac.th

ABSTRACT

Tobacco smoking is one of the leading causes of preventable death in the world. Malaysia aims to be a smoke-free nation in 2045. Despite many prevention and health promotion efforts made to develop awareness among the people, it is not only seeing more people smoking but the age of the smokers are getting younger. This paper aimed to identify factors associated with knowledge, attitude and practice toward smoking among students in the University of Malaya, Kuala Lumpur, Malaysia. A cross-sectional study include students aged 18 and above who enrolled at the time of the study. Overall, 843 students were assessed through self-administered questionnaire. The result revealed that there was moderate positive correlation between practice with knowledge ($r = +0.202$, $p < 0.001$), and attitude ($r = +0.239$, $p < 0.001$) toward smoking. There was no factor found to be associated with knowledge. However, statistical analysis indicated a significant association between attitude with gender, current level of study and family income. The results indicated that female had more negative attitude toward smoking than male, students who are studying in postgraduate had more negative attitude toward smoking than those who are studying in diploma or bachelor. Student with family income more than RM 2,500 per month had more negative attitude toward smoking than those with family income less than RM 2,500. In addition, the factors found to be associated with practice were gender, ethnicity, current level of study and family income. This study suggested that students who are non-smoker and were studying in health science or science subjects had negative attitudes higher than other group.

Keywords: knowledge; attitude; practice; smoking

Bayesian Inferences of Two-State Markov Switching Integer-Valued GARCH Models with Applications

Khemmanant Khamthong* and Cathy W.S. Chen

Department of Statistics, Feng Chia University, Taiwan
*Corresponding e-mail: khemmanant@gmail.com

ABSTRACT

This study proposes negative binomial Markov switching and threshold GARCH models for time series with non-negative counts. The characteristics of the proposed integer-valued GARCH models include over-dispersion, consecutive zeros, and switching coefficients for meteorological covariates. We perform parameter estimation and model selection within a Bayesian framework through a Markov chain Monte Carlo (MCMC) scheme. To solve a label switching problem for Markov switching models, we relabel the MCMC outputs to have a higher intensity in state 2, basing model selection is based on the Bayes factor. A simulation study checks the estimation performance of the Bayesian methods. For empirical analysis, we examine weekly time series data of dengue hemorrhagic fever (DHF) cases in four provinces of northeastern Thailand. Our findings reveal that the Markov switching integer-valued GARCH model with meteorological variables can describe DHF case counts better than the threshold integer-valued GARCH model.

Keywords: Bayes factor; over-dispersion, Markov switching; Negative binomial INGARCH model; Threshold INGARCH model; Bayesian inference; MCMC

Modelling the Land Surface Temperature and Its Related Factors: A Case Study in Peninsular Malaysia

Nur Arzilah Ismail[1,2*], Wan Zawiah Wan Zin[1], Choong-Yeun Liong[1], Zamira Hasanah Zamzuri[1], Kamarulzaman Ibrahim[1] and Don McNeil[3,4]

[1]School of Mathematical Sciences, Faculty of Science and Technology,
Universiti Kebangsaan Malaysia, Bangi, Malaysia
[2]Center for Engineering Education Research and Built Environment, Faculty of Engineering and Built Environment,
Universiti Kebangsaan Malaysia, Bangi, Malaysia
*Corresponding Email: nurarzilah@gmail.com
[3]Faculty of Science, Prince Songkhla University, Narathiwat, Thailand
[4]Department of Statistics, Macquarie University, Australia

ABSTRACT

It is vital to understand the climate changes and identifying related factors that contribute to the phenomenon. Elevation, land cover and Normalized Difference Vegetation Index (NDVI) are among factors that commonly studied in relation to the changes of Land Surface Temperature (LST). This paper focuses on explaining the relationship between LST changes and the NDVI. A multidimensional scaling procedure is performed first in order to match the two variables, as their dimensions are different. Then, the NDVI values are categorized into four groups using the recursive partitioning procedure. It is observed that most of the regions that exhibit an increasing pattern in temperature changes are the regions with high initial NDVI and low change in NDVI. It is suspected that the logging activities contributed to this scenario shown by the decrement on the NDVI values. The relationship between the LST changes and elevation, land cover and NDVI are also investigated by fitting the simple linear regression model. Results show that the variation in the LST trends is contributed mostly by the elevation (8%) followed by the NDVI pattern (7.3%) and the land cover (5.7%). By combining the effects of the elevation and NDVI patterns, the overall variation explained by these three variables is increased from 14.7% to 15.9%. We conclude that the elevation, land cover and NDVI patterns have a little contribution to the LST changes; supported by the fact that the overall variation explained is only around 16%.

Keywords: land surface temperature (LST); normalized difference vegetation index (NDVI); elevation; land cover

Learning Model Discrepancy for Dynamical Systems using Gaussian Processes

Kamonrat Suphawan[1*] and Richard Wilkinson[2]

[1]Department of Statistics, Chiang Mai University, Chiang Mai, Thailand
*Corresponding email: kamonrat.s@cmu.ac.th
[2]School of Mathematics and Statistics, Sheffield University, Sheffield, UK
Email: r.d.wilkinson@sheffield.ac.uk

ABSTRACT

Gaussian processes (GPs) can be used as non-parametric models of the simulator discrepancy in dynamical systems in order to correct existing models. Simultaneous inference of simulator parameters and the discrepancy is challenging in systems where the state is not directly observed, as there is confounding between the two different sets of quantities. Here we introduce a methodology to infer both the unknown state and the system parameters when the state dynamics consist of a parametric simulator plus a GP discrepancy model. We use a simulation study to show that we can estimate both the structural error in the dynamics (by modelling the simulator discrepancy) and estimate the unknown system parameter. This can lead to significant improvements in predictive skill.

Keywords: dynamical system**;** simulator discrepancy**;** Gaussian process

Elevation and Land Cover Impact on the Land Surface Temperature in Peninsular Malaysia

Choong-Yeun Liong[1*], Zamira Hasanah Zamzuri[1], Nur Arzilah Ismail[1], Wan Zawiah Wan Zin[1], Kamarulzaman Ibrahim[1] and Don McNeil[2,3]

[1]School of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia
*Corresponding Email: lg@ukm.edu.my
[2]Faculty of Science and Technology, Prince of Songkhla University, Pattani, Thailand
[3]Department of Statistics, Macquarie University, Australia

ABSTRACT

Typically temperature changes are linked to human activitities and the geographical factors on the studied area. The elevation and land cover are often associated with the trend of the Land Surface Temperature (LST). Focusing on these two variables, we map the trend of LST generated by a spline function to the corresponding elevation and land cover of the regions. Results show that 18 superregions in Peninsular Malaysia show a stable trend on the temperature changes (41.9-44.5%), whereas around 32.4 to 34.8% of the subregions show an increasing trend. It is also observed that Brinchang, Pahang, an area with high elevation exhibits an increment in the temperature changes. Other regions identified with the similar pattern are Kuala Krai, Kelantan and Karak, Pahang. Inspecting on the land cover, the forest covers most of the areas in Peninsular Malaysia and there is no apparent pattern observed at the land cover considered; suggesting that land cover may have negligible impact on the temperature changes. In conclusion, we observed that the association between elevation and land cover with the LST trends are described to be not apparent considering the fact that Malaysia's climate is categorized as equatorial, being hot and humid throughout the year.

Keywords: land surface temperature (LST); elevation; land cover

Trends and Patterns of the Land Surface Temperature in Peninsular Malaysia

Zamira Hasanah Zamzuri[1*], Choong-Yeun Liong[1], Nur Arzilah Ismail[1], Wan Zawiah Wan Zin[1], Kamarulzaman Ibrahim[1] and Don McNeil[2,3]

[1]School of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia
*Corresponding Email: zamira@ukm.edu.my
[2]Faculty of Science and Technology, Prince of Songkhla University, Pattani, Thailand
[3]Department of Statistics, Macquarie University, Australia

ABSTRACT

The phenomenon of warming temperature is alarming, resulting on the important need to understand the trends and impacts of this issue in our changing climate. By using the Land Surface Temperature (LST) data from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS), we study 3081 subregions in Peninsular Malaysia. A spline function was fitted to the data to estimate the LST trend for 15 years, from the year 2000. Using principal component analysis, we classify the curves representing the temperature changes into five categories: increasing-accelerating, increasing-decelerating, decreasing-accelerating, decreasing-decelerating and stable. Results show that about 23.5% of the subregions are classified as stable corresponds to the temperature changes. Most of the subregions are classified into group increasing-decelerating (33.7%). Subregions with increasing-accelerating and decreasing-decelerating have almost similar percentage around 19%. The pattern with least number of subregions is decreasing-accelerating with only 5.2%. Subregions with decreasing-decelerating pattern are obviously identified as the subregions with the presence of water nearby, for example located close to the sea, lakes or rivers. Other pattern categories show no distinct characteristics to be identified. The hot spot regions are identified at a few districts in Pahang and Kuala Krai, Kelantan as the temperature changes trend shows an abrupt increment. Overall, the temperature changes in Peninsular Malaysia are quite stable with exception on a few subregions that show accelerated increment over the 15 years.

Keywords: land surface temperature (LST); mapping; spline function

Modelling the Dynamics of the Nutritional Intake of Schoolchildrenin a City in the National Capital Region of the Philippines

Anthony Zosa[1*], Len Patrick Dominic Garces[1,2], Zarah Garcia[3], Normahitta Gordoncillo[3], Joselito Sescon[4], Eden Delight Miro[1], and Lean Frazl Yao[1]

[1]Department of Mathematics, Ateneo de Manila University, Quezon City, Philippines
*Corresponding email: azosa@ateneo.edu
Email: {lgarces, eprovido, lyao}@ateneo.edu
[2]School of Information Technology and Mathematical Sciences, University of South Australia, Adelaide, Australia
Email: lgarces@ateneo.edu
[3]Institute of Human Nutrition and Food, Laguna, University of the Philippines Los Baños, Philippines
Email: {zpgarcia, npgordoncillo}@up.edu.ph
[4]Ateneo de Manila University/Department of Economics, Quezon City, Philippines
Email: jsescon@ateneo.edu

ABSTRACT

Hunger and malnutrition are significant deterrents to school attendance and performance of school-age children. Studies have shown that children who suffer from poor nutrition perform less well in school and will be more likely to drop out of school. One of the more popular method to address this issue is through the implementation of school feeding programs. In 2012, a city in the Philippines launched the first city-wide school feeding program (CSFP) in the country. It now caters to more than 17,000 students daily through a single centralized kitchen. Its modality is in-school feeding for a duration of 120 days which targets undernourished children in Kindergarten to Grade 6. This study aims to analyze the nutritional intake of the beneficiary and non-beneficiary of CSFP. A stratified random sampling clustered respondents into CSFP beneficiaries and non-beneficiaries. Alongside a household survey of the households of randomly selected schoolchildren, the schoolchildren's food intake was assessed. A repeated and assisted 24-hour food recall was employed, which involves a structured interview intended to capture all food, including beverages, that the child-participant consumed the day before the interview. Each household was visited on three non-consecutive days in one week: twice on weekdays and once on a weekend. Data was collected from all households in one month. Two types of models are used to describe the nutritional dynamics of the children in this study. The first type of model consists of conventional methodologies relying on univariate GLM with 2x2x2 way ANOVA and multivariate GLM with MANOVA. The second type employs a regression model to measure the dietary impact of the CSFP. The strengths and weaknesses for each type are compared using the results derived from conventional univariate methods such as descriptive statistics (means and their confidence intervals), one sample t-test, t-test for dependent samples, and independent sample t-tests.

Keywords: malnutrition, school feeding; nutritional intake; impact evaluation

Comparison of Variance Estimation Methods for the Turing-based Geometric Estimator

Orasa Anan[1*] and Wanpen Chantarangsri[2]

[1]Mathematics and Statistics Department, Faculty of Science, Thaksin University, Phattalung Campus, Phattalung, Thailand
[2]Department of Mathematics, Faculty of Science, Naknon Pathom Rajabhat University, Nakhon Pathom, Thailand
[*]Corresponding email: aorasa@tsu.ac.th
Email: wanpen@webmail.npru.ac.th

ABSTRACT

Capture-recapture techniques are very powerful tool and widely used for estimating an elusive target population size. Capture-recapture count data are presented in form of the frequencies data. They consist of the frequency of units detected exactly once, twice, and so on, and the frequency of undetected unites is unknown. As consequence, the resulting distribution is a zero-truncated count distribution. In reality, counting occasions are not known in advance, therefore, the series of frequencies assumed to be the Poisson distribution. In fact, the target population might be heterogeneous, resulting in over or under dispersion based on the basic models. The mixed Poisson, which is the exponential-Poisson mixture model, have been widely used to construct population size estimator for capture-recapture data. The original Turing estimator provides a good performance under the Poisson distribution. Additionally, an extension of Turing estimator, called the Turing-based geometric distribution (TG) approach was proposed for the heterogeneous population. It was recommended as an easy way to estimate the target population size when Capture-recapture data follow the zero-truncated Geometric distribution. In this work, we derived uncertainty measures for the TG estimator by using a conditional technique mixed with the delta-method. It is emphasized that although the analytic approach to compute uncertainty measures can be easily used in practice, it is valid asymptotically and requires a large sample size. Therefore, resampling approaches, true bootstraps imputed bootstrap and reduced bootstrap, are proposed as alternative methods to get uncertain measures. The study compares performance of variance of analytic and resampling method by using the simulation study. Overall, the analytic approach remains successful to estimate variance in the case of large sample size. The imputed bootstrap is the most realistic for the small sample size.

Keywords: capture-recapture; Turing estimator; zero-truncated geometric distribution

Analysis of Entropy for Exponential Distribution under Multiply Type II Censored Competing Risks Data

Kyeongjun Lee[1], Jeayoung Gwag[2] and Nanhee Yun[2*]

[1]Division of Mathematics and Big Data Science, Daegu University, Gyeongsan, Republic of Korea
Email: indra_74@naver.com
[2]Department of Statistics, Daegu University, Gyeongsan, Republic of Korea
Email: gwag5999@naver.com
*Corresponding email: dnflsksl12@naver.com

ABSTRACT

Entropy, which is one of the important terms in statistical mechanics, was originally defined in physics especially in the second law of thermodynamics. It is generally known that the lifetimes of test items may not be recorded exactly. In this paper, therefore, we consider the classical and Bayes estimation of the entropy of an exponential distribution under multiply type II censored competing risks model. It is observed that the MLE of the entropy cannot be obtained in closed form as expected, and we have to be obtained by solving two non-linear equations simultaneously. We further consider the Bayes estimation of the entropy based on fairly flexible priors. The Bayes estimators for the entropy of exponential distribution based on the symmetric and asymmetric loss functions such as squared error loss function (SELF), precautionary loss function (PLF) and linex loss function (LLF) are provided. It is observed that the Bayes estimators cannot be obtained in closed form, and we provide the Lindley approximate method of the Bayes estimates. Monte Carlo simulations are conducted to compare the performances among different estimators considered, and real data sets under multiply type II censored competing risks model are analyzed for illustrative purposes.

Keywords: Bayes estimation; competing risks model; exponential distribution; Lindley approximation method; multiply type II; censored data

Sentiment Analysis of Thai Movie Reviews

Sasimaphon Phromphan, Phimphaka Taninpong* and Weerinrada Wongrin

Department of Statistics Faculty of Science Chiang Mai University, Chiang Mai, Thailand
Email: sasimaphon.p@gmail.com
*Corresponding email: p.taninpong@gmail.com
Email: weerinradaj@gmail.com

ABSTRACT

This research aims to analyze the sentiments of the movie reviews from people who watched the movie in the theater. The movie reviews are collected from the Thai public opinion website, Pantip.com. This study considered only the action film showed in Thailand during December 2017 to January 2018. There are five action movies which are 12 Strong, Black Panther, Den of Thieve, Maze Runner: The Death Cure and the commuter. The classification technique used in this study is Naïve Bayes method. The sentiments of the movie reviews are classified into two groups: positive opinion and negative opinion. The results show that the accuracy of the sentiments classification model of 12 Strong, Black Panther, Den of Thieve, Maze Runner: The Death Cure and the commuter movie reviews are 77.77%, 63.28%, 57.14%, 64.17% and 28.57%, respectively.

Keywords: opinion mining; movie review; pantip.com; sentiment analysis

Extreme Value Distribution for Drought on the Korean Peninsula based on Inter Amount Time

Mihye Kim[1], Hyeju Oh[2], and Sanghoo Yoon[1*]

[1]Division of Mathematics and Big Data Science, Daegu University, Gyeongsan, South Korea
Email: thn03105@naver.com
[2]Department of Statistics, Daegu University, Gyeongsan, South Korea
*Corresponding email: statstar@daegu.ac.kr

ABSTRACT

Drought refers to periods when the water supply is scarce and generally occurs in regions. A drought has an economic and environmental adverse impact on the interaction between water supply and demand. In previous research, drought was evaluated on the number of days without rain. Inter-amount time means the time to reach a certain amount of precipitation. Therefore, the data does not contain 0 values. In addition, it can be used to quantitatively evaluate both floods and droughts. This study assessed drought by region using the time data to fill a small amount between 10mm and 40mm. The data were collected by 63 weather stations operated from 1986 to 2016 and the resolution of time was 1 hour. The block maxima data was applied to the generalized extreme value distribution. The estimated parameters of the generalized extreme model estimate were estimated by L-moments estimation that is suitable for the small sample. The goodness of fit test was performed by Kolmogorov-Smirnov test and Camer-von-Mises test. Finally, the spatial interpolation map of return levels, 25 years, 50 years, and 100 years, were calculated to quantify the risk of drought respectively.

Keywords: inter amount time; drought; extreme value distribution; return level

The Method for Detect Thresholds for Heavy Rainfall Warning System

Yeongeun Hwang, Dayoung Kang and Sanghoo Yoon*

Division of Mathematics and Big Data Science, Daegu University, Gyeongsan, South Korea
Email: hye3775@naver.com
*Corresponding email: statstar@daegu.ac.kr

ABSTRACT

The Korean Meteorological Administration operates a heavy rainfall warning system which is based on the probability rainfall amount using the generalized extreme value distribution. That are 6-hour rainfall is more than 32mm attention, more than 70mm watch, more than 110mm warning, and more than 195mm severe. However, this risk level criterion does not reflect local characteristics. This study dealt with the thresholds for heavy rainfall based on regionality. The thresholds were revised by assessing the risk of heavy rainfall data and observed rainfall for 10 years between 2005 and 2015. In addition, the actual damage causes and observed rainfall data were used for detecting empirical values by the logistic regression model. We hope that the proposed thresholds will help to reduce human and property damage.

Keywords: heavy rainfall; logistic regression; quantile; threshold

Exact Maximum Likelihood Estimation of Parameter under Unified Progressive Hybrid Censored Exponential Model

Kyeongjun Lee[1], Minyeong Han[1] and Wonhee Lee[2*]

[1]Division of Mathematics and Big Data Science, Daegu University, Gyeongsan, Republic of Korea
Email: indra_74@naver.com
Email: gks7052@naver.com
[2]Department of Statistics, Daegu University, Gyeongsan, Republic of Korea
*Corresponding Email: xellos74@gmail.com

ABSTRACT

Recently, progressive hybrid censoring schemes have become quite popular in a life-testing problem and reliability analysis. Exact likelihood estimation methods have been developed under generalized Type I and Type II progressive hybrid censored exponential model. Though these two new schemes of censored sampling are improvements over the old ones, they still face some problems. In this article, we propose an unified progressive hybrid censoring scheme which includes many cases considered earlier as special cases. We then derive the exact distribution of the maximum likelihood estimator as well as exact confidence intervals for the mean of the exponential distribution under this general unified progressive hybrid censored exponential model. Finally, we present some examples to illustrate all the methods of inference developed here.

Keywords: exact maximum likelihood estimation; exponential data; moment generating function, unified progressive hybrid censoring

A Comparative Study of Sinusoidal Model for Oscillatory Component of SSA Decomposition Results on Electricity Load Data

Winita Sulandari[1,2*], Subanar[1], Suhartono[3], Herni Utami[1], and Muhammad Hisyam Lee[4]

[1]Department of Mathematics, Universitas Gadjah Mada, Yogyakarta, Indonesia
*Corresponding email: winita@mipa.ugm.ac.id
[2]Study Program of Statistics, Universitas Sebelas Maret, Surakarta, Indonesia
Email: subanar@ugm.ac.id
[3]Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
Email: suhartono@its.ac.id
Email: herni_utami@ugm.ac.id
[4]Department of Mathematical Sciences, Universiti Teknologi Malaysia, Skudai, Johor, Malaysia
Email: mhl@utm.my

ABSTRACT

Electricity load time series modeling is always interesting. The accurate electricity load forecasting that obtained from the best fit model will provide information to manage power generation efficiencies. Generally, the load data exhibit trend and seasonal behavior. This study considers SSA (Singular Spectrum Analysis)-based forecasting model for the hourly and half hourly electricity load data. The series is decomposed by SSA into signal and noise component. The signal is then decomposed into trend and some oscillatory series. This work focuses on the modeling of the oscillatory series. Since the SSA-based forecasting model is defined by first modeling each component separately and then combining them, the performance of each component model will influence the goodness of fit of the combination model. The purposes of this paper are to compare the performance of the oscillatory models obtained by the iterative OLS and LM method and compare the performance of the deterministic SSA-based signal model that constructed by the linear combination of the trend and several separable oscillatory models with the Polynomial-Fourier model. The performance of the oscillatory models are evaluated via RMSE (Root Mean Square Error) and coefficient of determination (R squared/$\boldsymbol{R^2}$) of the training data. The model of each oscillatory series with the minimum RMSE and $\boldsymbol{R^2}$ will be the chosen one and will be used for further evaluation. The results show that the performance of the sinusoidal models obtained by iterative OLS and LM method tend to be similar and the performance of the deterministic SSA-based signal model is better than the Polynomial-Fourier model.

Keywords: electricity load data; SSA; oscillatory; Fourier

The Optimal Network for Fire Stations in Seoul based on the Density of Fire Incidents

Daeseong Kim, Seungjae Kim, Sanghoo Yoon*

Division of Mathematics and Big Data Science, Daegu Universtiy
*Corresponding email: statstar@daegu.ac.kr

ABSTRACT

A fire station is an infrastructure for suppressing the fire. In the case of urban areas, it is necessary to regularly evaluate whether the location of fire stations are appropriate because the area under the jurisdiction can be changed through the development of large-scale residential land. This study dealt with the optimal network for fire stations between 2015 and 2017 in Seoul by visualization. The fire incidents were collected by public data portal from the Korean government and it was visualized by sampling technique. Because the exact locations of fire were not offered spatial density plot was used to evaluate the risk of fire. Then, the location of fire stations was mashed up for assessing the relevance of the current location. In addition, we identified the commercial, residential, and industrial districts of the 90m resolution land cover map, which is provided by the Ministry of Environment, and the optimal fire station network was suggested by both the distance from the fire station and the density of fire incidents.

Keywords: optimal network; spatial density

Goodness of Fit tests for Multiply Progressively Type II Censored Data from a Gumbel distribution

Subin Cho, Sujeong Chae and Kyeongjun Lee*

Division of Mathematics and Big Data Science, Daegu University, Gyeongsan, Republic of Korea
Email: whtnqls3955@naver.com
Email: chrystal0710@naver.com
*Corresponding email: indra_74@naver.com

ABSTRACT

The Gumbel distribution (type I generalized extreme value distribution) is used to model the distribution of the maximum of a number of samples. This distribution might be used to represent the distribution of the maximum level of a river in a particular year if there was a list of maximum values for the past ten years. The goodness of fit test for Gumbel distribution is very important in natural disaster data analysis. Therefore, we propose the two test statistics to test goodness of fit for the Gumbel distribution under multiply progressive type II censoring. Also, we propose new graphical method to goodness of fit test for the Gumbel distribution under multiply progressive type II censoring. We compare the new test statistic with the Pakyari and Balakrishnan (2013) test in terms of the power of the test through by Monte Carlo method. The new test statistics are more powerful than Pakyari and Balakrishnan (2013) test.

Keywords: goodness of fit test; Gumbel distribution; Lorenz curve; multiply progressive type II censoring

Investigation of Prevalence and Abundance of Fish Fingerling in the Vicinity Power Plant in Tropical Estuarine of Thailand

Sarawuth Chesoh*, Apiradee Lim, and Don McNeil

Faculty of Science and Technology, Prince of Songkla University, Pattani 94000, Thailand
*Corresponding email: sarawuth.c@psu.ac.th

ABSTRACT

Monthly fish fingerling was collected by bongo nets from January 2005 to December 2015 from the Na Thap tidal river in Thailand. Species and amount of fish fingerling were examined and categorized. Logistic model was used to identify factors associated with the prevalence and abundance of fingerling. Factor analysis was used to group fingerling species and hence produced six interpretable factors. Finally, linear regression was used to investigate the association between each of the six factors and determinants (sampling site, season and year) by using the R statistical system. A total of 58 aquatic animal fingerlings were collected in the average density of 60.2 (0–801) individuals per mega litres of water volume. The results of factor analysis revealed that factor 1 was represented by the largest group of 29 marine species. Factors 2 represented a medium-sized group of 13 freshwater fish. Factor 3 represented 6 euryhaline fish species. Factor 4 was a several low salinity brackish water species. Factor 5 was a solely mojarra fish, and factor 6 was ponyfishes and common glassfishes. The overall mean of prevalence of fingerling was 62.8% and associated with month, year and sampling site. The highest prevalence occurred in May (84.2%) and slightly decreased by upstream to the lowest in November (36.8%). The abundance of fingerling in factor 1 reached the maximum peak in April and then steadily decreased in December while species in factor 2 reached the minimum peak in April and steadily increased in December. Factor 3 and 4 were mostly above the overall mean during January to June. Both prevalence and abundance of fingerling were associated with month, year and sampling site and decreased by upstream sites. The finding confirmed seasonal effects of each fish group seasonal moving from coastal zone to downstream and upstream of the tropical tidal river.

Keywords: prevalence; abundance; fish fingerling; logistic model; factor analysis; estuarine

Equivalence of Measurements

Puntipa Wanitjirattikal[1*] and Joshua D. Naranjo[2]

[1]King Mongkut's Institute of Technology Ladkrabang, Statistics, Bangkok, Thailand
*Corresponding email: Puntipa.wa@kmitl.ac.th
[2]Western Michigan University/Statistics, MI, USA
Email: joshua.naranjo@wmich.edu

ABSTRACT

In pharmaceutical and medical studies, we would like to show any formulations or two treatments are equivalent. For example, Westergren ESR and STATplus ESR are two popular measurements of sedimentation rate, which are used to monitor disease severity in patients with rheumatoid arthritis and other inflammatory rheumatologic conditions. Westergren ESR is a well-known measurement that was developed by R. S. Fahraeus and A.V.A. Westergren in 1921, while STATplus ESR is an innovative measurement to accelerate turnaround time. Compared with Westergren ESR, the result from STATplus ESR is easier to understand. Since these two measurements can be used to test the same study, it is necessary to know if they can be switched. A null hypothesis for equality study is $H_0 : \mu_y = \mu_x$. Typically, a new measurement process Y is compared with an existing (or standard) measurement process X. Paired data of these two measurements occur because they are used on the same subject. Usually, paired t-test is appropriate for paired data, but it does not fit well for some situations because paired t-test can only be used to check significant differences from paired data. If the paired data have positive or negative association, the result from the paired t-test might be the same. For example, one paired dataset is $\boldsymbol{X_1}$ = (3,4,5)^T, $\boldsymbol{Y_1}$ = (-5, -2, 1)^T which has positive correlation, and the other one paired dataset is $\boldsymbol{X_2}$ = (3, 4, 5)^T, $\boldsymbol{Y_2}$ = (-1,-2,-3)^T which has negative correlation. But paired t-tests give the same conclusion because they have the same differences. Moreover, the paired t-test might have low power for scale-type relationships with the form $Y = \beta X$ when β is close to 1. In this paper, we propose a test that has reasonable power for both shift and scale-type relationships, which is based on shift-scale type relationships with the form $Y = \beta X + \alpha$. We consider an equivalence testing for hypothesis $H_1 : |\mu_y - \mu_x| = 0$ or $H_1 : |\mu_y - \mu_x| < \Delta$, where Δ is a small amount. It is an approach to swap the hypotheses so that statistical equivalence of the two measurements is the alternative hypothesis and bears the burden of proof. We conclude "equivalence" only if there is evidence to support the claim that the magnitude of disagreement between the two measurements lies within specified limits.

Keywords: equivalence testing; paired t test; two one-sided test

Trends in Coding Education using Social Network Analysis

Jongtae Kim*, Hyoeun Woo and Hyein Koo
Division of Mathematics and Big Data Science, Daegu University, Kyungbuk, South Korea
*Corresponding email: jtkim@daegu.ac.kr
Email: qkql3028@naver.com
Email: ghi3621@gmail.com

ABSTRACT

Recently, social network analysis has played an important role in analyzing big data that uses social networks and graph theory to investigate social structure, and has a characteristic that network structure has nodes and links connecting them. The purpose of this study is to analyze the trend of coding education on the subject of "Korean Coding Education in each field" using social network analysis. In this study, 326 papers on coding education presented in the "Coding Education in the Field" cited in the Korean Journal for the last 5 years were analyzed focusing on the main subject of research subjects and research subjects. The data were collected by classifying the major terms and fields of each paper (Humanities, Social sciences, Natural sciences, Engineering, Medical science, Agricultural and marine, Artistic and Complex sciences). This study would like to propose whether the field of coding education is different and how it can be converged. Furthermore, the recent study on coding education suggested future trends in coding education.

Keywords: coding education; graph; social network analysis; trend analysis.

A Study on the Factors Related to Depression in Adolescent in Korea

Jinseub Hwang*, Jihoon Lee, Hyemin Kwon, Dohyang Kim, Dayoung Yang and Sungmin Hong

Division of Mathematics and Big Data Science, Daegu University, Gyeongsangbuk-do 38453, Republic of Korea
*Corresponding email: hjs0409@daegu.ac.kr
Email: ljhoon01@naver.com
Email: mhk1022k@naver.com
Email: gj1705@naver.com
Email: ydy0093@naver.com
Email: daw0910@naver.com

ABSTRACT

Depression is the most common mental illness in the world, and it is increasing steadily. Depression in an adolescent who is the emotionally unstable is different from that of an adult. They tend to solve depression by suicide. In this study, we identify the factors related to depression in adolescent in Korea and we aim to reduce the suicide by the management of subjects exposed to depression. We use the national sample data which is 'Online Survey of Adolescent's Health Behavior' provided by the Centers for Disease Control and Prevention based on complex survey design. The number of subjects is 54,362 excluding missing value among 62,276 adolescent. To identify the factors related to depression we perform a survey logistic regression analysis considered strata, cluster and weight. As the results, many factors including gender, age, academic scores and so on are significantly related depression and the stress and happiness are the most relevant related factors (OR=4.258, 95% C.I.: 3.872-4.683 & OR=3.980, 95% C.I.: 3.657-4.332, respectively). Based on the results of this study, we hope to be used as basic data for policy development to prevent suicide for adolescents in Korea.

Keywords: adolescent complex survey design depression; logistic regression

Estimation of Stress-Strength Reliability using a Generalization of Power Transformed Half-Logistic Distribution

Thomas Xavier

Email: thomasmxavier@gmail.com

ABSTRACT

A new probability model obtained by generalizing the power transformed half logistic distribution is introduced by transforming the type II beta distribution. The basic properties of the distribution are studied and can be observed that the distribution can be used to model heavy tailed data. Further the expression for stress strength reliability of a single component system with strength following the proposed model and different cases for stress are obtained. Different methods of estimation of parameters, method of moments, quantile estimation and maximum likelihood estimation are also explained. The usefulness of the model is also studied by applying it for a real life data.

Keywords: logistic distribution; type II beta distribution; quantile estimation; maximum likelihood estimation

Organizers

Sponsors