13th Islamic Countries Conference on Statistical Sciences

PROCEEDINGS ICCS-13

Vol. 27

Editors:

Dr. Munir Ahmad & Dr. Shahjahan Khan

STATISTICS FOR BETTER LIFE

ISBN 978-969-8858-16-2

December 18-21, 2014

at

IPB International Convention Centre,

Bogor, Indonesia

Jointly organized by

Islamic Countries Society of Statistical Sciences

Lahore, Pakistan

IPB Bogor Agricultural University

Bogor, Indonesia

ORGANIZERS

44-A, Civic Centre, Sabzazar, Multan Road, Lahore, Pakistan

Email: secretary@isoss.net URL: www.isoss.net

Published by: ISOSS, Lahore, Pakistan.

"All papers published

in the Proceedings of

ICCS-13

were accepted after formal peer review

by the experts in the relevant field"

Dr. Munir Ahmad
Dr. Shahjahan Khan
Editors

CONTENTS

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 1-8

SELF-INVERSION-BASED MODIFICATION OF CROW AND SIDDIQUI'S COEFFICIENT OF KURTOSIS TO ACHIEVE GAINS IN EFFICIENCY

Saleha Naghmi Habibullah
Department of Statistics, Kinnaird College for Women,
Lahore, Pakistan. Email: salehahabibullah@gmail.com
and
Syeda Shan-E-Fatima
Department of Statistics, Government College University
Lahore, Pakistan

ABSTRACT

A random variable X is said to be 'Self-Inverse at Unity' ('SIU') when the reciprocal of X is distributed exactly as X. Habibullah & Saunders (2011) and Fatima et al. (2013) propose self-inversion-based modifications to well-known estimators of the cumulative distribution function and the cumulative hazard function respectively. Fatima & Habibullah (2013a & b) modify the formulae of some well-known estimators of *central tendency and dispersion* for reciprocal-invariant distributions. By performing simulation studies, they demonstrate that the modified formulae are likely to be more efficient than the original formulae when sampling from distributions self-inverse at unity. Habibullah & Fatima (2014a&b) focus on the phenomena of *kurtosis and skewness*, and propose self-inversion-based modifications to (i) the well-known Percentile Coefficient of Kurtosis and (ii) Kelley's Measure of Skewness, respectively. In this paper, we adopt a similar approach for proposing a modification to Crow and Siddiqui's Coefficient of Kurtosis, the proposed formula being applicable when the sample has been drawn from a distribution self-inverse at unity. By carrying out simulation studies based on 1000 samples of various sizes drawn from the SIU version of the Birnbaum Saunders distribution, we show that the proposed modification yields *gains in efficiency* which, obviously, has important implications for *accurate modelling.*

Keywords: Self-Inverse at Unity, Crow & Siddiqui's Coefficient.

1. INTRODUCTION

Habibullah and Saunders (2011) introduce the term "Self-inverse" asserting that a non-negative continuous random variable X will be said to be 'Self-Inverse at β ' when the probability density function of β / X is identical to that of X / β. The case $\beta = 1$ yields "self-inversion at unity" in which case the $(1\text{-}q)^{th}$ quantile of the distribution of the random variable X is the multiplicative inverse of the q^{th} quantile and, as such, the median of the distribution is unity. For some properties of such reciprocal-invariant distributions, see Seshadri (1965), Saunders (1974) and Habibullah et al. (2010).

Habibullah and Saunders (2011) utilize the property of self-inversion at unity to modify the formula of the well-known estimator of the *cumulative distribution function* and demonstrate the usefulness of the modified formula by showing that its sampling distribution is *narrower* than that of the original formula when a large number of samples of a particular size are drawn from a distribution that is self-inverse at unity; Fatima et al. (2013) obtain a similar result by modifying the Nelson Aalen estimator for estimating the *cumulative hazard function*.

Fatima and Habibullah (2013a) propose self-inversion-based modifications of *L-estimators* of three different measures of *central tendency* of reciprocal-invariant distributions, and Fatima and Habibullah (2013b) present self-inversion-based modifications to L-estimators of three different measures of dispersion. Habibullah and Fatima (2014a) focus on the property of peakedness of a distribution self-inverse at unity and propose a modification to the formula of the well-known *percentile coefficient of kurtosis*; Habibullah and Fatima (2014b) propose a modification to the formula of the *Kelley's measure of skewness.* By performing simulation studies based on well-known distributions, Fatima and Habibullah (2013a,b) and Habibullah and Fatima (2014a,b) demonstrate that the proposed modified formulae are *more efficient* than the well-known formulae in the estimation of the corresponding population parameters when sampling from a distribution self-inverse at unity (SIU).

In this paper, we propose a modification to Crow and Siddiqui's coefficient of kurtosis and, by carrying out simulation studies based on 1000 samples of various sizes drawn from the Birnbaum Saunders distribution, we demonstrate the proposed modification yields *gains in efficiency.*

2. MAIN RESULTS REGARDING RECENTLY DEVELOPED SIU-BASED MODIFICATIONS TO ESTIMATORS OF CENTRE, SPREAD, SKEWNESS & KURTOSIS

Fatima and Habibullah (2013a) take up the estimation of measures of *central tendency* of reciprocal-invariant distributions, and propose self-inversion-based modifications to three *L-estimators* of central tendency i.e. the mid-range, mid-hinge and arithmetic mean. Fatima and Habibullah (2013b) focus on the estimation of measures of *dispersion*, and present self-inversion-based modifications to the *range*, *inter-decile range* and *inter-quartile range*. Habibullah & Fatima (2014a) concentrate on the property of peakedness of a distribution self-inverse at unity and put forth a modification to the formula of the well-known *percentile coefficient of kurtosis*. Habibullah & Fatima (2014b) present a modification to the formula of *Kelley's measure of skewness*.

The well-known estimators as well as the proposed modifications are presented in Table 2.1.

Table 2.1
Well-Known Estimators as well as the SIU-Based Modifications to Estimators of Centre, Spread, Skewness & Kurtosis

Estimation of	Well-known formula	Modified Formula
Central Tendency	Mid-Range $= \frac{X_0 + X_m}{2}$	Mid-Range $_{SIU} = \frac{1}{4}\left[(X_0 + X_m) + \left({}^1/_{X_0} + {}^1/_{X_m}\right)\right]$
	Mid-Hinge $= \frac{Q_1 + Q_3}{2}$	Mid-Hinge $_{SIU} = \frac{1}{4}\left[(Q_1 + Q_3) + \left({}^1/_{Q_1} + {}^1/_{Q_3}\right)\right]$
	A.M. $= \frac{\sum_{i=1}^{n} X_i}{n}$	A.M.$_{SIU} = \frac{1}{2n}\left[\sum_{i=1}^{n} X_i + \sum_{j=1}^{n} {}^1/_{X_j}\right]$
Dispersion	Range $= X_m - X_o$	$\text{Range}_{SIU} = \frac{1}{2}\left[(X_m - X_o) + \left(\frac{1}{X_o} - \frac{1}{X_m}\right)\right]$
	$IDR = D_{90} - D_{10}$	$IDR_{SIU} = \frac{1}{2}\left[(D_{90} - D_{10}) + \left(\frac{1}{D_{10}} - \frac{1}{D_{90}}\right)\right]$
	$IQR = Q_3 - Q_1$	$IQR_{SIU} = \frac{1}{2}\left[(Q_3 - Q_1) + \left(\frac{1}{Q_1} - \frac{1}{Q_3}\right)\right]$
Kurtosis	$K = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})}$ $0<K<0.5$	$K_{SIU} = \frac{\left[\left(Q_3 - \frac{1}{Q_3}\right) - \left(Q_1 - \frac{1}{Q_1}\right)\right]}{2\left[\left(P_{90} - \frac{1}{P_{90}}\right) - \left(P_{10} - \frac{1}{P_{10}}\right)\right]}$
Skewness	$Sk_{Kelley} = \frac{P_{90} - 2\tilde{X} + P_{10}}{P_{90} - P_{10}}$	$Sk_{Kelley-SIU} = \frac{\left[P_{90(sample)} + \left(P_{90(sample)}\right)^{-1}\right] - 4 + \left[P_{10(sample)} + \left(P_{10(sample)}\right)^{-1}\right]}{\left[P_{90(sample)} - \left(P_{90(sample)}\right)^{-1}\right] - \left[P_{10(sample)} - \left(P_{10(sample)}\right)^{-1}\right]}$

Through simulation studies based on repeated sampling from well-known distributions possessing the self-inversion at unity property, the authors demonstrate that each of the proposed modified formulae is *more efficient* than the corresponding well-known formula when sampling from an SIU distribution. An outline of the simulation results in the case of the Kelley's Measure of Skewness and its modified version is given in Table 2.2. Here 2000 samples of size 10 each have been drawn from the Birnbaum Saunders distribution with $\alpha = \beta = 1$.

3. SIU-BASED MODIFICATION TO CROW AND SIDDIQUI'S COEFFICIENT OF KURTOSIS

In this paper, we focus on the estimation of the kurtosis of a distribution and propose a modification to the formula of Crow and Siddiqui's coefficient of kurtosis in the case of distributions self-inverse at unity. By carrying out simulation studies based on repeated

sampling from the Birnbaum Saunders distribution for three different choices of sample size, we demonstrate the proposed modification yields *gains in efficiency.*

Table 2.2
Coefficient of Range and Coefficient of Variation of the Sampling Distribution of the Well-Known Kelley's Measure of Skewness as well as the Sampling Distribution of the SIU-Based Modification to Kelley's Measure

	Well-known Formula of Kelley's Measure of Skewness	Modified Formula	Remarks
Minimum Value of Sampling Distribution	-0.2064	0.2549	-------
Maximum Value of Sampling Distribution	0.9315	0.8055	-------
Range of Sampling Distribution (Maximum value minus Minimum value)	1.1379	0.5505	Range of Sampling Distribution of modified formula ***less than*** Range of Sampling Distribution of well-known formula
Coefficient of Range of Sampling Distribution	1.5693 =156.93%	0.5192 =51.92%	Coefficient of Range of Sampling Distribution of modified formula *approximately one-third of* Coefficient of Range of Sampling Distribution of well-known formula
Variance of Sampling Distribution	0.0563	0.0080	Variance of Sampling Distribution of modified formula *less* than Variance of Sampling Distribution of well-known formula
Mean of Sampling Distribution	0.5333	0.6090	-------
Coefficient of Variation of Sampling Distribution	0.4448 =44.48%	0.1467 = 14.67%	Coefficient of Variation of Sampling Distribution of modified formula *approximately one-third of* Coefficient of Variation of Sampling Distribution of well-known formula

3.1 Crow and Siddiqui's Coefficient of Kurtosis and its Modification in the Case of SIU Distributions

The well-known Crow and Siddiqui's coefficient of kurtosis is given by

$$K_{Crow\_\&\_Siddiqui} = \frac{P_{97.5} - P_{2.5}}{Q_3 - Q_1} = \frac{P_{97.5} - P_{2.5}}{P_{75} - P_{25}} \qquad (3.1)$$

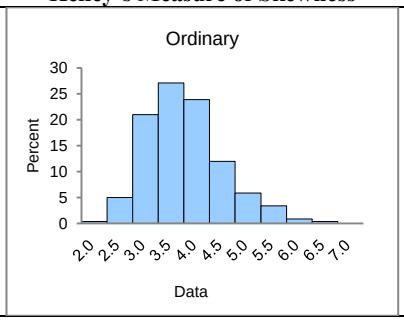
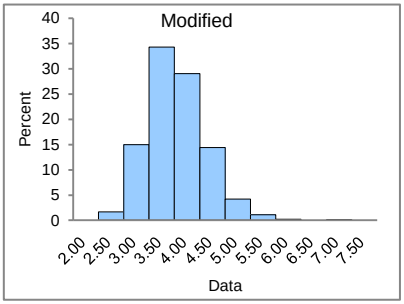
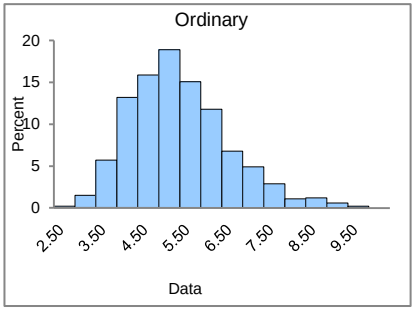
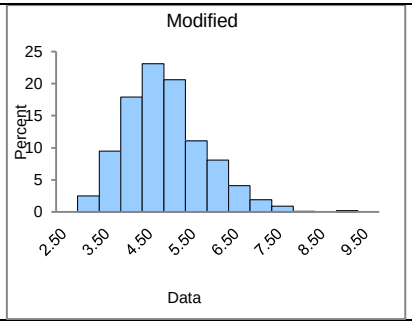
where $P_{97.5}$ denotes the "97.5th" percentile whereas $P_{2.5}$ denotes the "2.5th" percentile. Utilizing the property of self-inversion at unity, we propose the following modification to formula (3.1):

$$K_{SIU\_Crow\_\&\_Siddiqui} = \frac{\left[\frac{\left(1+P_{2.5(sample)}P_{97.5(sample)}\right)\left(P_{97.5(sample)}-P_{2.5(sample)}\right)}{P_{2.5(sample)}P_{97.5(sample)}}\right]}{\left[\frac{\left(1+P_{25(sample)}P_{75(sample)}\right)\left(P_{75(sample)}-P_{25(sample)}\right)}{P_{25(sample)}P_{75(sample)}}\right]} \quad (3.2)$$

3.2 Simulation Study

In this section, we present the results of a simulation study that has been carried out in order to demonstrate that the modified estimator provide *gains in efficiency*. Self-inverse at unity version of the Birnbaum Saunders distribution has been considered for this purpose.

We begin by presenting histograms of the sampling distributions of the Crow and Siddiqui's Coefficient of Kurtosis and the self-inversion-based modified estimator obtained by drawing 1000 samples of size 50 each, an equal number of samples of size 100 each as well as 1000 samples of size 150 each drawn from the Birnbaum Saunders distribution with $\alpha = \beta = 1$.

Sample Size 'n'	**Well-Known Kelley's Measure of Skewness**	**SIU-Based Modification to Kelley's Measure**
50		
100		

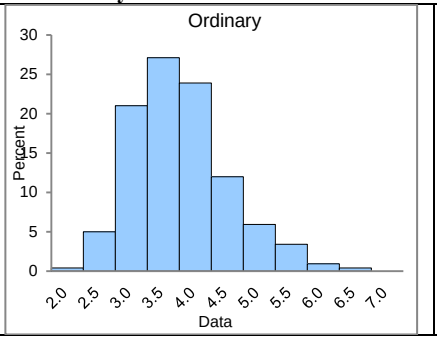
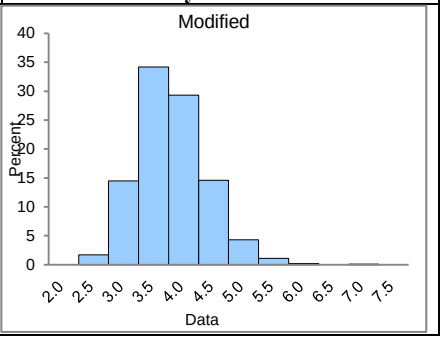
Sample Size 'n'	Well-Known Kelley's Measure of Skewness	SIU-Based Modification to Kelley's Measure
150		

Figure 2.3: Histograms of Sampling Distributions of the Well-Known Kelley's Measure of Skewness as well as the Sampling Distribution of the SIU-Based Modification to Kelley's Measure

Subsections 3.2.1 to 3.2.3 below present a comparison of the sampling distributions of the Crow and Siddiqui's Coefficient of Kurtosis and the self-inversion-based modified estimator based on the coefficients of range and coefficients of variation of the sampling distributions.

3.2.1 Comparison of Coefficients of Range:

Table 3.1 contains maximum, minimum and as well as values of ranges and coefficients of range of the sampling distributions of the well-known Crow and Siddiqui's Coefficient of Kurtosis and the modified estimator.

Table 3.1
Minimum and Maximum values, Ranges and Coefficients of Range of the sampling distributions of Crow and Siddiqui's Coefficient of Kurtosis and Modified when sampling from the Birnbaum Saunders distribution with $\alpha=\beta=1$.

Sample Size n	Well-Known Estimator				Newly Proposed Estimator			
	Minimum	Maximum	**Range**	**Coefficient of Range**	Minimum	Maximum	**Range**	**Coefficient of Range**
50	-1.02	11.51	12.53	1.1945 =119.45%	-1.99	10.98	12.98	1.4438 =144.38%
100	2.87	9.53	6.66	0.5371 =53.71%	3.09	9.41	6.32	0.5056 =50.56%
150	2.40	6.65	4.25	0.4696 =46.96%	2.56	7.17	4.60	0.4728 =47.28%

3.2.2 Comparison of Coefficients of Variation

Table 3.2 contains means, variances and coefficients of variation of the sampling distributions of the well-known Crow and Siddiqui's Coefficient of Kurtosis and the modified estimator.

3.2.3 Discussion and Interpretation

It is interesting to find that, contrary to expectations, the coefficient of range of the sampling distribution of the self-inverse-based modified estimator is not substantially less than that of the well-known Crow and Siddiqui's Coefficient of Kurtosis. However, for each choice of sample size, the histograms of the two sampling distributions

Table 3.2
Means, Variances and Coefficients of Variation of the sampling distributions of Crow and Siddiqui's Coefficient of Kurtosis and Modified Estimator when sampling from the Birnbaum Saunders distribution with $\alpha=\beta=1$.

Sample Size n	Well-Known Estimator			Newly Proposed Estimator		
	Mean	Variance	Coefficient of Range	Mean	Variance	Coefficient of Range
50	4.574	2.324	0.333 =33.3%	4.424	1.173	0.245 =24.5%
100	5.495	1.362	0.212 =21.2%	5.021	0.853	0.184 =18.4%
150	4.001	0.561	0.187 =18.7%	4.035	0.334	0.143 =14.3%

Testify to the well-known assertion that the range is unduly affected by extreme values. As far as the coefficient of variation is concerned, we find that, for each of the three choices of sample sizes, the coefficient of variation of the modified estimator is smaller than that of the well known estimator.

4. CONCLUDING REMARKS

In this paper, we have utilized the property of self-inversion at unity to propose a modification to the formula of Crow and Siddiqui's Coefficient of Kurtosis and, through a simulation study based on repeated sampling from the Birnbaum Saunders distribution, have demonstrated that the modified formula yields an estimator the sampling distribution of which is *narrower* than that of the original estimator in the case when one is sampling from a distribution that is self-inverse at unity. It appears that it may be useful to adopt the proposed formula as an estimator of the kurtosis of the distribution when one has reasons to believe that a single-parameter distribution defined on the positive half-line and invariant under the reciprocal transformation is a suitable probability model for the data at hand, and one is interested in efficient estimation of the distribution-parameter on the basis of the *kurtosis* of the distribution.

REFERENCES

1. Fatima, S.S. and Habibullah, S.N. (2013b). "On Modifications of L-Estimators of Dispersion in the Case of Self-Inverse Distributions", Proceedings of 11th International Conference on Statistical Sciences: Social Accountability, Global Economics and Human Resource Development with Special Reference to Pakistan (Indus International Institute (NCBA&E Sub-Campus), Dera Ghazi Khan, Pakistan, October 21-23, 2013).

2. Fatima, S.S. and Habibullah, S.N. (2013a). "Self-Inversion-Based Modifications of L-Estimators of Central Tendency for Probability Distributions in the Field of Reliability and Safety" *International Conference on Safety, Construction Engineering and Project Management (ICSCEPM 2013), "Issues, Challenges and Opportunities in Developing Countries"* organized by NUST August 19-21, 2013, Islamabad, Pakistan.
3. Fatima, S.S., Habibullah, S.N. and Saunders, S.C. (2013). "Some Results Pertaining to the Hazard Functions of Self-Inverse Life-Distributions"; S. N. Habibullah invited to render Oral Presentation of this paper as KEYNOTE SPEAKER at the Third International Conference on Aerospace Science and Engineering (ICASE 2013) (Islamabad, Pakistan, August 21-23, 2013) organized by Institute of Space Technology (IST) Islamabad, Pakistan.
4. Habibullah, S.N. and Fatima, S.S. (2014b). "SIU-Based Modification in Kelley's Measure of Skewness to Achieve Gains in Efficiency" presented at the Second ISM International Statistical Conference 2014 with Applications in Sciences and Engineering (ISM II) organized by Faculty of Industrial Sciences and Technology, Universiti Malaysia Pahang, Kuantan, Malaysia on Aug 12-14, 2014; *Sponsoring Agency: Higher Education Commission, Pakistan.*
5. Habibullah, S.N. and Fatima, S.S. (2014a). "On Modification of Estimator of the Percentile Coefficient of Kurtosis in the Case of Distributions Self-Inverse at Unity" submitted to 12th International Conference on Statistical Sciences: Application of Statistics in Policy Development and Monitoring of Health, Finance, Education, Information Technology and Economics held at Dow University of Health Sciences, Karachi, Pakistan, March 24-26, 2014.
6. Habibullah, S.N. and Saunders, S.C. (2011). "A Role for Self-Inversion", *Proceedings of International Conference on Advanced Modeling and Simulation (ICAMS, Nov 28-30, 2011)* published by Department of Mechanical Engineering, College of Electrical and Mechanical Engineering, National University of Science and Technology (NUST), Islamabad, Pakistan, Copyright 2011, ISBN 978-869-8535-11-7.
7. Habibullah, S.N., Memon, A.Z. and Ahmad, M. (2010). *On a Class of Distributions Closed Under Inversion*, Lambert Academic Publishing (LAP), ISBN 978-3-8383-4868-1.
8. Saunders, S.C. (1974), "A Family of Random Variables Closed Under Reciprocation", *J. Amer. Statist. Assoc.*, 69(346), 533-539.
9. Seshadri, V. (1965), "On Random Variables which have the Same Distribution as their Reciprocals", *Can. Math. Bull.*, 8(6), 819-824.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 9-16

MODIFICATION IN KAPPA STATISTICS-A NEW APPROACH

Sundus Iftikhar
Karachi University, Karachi, Pakistan
Email: sundusiftikhar@gmail.com;

and

Nazeer Khan
Jinnah Sindh Medical University, Karachi, Pakistan
Email: nazeerkhan54@gmail.com

ABSTRACT

Assessing the strength of agreement among clinicians is one of the main concerns of medical researchers. In medical science scoring methods are commonly used for assessing certain deformities. Development of valid scoring systems requires substantial agreement between the raters. The Kappa statistics is one of the most widely used statistical technique for assessing the degree of inter-rater agreement for qualitative items. Despite of its popularity, kappa statistics has some serious weaknesses due to which it has been replaced by AC1 statistics proposed by Kilem Gwet in 2001. It has been proved that Gwet's AC1 is more stable and less affected by prevalence and marginal probability than Cohen's Kappa.

It is noted by the authors of this paper that for same sample size, same overall agreement and same off diagonal numbers but different diagonal numbers the magnitude of Gwet'sAC1 statistics changes considerably. However, any agreement statistics should provide approximately similar results. On this basis we have proposed a new and simpler formula ***"SNI statistics"*** based on minimum expected agreement which is more stable in comparison to both agreement statistics.

Key words: Examiners' reliability, Cohen's Kappa, Gwet'sAC1, Adjusted Kappa,

INTRODUCTION

Assessing agreement between independent raters appraising same rating system is of very much important concern in medical and social sciences. Scoring or rating systems are widely use in medical sciences in order to judge whether patient has particular disease or not. Inter-rater agreement also helps in evaluation of reliability of such scoring and rating systems. Cohen's Kappa is the most widely used statistical method for evaluating inter-rater agreement. But it has already been shown that Kappa statistics is not satisfactory. In 2002, Kilem Gwet has shown that Kappa statistics is greatly affected by marginal probabilities and prevalence (Gwet 2002). Considering, unreliability of Kappa statistics Gwet proposed alternative method AC1 statistics which is proved to be more stable and less affected by prevalence and marginal probabilities in comparison to Kappa statistics(Gwet 2002, Wongpakaran, Wongpakaran et al. 2013).

Kilem Gwet recommended the use of chance-independent agreement which inflates the overall probability of agreement. But we noted that for same sample size, same overall agreement and same off-diagonal numbers but different diagonal numbers, the magnitude of Gwet'sAC1 statistics changes considerably though it should not.

With the intention offixing this issue we have proposed a very simple and easy to use formula named ***"SNI statistics"*** based on minimum expected agreement which is more stable and remains approximately the same for same sample size and for same overall agreement but different diagonal numbers. In this paper we have compared the results of this new formula with both Cohen's Kappa and Gwet'sAC1 statistics.

STATISTICAL ANALYSIS

Cohen's Kappa and Gwet'sAC1 statistics were calculated using STATSDIRECT software. Manual calculation was also performed.

Cohen's kappa:

$$\mathcal{K} = \frac{P - e_K}{1 - e_K}$$

where P is the percent of observed agreement

$$P = \frac{\sum_{i=1}^{q} the\ number\ of\ times\ both\ raters\ classify\ a\ subject\ into\ category\ i}{Total\ numebr\ of\ obsertavions(N)}$$

and e_k is chance agreement probability calculated as

$$\sum_{i=1}^{q} \left(\frac{Total\ no.of\ obs.in\ category\ i\ of\ rater\ A}{Total\ numebr\ of\ obsertavions(N)}\right) X \left(\frac{Total\ no.of\ obs.in\ category\ i\ of\ rater\ B}{Total\ numebr\ of\ obsertavions(N)}\right)$$

AC1 Statistics:

$$\mathcal{K} = \frac{P - e_\gamma}{1 - e_\gamma}$$

where, e_γ is the chance-independent agreement

FOR 2X2 CONTINGENCY TABLE

e_γ is calculated as 2Q(1-Q)

$$Q = \left(\frac{Total\ no.of\ obs.in\ category\ 1\ of\ rater\ A + Total\ no.of\ obs.in\ category\ 1\ of\ rater\ B}{2N}\right)$$

FOR 3X3 CONTINGENCY TABLEOR MORE

e_γ is calculated as $\frac{\sum_{i=1}^{q} Q_i(1-Q_i)}{Total\ number\ of\ categories\ (q)-1}$

where *i* is the *i*th number of category and *i*=1,2,….,q

$$Q_i = \left(\frac{Total\ no.of\ obs.in\ category\ i\ of\ rater\ A + Total\ no.of\ obs.in\ category\ i\ of\ rater\ B}{2N}\right)$$

SNI STATISTICS:

$$\mathcal{K}_{adj} = \frac{P - e_v}{1 - e_v}$$

and e_v is the expected minimum agreement calculated as

$$\frac{Average(Min.sample\ size\ of\ cat1, Min.sample\ size\ cat2, \ldots\ldots Min.sample\ size\ of\ catq)}{Total\ numebr\ of\ obsertavions(N)}$$

RESULTS

We have tried to explain the differences between the three agreement statistics with the help of examples. In example 1, 2 and 3 we have taken four 2x2 tables, 3x3 tables and 4x4 tables respectively with same sample size, same overall observed agreement and same off-diagonal elements but different diagonal numbers. From both the examples it can be seen that the magnitude of Gwet'sAC1 statistics changes noticeably but the SNI statistics remained the same.

Statisticians are always interested in parsimonious models and formulas. SNI statistics is parsimonious formula and quite easy to use.

EXAMPLE 1:2X2 CONTINGENCY TABLES

Consider the following four 2x2 contingency tables with same sample size, same overall observed agreement but different diagonal numbers

		Rater B		
		1	2	Total
Rater A	1	15	1	16
	2	1	2	3
	Total	16	3	19
P	0.8947	Cohen's Kappa		0.6042
e(ϒ)	0.2659	Gwet'sAC1		0.8566
e(k)	0.7341	SNI statistics		0.7895
e(ν)	0.5000			

		Rater D		
		1	2	Total
Rater C	1	5	1	6
	2	1	12	13
	Total	6	13	19
P	0.8947	Cohen's Kappa		0.7564
e(ϒ)	0.4321	Gwet'sAC1		0.8146
e(k)	0.5679	SNI statistics		0.7895
e(ν)	0.5000			

	Rater F			
		1	2	Total
Rater E	1	10	1	11
	2	1	7	8
	Total	11	8	19
P	0.8947	Cohen's Kappa		0.7841
e(ϒ)	0.4875	Gwet'sAC1		0.7946
e(k)	0.5125	SNI statistics		0.7895
e(ν)	0.5000			

	Rater H			
		1	2	Total
Rater G	1	16	1	17
	2	1	1	2
	Total	17	2	19
P	0.8947	Cohen's Kappa		0.4412
e(ϒ)	0.1884	Gwet'sAC1		0.8703
e(k)	0.8116	SNI statistics		0.7895
e(ν)	0.5000			

EXAMPLE 2: 3X3 CONTINGENCY TABLES

	Rater B				
		1	2	3	Total
Rater A	1	5	3	2	10
	2	2	3	4	9
	3	0	2	3	5
	Total	7	8	9	24
P	0.4583	Cohen's Kappa			0.1979
e(ϒ)	0.3320	Gwet'sAC1			0.1891
e(k)	0.3247	SNI statistics			0.2500
e(ν)	0.2778				

	Rater D				
		1	2	3	Total
Rater C	1	0	3	2	5
	2	2	11	4	17
	3	0	2	0	2
	Total	2	16	6	24
P	0.4583	Cohen's Kappa			-0.1064
e(ϒ)	0.2391	Gwet'sAC1			0.2881
e(k)	0.5104	SNI statistics			0.2500
e(ν)	0.2778				

	Rater F				
		1	2	3	Total
Rater E	1	1	3	2	6
	2	2	10	4	16
	3	0	2	0	2
	Total	3	15	6	24
P	0.4583	Cohen's Kappa			-0.0196
e(ϒ)	0.2600	Gwet'sAC1			0.2680
e(k)	0.4688	SNI statistics			0.2500
e(ν)	0.2778				

	Rater H				
		1	2	3	Total
Rater G	1	11	3	2	16
	2	2	0	4	6
	3	0	2	0	2
	Total	13	5	6	24
P	0.4583	Cohen's Kappa			0.0429
e(ϒ)	0.2773	Gwet'sAC1			0.2505
e(k)	0.4340	SNI statistics			0.2500
e(ν)	0.2778				

EXAMPLE 3: 4X4 CONTINGENCY TABLES

		Rater B				
		1	2	3	4	Total
Rater A	1	21	12	0	0	33
	2	4	17	1	0	22
	3	3	9	15	2	29
	4	0	0	0	1	1
	Total	28	38	16	3	85
P	0.6353	Cohen's Kappa				0.4728
e(v)	0.1971	Gwet'sAC1				0.5292
e(k)	0.3082	SNI statistics				0.5458
e(ϒ)	0.2254					

		Rater B				
		1	2	3	4	Total
Rater A	1	8	12	0	0	20
	2	4	2	1	0	7
	3	3	9	16	2	30
	4	0	0	0	28	28
	Total	15	23	17	30	85
P	0.6353	Cohen's Kappa				0.5133
e(v)	0.1971	Gwet'sAC1				0.5172
e(k)	0.2507	SNI statistics				0.5458
e(ϒ)	0.2445					

		Rater B				
		1	2	3	4	Total
Rater A	1	26	12	0	0	38
	2	4	0	1	0	5
	3	3	9	28	2	42
	4	0	0	0	0	0
	Total	33	21	29	2	85
P	0.6353	Cohen's Kappa				0.4331
e(v)	0.1971	Gwet'sAC1				0.5388
e(k)	0.3567	SNI statistics				0.5458
e(ϒ)	0.2092					

		Rater B				
		1	2	3	4	Total
Rater A	1	0	12	0	0	12
	2	4	54	1	0	59
	3	3	9	0	2	14
	4	0	0	0	0	0
	Total	7	75	1	2	85
P	0.6353	Cohen's Kappa				0.0248
e(v)	0.1971	Gwet'sAC1				0.5858
e(k)	0.6260	SNI statistics				0.5458
e(ϒ)	0.1194					

CONCLUSION

SNI statistics is a parsimonious, more stable, simpler and easy to use formula as compared to the other two agreement statistics.

REFERENCES

1. Gwet, K. (2002). "Kappa statistic is not satisfactory for assessing the extent of agreement between raters." Statistical Methods for Inter-rater Reliability Assessment, 1(6), 1-6.
2. Wongpakaran, N., et al. (2013). "A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples." BMC medical research methodology, 13(1), 61.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 17-28

DEVELOPMENT OF GROWTH CHARTS OF PAKISTANI CHILDREN USING QUANTILE REGRESSION

Sundus Iftikhar
Karachi University, Karachi, Pakistan
Email: sundusiftikhar@gmail.com

and

Nazeer Khan
Jinnah Sindh Medical University, Karachi, Pakistan
Email: nazeerkhan54@gmail.com

ABSTRACT

BACKGROUND: Different countries have developed their separate growth charts so that health practitioners can monitor growth pattern of the children using only one national standard. ***OBJECTIVE:*** To develop the growth charts of children of Pakistan using quantile regression and Cole's LMS method and compare the results with WHO anthropometric standards. ***METHODOLOGY:*** This study is a part of a larger national clinical survey for *Time of eruption of permanent teeth*, covering all the provincial capitals of Pakistan and Larkanacity. A total sample of 9515 students was collected; 4370 from Karachi, 1324 from Larkana, 1267 from Quetta and 2554 from Peshawar. The data of height, weight and age along with other required dental information were collected from each child. Quantile regression, Cole's LMS method and WHO anthropometric system were used to develop the growth chart of height-, weight- and BMI-for-age. ***RESULTS:*** The percentile values for BMI-for-age, Height-for-age and Weight-for-age are comparable for both Quantile regression and Cole's LMS method; however the WHO percentiles differ widely with these two methods resulting that the WHO growth standards are not applicable for Pakistani children. ***CONCLUSION:*** Study showed that WHO growth chart is not suitable for Pakistani children. Furthermore, quantile regression can be used in developing a nation growth charts for Pakistani children.

KEYWORDS

Quantile regression, Cole's LMS method, Growth charts, WHO growth charts, Pakistani children

INTRODUCTION

Historically Anthropometry (measurement of the human individual) was the interest of scientists from very early age. Normal growth pattern was measured and plotted in the form of growth chart. However, continuous update of these growth charts is essential to see any change due to dieting habits and environmental factors (Karlberg, 1999). These growth charts are very useful to determine the physiological needs of the children

(Borghi, 2006) and help in academic and planning to the field of medicine, education, pharmaceutical industries and government agencies (Hensinger 1998). Studies have been conducted to develop the growth charts almost all over the world, e.g. USA (Hedley, 2004), UK (Wright, 2008), Italy (Cacciari, 2002), Turkey (Neyazi, 2006), India (Tarozzi, 2008) and Pakistan (Kelly, 1997; Kamal, 2004; Aziz, 2012; Mushtaq, 2012). In addition to national studies multi-countries studies have also been conducted to develop combined growth chart, such as de Onis(2006)article covered Norway, Brazil, Ghana, India, Oman and USA; Onayngo (2007)article involved Argentina, Italy, Maldives and Pakistan; and Bonthuis (2012) article comprised the data of 28 European countries.

As so many studies have been conducted for growth charts for children, so as many different statistical techniques have been applied to develop the growth charts and smoothing for the irregularities in the curves. A detailed review has be presented by Borghi (2006).Methods have been developed of taking care of first four moments (mean, SD, skewness and kurtosis) for estimating the centiles. However, quantile regression method has not been given much attention. In this paper authors have used a multicenter countrywide data of Pakistan to develop the growth curve by using quantile regression and compared the results with WHO standard percentile and LMS method [Box-Cox transformation (L), the median (M), and the generalized coefficient of variation (S)] (Flagel, 2013).

METHODOLOGY

This study is a part of a larger national clinical survey for "*Time and sequence of eruption of permanent teeth*". This cross-sectional study is approved and funded by Higher Education Commission, Islamabad. A total sample of 9515 students was collected using systematic random sampling; 4370 from Karachi, 1324 from Larkana, 1267 from Quetta and 2554 from Peshawar. The data was entered in MS Excel. Quantile regression was applied using SAS 9.2 software and Cole's LMS method was applied using STATA 12. Investigator-administered questionnaire was used to obtain information from the children who had 'just erupted teeth' about age, gender, height and weight along with other required dental information. The age range was 4 to 15 years. Permission to conduct this survey in the respective schools was obtained from Principal. Institutional Review Board of Dow University of Health Sciences has approved this study for ethical consideration.

Procedure: Same methodology has been applied in all the four study centers. Detailed methodology for Karachi center has been given in Khan (2011) and Khan (2012). Briefly, 102 schools were randomly selected from 6,508 schools of Karachi, 24 schools were selected from Quetta and 15 schools were selected from Larkana and 20 schools were chosen from Peshawar. The clinicians was trained and calibrated by a senior Pedodontist, who has been involved in many such studies (Chohan 2007). Kappa statistic was used to find inter-examiner reliability between examiners (field dentists) and with the reference (trainer) examiner.

A team of dentists (1 male and 1 female) and assistants (1 male and 1 female) were hired and the objectives and methodology were explained in detail. Consent form was sent to the parents before the day of examination. All the present students were examined

for the general checkup, if the parents have given the consent and the child is Pakistani. Among those students, if a child has have just erupted tooth, then the child was taken away from the class. The selected child was clinically examined for the eruption of teeth, caries experience and hygiene index and some other information regarding the dieting habits was collected. The height and weight are also measured for each child. The date of birth was obtained from the school record. The criterion of just erupted teeth was defined as: a tooth deemed to have emerged if any part of it was visible in the mouth.

RESULTS

Among 8373 school going children, 4379(52.3%) were males and 3976(47.5%) were females. Mean (SD) age was 9.3 (2.3) years. The Mean (SD) height, weight and BMI were, 130.56 (13.86) cm, 27.88(9.87) Kg and 15.97 (3.64) Kg/m^2 respectively (Table 1). BMI-for-Age, Height-for-Age and Weight-for-Age were computed for boys and girls, using quantile regression, LMS method, WHO-percentile and WHO-standard percentile (Table 2 – Table 19).

The WHO percentiles calculated using free software WHO anthroplus differs significantly from the centiles charts developed using Quantile regression and Cole's LMS method; whereas visually small differences exists between the centiles of Quantile regression and Cole's LMS method. It has been observed that children who fall in extreme categories of WHO are in normal ranges of the other two methods.

Due to significant differences in WHO growth references and the centiles developed by the other two methods revealed that using WHO standard growth references for our population may place children at risk of misdiagnosis; inferring that WHO growth charts are not suitable for Pakistani children.

CONCLUSION

Study showed that the anthropometric measurements of Pakistani children are incomparable with WHO standard references and are not suitable for our population. Furthermore, quantile regression can be used as an alternative method to develop growth charts.

REFERENCES

1. Aziz S, Noor-ul-Ain W, Majeed R, Khan MA, Qayum I, Ahmed I, Hosain K (2012). Growth centile charts (anthropometric measurement) of Pakistani pediatric population. *J Pak Med Assoc*, 62(4): 347-77.
2. Bonthuis M, van Stralen KJ, Verrina E, Edefonti A, Molchanova EA, et al (2012). Use of National and International Growth Charts for Studying Height in European Children: Development of Up-To-Date European Height-For-Age Charts. *PLoS ONE*, 7(8): e42506. doi: 10.1371/journal.pone.0042506,
3. Borghi E, de Onis M, Garza C, Van den Broeck J, Frongillo JE et al (2006). Construction of the World Health Organization child growth standards: selection of methods for attained Growth curves. *Statist. Med*.; 25:247–265.

4. Cacciari E, Milani S, Balsamo A, Dammacco F, De Luca F, Chiarelli F, Pasquino AM, Tonini G, Vanelli M (2002). Italian cross-sectional growth charts for height, weight and BMI (6–20y). *Eur J Clin Nut,* 56(2):171-80.
5. de Onis M, Onayango AW, Borghi E, Garza C, Yong H (2006). Comparison of the World Health Organization (WHO) child growth standards and the National Center for Health Statistics/WHO international growth reference: implications for child health programmes. *Public Health Nutr*, 9(7): 942-7.
6. Flegal KM, Cole TJ (2013). Construction of LMS Parameters for the Centers for Disease Control and Prevention 2000 Growth Charts. *National Health Statistics* Report; 63 (198.246.124.22):
7. Hedley AA, Ogden CL, Johnson CL, Carroll MD, Curtin LR, Flegal KM (2004). Prevalence of Overweight and Obesity among US Children, Adolescents, and Adults, 1999-2002. J Am Med Assoc, 291(23): 2847-50.
8. Hensinger RN (1998). The challenge of growth: the fourth dimension of pediatric care, *J PedOrthop,* 18: 141-144.
9. Kamal SA, Firdous S, Alam SJ (2004). An investigation of the growth profiles of Pakistani children. *Int J Biol Biotech*, 1(4): 709-717.
10. Karlberg J, Cheung YB, Luo ZC (1999). An update on the update of the growth charts. *Acta Pedistrica*, 88: 797-802.
11. Kelly AM, Shaw NJ, Thomas AMC, Pynsent PB, Baker DJ (1997). Growth of Pakistani children in relation to the 1990 growth standards. *Arch Dis Child*, 77(5): 401-405.
12. Mushtaq MU, Gull S, Mushtaq K, Abdullah HM, Khurshid U, Shahis U, Shad MA, Akram J (2012). Height, weight, and BMI percentile and nutritional status relative to the international growth references among Pakistani school-aged children. BMC Pediatric, 12; 31-41.
13. Neyazi O, Furman A, Bundak R, Gunoz H, Darendeliler F, Bas F (2006). Growth references for Turkish children aged 6 to 18 years. Acta Paediatr, 95(12), 1635-41.
14. Onyango AW, de Onis M, Caroli M, Shah U, Sguassero Y, Redondo N, Berenise B (2007). Field-Testing the WHO child growth standards in four countries. J. Nutr, 137: 149-152.
15. Tarozzi A (2008). Growth reference charts and the nutritional status of Indian children. *Econ Hum Biol*, 6(3): 455-68. doi: 10.1016/j.ehb.2008.07.004. Epub 2008 Jul 26.
16. Wright C, Lakshman R, Emmett P, Ong KK (2008). Implications of adopting the WHO 2006 Child Growth Standard in the UK: two prospective cohort studies. *Arch Dis Child*, 93: 566-569. doi:10.1136/adc.2007.126854.
17. Khan N (2011). Eruption time of permanent teeth in Pakistani children. Iranian J Publ Health. 40(4): 63-73.
18. Khan N (2012). Time and sequence of eruption of permanent teeth in Pakistani children: First database of Pakistani children. Lambert Academic Publication, Deutschland, Germany.
19. Chohan AN, Khan NB, Al Nahedh L, Bin Hassan M, Al Sufyani N. Eruption time of permanent first molars and incisors among female primary school children of Riyadh. J Dow Univ Health Sc 2007; 1(2): 53-58.

Table 1
Descriptive Statistics for Demographic Information

	Overall	Male	Female
	n=8373	**n=4397**	**n=3976**
Age (years)			
Mean±SD	9.3±2.30	9.14±2.21	9.14±2.21
Median(IQR)	9(7-11)	9(7-11)	9(7-11)
Min-Max	4-15	4-15	4-15
BMI (kg/m$^{2)}$			
Mean±SD	15.97±3.64	16.33±4.04	15.54±3.09
Median(IQR)	15.31 (13.85-17.36)	15.53 (13.89-17.84)	15.09 (13.72-16.83)
Min-Max	5.94-59.49	7.06-59.49	5.94-51.44
Height (cm)			
Mean±SD	130.56±13.86	131.63±14.16	129.37±13.42
Median(IQR)	130 (121-140)	131 (121-141)	129 (120-139)
Min-Max	78-192	78-192	78-192
Weight (kg)			
Mean±SD	27.88±9.87	29.02±10.63	26.62±8.80
Median(IQR)	26 (20-33)	27 (21-35)	25 (20-31)
Min-Max	7-95	7-95	7-94

Table 2:
BMI-FOR-AGE using Quantile regression
(Total sample size=8371; 3976 Females and 4395 males)

Age in years	Age in months	Quantile regression-Percentiles (BMI kg/m^2) For Girls										
		1%	**3%**	**5%**	**15%**	**25%**	**50%**	**75%**	**85%**	**95%**	**97%**	**99%**
4	48	7.6	9.3	9.9	11.8	12.4	13.5	14.6	15.5	17.1	17.7	20.0
5	60	8.1	9.6	10.3	12.0	12.6	13.8	15.0	16.0	17.8	18.5	20.9
6	72	8.5	10.0	10.6	12.3	12.9	14.1	15.5	16.5	18.5	19.2	21.8
7	84	8.9	10.4	11.0	12.5	13.2	14.5	15.9	17.0	19.2	20.0	22.7
8	96	9.3	10.8	11.3	12.8	13.5	14.8	16.3	17.4	19.8	20.8	23.6
9	108	9.8	11.2	11.7	13.1	13.7	15.1	16.7	17.9	20.5	21.5	24.5
10	120	10.2	11.6	12.1	13.3	14.0	15.4	17.2	18.4	21.2	22.3	25.3
11	132	10.6	12.0	12.4	13.6	14.3	15.8	17.6	18.9	21.9	23.1	26.2
12	144	11.1	12.4	12.8	13.9	14.5	16.1	18.0	19.4	22.6	23.8	27.1
13	156	11.5	12.8	13.2	14.1	14.8	16.5	18.5	19.8	23.2	24.6	28.0
14	168	12.0	13.2	13.5	14.4	15.1	16.8	19.0	20.4	24.0	25.5	29.0
15	180	12.3	13.5	13.9	14.6	15.4	17.1	19.3	20.8	24.6	26.1	29.8

Table 3:
BMI-FOR-AGE using LMS method

Age in years	Age in months	Percentiles using LMS Method for BMI (Kg/m^2)-Girls										
		1%	3%	5%	15%	25%	50%	75%	85%	95%	97%	99%
4	48	-	-	-	-	-	-	-	18.0	21.0	-	-
5	60	9.0	10.0	10.5	-	-	15.0	-	17.0	-	19.0	-
6	72	8.7	10.0	10.0	11.0	12.0	14.0	16.0	18.0	20.0	21.0	24.0
7	84	9.0	10.0	11.0	12.0	13.0	14.0	16.0	17.0	19.0	19.6	21.0
8	96	10.0	11.0	11.0	12.0	13.0	15.0	16.0	17.0	19.0	19.4	20.5
9	108	11.0	12.0	12.0	13.0	14.0	15.0	17.0	18.0	20.0	21.0	23.0
10	120	11.0	12.0	12.0	13.0	14.0	16.0	18.0	19.0	21.6	23.0	25.5
11	132	11.0	12.0	12.0	13.3	14.0	16.0	18.0	19.7	22.4	23.6	26.7
12	144	-	12.0	13.0	14.0	15.0	16.0	19.0	-	23.3	25.3	28.3
13	156	12.0	13.0	13.0	14.0	15.0	17.0	-	20.0	23.0	-	26.7
14	168	-	11.0	11.0	-	-	-	-	-	-	25.0	-
15	180	-	-	12.0	-	-	16.0	-	-	-	-	26.0

Table 4:
BMI-FOR-AGE using WHO-percentile

Age in years	Age in months	WHO-Percentiles (BMI Kg/m^2) For Girls										
		1%	3%	5%	15%	25%	50%	75%	85%	95%	97%	99%
4	48	12.3	13.0	-	-	-	-	-	-	-	-	19.1
5	60	12.3	12.8	13.1	-	-	15.3	-	16.9	18.1	-	19.2
6	72	12.3	12.8	13.1	13.8	14.3	15.3	16.4	17.1	18.3	18.9	20.0
7	84	12.3	12.9	13.1	13.9	14.4	-	16.6	17.4	18.9	19.5	21.0
8	96	12.5	13.0	13.3	14.1	14.6	15.7	17.0	17.8	19.5	20.1	21.8
9	108	12.7	13.2	13.6	14.4	14.9	16.1	17.5	18.4	20.1	21.1	22.7
10	120	13.0	13.6	13.9	14.8	15.4	-	18.1	19.1	21.2	22.1	24.2
11	132	13.4	14.0	14.4	15.3	15.9	17.3	18.9	20.0	22.2	23.4	25.8
12	144	13.9	14.5	14.9	15.9	16.5	18.0	19.8	-	23.3	24.4	27.9
13	156	14.4	15.1	15.5	16.6	-	-	-	-	24.4	25.9	28.4
14	168	15.0	15.7	16.0	-	-	19.6	-	-	25.3	-	-
15	180	15.2	16.0	-	-	-	-	-	-	-	-	-

Table 5:
BMI-FOR-AGE for boys using Quantile regression

Age in years	Age in months	Quantile regression-Percentiles (BMI kg/m^2) For Boys										
		1%	3%	5%	15%	25%	50%	75%	85%	95%	97%	99%
4	48	7.4	9.0	9.8	11.3	12.0	12.8	13.8	14.3	16.0	16.7	17.5
5	60	8.0	9.5	10.2	11.7	12.4	13.4	14.6	15.2	17.2	18.0	19.6
6	72	8.5	10.0	10.6	12.1	12.8	13.9	15.3	16.1	18.4	19.4	21.6
7	84	9.0	10.4	11.1	12.4	13.2	14.4	16.0	17.0	19.5	20.7	23.7
8	96	9.6	10.9	11.5	12.8	13.6	14.9	16.8	18.0	20.7	22.1	25.8
9	108	10.1	11.4	11.9	13.2	14.0	15.5	17.5	18.9	21.9	23.4	27.8
10	120	10.7	11.8	12.3	13.5	14.4	16.0	18.2	19.8	23.1	24.7	29.9
11	132	11.2	12.3	12.8	13.9	14.8	16.5	19.0	20.7	24.3	26.1	32.0
12	144	11.8	12.8	13.2	14.3	15.2	17.0	19.7	21.6	25.4	27.4	34.1
13	156	12.3	13.2	13.6	14.6	15.6	17.6	20.5	22.5	26.6	28.8	36.2
14	168	12.9	13.3	14.1	15.0	16.0	18.1	21.2	23.5	27.9	30.2	38.3
15	180	13.2	14.1	14.4	15.3	16.3	18.5	21.8	24.1	28.7	31.1	39.8

Table 6:
BMI-FOR-AGE for boys using LMS method

Age in years	Age in months	Percentiles using LMS Method for BMI (Kg/m^2)-Boys										
		1%	3%	5%	15%	25%	50%	75%	85%	95%	97%	99%
4	48	-	-	-	-	-	-	17.0	-	-	-	-
5	60	7.0	9.3	-	12.0	13.0	-	16.0	17.0	-	-	-
6	72	8.4	9.0	-	12.0	12.8	14.0	16.0	17.0	-	20.0	21.0
7	84	9.5	10.0	11.0	12.0	13.0	14.0	16.0	17.0	19.0	20.0	21.7
8	96	10.8	11.0	12.0	13.0	13.0	15.0	16.0	17.0	19.0	20.0	22.1
9	108	11.0	12.0	12.0	13.0	14.0	15.0	17.0	18.0	21.0	22.0	24.8
10	120	11.0	12.0	12.0	14.0	14.0	16.0	19.0	20.0	23.0	24.3	28.2
11	132	11.2	12.0	13.0	14.0	15.0	17.0	19.4	21.0	24.8	26.0	29.8
12	144	12.0	12.5	13.0	14.0	15.0	18.0	21.0	23.0	27.0	29.6	36.8
13	156	12.3	13.0	14.0	15.0	16.0	18.0	21.0	24.0	-	31.0	43.0
14	168	12.0	12.0	13.0	15.0	16.0	18.0	-	-	29.0	-	37.5
15	180	13.0	14.0	-	16.0	17.0	19.0	-	-	-	-	33.0

Table 7:
BMI-FOR-AGE for boys using WHO-percentile

Age in years	Age in months	WHO-Percentiles (BMI Kg/m^2) For Boys										
		1%	3%	5%	15%	25%	50%	75%	85%	95%	97%	99%
4	48	-	-	-	-	-	-	-	-	-	-	-
5	60	-	-	-	-	-	-	-	-	-	-	-
6	72	12.7	-	13.4	14.0	-	-	16.3	-	17.9	18.3	18.9
7	84	12.8	13.2	13.5	-	14.6	15.5	16.5	17.1	18.3	18.8	19.8
8	96	12.9	13.4	13.7	14.4	14.8	15.8	16.8	17.5	18.9	19.4	20.6
9	108	13.1	13.6	13.9	14.6	15.0	16.1	17.3	18.0	19.5	20.1	21.5
10	120	13.3	13.9	14.1	14.9	15.4	16.4	17.8	18.6	20.2	21.0	22.9
11	132	13.6	14.2	14.5	15.3	15.8	17.0	18.4	19.3	21.1	22.1	24.3
12	144	14.1	14.6	14.9	15.8	16.3	17.5	19.1	20.0	22.0	23.0	25.3
13	156	14.5	15.1	15.4	16.3	16.9	-	19.9	21.0	23.2	24.2	26.8
14	168	14.9	15.6	16.0	-	-	-	20.8	-	24.3	25.2	28.1
15	180	15.5	16.2	-	-	-	19.8	-	-	25.2	-	-

Table 8:
HEIGHT-FOR-AGE for girlsusing Quantile regression

Age in years	Age in months	Quantile regression-Percentiles (Height in cm) For Girls										
		1%	3%	5%	15%	25%	50%	75%	85%	95%	97%	99%
4	48	91.4	93.2	94.0	98.5	101.3	106.0	111.4	115.8	123.8	127.0	132.0
5	60	94.6	96.8	98.0	102.8	105.7	110.5	116.0	120.2	128.0	131.0	136.0
6	72	97.7	100.4	102.0	107.0	110.0	115.0	120.6	124.6	132.2	135.0	140.0
7	84	100.9	104.0	106.0	111.3	114.4	119.5	125.2	129.0	136.4	139.0	144.0
8	96	104.0	107.6	110.0	115.5	118.7	124.0	129.8	133.4	140.5	143.0	148.0
9	108	107.2	111.2	114.0	119.8	123.0	128.5	134.4	137.8	144.7	147.0	152.0
10	120	110.3	114.8	118.0	124.0	127.4	133.1	139.1	142.2	148.9	151.0	156.0
11	132	113.5	118.5	122.1	128.3	131.7	137.6	143.7	146.7	153.1	155.1	160.1
12	144	116.7	122.1	126.1	132.6	136.1	142.1	148.3	151.1	157.3	159.1	164.1
13	156	119.9	125.8	130.2	137.0	140.6	146.7	153.0	155.6	161.6	163.2	168.2
14	168	123.3	129.7	134.6	141.6	145.3	151.7	158.1	160.5	166.1	167.6	172.6
15	180	126.0	132.8	138.0	145.3	149.0	155.5	162.0	164.2	169.7	171.0	176.0

Table 9:
HEIGHT-FOR-AGE for girls using LMS method

Age in years	Age in months	Percentiles using LMS Method (Height in cm) for Girls										
		1%	3%	5%	15%	25%	50%	75%	85%	95%	97%	99%
4	48	-	-	106.0	-	-	-	-	-	-	-	-
5	60	97.0	99.5	-	106.0	-	114.0	119.0	122.0	-	128.0	131.0
6	72	91.8	-	-	-	108.0	-	-	-	129.0	131.0	135.8
7	84	101.6	104.0	106.0	-	113.0	-	-	-	134.0	136.0	141.0
8	96	107.3	110.0	-	-	-	124.0	-	134.0	141.0	144.5	151.0
9	108	111.0	114.0	116.0	-	123.0	129.0	-	-	146.0	149.7	155.7
10	120	112.0	115.5	118.0	-	128.0	135.0	-	145.0	151.0	153.0	157.1
11	132	116.3	121.0	123.0	129.0	-	-	145.0	-	154.0	157.0	161.0
12	144	118.7	124.0	-	-	-	-	-	-	158.0	159.0	-
13	156	-	127.3	130.7	137.0	-	-	-	-	160.0	-	164.0
14	168	121.0	125.3	129.0	137.0	-	-	155.0	-	-	-	-
15	180	-	118.0	-	-	-	-	-	-	168.0	-	-

Table 10:
HEIGHT-FOR-AGE for girls using WHO-percentile

Age in years	Age in months	WHO-Percentiles (Height in cm) For Girls										
		1%	3%	5%	15%	25%	50%	75%	85%	95%	97%	99%
4	48	-	-	-	-	-	-	-	-	-	111.00	113.00
5	60	99.00	-	-	-	-	-	-	-	-	-	120.09
6	72	102.44	-	-	-	-	-	-	-	-	125.00	128.00
7	84	107.58	-	112.00	-	-	-	-	-	130.00	131.00	133.20
8	96	112.57	116.00	117.00	-	-	-	-	-	136.00	-	140.38
9	108	118.09	121.00	-	-	-	-	-	-	-	144.00	146.86
10	120	123.48	127.00	128.00	132.00	-	-	143.00	-	149.00	151.00	154.09
11	132	129.43	132.00	134.00	138.00	-	145.00	-	152.00	-	-	161.00
12	144	134.97	138.00	140.00	-	-	-	-	-	-	-	-
13	156	140.37	143.00	145.00	-	-	-	161.00	-	-	-	-
14	168	142.86	147.00	-	-	155.00	-	-	167.00	-	-	-
15	180	146.00	-	-	-	157.00	-	-	-	-	-	-

Table 11:
HEIGHT-FOR-AGE for boys using Quantile regression

Age in years	Age in months	Quantile regression-Percentiles (Height in cm) For Boys										
		1%	3%	5%	15%	25%	50%	75%	85%	95%	97%	99%
4	48	93.3	94.2	95.7	100.0	104.0	108.0	113.8	117.0	120.0	124.9	130.3
5	60	95.7	97.6	99.4	104.0	108.0	112.4	118.2	121.6	125.1	129.6	135.1
6	72	98.0	101.0	103.0	108.0	112.0	116.7	122.6	126.0	130.0	134.3	139.8
7	84	100.4	104.4	106.7	112.0	116.0	121.0	127.0	130.5	135.0	139.0	144.5
8	96	102.7	107.8	110.4	116.0	120.0	125.4	131.4	135.0	140.0	143.7	149.3
9	108	105.0	111.2	114.0	120.0	124.0	129.7	135.8	139.5	145.1	148.5	154.0
10	120	107.4	114.6	117.7	124.0	128.0	134.1	140.3	144.1	150.1	153.2	158.8
11	132	109.7	118.0	121.4	128.0	132.0	138.4	144.7	148.6	155.1	157.9	163.6
12	144	112.0	121.5	125.1	132.1	136.1	142.7	149.1	153.1	160.1	162.7	168.3
13	156	114.4	124.9	128.8	136.1	140.1	147.1	153.5	157.6	165.2	167.4	173.2
14	168	116.8	128.4	132.6	140.3	144.3	151.6	158.1	162.3	170.4	172.3	178.1
15	180	118.5	130.9	135.2	143.1	147.1	154.7	161.3	165.5	173.9	175.7	181.5

Table 12:
HEIGHT-FOR-AGE for boys using LMS method

Age in years	Age in months	Quantile regression-Percentiles (Height in cm) For Boys										
		1%	3%	5%	15%	25%	50%	75%	85%	95%	97%	99%
4	48	-	-	-	-	-	-	-	-	-	144.0	-
5	60	93.0	99.0	101.0	-	110.0	116.0	122.0	125.0	130.0	-	135.3
6	72	91.8	96.3	99.0	-	-	-	123.0	127.0	133.0	135.3	140.0
7	84	103.4	107.0	108.0	-	115.0	-	-	130.0	137.0	139.0	145.0
8	96	105.8	110.0	-	116.0	119.0	-	-	134.0	140.0	142.3	147.0
9	108	111.0	115.0	117.0	122.0	125.0	-	137.0	-	147.0	149.0	154.0
10	120	113.7	118.4	121.0	-	130.0	136.0	142.0	145.0	150.0	152.0	155.8
11	132	115.7	120.0	122.0	-	-	-	145.0	-	155.0	157.0	162.0
12	144	112.0	118.5	122.0	-	134.0	142.0	-	-	160.0	162.5	167.0
13	156	108.0	119.8	-	133.0	-	-	-	-	170.0	173.0	179.0
14	168	114.5	123.0	-	138.0	-	-	162.0	-	-	177.0	180.0
15	180	-	129.0	138.0	-	-	-	-	-	-	-	-

Table 13:
HEIGHT-FOR-AGE for boys using WHO-percentile

Age in years	Age in months	WHO-Percentiles (Height in cm) For Boys										
		1%	3%	5%	15%	25%	50%	75%	85%	95%	97%	99%
4	48	-	-	-	-	-	-	-	-	-	-	-
5	60	-	-	-	-	-	-	-	-	-	-	-
6	72	105.00	-	108.00	-	-	116.00	-	121.00	124.00	125.00	127.40
7	84	109.56	112.00	113.00	-	-	-	-	-	-	132.00	134.45
8	96	113.59	117.00	118.00	-	-	-	131.00	133.00	-	138.00	140.13
9	108	118.47	121.00	-	-	-	-	-	-	-	144.00	147.00
10	120	122.10	126.00	127.00	-	-	-	142.00	-	148.00	150.00	152.78
11	132	127.28	130.00	132.00	-	-	-	-	150.00	154.00	156.00	159.70
12	144	131.90	136.00	-	-	-	149.00	154.00	-	161.00	162.00	165.60
13	156	138.00	142.00	144.00	-	151.00	156.00	161.00	-	168.00	170.00	174.00
14	168	145.56	149.00	-	-	158.00	-	-	-	-	178.00	180.00
15	180	150.00	154.00	156.00	-	-	-	-	-	-	-	-

Table 14:
WEIGHT-FOR-AGE for girls using Quantile regression

Age in years	Age in months	Quantile regression-Percentiles (Weight in Kg) For Girls										
		1%	3%	5%	15%	25%	50%	75%	85%	95%	97%	99%
4	48	6.5	8.4	9.2	11.3	12.0	13.2	16.0	18.0	22.0	22.5	25.8
5	60	8.0	10.0	10.8	13.0	14.0	15.6	18.7	21.0	25.4	26.3	30.0
6	72	9.5	11.6	12.4	14.8	16.0	18.0	21.4	24.0	28.7	30.0	34.2
7	84	11.0	13.2	14.0	16.5	18.0	20.4	24.2	27.0	32.0	33.8	38.4
8	96	12.5	14.7	15.6	18.3	20.0	22.8	26.9	30.0	35.4	37.5	42.5
9	108	14.0	16.3	17.2	20.0	22.0	25.2	29.6	33.0	38.7	41.3	46.7
10	120	15.5	17.9	18.8	21.8	24.0	27.6	32.3	36.0	42.0	45.0	50.9

Table 15:
WEIGHT-FOR-AGE for girls using LMS method

Age in years	Age in months	Percentiles using LMS Method (Weight in Kg) for Girls										
		1%	3%	5%	15%	25%	50%	75%	85%	95%	97%	99%
4	48	-	-	-	16.0	-	-	-	-	-	-	-
5	60	9.0	11.0	12.0	-	-	-	22.0	-	-	28.0	30.0
6	72	9.0	-	-	-	-	-	-	-	28.0	30.0	32.0
7	84	11.0	-	-	-	-	20.0	-	-	30.0	32.0	35.2
8	96	13.3	15.0	-	-	-	-	-	-	33.0	35.2	39.6
9	108	16.0	17.0	18.0	-	-	-	-	33.0	39.0	42.3	52.3
10	120	17.0	19.0	-	-	24.0	-	-	36.0	42.0	45.0	51.9

Table 16:
WEIGHT-FOR-AGE for girls using WHO-percentile

Age in years	Age in months	WHO-Percentiles (Weight in Kg) For Girls										
		1%	3%	5%	15%	25%	50%	75%	85%	95%	97%	99%
4	48	-	-	-	-	-	-	-	-	-	21.00	-
5	60	13.00	14.00	-	-	-	-	-	-	-	-	26.00
6	72	-	-	16.00	-	-	-	-	-	26.00	27.00	29.40
7	84	16.00	17.00	-	-	-	-	25.00	-	-	31.00	34.60
8	96	18.00	19.00	-	-	-	25.00	28.00	-	-	35.00	39.18
9	108	20.00	21.00	-	-	-	-	-	-	38.00	40.00	44.40
10	120	22.00	-	-	-	-	-	36.00	-	-	-	50.89

Table 17:
WEIGHT-FOR-AGE for boys using Quantile regression

Age in years	Age in months	Quantile regression-Percentiles (Weight in kg) For Boys										
		1%	3%	5%	15%	25%	50%	75%	85%	95%	97%	99%
4	48	5.0	7.0	8.0	11.0	10.0	12.8	15.0	16.7	20.3	21.0	25.0
5	60	7.0	9.0	10.0	13.0	12.5	15.5	18.4	20.4	24.5	25.8	30.1
6	72	9.0	11.0	12.0	15.0	15.0	18.3	21.7	24.0	28.8	30.5	35.0
7	84	11.0	13.0	14.0	17.0	17.5	21.0	25.0	27.7	33.0	35.3	40.0
8	96	13.0	15.0	16.0	19.0	20.0	23.8	28.4	31.4	37.3	40.0	45.0
9	108	15.0	17.0	18.0	21.0	22.5	26.5	31.7	35.0	41.5	44.8	50.1
10	120	17.0	19.0	20.0	23.0	25.0	29.3	35.0	38.7	45.8	49.6	55.1

Table 18:
WEIGHT-FOR-AGE for boys using LMS method

Age in years	Age in months	Percentiles using LMS Method (Weight in Kg) for Boys										
		1%	3%	5%	15%	25%	50%	75%	85%	95%	97%	99%
4	48	-	16.0	-	-	-	-	-	-	-	50.0	-
5	60	8.0	10.0	-	-	16.0	-	-	-	28.0	29.0	-
6	72	9.0	-	-	-	-	-	-	-	30.0	32.0	36.3
7	84	12.0	-	14.0	-	-	-	-	27.0	32.0	34.0	39.3
8	96	14.0	-	16.0	-	-	-	-	-	35.0	37.6	43.0
9	108	15.4	17.0	18.0	-	-	-	-	34.0	40.0	-	48.9
10	120	17.0	-	-	-	-	-	35.0	-	45.0	48.2	57.8

Table 19:
WEIGHT-FOR-AGE for boys using WHO-percentile

Age in years	Age in months	WHO-Percentiles (Weight in Kg) For Boys										
		1%	3%	5%	15%	25%	50%	75%	85%	95%	97%	99%
4	48	-	-	-	-	-	-	-	-	-	-	-
5	60	-	-	-	-	-	-	-	-	-	-	-
6	72	15.00	-	-	-	-	-	-	-	-	-	28.43
7	84	17.00	18.00	-	-	-	-	-	-	29.00	30.00	32.55
8	96	19.00	-	-	22.00	-	-	28.00	-	-	34.00	37.33
9	108	20.00	-	-	-	-	-	-	-	37.00	39.00	42.20
10	120	22.00	-	-	-	28.00	-	35.00	-	42.00	44.00	49.08

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 29-40

D-OPTIMAL DESIGNS FOR MORGAN MERCER FLODIN (MMF) MODELS AND ITS APPLICATION

Tatik Widiharih[1,2], Sri Haryatmi[2] and **Gunardi[2]**
[1]Department of Statistics, Diponegoro University, Semarang, Indonesia.
[2]Department of Mathematics, Gadjah Mada University, Yogyakarta, Indonesia
Email: widiharih@gmail.com
s_kartiko@yahoo.com
gunardi@ugm.ac.id

ABSTRACT

Locally D-optimal designs for Morgan Mercer Flodin (MMF) models with homoscedastic error are investigated. These models are restricted without intercept with two and three parameters . D-optimal criteria is based on Equivalence Theorem of Kiefer Wolfowitz (1960). Determination whether the design that meets the specified model is minimally supported design is based on Theorem 1 of Li and Majumdar (2008) which examines the behaviour of the standardized variance function in a vertical neighborhood of zero. Tchebysheff system and their properties plays a critical role on it. The results show that for design region in the interval [0, b],where b is selected such that the curve is relatively constant, the design is a minimally supported and one of the design point is boundary point.Application of these models are a function in pharmacokinetics which is Michaelis Menten model and in pharmacodynamics which is EMAX model.

Keywords: D-optimal; Tchebyshev system; Minimally Supported designs; Michaelis Menten, EMAX

1. INTRODUCTION

Morgan, et al. (1975) introduced the S-shaped growth curve models to explain the relationship between nutrient intake and an appropriate response. Originally this model can be applied in biology and animal husbandry areas, such as Sengul and Kiraz (2005), Tariq, et al. (2013), Topal, et al. (2008) , Utomo, et al(2012) and Tjorve (2003). Jericevic and Kuster (2005) used the specific case of Morgan Mercer Flodin model with two parameters which is Michaelis Menten model to describe the relationship between concentration of substrate and velocity of reaction. Knudsen (2001)used another specific case of Morgan Mercer Flodin model with three parameters which is EMAX model to describe the relationship between concentration andeffect of the drug. Morgan Mercer Flodin (MMF) model is a very specific growth curve model. There are two limits of response in this model (lowest and highest values of response) and after a certain boundary (maximum point) its response will be constant. The MMF function is:

$$y = \frac{\theta_1\theta_2+\theta_3 x^{\theta_4}}{\theta_2+x^{\theta_4}} \ , \theta_2, > 0, \theta_3 > \theta_1 > 0, \theta_4 \geq 1, x \in [a, b] \qquad (1)$$

where:

y : observed response of the organism (i.e; weight gain, plasma concentration of metabolities, etc.),
x : nutrient intake,
θ_1 : calculated ordinate intercept of the nutrient response curve,
θ_2 : nutrition constant,
θ_3: asymptotic or maximum response of organisme,
θ_4: apparent kinetic order of the response with respect to x as x approached zero.

The most important design criterion in applications is that of D-optimality, in which the generalized variance of the parameter estimates, or its logarithm $-log|M(\xi,\theta)|$, is minimizes (Atkinson, et al. (2007). Determination of optimal designs for the MMF model is particularly difficult because, unlike for linear models, the Fisher information matrix and the optimal design depend on the values of the unknown parameters. A method that is most widely used dealing with this problem is the local optimality approach, in which the optimality criterion function is evaluated using assumed values of the parameters. Chernoff (1953) proposed to adopt an initial guess $\boldsymbol{\theta} = \boldsymbol{\theta_0}$ for the unknown parameter vector and maximized the criterion function evaluated at the guessed value of the parameters. The resulting design is termed locally optimal design.

Some authors have been investigated D-optimal designs for growth curve in some models including Chang and Lay (2002) who used polynomials models, Dette and Pepelyshev (2008) also used sigmoidal growth models, Li (2011) used gompertz function, Li and Balakrishan (2011) used double exponential regrowth and LINEX regrowth, Widiharih, et al. (2012) used exponential models for weighted mean, and then extended it to generalized exponential and weighted exponential models with two parameters (Widiharih, et al. (2013)). Numerical approach for generalized exponential model with three parameters also has been investigated (Widiharih et al. (2013)).

MMFmodels without intercept by taking$\theta_1 = 0$ in equation (1).There are two models, that are MMF withtwo-parameter and MMF with three parameters as follows:

$$y = \frac{\theta_3 x}{\theta_2 + x} + \varepsilon, \theta_2, \theta_3 > 0, x\epsilon[a,b] \quad (2)$$

$$y = \frac{\theta_3 x^{\theta_4}}{\theta_2 + x^{\theta_4}} + \varepsilon, \theta_2, \theta_{3,} > 0, \theta_4 \geq 1, x\epsilon[a,b] \quad (3)$$

Michaelis Menten model is a special form of equation (2) by taking$x\epsilon[0,1]$ (Dette and Kiss (2012)). In their paper introduced optimal design for rational regression models, Michaelis Menten model as a special case of rational regression models. So that D-optimal design for Michaelis Menten of Dette and Kiss (2012) is a special case of D-optimal design model (2). Emax model is a special form of equation (3) by taking $x\epsilon[0,1]$ and $\theta_4 = 1$ (Dette, et al. (2004)) , so that the D-optimal design for Emax model of Dette, et al. (2004) is a special case of D-optimal design model (3). In this paper we determine locally D-optimal designs for two MMF models without intercept in equation (2) and (3) with Tchebysheff system approached and applied it for Michaelis Menten and EMAX models.

2. PRELIMINARIES

Suppose the design space is denoted by $\chi\chi$. Let H$\mathcal{H}$ be the class of probability distribution on the Borel set of$\chi\chi$, then any $\xi \in \mathcal{H}$His called an approximate design measure (Kiefer (1961)). The Equivalence Theorem provides an important tool in the theory of optimum designs. It was originally established for linear models by Kiefer and Wolfowitz (1960), and then it was extended to nonlinear models by White [23]. In order to state the Equivalence Theorem, we need the quantity $d(\xi, x)$ which denotes the standardized variance of the model based predicted response at *x*.
Consider the nonlinear model:

$$E(Y|x) = \eta(x, \theta) \tag{4}$$

Designs of *p* point is denoted by:

$$\xi = \begin{pmatrix} x_1 & x_2 & \dots & x_p \\ w_1 & w_2 & \dots & w_p \end{pmatrix} \tag{5}$$

where: $w_i = \frac{r_i}{n}$, r_i : number of observation at the point x_i, *n*: number of observation, $n = \sum_{i=1}^{p} r_i$ and $\sum_{i=1}^{p} w_i = 1$. The information matrix of designs ξ for model (4) is:

$$M(\xi, \theta) = \sum_{i=1}^{p} w_i h(x_i, \theta) h^T(x_i, \theta) \tag{6}$$

where: $h(x, \theta) = \frac{\partial \eta(x,\theta)}{\partial \theta} = (\frac{\partial \eta(x,\theta)}{\partial \theta_1}, \frac{\partial \eta(x,\theta)}{\partial \theta_2}, \dots, \frac{\partial \eta(x,\theta)}{\partial \theta_k})^T$ is the vector of partial derivatives of the conditional expectation E(Y|x) with respect to the parameters θ (*k* : is number of parameters in the model). A D-optimal designs maximizes $|M(\xi, \theta)|$, which is the determinant of the information matrix. The standardized variance $d(\xi, x)$ is:

$$d(\xi, x) = h^T(x, \theta) M^{-1}(\xi, \theta) h(x, \theta) \tag{7}$$

The Equivalence Theorem states that in the class of design measure $\mathcal{H}$, if a design measureξ^* satisfies any one of the following three conditions, then it satiesfies the other two:

1. ξ^*maximizes $|M(\xi, \theta)|$
2. ξ^* minimizes $max_{x \epsilon \chi} d(\xi, x)$
3. $max_{x \epsilon \chi} d(\xi^*, x) = k$ where *k* is number of parameters

It follows from Caratheodory's theorem that determination of D-optimal design is restricted with *n* support points $x_1,\dots,x_n$ and corresponding probabilities $w_1,\dots,w_n$ for *n=k,....,k(k+1)/2*, where *k* is the number of parameters. If*n=k*, the design is minimally supported. If the D-optimal design is minimally supported then its support are uniform weight, i.e., $w_i = \frac{1}{k}$for *i=1,...,k*. This will considerably reduce the difficulty of the problem since the D-optimality criterion becomes a function of the *k* unknown support points $x_1,\dots,x_k$ only. Li and Majumdar (2008) in Theorem 1 result in a sufficient condition to ensure that the D-optimal design is minimally supported. Here, we will adopt this approach. Tchebysheff system plays a critical role on it. We use definition and properties of Tchebysheff system which introduced by Karlin and Studden (1966).

3. MAIN RESULTS

3.1 D-optimal Design for Two Parameters MMF Model

Two parameters MMF model without intercept is known as Michaelis Menten model as in equation (2):

$$y = \frac{\theta_3 x}{\theta_2 + x} + \varepsilon, \theta_2, \theta_3 > 0, x\epsilon[a,b]$$

with homoscedastic error. Here $\theta = (\theta_2, \theta_3)$ is the parameter of interest. A straightforward calculation yields for the vector of partial derivative:

$$h(x,\theta) = (-\frac{\theta_3 x}{(\theta_2+x)^2}, \frac{x}{(\theta_2+x)})^T \quad (8)$$

With our first result we establish the basic properties of locally D-optimal design of model (2).

Theorem 3.1

Support points of D-optimal design of model (2) do not depend on θ_3 which are :

$$x_1 = \frac{\theta_2 b}{2\theta_2 + b}, (a < x_1), x_2 = b$$

with uniform of its support i.e 0.5.

Proof.

Let m^{ij} denote *(i,j)*th element of $M^{-1}(\xi,\theta)$, then : $d(\xi,x) = \frac{x^2}{(\theta_2+x)^2}\left[m^{11}\frac{\theta_3^2}{(\theta_2+x)^2} + m^{22} - 2m^{12}\frac{\theta_3}{(\theta_2+x)}\right]$. Based on Theorm 1 of Li and Majumdar [10], $d(\xi,x) - 2 + c = \frac{x^2}{(\theta_2+x)^2}.g(x)$. The roots of $d(\xi,x) - 2 + c$ are same as the roots of $g(x)$. $g(x)$ is a linear combination of: $\left\{1, \frac{1}{(\theta_2+x)^2}, \frac{1}{(\theta_2+x)^2}, \frac{x^2}{(\theta_2+x)^2}\right\}$ that is a Tchebysheffsystem, thus $g(x)$ has at most *3=2k-1* roots. From part 3 of Theorem 1 of Li and Majumdar [13], for $\chi = [a,b]$ if D-optimal designs exist then it is minimally supported design with either *a* or *b*as a design point. In this case *b* is the design point and the designs ξ is :

$$\xi = \begin{pmatrix} x_1 & b \\ 1/2 & 1/2 \end{pmatrix} \quad (9)$$

Element of the information matrix $M(\xi,\theta)$ are:

$$m_{11} \propto \sum_{i=1}^{2}\frac{\theta_3^2 x_i^2}{(\theta_2+x_i)^4} m_{22} \propto \sum_{i=1}^{2}\frac{x_i^2}{(\theta_2+x_i)^2} m_{12} \propto \sum_{i=1}^{2}\frac{-\theta_3 x_i^2}{(\theta_2+x_i)^3}$$

The determinant of information matrix is :$|M(\xi,\theta)| \propto \frac{\theta_3^2 x_1^2 b^2}{(\theta_2+x_1)^4(\theta_2+b)^4}(x_1-b)^2$

x_1 is the maximizes of $|M(\xi,\theta)|$, that is

$x_1 = \frac{\theta_2 b}{2\theta_2 + b}$. Clear that x_1 and x_2 do not depend on θ_3 ■

Application.

Jericevic and Kuster [8] used the Michaelis Menten model to describe the relationship between concentration of substrate (s) and velocity of reaction (v). Data set is presented in Table 1.

Table 1.
Data Set of Concentration Substrate (s) and Velocity (v)

S ($mmol\ dm^{-3}$)	V ($\mu mol\ dm^{-1}min^{-1}$)	S ($mmol\ dm^{-3}$)	V ($\mu mol\ dm^{-1}min^{-1}$)
0.25	2.40	0.70	6.20
0.30	2.60	1.00	7.40
0.40	4.20	1.40	10.20
0.50	3.80	2.00	11.40

Based data set in Table 1, nonlinear OLS parameters estimate and t test of parameters in equation (2) is presented in Table 2. The value of R_{sq} and $Adj.R_{sq}$ are 0.9814 and 0.9784 respectively.

Table 2.
Nonlinear OLS parameters Estimate and t Test of Data Set in Table 1.

Parameter	Estimate	Approx Std Err	T value	Approx P_{value}
θ_3	25.39759	3.9612	6.41	0.0007
θ_2	2.326236	0.5526	4.21	0.0056

Graph of the estimate model of equation (2) is presented in Figure 1.

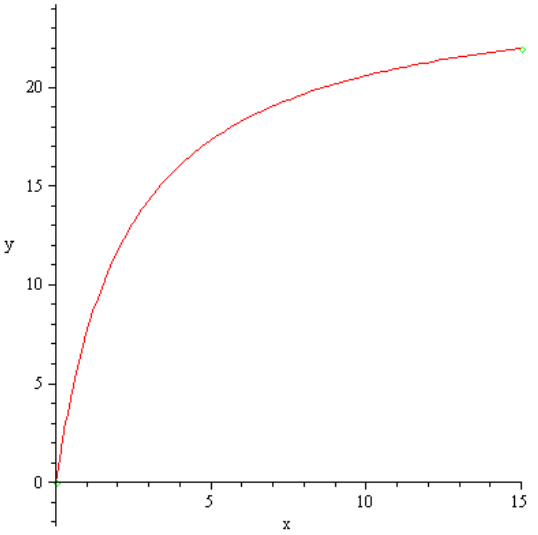


Figure 1. Estimate Curve of Model (2)

The value of parameters estimate can be used as the prior information to determine the support points such that the design satisfy the D-optimal design. Based on Figure 1, we suggest the design region and support points as in Table 3 with uniform of its support.

Table 3
Design Region and Support points of Model in Equation (2) with $\boldsymbol{\theta_3}$ =25.39759, $\boldsymbol{\theta_2}$ = 2.326236

Design Region	Design support		Design Region	Design support	
	x_1	x_2		x_1	x_2
[0, 1.50]	0.56715	1.50	[0, 3.75]	1.03819	3.75
[0, 1.75]	0.63583	1.75	[0, 4.00]	1.07541	4.00
[0, 2.00]	0.69936	2.00	[0, 4.25]	1.11053	4.25
[0, 2.25]	0.75828	2.25	[0, 4.50]	1.14374	4.50
[0, 2.50]	0.81309	2.50	[0, 4.75]	1.17518	4.75
[0, 2.75]	0.86419	2.75	[0, 5.00]	1.20499	5.00
[0, 3.00]	0.91195	3.00	[0, 5.25]	1.23330	5.25
[0, 3.25]	0.95670	3.25	[0, 5.50]	1.26022	5.50
[0, 3.50]	0.99869	3.50	[0, 5.75]	1.28583	5.75

For design region [0,1.5] the support points are x_1=0.69936 and x_2 = 1.5, and the standardized variance ($d(\xi, x)$) of the support points are 2.000000069 and 2.0000001 respectively. The curve of the standardized variance is presented in Figure 2.

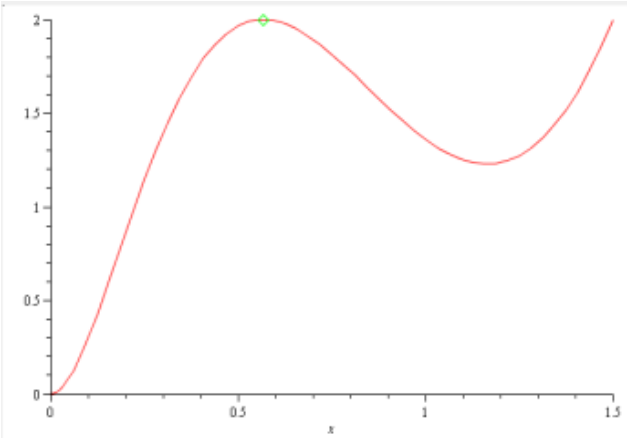


Figure 2. Standardized Variance Model (2) with design region [0, 1.5] and $\boldsymbol{\theta_3}$ =25.39759, $\boldsymbol{\theta_2}$ = 2.326236

3.2 *D-optimal Designfor Three parameters MMF model*.

Three parameters MMF models without intercept is known as EMAX model in equation (3):

$$y = \frac{\theta_3 x^{\theta_4}}{\theta_2 + x^{\theta_4}} + \varepsilon, \theta_2, \theta_3, > 0, \theta_4 \geq 1, x\epsilon[a, b]$$

with homoscedastic error. Here $\theta = (\theta_2, \theta_3, \theta_4)$ is the parameter of interest. A straightforward calculation yields for the vector of partial derivative :

$$h(x,\theta) = (-\frac{\theta_3 x^{\theta_4}}{(\theta_2+x^{\theta_4})^2}, \frac{x^{\theta_4}}{(\theta_2+x^{\theta_4})}, \frac{\theta_2\theta_3 x^{\theta_4}\ln(x)}{(\theta_2+x^{\theta_4})^2})^T \quad (10)$$

With our second result we establish the basic properties of locally D-optimal design of model (3).

Theorem 3.2

Support points of D-optimal design of model (3) do not depent on θ_3 wich are x_1, x_2 and$x_3 = b$ with uniform of its support i.e 1/3, x_1, x_2 are maximizes :

$|M(\xi,\theta)| \propto (x_1x_2x_3)^{2\theta_4}(A+B)$, where :

$A = \sum_{i=1}^{3} \frac{(\ln(x_i)-\ln(x_j))^2}{(\theta_2+x_i{}^{\theta_4})^4(\theta_2+x_j{}^{\theta_4})^4(\theta_2+x_k{}^{\theta_4})^2}$,

$B = \sum_{i=1}^{3} \frac{\ln(x_i)\ln(x_j)+\ln(x_i)\ln(x_k)-\ln^2(x_i)-\ln(x_j)\ln(x_k)}{(\theta_2+x_i{}^{\theta_4})^4(\theta_2+x_j{}^{\theta_4})^3(\theta_2+x_k{}^{\theta_4})^3}$, $i \neq j \neq k, j, k = 1,2,3,\ x_3 = b$

Proof.

Let m^{ij} denote *(i,j)th* element of $M^{-1}(\xi,\theta)$, then :

$$d(\xi,x) = \frac{x^{2\theta_4}}{(\theta_2+x^{\theta_4})^2}\left[m^{11}\frac{\theta_3^2}{(\theta_2+x^{\theta_4})^2} + m^{22} + m^{33}\frac{\theta_2^2\theta_3^2 ln^2(x)}{(\theta_2+x^{\theta_4})^2} - 2m^{12}\frac{\theta_3}{(\theta_2+x^{\theta_4})} - 2m^{13}\frac{\theta_2\theta_3^2 ln(x)}{(\theta_2+x^{\theta_4})^2} + 2m^{23}\frac{\theta_2\theta_3 ln(x)}{(\theta_2+x^{\theta_4})}\right]$$

Based on Theorm 1 of Li and Majumdar [10], $d(\xi,x) - 3 + c = \frac{x^{2\theta_4}}{(\theta_2+x^{\theta_4})^2}.g(x)$.

The roots of $d(\xi,x) - 3 + c+c$ are same as the roots of *g(x)*. *g(x)* is a linear combination of:

$$\left\{1, \frac{\ln(x)}{(\theta_2+x^{\theta_4})}, \frac{1}{(\theta_2+x^{\theta_4})^2}, \frac{\ln(x)}{(\theta_2+x^{\theta_4})^2}, \frac{\ln^2(x)}{(\theta_2+x^{\theta_4})^2}, \frac{x^{2\theta_4}}{(\theta_2+x^{\theta_4})^2}\right\}$$

that is a Tchebysheff system, thus $g(x)$ has at most *5=2k-1* roots. From part 3 of Theorem 1 of Li and Majumdar (2008), for $\chi = [a,b]$ if D-optimal designs exist then it is minimally supported design with either a or b as a design point. In this case b is the design point and the designs ξ is :

$$\xi = \begin{pmatrix} x_1 & x_2 & b \\ 1/3 & 1/3 & 1/3 \end{pmatrix} \quad (11)$$

Element of the information matrix $M(\xi,\theta)$ are:

$$m_{11} \propto \sum_{i=1}^{3} \frac{\theta_3^2 x_i^{2\theta_4}}{(\theta_2 + x_i{}^{\theta_4})^4} m_{22} \propto \sum_{i=1}^{3} \frac{x_i^{2\theta_4}}{(\theta_2 + x_i{}^{\theta_4})^2}$$

$$m_{33} \propto \sum_{i=1}^{3} \frac{\theta_2^2\theta_3^2 x_i^{2\theta_4} ln^2(x_i)}{(\theta_2 + x_i{}^{\theta_4})^4} m_{12} \propto \sum_{i=1}^{3} \frac{-\theta_3 x_i^{2\theta_4}}{(\theta_2 + x_i{}^{\theta_4})^3} m_{13}$$
$$\propto \sum_{i=1}^{3} \frac{\theta_2\theta_3^2 x_i^{2\theta_4} ln(x_i)}{(\theta_2 + x_i{}^{\theta_4})^4} m_{23} \propto \sum_{i=1}^{3} \frac{\theta_2\theta_3 x_i^{2\theta_4} ln(x_i)}{(\theta_2 + x_i{}^{\theta_4})^3}$$

The determinant of information matrix is **:**

$$|\boldsymbol{M}(\boldsymbol{\xi}, \boldsymbol{\theta})| \propto (\boldsymbol{x_1 x_2 x_3})^{\boldsymbol{2\theta_4}} (\boldsymbol{A} + \boldsymbol{B}),$$

where :

$$A = \sum_{i=1}^{3} \frac{\left(\ln(x_i) - \ln(x_j)\right)^2}{\left(\theta_2 + x_i{}^{\theta_4}\right)^4 \left(\theta_2 + x_j{}^{\theta_4}\right)^4 \left(\theta_2 + x_k{}^{\theta_4}\right)^2},$$

$$B = \sum_{i=1}^{3} \frac{\ln(x_i)\ln(x_j) + \ln(x_i)\ln(x_k) - \ln^2(x_i) - \ln(x_j)\ln(x_k)}{(\theta_2 + x_i{}^{\theta_4})^4 \left(\theta_2 + x_j{}^{\theta_4}\right)^3 (\theta_2 + x_k{}^{\theta_4})^3}$$
$$i \neq j \neq k, j, k = 1,2,3,\ x_3 = b$$

$x_1, x_1$are the maximizes of $|M(\xi, \theta)|$. Clear that x_1 and x_2 do not depend on θ_3 ■

Application.

Holford (2013) used the sigmoid EMAX to describe the relationship between concentration at the receptor (c) and effect of drug (e). Data set is presented inTable 4.

Table 4
Data Set of Concentration at th Receptor (c) and Efect of the drug (e)

c	e
160.000	94.0
80.000	89.0
40.000	80.0
20.000	67.0
10.000	50.0
5.000	33.0
2.500	20.0
1.250	11.0
0.625	5.9

Based data set in Table 4, nonlinear OLS parameters estimate and t test of parameters in equation (3) is presented in Table 5. The value of R_{sq} and Adj.R_{sq} are 1.00 and 1.00 respectively. Graph of the estimate model of equation (3) is presented in Figure 3.

Table 5
Nonlinear OLS parameters Estimate and t Test of Data Set in Table 1.

Parameter	Estimate	Approx Std Err	T value	Approx P_{value}
θ_3	99.77242	0.2355	423.66	<0.0001
θ_2	10.13847	0.0896	113.20	<0.0001
θ_4	1.008629	0.00529	190.70	<0.0001

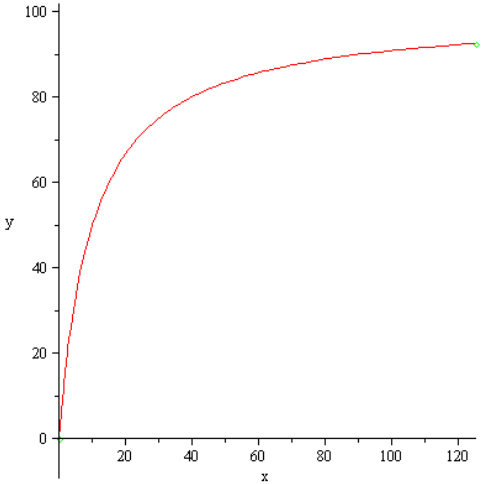


Figure 3. Estimate Curve of Model (3)

The value of parameters estimate can be used as the prior information to determine the support points such that the design satisfy the D-optimal design. Based on Figure 3, we suggest the design region and support points as in Table 6 with uniform of its support.

Table 6
Design Region and Support points of Model in Equation (3) with θ_3 =99.77242, $\theta_2 = 10.13847$ and $\theta_4 = 1.008629$

Design Region	Design support		
	x_1	x_2	x_3
[0, 60]	2.2682	14.0930	60
[0, 55]	2.2101	13.5704	55
[0, 50]	2.1451	12.9971	50
[0, 45]	2.0717	12.3648	45
[0, 40]	1.9878	11.6625	40
[0, 35]	1.8909	10.8762	35
[0, 30]	1.7772	9.9879	30
[0, 25]	1.6414	8.9731	25
[0, 20]	1.4753	7.7979	20
[0, 15]	1.2657	6.4138	15
[0, 10]	0.9897	4.7472	10

For design region [0,50] the support points are x_1=2.1415, x_2 = 12.9971, x_3=50 and the standardized variance of the support points are 2.999999 , 2.999999 and 2.999999. The curve of the standardized variance is presented in Figure 4.

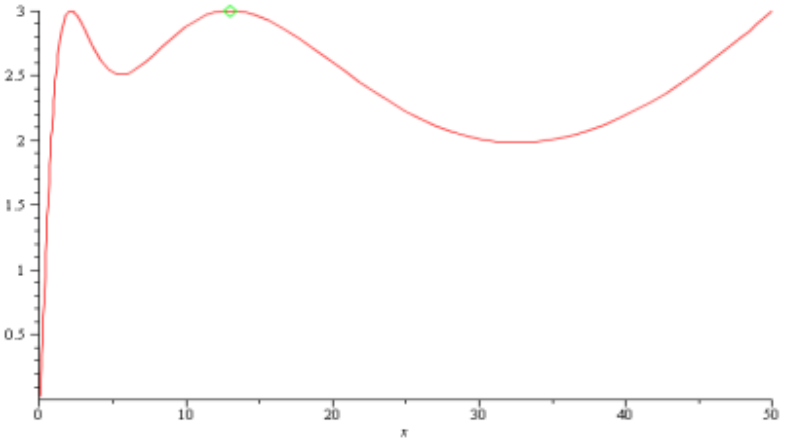


Figure 4. Standardized Variance Model (3) with design region [0, 1.5] , θ_3 =99.77242, $\theta_2 = 10.13847$and $\theta_4 = 1.008629$

6. CONCLUSION

D-optimal designs for Morgan Mercer Flodin models with two and three parameters without intercept and homoscedastic variance has introduced. Our tools are result which are derived from the Kiefer-Wolowitz Equivalence Theorem, we adopt Theorem 1 of Li and Majumdar (2008), Lemma 3.2 of Li and Balakrishnan (2011) and properties of Tchebyshev system. The formers, Theorem 3.1 and Theorem 3.2 are general result that gives D-optimal designs for some design region. Michaelis Menten and EMAX models are specified cases of Morgan Mercer Flodin model.

ACKNOWLEDGEMENTS

This work was supported by BOPTN DIPA UNDIP No. 186-19/UN7.5.1/PG/2014. The authors would like to thank the Referees and the Associate Editor for their valuable comments and suggestions which improve the paper.

REFERENCES

1. Atkinson, A.C., Donev, A.N. and Tobias, R.D. (2007). *Optimum Experimental Designs, with SAS*. OXFORD University Press.
2. Chernoff, H. (1953). Locally optimal designs for estimating parameters. *The Annals of Statistics*, 24, 586-602.
3. Chang, F.C. and Lay, C.F. (2002). Optimal Designs for a Growth Curve Models. *Journal of Statistical Planning and Inference*,104, 427-438.

4. Dette, H., Melas, V.B. and Wong, W.K. (2004). Optimal design for Goodness of Fit of the Michelis Menten Enzyme Kinetic Function. working paper. Ruhr-Universitat Bochum, Fakultat furr Mathematik, 44780 Bochum, Germany.
5. Dette, H. and Pepelyshev, A. (2008). Efficient Experimental Designs for Sigmoidal Growth Models. *Journal of Statistical Planning and Inference*, 138, 2-17.
6. Dette, H. and Kiss, C. (2012). Optimal Designs for Rational Regression Models. *Working Paper*. Fakultat fur Mathematik, Ruhr-Universitat Bochum, 44780 Bochum, Germany.
7. Holford, N. (2013). *Pharmacodynamic Principles and the Time Course of Immediate Drug Effect.* Department Pharmacology and Clinical Pharmacology, University of Auckland, New Zealand.
8. Jericevic, Z. and Kuster, Z. (2005). Nonlinear Optimization of Parameters in Michaelis Menten Kinetics. *Croatia Chemica Acta CCACAA*,78(4),519-523.
9. Karlin, S. and Studden, W.J. (1966).*Tchebyshev System: With Application in Analysis and Statistics.* Wiley, New York.
10. Kiefer, J. and Wolfowitz, J. (1960). The Equivalence of Two Extremum Problems. Can. Jnl. Math., 12, 363-366.
11. Kiefer, J. (1961). Optimum Designs in Regression Problems II. *Annals of Mathematical Statistics*, 32, 298-325.
12. Knudsen, J.O. (2001). *General Concepts of Pharmacodynamics*. Department of Clinical Microbiology, Rigshospitalet Copenhagen Denmark.
13. Li, G., and Majumdar, D. (2008). D-optimal designs for logistic models with three and four parameters. *Journal of Statistical Planning and Inference*, 138, 190-1959.
14. Li, G. (2011). Optimal and Eficient Designs for Gompertz Regression Models *Ann Inst Stat Math*, DOI 10.1007/s10463-011-03040-y.
15. Li, G. and Balakrishnan, N. (2011). Optimal Designs for Tumor Regrowth Models. *Journal of Statistical Planning and Inference* ,141,644-654.
16. Lopez, J., et al. (2000). A Generalized Michaelis Menten Equation for the Analysis of Growth. *JANIM SCI*, 78, 1816-1828.
17. Morgan, P.H., Mercer. L.P. and Flodin, N. (1975). General model for nutritional responses of higher organisms. *Proc. Nat. Acad. Sci. USA*, 72, 4327-4331.
18. Sengul, T. and Kiraz, S. (2005). Nonlinear models for growth curves in large white tukeys.*Turk J Vet Anim Sci*, 29, 331-337.
19. Tariq, M.M., Iqbal, F., Eyduran, E., Bajwa, M.A., Huma, Z.E. and Waheed, A. (2013). Comparisson of nonlinear function to describe the growth in Mengali sheep breed of Balochistan. *Pakistan J. Zool*, 45(3), 661-665.
20. Tjorve, E. (2003). Shapes and Function of Species Area Curve: a review of possible model. *Journal of Biogeography*, 30, 827-835.
21. Topal, M. and Bolukbasi, S.C. (2008). Comparison of Nonlinear Curve Models in Broiler Chickens. J.Appl.Anim.Res, 34,149-152.
22. Utomo, P.M., Suhendang, E., Syafii, W. and Simangunsong, B.C.H. (2012). Model Produksi Daun Pada Hutan Tanaman Kayu Putih (*Melaleuca cajuputi* Subsp. *cajuputi* Powell) Sistem Pemanenan Pangkas Tunas. *Jurnal Penelitian Hutan Tanaman* 9(4), 195-208.
23. White, L. (1973). (1973). An Extension of the general equivalence theorem to nonlinear models. *Biometrika*, 60, 345-348.

24. Widiharih, T., Haryatmi, S. and Gunardi. (2012). Rancangan D-Optimal Untuk Model Regresi Eksponensial Dengan Mean Terboboti. Prosiding KNM XVI UNPAD.
25. Widiharih, T., Haryatmi, S. and Gunardi. (2013). D-optimal designs for weighted exponential and generalized exponential models. *Applied Mathematical Sciences*, 7, 1067-1079.
26. Widiharih, T., Haryatmi, S. and Gunardi. (2013). Pendekatan Numeris Rancangan D-optimal Untuk Model Regresi Eksponensial Tergeneral Tiga Parameter. *Prosiding Seminar Nasional Statistika UNDIP*.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 41-54

THE PPERFORMANCE OF LS, LAD, AND MLAD REGRESSION ON THE STACK LOSS DATA

Setyono[1], I Made Sumertajaya[2], Anang Kurnia[3] and **Ahmad Ansori Mattjik[4]**
Department of Agrotechnology, Bogor Djuanda University,
Jl. Tol Ciawi 1 Bogor 16720, Indonesia
Email: onoytes@yahoo.co.id,
imsjaya@yahoo.com,
akstk29@yahoo.com,
aamattjik@gmail.com

ABSTRACT

Estimation of regression coefficients based on residuals optimization which are commonly known, are by minimizing the residual sum of squares (LS) and by minimizing the sum of absolute residual (LAD). Estimation by minimizing the maximum absolute residual (MLAD) has not been developed. The purpose of this study was to determine whether the linear programming can be used to estimate regression coefficients that minimize the maximum absolute residual and compare its results with the results of the LS and LAD. The data used was the Stack Loss data that commonly used for regression method testing. The study used a simulation experiment with 1000 replications using the error generated from the normal distribution. The results showed that linear programming can be used to estimate regression coefficients that minimize the maximum absolute residual, the LAD regression is the best for cross-validation criteria, whereas LS regression is the most stable according to the criteria of residual sum of squares, sum of absolute residual, and the maximum absolute residual.

KEYWORDS

absolute residual, cross validation, linear programming, MLAD, minimax regression.

1. INTRODUCTION

The theory of classical linear model is a theory for conditional estimation, i.e., $\mu(Y|\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$. That regression is commonly known as mean model, ie modeling the mean of Y at a particular value of x. Regression coefficients estimation can be done based on error distribution or based on residual optimization. Estimation based on error distribution is using maximum likelihood method, while the estimation based on residual optimization can be done using several ways. The way that is commonly known is by minimizing the residual sum of squares (LS) and the other is by minimizing the sum of absolute residual (LAD). Least squares method has been implemented on all computer program packages for statistics, while the LAD method implementation can be done using the computer program that provides quantile regression package. Meanwhile, residual optimization by minimizing the maximum absolute residual (MLAD) has been

pioneered by Rudolf *et al.* (1999), but has not been developed in a computer program package for statistics.

Least squares regression can be implemented earliest because the regression coefficient estimators and their properties can be obtained analytically. Therefore, regression learning for undergraduate students is using this method. The LS estimator is often used as an initial value for the regression coefficient estimation that require iterative solution, for example Student's t-regression.

LAD method can be done in several ways, among others are with the median regression and iteratively weighted regression (IWLS). LAD method produces a robust estimator because it gives weight that the magnitude is inversely proportional to the size of residual. Therefore LAD regression is not susceptible to outliers. LAD estimator is not unique and Hao and Naiman (2007) have shown that the median is one solution. Therefore LAD regression can be done with a median (quantile-0.5) regression.

MLAD regression used when the desired model has never had a large residual, or the maximum residual is minimized. This is important because the issue of public interest, the case with a large deviation is in the spotlight, even though other cases safe. As a simple illustration is setting the wheels, either too tilted to the left or too tilted to the right at a point can be a problem even though at most of points is in the middle. In a such case illustration, the LS regression and LAD regression is less suitable, because even though at almost all of other points have a small residual, not close the possibility of large residual at one point.

At this time, not yet available computer program package for MLAD regression. In this study, the MLAD regression coefficient estimation is done using linear programming. The first objective of this study is to examine whether the residual optimization using linear programming is successfully minimizing the maximum absolute residual. For that objective, the maximum absolute residual obtained from MLAD regression compared with the maximum absolute residual obtained from other regression methods.

In parameter estimation based on the residual optimization, the method used is also as a goodness criteria. For example, if the criteria used is the residual sum of squares, the best method is least squares. The second objective of this study is to examine whether the MLAD method is the best in terms of maximum absolute residual, and also is quite good in terms of residual sum of squares and the sum of absolute residual.

2. DATA AND METHOD

This study consists of three subjects, namely: computational techniques, applications on the Stack Loss data, and simulation experiments. In the computation techniques will be presented how to calculate MLAD regression. On the data application on Stack Loss data, will be presented the standard error estimation through residual bootstrap, also will be presented the goodness of estimators evaluation through cross validation. Meanwhile, on the simulation experiments will be done residual evaluation when the response data have normal distribution.

Computation of the MLAD Regression Coefficients

In the regression model $y_i=\mathbf{x}_i'\boldsymbol{\beta}+\varepsilon_i$, suppose $\mathbf{b}$ is the estimator of $\boldsymbol{\beta}$, then the model estimator is $y_i=\mathbf{x}_i'\mathbf{b}+e_i$ or in matrix notation is $\mathbf{y}=X\mathbf{b}+\mathbf{e}$. Regression method that minimizes the maximum absolute residual (MLAD) can be written in the arguments $\min\{\max|\, y_i-\mathbf{x}_i'\mathbf{b}\,|\}$.

MLAD estimators can be obtained with the following guidelines. Suppose y_i is the response of the i^{th} observation, $\mathbf{x}_i'$ is the i^{th} observation of the covariates vector, $\mathbf{b}$ is the regression coefficient vector, and $z\geq 0$ is the upper boundary of absolute residual so that $0\leq|y_i-\mathbf{x}_i'\mathbf{b}|\leq z$ for all i^{th} observation. When $e_i\geq 0$ then $0\leq y_i-\mathbf{x}_i'\mathbf{b}\leq z$ or $\mathbf{x}_i'\mathbf{b}+z\geq y_i$. When $e_i<0$ then $-z\leq y_i-\mathbf{x}_i'\mathbf{b}<0$ or $\mathbf{x}_i'\mathbf{b}-z\leq y_i$. Therefore, in MLAD regression the value of z is minimized using constraint $\mathbf{x}_i'\mathbf{b}+z\geq y_i$ and $\mathbf{x}_i'\mathbf{b}-z\leq y_i$. A more detailed study of linear programming can refer to McCarl and Spreen (1977), while to implement it in the R language can refer to Rizzo (2008).For the first purpose, the maximum absolute residual obtained by procedure above is compared with the maximum absolute residual obtained by other methods.

The data will be used is the Stack Loss data (Montgomery, Peck 1992), which once used by normal modeling, normal modeling with outliers diagnosis, Huber method, Andrews method, and t modeling (Setyono *et al.* 1996). This data consists of 21 observations of four variables, namely the stack loss (Y), water flow (X1), water temperature (X2), and acid concentration (X3). Stack Loss Data are presented in Table 1.

Table 1
Stack Loss Data

No	Y	X1	X2	X3	No	Y	X1	X2	X3
1	42	80	27	89	11	14	58	18	89
2	37	80	27	88	12	13	58	17	88
3	37	75	25	90	13	11	58	18	82
4	28	62	24	87	14	12	58	19	93
5	18	62	22	87	15	8	50	18	89
6	18	62	23	87	16	7	50	18	86
7	19	62	24	93	17	8	50	19	72
8	20	62	24	93	18	8	50	19	79
9	15	58	23	87	19	9	50	20	80
10	14	58	18	80	20	15	56	20	82
					21	15	70	20	91

Standard Error Estimation Using Bootstrap

In the LS method standard error of regression coefficients can be derived mathematically into a formula (closed form). In MLAD method, regression coefficients and their standard errors cannot be expressed in closed form. For the purposes of inference, it is required standard error estimates for regression coefficients based on a set of data. One way to estimate standard error is through bootstrapping. Bootstrap is taking a sample with replacement repeatedly. Bootstrap done in two ways, namely bootstrap of

observations and bootstrap of residuals (Givens and Hoeting 2005). Bootstrap of observation means considering the value of observation pairs (x, y) is a random sample of the population of observation pairs (x, y). Bootstrap of observations has been done by Setyono *et al.* (1996). While the bootstrap of residuals means consider the design matrix is fixed, while the error is random. Bootstrap of residuals has been done by Zhu and Jing (2010).

Estimation standard error of the regression coefficient is more appropriate to use the bootstrap of residuals. For that purpose, it is assumed that the distribution of residual e_i represent the distribution of error ε_i, so it can be done bootstrap based on e_i size n. Step details as follows:

1. Performed regression on the data to be analyzed, in order to obtain the regression coefficients **b** and residual **e**.
2. Calculated $\hat{\mathbf{u}} = \mathrm{X}\mathbf{b}$
3. Taken samples with replacement of e_i result of step 1 size of n, as value of d_i
4. Calculated value of $z_i = \hat{u}_i + d_i$
5. Performed regression **z** on X, thus obtained regression coefficient a
6. Be repeated 1000 times to steps 3-6

The standard deviation of a is considered as standard error of b.

Cross Validation

One of goodness of the method on a set of data is evaluated using cross validation criteria. Cross validation in this study is done with the following steps:

1. Starting from i=1
2. The i^{th} observation is dropped
3. Do the regression coefficient estimation exclude the i^{th} observation
4. Estimate the value of the i^{th} response based on the model of third step
5. Calculate difference between the value of the i^{th} response observed with the value of i^{th} response estimated, and then recorded as e_i
6. Perform steps 2 through 5 for i=2, 3, …, n

After that, it is calculated cross validation (CV) value in three ways, namely:

$$\text{Mean of squared prediction error} = \mathrm{CV}(1) = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

$$\text{Mean of absolute prediction error} = \mathrm{CV}(2) = \frac{1}{n}\sum_{i=1}^{n} |e_i|$$

$$\text{Maximum of absolute prediction error} = \mathrm{CV}(3) = max(|e_i|)$$

Simulation Study

Study to examine whether the MLAD regression also quite good in terms of residual sum of squares and sum of absolute residual held by the regression coefficient estimation for 1000 sets of data with methods MLAD, LAD, and LS. In each method be measured the residual sum of squares (SSR), the sum of absolute residual (SAR), and the maximum absolute residual (MAR). After that, their value of SSR, SAR, and MAR be compared.

The simulation was performed with the following steps:

1. Assume that the Stack Loss data is population data
2. Calculated $\beta 0$, $\beta 1$, $\beta 2$, $\beta 3$ by LS then calculated $\mu_i=\beta 0+\beta 1X1_i+\beta 2X2_i+\beta 3X3_i$
3. Generated 21 random number ei from standard normal distribution
4. Calculated yi=μi+ei
5. Regressed Y on X by LS, LAD, and MLAD method, then from each method is obtained b0, b1, b2, b3, residual sum of squares, sum of absolute residual, and maximum absolute residual
6. Be repeated 1000 times to steps 3-5
7. Calculated the average of b0, b1, b2, b3, residual sum of squares, sum of absolute residual, and maximum absolute residual, each of the LS, LAD, and MLAD methods

3. RESULT AND DISCUSSION

Computation Techniques

Suppose that on the data set {2,3,5,7,11,13} will be determined the measure of central tendency, which the maximum absolute residual is minimized, in other words, will be determined k that makes $\max(|y_i - k|)$ is minimized. The relationship between measures of central tendency with a maximum absolute residual is forming a curve concave upward. (Figure 1). From that figure it can be seen that the minimum of residual sum of squares is unique and differentiable at its minimum point, minimum of maximum absolute residual is unique but not differentiable at its minimum point, while the minimum of sum of absolute residual is not unique. Thus the LS and MLAD estimators are unique, but the LAD estimator is not unique.

Because not differentiable at its minimum point, the MLAD estimators cannot be obtained by calculus, but can be obtained through a linear programming with the objective function of minimizing z with constraints:

- $k+z\geq 2$, $k+z\geq 3$, $k+z\geq 5$, $k+z\geq 7$, $k+z\geq 11$, $k+z\geq 13$ (can be represented by $k+z\geq 13$)
- $k-z\leq 2$, $k-z\leq 3$, $k-z\leq 5$, $k-z\leq 7$, $k-z\leq 11$, dan $k-z\leq 13$ (can be represented by $k-z\leq 2$)

The point of intersection between the k+z= 3 with k-z=2 occurs at k=7.5 and z=5.5. The minimum value of z occurs at that intersection point. In the intercept model regression, MLAD estimators value is the mid-range, (2 + 13) /2=7.5, as published by Akcay and At (2006), and the maximum value of absolute residual is the half of range, (13-2) / 2 = 5.5.

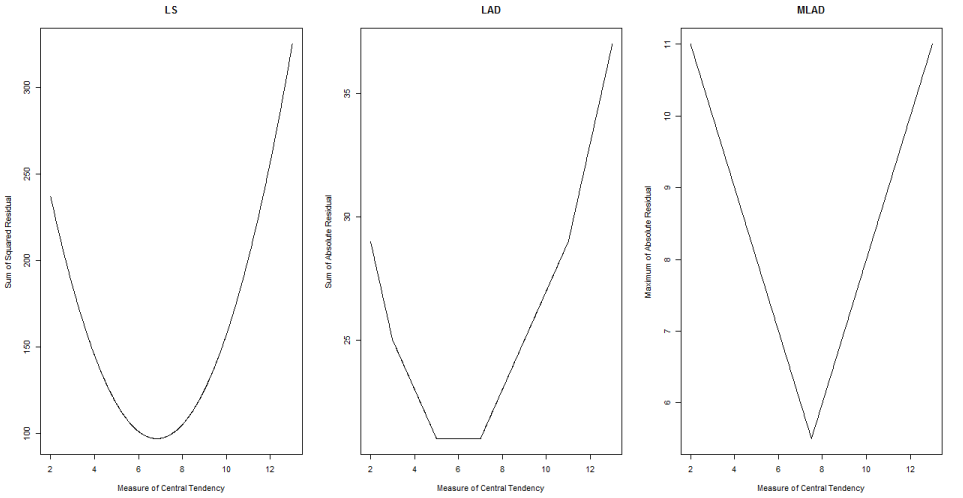


Figure 1: The Relationship between Measures of Central Tendency with Maximum of Absolute Residual, Sum of Absolute Residual, and Sum of Squared Residual

Suppose that the group of pairs data (x, y) to be regressed is {(2,2), (4,3), (6,5), (8,7), (10,11)}. Simple linear regression y = a + bx using MLAD is done by minimizing the objective function z with constraints:

- $a+2b-z\leq 2$, $a+4b-z\leq 3$, $a+6b-z\leq 5$, $a+8b-z\leq 7$, $a+10b-z\leq 11$
- $a+2b+z\geq 2$, $a+4b+z\geq 3$, $a+6b+z\geq 5$, $a+8b+z\geq 7$, $a+10b+z\geq 11$

The resulting regression equation is y = -1 125 + 1.125x with maximum absolute residual (z) is 0.875.

Applications on the Stack Loss Data

Regression analysis on the Stack Loss data has been done using several methods, namely LS, Huber, Andrew (Montgomery, Peck 1992), t_3, t_5 (Setyono *et al.* 1996), and at this time also used LAD and MLAD. The estimated value of regression coefficient, the maximum value of the absolute residual (MAR), the sum of the absolute residual (SAR), and the sum of squared residual (SSR) of several methods for the Stack Loss data are presented in Table 2.

Table 2
Value of Maximum of Absolute Residual, Sum of Absolute Residual, and Sum of Squared Residual from Some Methods on Stack Loss Data

Metode	b0	b1	b2	b3	MAR	SAR	SSR
LS	-39.92	0.72	1.30	-0.15	7.24	49.70	178.83
LAD	-39.69	0.83	0.57	-0.06	9.48	42.08	227.47
MLAD	-27.18	0.58	1.86	-0.34	4.74	61.69	223.10
Huber	-41.00	0.83	0.91	-0.13	8.47	46.28	194.23
Andrew	-37.20	0.82	0.52	-0.07	9.23	43.84	240.58
t(v=3)	-39.96	0.86	0.69	-0.11	9.03	44.89	212.39
t(v=5)	-39.92	0.84	0.88	-0.13	9.65	47.20	202.26

It appears that on the Stack Loss data, MLAD estimates obtained from linear programming produces the smallest maximum of absolute residuals, LS estimates produces the smallest sum of squared residuals, while the LAD estimates produces the smallest sum of absolute residuals. Thus the linear program has been made already successfully choose the regression coefficient that minimize the maximum of absolute residual.

Based on the proximity of the resulting regression coefficient, a number of the above methods can be classified into three groups. The first group is the LS method, the second group is MLAD method, and the third group is the LAD, Huber, Ramsay, Anrew, Hampel, and t methods. The third group is known as a robust method that is not easily affected by outliers. Its characteristics are: give great weight to the small residual and give little weight to the large residual.LAD method can be solved through an iterative weighted regression that give great weight to the small residual and give little weight to the large, so it is worth entering that category.

Standard error of statistic reflects the efficiency measure of the parameter estimator and useful for inference purposes. Standard error of the LS estimator can be obtained analytically, but the standard error of the MLAD estimator cannot. This time the standard error of the three methods obtained through residual bootstrap approach as described in the methodology. The magnitude of the standard error of the regression coefficient of the three methods are presented in **Table 3**.

Table 3
Mean and Standard Error of Regression Coefficients Using Bootstrap

Method	b0	b1	b2	b3
MLAD	-27.0316	0.5765	1.8612	-0.3395
	(9.8465)	(0.1106)	(0.2950)	(0.1288)
LAD	-39.3659	0.8286	0.5820	-0.0653
	(8.1736)	(0.0896)	(0.2394)	(0.1043)
LS	-40.0537	0.7147	1.2944	-0.1498
	(10.4007)	(0.1205)	(0.3278)	(0.1402)

The smallest value of the standard error obtained through residual bootstrap method is achieved by the LAD, followed by MLAD, then LS. Thus according to this criteria, LAD is the best for Stack Loss data.

Goodness of the method can be evaluated based on its ability to predict, one of them is done using cross validation criteria. Cross validation is performed by removing an observation, perform regression coefficient estimates based on the remaining observations, and then estimate the response of the discarded observation and calculate the difference (error). Cross validation on the Stack Loss data is initially carried out on the first observation, while other observations are as trained. Then carried out cross validation of the second observation, while other observations as trained. And so on until the last observation. Based on the prediction error value of the first to the last observation will be obtained the mean of squared prediction error (CV1), the mean of absolute prediction error (CV2), and the maximum of absolute prediction error (CV3).

The recapitulation of the cross-validation value that obtained by LS, LAD, and MLAD estimators on Stack Loss data are presented in Table 4.

Table 4
Cross Validation of MLAD, LAD, and LS Regression on Stackloss Data

Cross Validation Type	MLAD	LAD	LS
Mean of Squared Prediction Error	15.2375	11.0005	13.8985
Mean of Absolute Prediction Error	3.3114	2.0642	2.9780
Maximum of Absolute Prediction Error	9.0630	9.5366	10.1161

It appears that the smallest maximum of absolute prediction error is obtained by MLAD regression, followed by LAD and the last is LS. Meanwhile, the smallest mean of absolute prediction error and the smallest mean of squared prediction error are achieved by LAD regression, followed by LS, and then MLAD. Thus in general the best regression method using cross validation criteria for Stack Loss data is LAD regression.

Stackloss data is known as data "disliked" by LS regression, so many robust regression which try these data as an alternative to the LS regression. Even if the LS regression is applied to this data, it is usually accompanied by a diagnosis of outliers. LAD regression including robust regression, or at least, more robust than LS and MLAD. Therefore it is understandable if the prediction error generated by the LAD is relatively better.

SIMULATION STUDY

Furthermore, it is be done the simulation with 1000 replications using Stack Loss data. As covariates are air flow (X1), water temperature (X2), and acid concentration (X3), whereas as the response variable is predicted stack loss value according to LS regression (Y) with Y= -39,9197+ 0,7156 X1+1.2953 X2 - 0,1521 X3 + random number Normal(0,1). The distributions of maximum of absolute residual (MAR), sum of absolute residual (SAR), and sum of squared residual (SSR) generated by each of the regression equations are presented in Figure 2. The narrowest distribution of maximum of absolute residual is achieved by MLAD method, followed by LS, then LAD. The narrowest distribution of sum of absolute residual is achieved by LAD, followed by LS, than MLAD. The narrowest distribution of residual sum of squares is achieved by LS, followed by LAD, then MLAD.

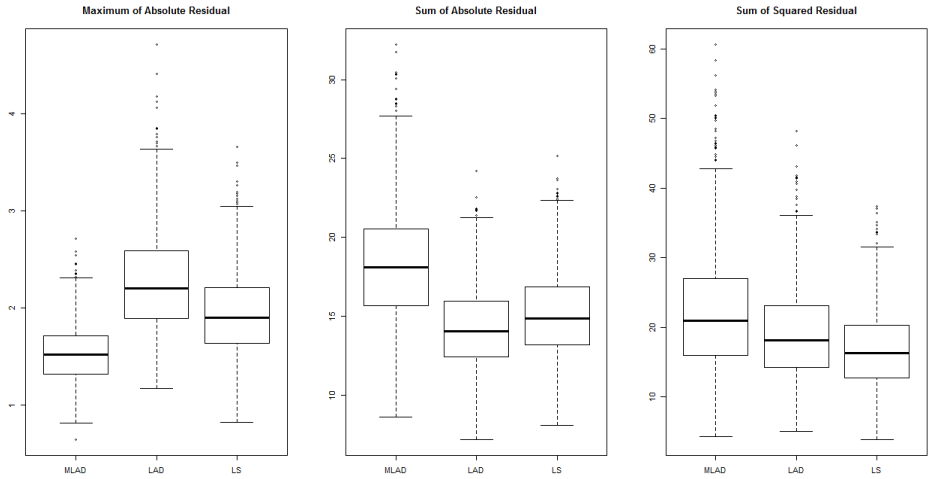


Figure 2: Boxplot of Maximum of Absolute Residual, Sum of Absolute Residual, and Sum of Squared Residual on MLAD, LAD, and LS

Descriptive statistics of maximum of absolute residual, sum of absolute residual, and sum of squared residual, result of simulation on Stack Loss data, are presented in Table 5. The smallest average of maximum of absolute residual is achieved by MLAD method, followed by LS, then LAD. The smallest average of sum of absolute residual is achieved by LAD, followed by LS, than MLAD. The smallest average of residual sum of squares is achieved by LS, followed by LAD, then MLAD. Thus, each method is superior in accordance with its criteria of goodness, and in general the LS method is relatively well on all criteria of goodness.

Table 5
Descriptive Statistics Maximum of Absolute Residual, Sum of Absolute Residual, and Sum of Squared Residual Result of Simulation on Stack Loss Data

Method		MAR	SAR	SSR
MLAD	minimum	0,6402	8,6039	4,3329
	maximum	2,7096	32,1995	60,6058
	mean	1,5298	18,3253	22,4346
	standar deviation	0,2958	3,7078	8,7871
LAD	minimum	1,1656	7,1781	5,0254
	maximum	4,7120	24,1960	48,2055
	mean	2,2637	14,2244	19,1623
	standar deviation	0,5299	2,6261	6,7976
LS	minimum	0,8141	8,0723	3,8718
	maximum	3,6594	25,1820	37,4102
	mean	1,9364	15,0759	17,0002
	standar deviation	0,4200	2,7702	5,8166

The frequency of the LS method, LAD, and MLAD ranked first, second, and third according to the sum of squared residual (SSR), the sum of absolute residual (SAR), and the maximum of absolute residual (MAR) are presented in Table 6.

Table 6
Frequency of ranked 1, 2, 3 according to the Maximum of Absolute Residual, Sum of Absolute Residual, and Sum of Squared Residual

Method	Ranked MAR			Ranked SAR			Ranked SSR		
	1	2	3	1	2	3	1	2	3
MLAD	1000	0	0	0	10	990	0	235	765
LAD	0	101	899	1000	0	0	0	765	235
LS	0	899	101	0	990	10	1000	0	0

It appears that the MLAD method always ranked first by the maximum of absolute residual, LAD method always ranked first by sum of absolute residual, and LS method always ranked first according to sum of squared residual. According to the criteria of maximum of absolute residual and sum of absolute residual, LS method is generally ranked second. According to the criteria of sum of squared residual, LAD method is better than MLAD in terms of ranked second. Thus the LS method is moderate, ie obtaining the best category according to criteria of sum of squared residual and obtain good categories according to the criteria of maximum of absolute residual and sum of absolute residual.

Distribution of b0, b1, b2, and b3 are presented in Figure 3. Judging from the boxplot can be seen that in general the distribution b0, b1, b2, and b3 are symmetric. From Kolmogorov-Smirnov test results using the R program (Table 7) it can be concluded that the distribution of b0, b1, b2, and b3 are not rejected as a normal distribution. This result is a good information and useful for inference.

Table 7
Kolmogorov-Smirnov Test of Regression Coefficients

Method	Regression Coefficient	Statistic	p-value
MLAD	b0	0,0207	0,7839
	b1	0,0241	0,6065
	b2	0,0201	0,8132
	b3	0,0127	0,9969
LAD	b0	0,0162	0,9558
	b1	0,0196	0,8384
	b2	0,0171	0,9317
	b3	0,0179	0,9041
LS	b0	0,0239	0,6172
	b1	0,0137	0,9916
	b2	0,0187	0,8750
	b3	0,0208	0,7799

Descriptive statistics of b0, b1, b2, and b3 are presented in Table 8. Based on the distribution of b0, b1, b2, and b3 in Figure 3 and the standard error of b0, b1, b2, and b3 in Table 8, the best method is LS, followed by LAD, then MLAD. This is understandable because in this simulation the data are generated from a normal distribution, while the LS method obtains parameter estimator similar to its obtained by the maximum likelihood for normal distribution.

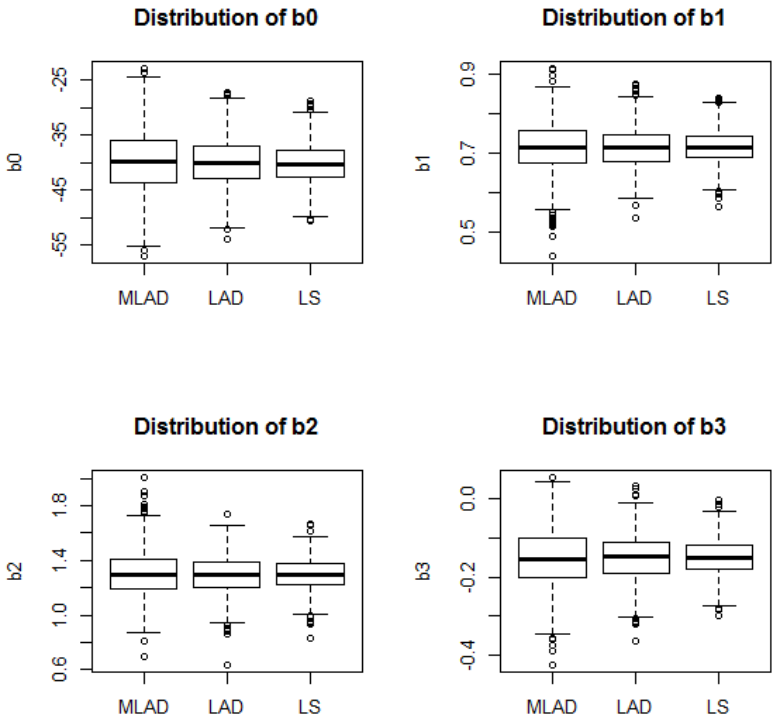


Figure 3: Distribution of b0, b1, b2, and b3

Parameters of the regression coefficients those are used to generate these simulation data are $\beta 0 = -39.9197$, $\beta 1 = 0.7156$, $\beta 2 = 1.2953$, and $\beta 3 = -0.1521$. From Table 8 appears that the mean value of b0, b1, and b3 of the three methods already converging to its parameters. This indicates that these three methods empirically produce no statistical bias. This is supported by the ratio between the mean error $(b-\beta)$ with root mean squared error, the value (if calculated) is under 1.

Table 8
Descriptive statistics of b0, b1, b2, and b3

		b0	b1	b2	b3
MLAD	minimum	-57,0016	0,4392	0,7002	-0,4264
	maximum	-22,9071	0,9121	2,0053	0,0555
	Mean	-39,9272	0,7144	1,3017	-0,1528
	standard deviation	5,6163	0,0644	0,1753	0,0749
	mean error	-0,0075	-0,0013	0,0064	-0,0007
	mean of absolute error	4,5180	0,0504	0,1381	0,0598
	maximum of absolute error	17,0820	0,2765	0,7100	0,2743
	mean of squared error	31,5119	0,0041	0,0307	0,0056
LAD	minimum	-53,8980	0,5334	0,6364	-0,3659
	maximum	-27,2071	0,8729	1,7358	0,0341
	mean	-40,0060	0,7138	1,2968	-0,1504
	standard deviation	4,5440	0,0505	0,1369	0,0591
	mean error	-0,0863	-0,0018	0,0015	0,0018
	mean of absolute error	3,6208	0,0402	0,1079	0,0472
	maximum of absolute error	13,9783	0,1822	0,6589	0,2137
	mean of squared error	20,6346	0,0026	0,0187	0,0035
LS	minimum	-50,6775	0,5628	0,8344	-0,3009
	maximum	-28,8430	0,8366	1,6659	-0,0006
	mean	-40,1205	0,7149	1,2969	-0,1497
	standard deviation	3,6328	0,0417	0,1126	0,0478
	mean error	-0,2009	-0,0007	0,0016	0,0024
	mean of absolute error	2,8900	0,0331	0,0887	0,0383
	maximum of absolute error	11,0767	0,1529	0,4609	0,1515
	mean of squared error	13,2244	0,0017	0,0127	0,0023

Although MLAD regression produces the smallest maximum absolute residual, does not mean that MLAD regression also produces the smallest maximum error in each regression coefficient generated. The result of the simulation using 1000 replicates that are generated from a normal distribution shows that the smallest maximum error is obtained by LS, followed by LAD, and the last is MLAD (Table 8). Thus MLAD regression is more suitable to estimate the response value than to estimate the functional relationship between the covariates with the response.

4. COMMENTS AND CONCLUSION

Measures of central tendency of a data set can be viewed as the result of residual optimization, while the residual optimization itself produce a measure of dispersion. Residual optimization by minimizing the residual sum of squares produces measure of central tendency in the form of mean and measure of dispersion in the form of variance. Residual optimization by minimizing the sum of absolute residual produces measure of central tendency as form of median and measure of dispersion as form of mean of absolute deviation. Meanwhile residual optimization by minimizing the maximum absolute residual produces measure of central tendency as form of mid-range and measure of dispersion as form of range.

Residual optimization by minimizing the maximum absolute residual can be developed into a method for estimating the regression coefficients. Regression obtained by minimizing the maximum absolute residual (MLAD), minimizing the sum of absolute residual (LAD), and minimizing the residual sum of squares (LS) is empirically produces no statistically biased. If it is used cross validation criteria, LAD regression is the best method for Stack Loss data. Meanwhile, LS is a method which is the most stable of all the residual optimization criteria. When the response variable is modeled as normal distribution, regression that minimizing the maximum absolute residual does not automatically become a regression that minimizing the maximum error for the regression coefficients. Therefore it is necessary to study more detailed simulations to obtain characteristics.

The MLAD regression coefficients obtained by linear programming. The nature of the linear program is only concerned with the different constraints and constraints which provide more stringent restrictions. This implies that MLAD regression disregard repeated observations. On the other hand these properties open up opportunities to obtain regression coefficient resulted from a subset of observations that equal with the regression coefficients resulted from the whole observation.

During this time statistical modeling only focuses on the mean response modeling. In the food supply modeling, which needed is to provide the minimum amount, but still sufficient. Desired model is not a mean model, but the maximum model or max $(Y|\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$. The resulting regression line is located at the top of the observations, so there is no positive residual. The fact, that MLAD regression ignoring repeated observations, also opens the opportunity to do the maximum or minimum response modeling. It will be a special case of quantile regression (Koenker and Bassett 1978).

In multiple regression, sometimes it is obtained regression coefficient with sign that do not make sense. MLAD regression performed with a linear program, so it is possible to control the signs and the range of values of regression coefficients that are in a reasonable range. In future, statistical modeling should not only be a covariate modeling and modeling of functional relationship between covariates and response, but also need to modeling of regression coefficients.

REFERENCES

1. Akcay, H. and At, N. (2006). Convergence analysis of central and minimax algorithms in scalar regressor models. *Mathematics of Control, Signals and Systems*. 18(1), 66-99.
2. Hao, L., Naiman, D.Q. (2007). *Quantile Regression*. California: Sage Publications, Inc.
3. Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*. Vol. 46, No. 1 (Jan 1978): 33-50
4. McCarl, B.A. and Spreen, T.H. (1997). *Applied Mathematical Programming Using Algebraic Systems*. Copyright Bruce A. McCarl and Thomas H. Spreen
5. Montgomery, D.C. and Peck, E,A. (1992). *Introduction to Linear Regression Analysis*. New York: John Wiley & Sons.
6. Rizzo, M.L. (2008). *Statistical Computing with R*. Chapman & Hall. London
7. Rudolf, M., Wolter, H., Zimmermann, H. (1999). A linear model for tracking error minimization. *Journal of Banking & Finance* 23 (1999) 85-103
8. Setyono, Notodiputro, K., Aunuddin and Mattjik, A.A. (1996). Pemodelan statistika atas dasar sebaran t student. *Forum Statistika dan Komputasi*, 1(2), 10-16
9. Zhu, J. and Jing, P. (2010). The analysis of bootstrap method in linear regression effect. *Journal of Mathematics Research*. 2(4); November 2010: 64-69.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 55-62

STRUCTURAL EQUATION MODELING WITH PARTIAL LEAST SQUARE ESTIMATOR (CASE STUDY IN MEASUREMENT OF THE LEVEL OF SATISFACTION AUDITOR OF GREEN AUDITS PROGRAMS)

Firmina Adlaida
Perwira No.44 Street, Dramaga, Bogor, West Java, Indonesia.
Email: firminadlaida@gmail.com

ABSTRACT

Statistical analysis with latent variables that are not observed directly uses structural equation modeling. The development of structural equation modeling gives some parameter estimation methods one of them partial least squares method. In this method, Parameter estimation and model's feasibility do not require assumptions distribution (distribution-free) from observable variables and the sample size should not be large. This method was applied to measure the satisfaction levels of the auditee to the auditor's green audits program and the factors that influence the satisfaction were analyzed. Structural model on partial least square gave exogenous variables that significantly affected endogenous variables (satisfaction), tangible variable. The variable had positive influence on the satisfaction which means better physical form, facilities, processes and results of the internal audit would make the auditee satisfied with the performance of the auditor.

Keywords: latent variables, structural equation modeling, partial least square, satisfaction level, green audits programs

1. INTRODUCTION

The Ministry of Agriculture Indonesia Republic as one of the state institutions is required to have General Inspectorate in accordance with the publication of the Presidium of the Cabinet decision No. 15 of 1966. General Inspectorate of the Ministry is assigned as the internal supervisor of The Ministry is required to create an atmosphere based supervision of professional competence, capability, and integrity. This is in accordance to reform the civil service in The Ministry is directed to the achievement of good governance, transparent, and accountable.

General Inspectorate of the Ministry launched green audits program in order to create professional supervision and to realize the vision and mission. The green audits program seeks to auditors carries on supervisory activities are more focused so that effective, efficient, economical, and obey the rules. Evaluation of the performance of the auditor who has conducted the audit process performed after the audit is completed. This analysis concerns about the auditee satisfaction level on the performance of the auditor who has worked. Factors that influence the auditee's satisfaction level will also be investigated. Satisfaction is variable that can't be observed directly. That is, measurement of the

variables can not be measured directly but through their indicators. Therefore, such variables are called latent variables, while the indicators are measurable variables called indicators (Schumacker dan Lomax 1996).

Factors that affect the auditee satisfaction can be observed through tangible variables (physical manifestation facilities, personnel, processes and results of the internal audit), reliability (professionalism of auditors is a blend of knowledge, skills and abilities so that auditors can carry out the assignment in accordance with auditing standards), responsiveness (speed of auditors response to assist and provide auditing and consulting services), assurance (guarantees of certainty that can be provided by the auditor to the auditee foster trust audit services, related to the independence and integrity of auditors) (Zeithaml *et al.* 1990). These four factors are considered to affect the auditee's satisfaction is also a latent variable. Bollen (1989) used structural equation modeling that can simultaneously analyze the complex relationship with some or all of the variables are latent variables. The relationship among the latent variables estimated by the structural model built by the measurement model containing relationships between explanatory variables with latent variables.

Complete structural equation modeling consists of the structural model and the measurement model. Measurement model is used to confirm the dimensions developed in the latent variables and structural model of the relationships that form the causality between the latent variables.

The development of structural equation modeling analysis produced various parameter estimation methods that can be used include maximum likelihood estimator, weighted least square, unweighted least square, general least square, partial least square. Each of these methods has its own criteria in the estimation process. Partial least squares (PLS) doesn't require assumptions (distribution-free) of the observation variables and the number of samples that do not have to be big. This is interesting to study empirically estimates for parameters use partial least squares estimators in structural equation modeling.

2. METODOLOGI

The data used in this study was auditee satisfaction level data on the performance of auditors and the factors that influence it. This data was sourced from Winniasri (2013). Data was obtained from the respondent's financial management officer at the 12 units Echelon 1 in the Ministry of Agriculture and Technical Implementation Unit. Selection of respondents by purposive sampling is a sampling technique that intentionally, researchers determine their own samples taken, not at random because there are certain considerations determined by researcher. Consideration in this case involved many personnel echelon 1 that managing finances in each office and the willingness of employees to fill out questionnaires. The number of respondents in the study collected a total of 110 respondents in accordance with the rule of thumb of structural equation analysis, namely the number of samples minimal five times the number of parameters / indicators (Ferdinand 2002).

Latent variables used the four major dimensions from the service quality concept of Zeithaml et al. (1990) that were namely tangible, reliability, responsiveness, and assurance. Variables associated with a tangible physical form facilities, personnel, processes and results of the internal audit. Reliability included professionalism of auditors. Responsiveness variable means speed of auditor response to assist and provide auditing and consulting services. Variables assurance regarding assurance can be given by the auditor to the auditee foster trust audit services.

Analysis Data Method

In general, the stages of data analysis in this study consisted of two stages: a descriptive analysis of the data and then fitted the model with partial least square (PLS) estimation method.

Descriptive analysis

Data analysis began with descriptive analysis aimed at to obtain a picture of the data and its properties then checked existing causes of certain symptoms.

Partial least squares estimation method

Stages of data processing samples at KTP method was as follows:

1. Model's specification

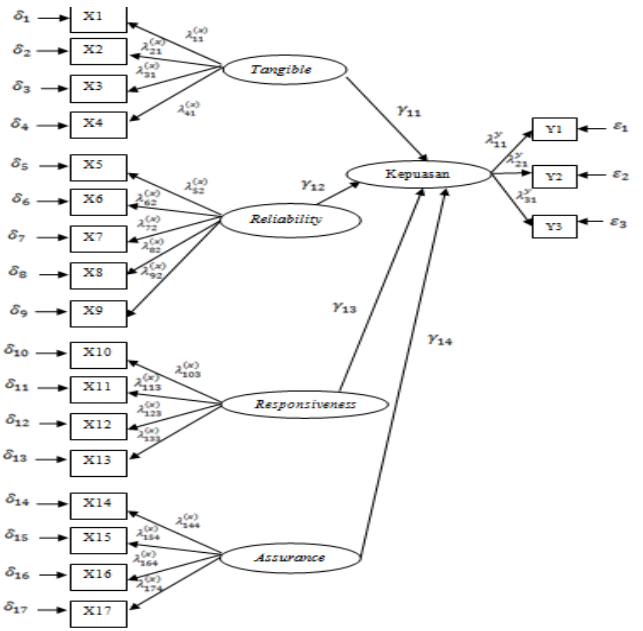


Figure 1: Forms relationship between auditee satisfaction levels with the factors that influence

2. Determined weighted, cross coefficients and the values of latent variables used PLS algorithms for the analysis of latent variable with the following steps:

Stage 1: iterative estimation of the initial weighted values and early latent variables started in step d and then steps a through d are repeated until convergent with the limit specified (Wold 1982).

a. Weighted the structural model

$$v_{ji} = \begin{cases} signcov(\eta_j, \eta_i) & if\ \eta_j\ and\ \eta_i\ adjacent, \\ 0, others \end{cases}$$

b. Estimation of structural models:

$$\hat{\eta}_j = \sum_i v_{ij}\eta_i$$

c. Weighted the measurement model

$$y_k = \hat{w}_k\eta_j + e_k \qquad (outward)$$

d. Estimation of the measurement model

$$\eta_j = f_i \sum_k \hat{w}_k y_k$$

Stage 2: Estimation of the across coefficients

3. Tested the validity of the convergence
4. Tested the validity of discrimination
5. Tested coefficients across models
6. Tested the magnitude of variability (R^2)
7. Interpreted the results

3. RESULT AND DISCUSSION

Description of Data

From the initial description that most of the information obtained echelon 1 under the Ministry of Agriculture had a good enough level of satisfaction or mediocre on the performance of the auditor who has worked. If the terms of the factors considered to influence the four variables we could saw that the components of both the auditor and the physical appearance of the facility personnel (tangible), professionalism of auditors (reliability), speed of auditors response to assist and provide auditing and consulting services (responsiveness), as well as the certainty that can be provided by the auditor (assurance) has been on a good level.

Differences in the level of satisfaction of some units of echelon 1 with most of the others were generally influenced by the increase / decrease in the level of tangible, reliability, and assurance variables. To further determine the structural variables which are not really affected the structural equation model would be used in the discussion of this chapter.

Model Parameter Estimation with Partial Least Squares Method

Coefficient estimator for the measurement model with the PLS method at the beginning of the model was presented through below. From these results we could set the measurement model and structural model.

For the endogenous variables

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} 0.71 \\ 0.93 \\ 0.93 \end{bmatrix} satisfaction + \begin{bmatrix} 0.21 \\ 0.05 \\ 0.41 \end{bmatrix}$$

For the exogenous variables

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} = \begin{bmatrix} 0.72 \\ 0.79 \\ 0.81 \\ 0.77 \end{bmatrix} tangible + \begin{bmatrix} 0.18 \\ 0.08 \\ 0.09 \\ 012 \end{bmatrix}$$

$$\begin{bmatrix} X_5 \\ X_6 \\ X_7 \\ X_8 \\ X_9 \end{bmatrix} = \begin{bmatrix} 0.74 \\ 0.84 \\ 0.86 \\ 0.89 \\ 0.88 \end{bmatrix} reliability + \begin{bmatrix} 0.17 \\ 0.29 \\ 0.24 \\ 0.25 \\ 0.26 \end{bmatrix}$$

$$\begin{bmatrix} X_{10} \\ X_{11} \\ X_{12} \\ X_{13} \end{bmatrix} = \begin{bmatrix} 0.67 \\ 0.87 \\ 0.81 \\ 0.75 \end{bmatrix} responsiveness + \begin{bmatrix} 0.09 \\ 0.06 \\ 0.07 \\ 0.08 \end{bmatrix}$$

$$\begin{bmatrix} X_{14} \\ X_{15} \\ X_{16} \\ X_{17} \end{bmatrix} = \begin{bmatrix} 0.86 \\ 0.77 \\ 0.86 \\ -0.06 \end{bmatrix} assurance + \begin{bmatrix} 0.30 \\ 0.23 \\ 0.26 \\ 0.27 \end{bmatrix}$$

Structural model was:

$$satisfaction = 0.50\ tangible - 0.10\ reliability - 0.10\ responsiveness + 0.22\ assurance + 0.18$$

Table 1
T-Value for exogenous latent variables of initial model

Variables		T-value
Tangible	-> Satisfaction	2.25
Responsiveness	-> Satisfaction	0.82
Reliability	-> Satisfaction	0.55
Assurance	-> Satisfaction	1.14

T-test values of each coefficient indicated only one exogenous variable whose value was greater than 1.96, which mean that only one latent exogenous variable (tangible) that significantly affected to the satisfaction.

Based on table 2, known that the cross coefficient measurement the resulting models with PLS method were one indicator variables not significant (loading factor value <0.4): X_{17}, remaining significant in reflected the latent variables.

Table 2
Loading Factor values for the structural models

latent variables	Indicators variables	loading factor initial model	Significance	loading factor final model	Significance
Tangible	x1	4.965	√	4.265	√
	x2	10.116	√	8.088	√
	x3	7.297	√	7.035	√
	x4	6.996	√	6.16	√
Reliability	x5	3.854	√	3.713	√
	x6	2.767	√	2.803	√
	x7	3.223	√	4.608	√
	x8	3.179	√	4.657	√
	x9	3.111	√	4.473	√
Responsiveness	x10	7.135	√	6.247	√
	x11	16.194	√	12.355	√
	x12	10.446	√	8.43	√
	x13	9.437	√	7.879	√
Assurance	x14	4.079	√	8.423	√
	x15	3.101	√	2.689	√
	x16	3.747	√	6.495	√
	x17	0.224	x		

Indicators were not significant would be dropped or removed and then would be the establishment of a new model that results are obtained as followed:
For the endogenous variables

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} 0.71 \\ 0.93 \\ 0.93 \end{bmatrix} satisfaction + \begin{bmatrix} 0.19 \\ 0.04 \\ 0.04 \end{bmatrix}$$

For the exogenous variables

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} = \begin{bmatrix} 0.72 \\ 0.79 \\ 0.81 \\ 0.77 \end{bmatrix} tangible + \begin{bmatrix} 0.14 \\ 0.08 \\ 0.09 \\ 0.10 \end{bmatrix}$$

$$\begin{bmatrix} X_5 \\ X_6 \\ X_7 \\ X_8 \\ X_9 \end{bmatrix} = \begin{bmatrix} 0.74 \\ 0.84 \\ 0.86 \\ 0.89 \\ 0.88 \end{bmatrix} reliability + \begin{bmatrix} 0.15 \\ 0.19 \\ 0.20 \\ 0.20 \\ 0.20 \end{bmatrix}$$

$$\begin{bmatrix} X_{10} \\ X_{11} \\ X_{12} \\ X_{13} \end{bmatrix} = \begin{bmatrix} 0.67 \\ 0.87 \\ 0.81 \\ 0.75 \end{bmatrix} resp\ \ onsiveness + \begin{bmatrix} 0.10 \\ 0.09 \\ 0.08 \\ 0.09 \end{bmatrix}$$

$$\begin{bmatrix} X_{14} \\ X_{15} \\ X_{16} \end{bmatrix} = \begin{bmatrix} 0.86 \\ 0.79 \\ 0.86 \end{bmatrix} assurance + \begin{bmatrix} 0.09 \\ 0.23 \\ 0.13 \end{bmatrix}$$

Structural model was:

$$satisfaction = 0.51\ tangible - 0.10\ reliability - 0.10\ responsiveness + 0.20\ assurance + 0.16$$

Based on the loading factor values shown in table 2, known that all the variables used were significant indicators reflected the latent variables.

Table 3
T-Value for exogenous latent variables of the model

Criteria		Value
Tangible	-> Satisfaction	2.46
Responsiveness	-> Satisfaction	0.89
Reliability	-> Satisfaction	0.53
Assurance	-> Satisfaction	1.22

The new model which formed still produced only one latent exogenous variable (tangible) that significant effected to the endogenous latent variables (satisfaction). That Variable was positive influence on the satisfaction variables mean the better physical form, facilities, processes and results of the internal audit, the auditee increasingly dissatisfied with the performance of the auditor who has been working. Variable reliability, responsiveness, assurance and t-test values that are positive but not significant.

Having created a new model in the variables measured levels of reliability reflected the latent variables. Testing was done by using the Value Composite Reliability (ρc) and Average Variance Extracted (AVE) which results are presented through the following table.

Table 4
Composite Reliability (ρ_c) **dan** ***Average Variance Extracted*** **(AVE)**

Latent Variable	ρ_c	*AVE*
Tangible	0.8544	0.5951
Reliability	0.9235	0.7081
Responsiveness	0.8592	0.6061
Assurance	0.8748	0.7001
Satisfaction	0.8964	0.7453

Pc value which was around 0.8 indicated that all latent variables used had a combined value of a good reliability. AVE values of all latent variables more than 0.5 indicated all latent variables could accommodate the diversity of indicator variables well.

R^2 = 0.2915 mean that 29.15% of satisfaction variables can be explained by the four latent exogenous variables, the rest was explained by other factors not described in the model.

These results suggested a model with PLS method accommodated the data used and only one latent exogenous variable (tangible) that was significant effected to the endogenous latent variables, namely the satisfaction of 4 latent variables proposed by existing theories.

6. COMMENTS AND CONCLUSION

Based on the analysis that had been presented previously could be concluded that:

1. initial measurement model was formed by PLS method showed that there was an indicator variable that was not significantly so that made the formation of a new model used indicator variables were significant.
2. Structural model formed by the PLS method gave exogenous latent variables that significant effected to the endogenous latent variables (satisfaction) was tangible variable. The variable had positive influence on the satisfaction which means better physical form, facilities, processes and results of the internal audit would make the auditee satisfied with the performance of the auditor.

REFERENCES

1. Bachrudin A, Tobing. (2003). *Analisis data untuk penelitian survei dengan menggunakan LISREL 8.30*. Bandung : Jurusan Statistika, FMIPA, UNPAD.
2. Bollen KA. (1989). *Structural equations with latent variables*. Canada: A Wiley-Interscience Publication.
3. Ferdinand. (2002). *Structural Equation Modeling (SEM) dalam penelitian Manajemen*. Badan Penerbit Universitas Diponegoro.
4. Kusnendi MS. (2008). *Model-model persamaan struktural satu dan multigrup sampel dengan lisrel*. Bandung: Alfabeta.
5. Schumacker RE, Lomax, RG. (1996). *A beginner's guide to SEM*. Jew Jersey : Lawrence Erlbaum Associates, inc. Pub.
6. Winniasri EF. (2013). *Tingkat satisfaction auditi terhadap kinerja Inspektoral Jenderal Kementerian Pertanian [tesis]*. Bogor (ID): Institut Pertanian Bogor.
7. Wold, H. 1982.*Partial least Square, Encyclopedia of statistical sciences Vol. VI*. New York: John Wiley and Sons.
8. Zeithaml VA, Parasuraman A, and Berry LL. (1990). *Delivering quality service: balancing costumer perceptions and expectations*. New York: The Free Press.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 63-72

A SYNTHETIC CONTROL CHART REEXPRESSION VECTOR VARIANCE FOR PROCESS MULTIVARIATE DISPERSION

Suwanda
Department of Statistics, Bandung Islamic University, Indonesia
Email: wanda_100358@yahoo.co.id

ABSTRACT

A control charts Reexpression Vector Variance (RVV) can be used to monitoring the dispersion of multivariate process as an alternative to the control charts Generalized Variance (GV) is commonly used. The synthetic control chart RVV is built in a combination the control charts RVV traditional (Shewhart class control chart) with the control chart of conforming run length (CRL). Average run length of the traditional control charts VV and the control charts VV synthetic calculated in case p=3 and n=5. The result, the synthetic control chart VV can give signal out of control more quickly when there is a small shift of the value of multivariate dispersion process.

KEYWORDS

Multivariate dispersion, vector variance, conforming run length, average run length.

1. INTRODUCTION

In the field of manufacturing industry, monitoring in a process of becoming an inevitability. This activity is done to continuous quality improvement.There are two phase in monitoring the process.Phase I consisting of the use of a control chart for (i) stage 1 ‘start-up stage’ in retrospective testing what 's the process in control when subgrup-sugrup first drawn; and (ii) stage 2, ‘future control stage’ which is testing whether the process remain in control when subgrup-subgrup future taken. In the multivariate characteristic, standard value with regard to the mean vector μ_0 and a covariance matrix Σ_0 (Alt and Bedewi, 1986).

The focus of this paper, development the control chart for monitoring multivariate dispersion. One such the measure of multivariate dispersion of that can be used is a reexpresion of vector variance (RVV). A RVV control chart built can be clasified in a class of a Shewhart control chart. In the previous researchers, a Shewhart control chart less sensitive in detecting small shifts of the process To improve its drawback Wu and Spedding (2000) have made a univariat control chart of the synthesis of cases is to control an average of the process. A control chart synthesis built by combining a Shewhart control chart and control chart conforming run length (CRL). The results show that the performance of a control chart synthetic average become more sensitive in detecting small shift. The idea of control chart synthesis, it has been adopted by some researchers as Huang and Chen (2005) to monitoring deviation standard, Ghute and Shirke (2008a) to controlling mean vector process and Ghute and Shirke (2008b) to

monitoring the multivariate dispersion of the process with gernelaized variance (GV) statistics. In this paper the synthetic control chart to be built in to controling the multivariate dispersion process with statistics rvv.

To this intention, in this paper, first be be drawn about a control chart RVV. The next be drawn also a control chart CRL in general. As basic subjects of in this paper is developing an algorithm to synthetic control chart RVV for monitoring multivariate dispertion. For example cases will be gave at the end.

2. THE CONTROL CHART RVV TRADISIONAL

Suppose a multivariate process (p variables) with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ positive definite, in a notation of random vector X, it is assumed $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The multivariate dispersion process will be controlled by RVVB defined by:

$$RVV = \left[Tr(S^2) \right]^{\frac{1}{2p}} \tag{1}$$

where, S is covariance with sample size n.
In control processes, $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$.

Berdasarkan distribusi limit (1), batas-batas kontrol bagi RVV fase II (Suwanda, 2014) adalah:

Based on the limiting distribution (1), control limits RVV for phase II (Suwanda, 2014) is:

$$\begin{aligned} UCL &= \upsilon_{\Sigma_0} + k\tau_{\Sigma_0} \\ Central &= \upsilon_{\Sigma_0} \\ LCL &= \upsilon_{\Sigma_0} - k\tau_{\Sigma_0} \end{aligned} \tag{2}$$

where,

$$\upsilon_{\Sigma_0} = \left[Tr(\Sigma_0^2) \right]^{1/(2p)} \tag{3}$$

$$\tau^2_{\Sigma_0} = \left[\frac{1}{4p^2} \frac{\sigma^2_{\Sigma_0}}{\left[Tr(\Sigma_0^2) \right]^{(2p-1)/p}} \right] \tag{4}$$

$$\sigma^2_{\Sigma_0} = \frac{8n}{(n-1)^2} Tr(\boldsymbol{\Sigma}_0^4) \tag{5}$$

and k = the constant to determine the probability that process out of control with the actual state process in control (α)

Dapat ditunjukkan (lihat apendiks) bahwa bahwa *average run length in control* untuk (2) adalah

Can be shown (see the appendix) that the average run length in control (2) is

$$ARL_0 = 1/\alpha = 1/2\Phi(-k) \tag{6}$$

while that the average run length out of control:

$$ARL_1 = 1/P \tag{7}$$

where

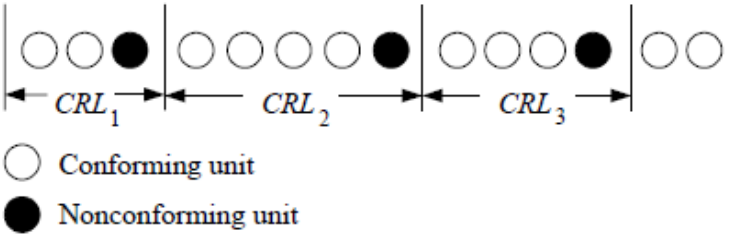
$$P = 1-\beta = 1-\left[\Phi(b)-\Phi(a)\right]$$

$$a = \frac{\nu_{\Sigma_0} - k\tau_{\Sigma_0} - \nu_{\Sigma_1}}{\tau_{\Sigma 1}}$$

$$b = \frac{\nu_{\Sigma_0} + k\tau_{\Sigma_0} - \nu_{\Sigma_1}}{\tau_{\Sigma 1}}$$

3. THE CRL CONTROL CHART

The CRL control chart proposed by Bourke (1991) developed originally to monitoring the quality of attributes to detect shift in fractions defective, proportions when the inspection 100 % is used. The runs length of conforming items between to items nonconforming successive taken as a basis for a control chart.The CRL defined as corresponding quantity of units is between the two nonconforming successive including the last unit is not conforming. The CRL can be explained in the next picture:



Pigure 1. Calculate of CRL

Mean and the cumulative probability function of CRL are:

$$\mu_{CRL} = \frac{1}{\theta}$$

$$F_\theta(CRL) = 1-(1-\theta)^{CRL},\ CRL = 1, 2, \ldots$$

In the case of detecting increasing of θ, the only of lower control limit used as follows:

$$LCL_{CRL} = L = \frac{\ln(1-\alpha_{CRL})}{\ln(1-\theta_0)} \tag{8}$$

where

$\alpha_{CRL} = 1-(1-\theta_0)^L$ = error type I

θ_0 = proportion of nonconforming *in control.*

L will be must integer.

If the sample CRL less than L,, should be suspected that the proportion of damage has been increased.

The Control chart synthetic for average can be develoved by combining the control chart average with control chart CRL.

The Average Run Length (ARL) for the control chart CRL is:

$$ARL_{CRL} = \frac{1}{F_\theta(L)} = \frac{1}{1-(1-\theta)^L}. \quad (9)$$

4. THE SYNTHETIC CONTROL CHART RVV

This chart was created with the combining the control chart RVV and CRL. The mechanism the state of process as follows. If the RVV fall in the control limits, note as the state conforming with the symbol circle empty and if RVV falling in outside the control limits, note as the state of the nonconforming with a symbol of black circle (see Figure 1). In the synthetic control chart, L and k determined that filled ARL_0 (10) as expected (e.g. 370 to control chart 3 sigma) and the ARL_1 (11) minimum. In this ARL_0 for the synthetic control chart is the multiplication of (6) and (9), namely:

$$ARL_{0S} = ARL_{0RVV} \times ARL_{0CRL} = \frac{1}{2\Phi(-k)} \frac{1}{1-(1-2\Phi(-k))^L} \quad (10)$$

dan ARL_1 untuk digaram kontrol sitesis adalah perkalian (7) dan (9) *out of control*, yaitu: and ARL_1 for Synthetic control chart is the multiplication (7) and (9), namely:

$$ARL_{1S} = ARL_{1RVV} \times ARL_{1CRL} = \frac{1}{P} \frac{1}{1-(1-P)^L} \quad (11)$$

The operation of synthetic control chart RVV are summarized as follows:

1. Set UCL and LCL from sub chart RVV/Sin and LCL L from sub chart CRL/Sin (determination of the control limits will be explained later).
2. On each inspection point, get a sample with size n observations, x_i and calculate the sample matrix covariance S_i and RVV. This RVV sample plotted on a sub chart RVV/Sin.
3. If the RVV/Sin greater than $LCL_{RVV/Sin}$ and less than $UCL_{RVV/Sin}$, this sample is called a appropriate sample (conforming) and monitoring returns to step 2.

Otherwise sample referred to is not appropriate (nonconforming) and controlling continues to the next step.

4. Check the number of sample *RVV* between sample now and the sample non-conforming. The number is extracted as a sample CRL from of sub-chart *CRL* / *Sin* and the synthetic chart.
5. If the sample CRL is larger than the lower limit sub-diagram *CRL* / *Sin*, the process in control and controlling next to step 2. Otherwise process it is out of control and controlling for the next step.
6. *Signal out of control* condition.
7. Make the act to obtain and remove the factor causes of out of control. And than summarize the process and then back to step 2.

To design a synthetic control chart, users should provide specifications follows:

The covariance matrix *in control* process, Σ_0.
Standard deviation of RVV, τ.
Sample sizel, n.
Covariance matrix out of control, Σ_1.
Average run length in control, ARL_0.

Usually, μ and Σ estimated from the data observations on the pilot runs and the size of the n = 4.5, or 6. Mean shift design is the distance through which it considered to be serious enough its impact on the quality of the product; then pertaining to a value ARL_1 that is as small as possible. The ARL_0 decided by the fulfillment of the level of any alarm. If handling error alarm is difficult, ARL_0 must large, other ARL_0 have to be set with a value of less to enhance the effectiveness of the detection.

A description of the program

This program can use to design parameters controlling $BKA_{RVV/Sin}$ $BKB_{RVV/Sin}$ and L that ensure the minimum for ARL to shift VV or change of covariance matrix for the synthetic described above. Data entry for flexibility, program determined by type error I $\alpha\ (1/ ARL_0)$.

The program design a synthetic chart based on the model optimization as follows:

The objective:

$$ARL_1 = \text{minimum} \tag{12}$$

$$\text{Constraint } ARL_0 = \frac{1}{2\Phi(-k)} \times \frac{1}{1-\left[1-2\Phi(-k)\right]^L} \tag{13}$$

The two desigm variable:

L, k

where $ARL_1(\Sigma_1)$ = ARL *out of control.*

The optimization identify the set of L and k so $ARL_1(\Sigma_1)$ minimum with ARL_0 be fixed.

The procedures complete of design programs can be summarized as follows:

1. Set μ, Σ_0, n, Σ_1 and ARL_0 or α.
2. An initial value L as 1.
3. Determine k with solving (2) numerically, as follows:

Let $ARL(0)=a$ and $\Phi(-k)=b$, Equation (2) into:

$$a=\frac{1}{2b}\times\frac{1}{1-[1-2b]^L}$$

$$2b-[1-2b]^L\ 2b-\frac{1}{a}=0$$

$$f(b)=2b-[1-2b]^L\ 2b-\frac{1}{a}=0$$

With Newton's method the roots of b can be determined with the following procedure: Taylor series $f(b)$ at this point b_0 are:

$$f(b)=f(b_o)+\frac{f^{'}(b_o)}{1!}(b-b_0)+\frac{f^{''}(b_o)}{2!}(b-b_0)^2+\frac{f^{'''}(b_o)}{3!}(b-b_0)^3+....=0$$

The linear approximation is:

$$f(b_o)+\frac{f^{'}(b_o)}{1!}(b-b_0)\approx 0$$

$$(b-b_0)=-\frac{f(b_o)}{f^{'}(b_o)}$$

$$b_1=b_0-\frac{f(b_o)}{f^{'}(b_o)}$$

In this case $0<\Phi(-k)=b<0.5$, because the initial value b_0 can be between 0 and 0.5.

4. Calculate ARL_1 for the values k and L using the following equation,

$$ARL_1=\frac{1}{P}\times\frac{1}{1-(1-P)^L}$$

where P as at Equation 7.

5. If ARL_1 being reduced, then stepped up l by the addition of one and go back to step-3. Other to the next step.
6. Let L and k as final value for the shyntetic chart.
7. Use the final value of k, calculate $UCL_{RVV/Sin}$ and $LCL_{RVV/Sin}$ as follow:

$$LCL_{RVV/Sin} = \nu_0 - k\tau_0, UCL_{RVV/Sin} = \nu_0 + k\tau_0$$

Illustration:

To better understand the process of a synthetic chart RVV, this following will be given an example case. Suppose a process involving three important variables with a covariance matrix,

$$\Sigma_0 = \begin{bmatrix} 1 & 0 & 0 \\ & 1 & 0 \\ & & 1 \end{bmatrix}$$

The monitoring done during the process of continuing with the use of the sample size n = 5. A syhthetic control chart RVV will be made to control to multivariate dispersion so sensitive to change all variable variance, increased to $\sigma_{ii} = 1.5, i = 1, 2, 3$ and the correlation between variables fixed equal to zero with ARL_0 = 370.

Dengan menggunakan Matlab R2008b diperoleh nilai-nilai L, k, BKB, BKA, dan ARL_1 untuk diagram RVV sintesis, seperti yang tercantum pada table berikut:

By using matlab R2008b obtained the values of L, k, LCL, UCL, and ARL_1 for the synthetic chart RVV, as articulated in the following table:

Table 1
The values k, LCL, UCL dan ARL_1 for
L 1 s/d 20 and ARL_0=370

L	k	LCL	UCL	ARL_1
1	1.943	0.846	1.556	25.305
2	2.085	0.820	1.582	20.005
3	2.164	0.806	1.596	17.716
4	2.219	0.796	1.606	16.408
5	2.260	0.788	1.614	15.561
6	2.294	0.782	1.620	14.972
7	2.322	0.777	1.625	14.546
8	2.346	0.772	1.630	14.228
9	2.366	0.769	1.633	13.987
10	2.385	0.765	1.637	13.802
11	2.402	0.762	1.640	13.660
12	2.417	0.759	1.643	13.552
13	2.430	0.757	1.645	13.470
14	2.443	0.755	1.647	13.409
15	2.455	0.752	1.649	13.365
16	2.466	0.750	1.651	13.336
17	2.476	0.749	1.653	13.319
18	**2.486**	**0.747**	**1.655**	**13.312**
19	2.495	0.745	1.657	13.314
20	2.503	0.744	1.658	13.323

Pada Tabel 1, untuk L=18 diperoleh ARL1=13.312, sedangkan pada L=19 memberikan ARL1= 13.314. Oleh karena itu ARL1 optimal dari diagram kontrol RVV sintesis terjadi pada L=18. Jadi batas-batas kontrol RVV sintesis adalah:

Sub diagram kontrol RVV/Sin:

In the Table 1, for L = 18 obtained ARL_1 = 13.312, while in L = 19 give ARL_1 = 13.314. Hence the optimal ARL_1 for synthetic chart RVV happened in L = 18. So the limits of control RVV synthtetic:

Sub control chart RVV:

$$UCL_{RVV/Sin} = 1.655$$

$$LCL_{RVV/Sin} = 0.747$$

Sub control chart CRL/Sin:

$$LCL_{CRL/Sin} = L = 18$$

with ARL_1=13.314.

Whereas the limits of the control chart RVV standard are:

$$UCL_{RVV} = 1.7491$$

$$Central_{RVV} = 1.2009$$

$$LCL_{RVV} = 0.5628$$

with ARL_1=27.0129.

It appears that in cases of the number of variables p = 3 and sample size n=5, a synthetic control chart RVV on average in the period of 13 or 14 will give the signal has happened changes RVV which was originally RVV_0 = 1.2009 be RVV_1 = 1.3747 or the ratio change PRVV = RVV_1 / RVV_0 = 1.14.The provision of this signal more quickly compared with an average length of the period of the provision of signals out of control the first time by a control chart RVV standard that is on the period 27 since the occurrence of a shift RVV to 27. Ghute and Shirke (2008) has calculated ARL1 for control chart GV standard ang synthetic control chat GV. To the case of p= 3, n = 5 and the ratio of change GV 1.2, standard give ARL1 = 75.56 fot control chart GV standar and give ARL1 = 56.84 for synthetic control chart GV.So a synthetic control chart RVV have the best performance to this case.

5. CONCLUSION

In this paper have been introduced a synthetic control chart to monitoring covariance matrix for the process distributed multivariate normal. A synthetic control chart RVV is a combination of control chart RVV and control chart CRL. In the case of number of variables p = 3 and sample size n = 5, a synthetic control chart RVV faster give the signal out of control that is in the period 14th (on average) of a diagram control standard rvv who gave the signal on period 27^{th}. Whereas a synthetic control GV in the same case give the first sinyal in period 57^{th}. control rvv peride for change. So a synthetic control chart RVV have the best performance to this case.

ACKNOWLEDGMENT

The support of research grants " Hibah Bersaing" from Director of General Higher Education Indonesia the year 2014 (2th year). Hence on this occasion we would like a lot of thank to Director of General Higher Education Indonesia Depdikbud, Ketua LPPM and Rector of the convention of this research activity. Hopefully this effort continues to be improved both in terms of quality and quantity

REFERENCES

1. Alt, F.B. Bedewi, G.E. (1986), SPC of Dispersion for Multivariate Data. In *ASQC – Anaheim*, h.248-254. Anaheim: American Society for Quality Control.
2. Ghute, V.B dan Shirke, D.T. (2008a). A Multivariate Synthetic Control Chart for Monitoring Process Vector Mean. *Communications in Statistics-Theory and Methods*, 37; 2236-2148.
3. Ghute, V.B dan Shirke, D.T. (2008b). A Multivariate Synthetic Control Chart for Process Dispersion. *Quality Technology of Quantitative Management*, Vol. 5, No. 3; 271-288.*Computer & Industrial Engineering.* 49: 221-240.
4. Huang, H.J dan Chen, F.L. (2005). A Synthetic Control Chart for Monitoring Process Dispersion with Sample Standard Deviation.
5. Suwanda (2014), Diagram Kontrol Reekspresi Variansi Vektor dan Implementasinya, *Prosiding Seminar Nasional Matematika dan Statistika*, Universitas Tanjungpura.
6. Wu, Z. dan Spedding, T.A. (2000). A Synthetics Control Chart for Detecting Small Shift in the Process Mean. *Journal of Quality Technology* ; **32**:1; 32–38.

APENDIKS:

ARL Diagram Kontrol RVV:

Average Run Length didefinisikan sebagai ekspektasi dari variable acak geometrik dengan parameter merupakan probabilitas menyatakan out of control (proporsi kerusakan), missal P.
Jadi,

$$ARL = \frac{1}{P}$$

Dibawah proses sebenarnya incontrol, $P = \alpha$
Jadi,

$$ARL_0 = \frac{1}{\alpha}$$

Dibawah proses sebenarnya out of control, $P = 1 - \beta$
Jadi,

$$ARL_1 = \frac{1}{1-\beta}$$

Sekarang akan ditentukan nilai β ,

$$\begin{aligned}\beta &= P\left(BKB < RVV < BKA \middle| \Sigma = \Sigma_1, \Sigma_1 = \Sigma_0\right)\\ &= P\left(\frac{BKB - \nu_{\Sigma_1}}{\tau_{\Sigma_1}} < Z < \frac{BKA - \nu_{\Sigma_1}}{\tau_{\Sigma_1}}\right)\\ &= P\left(\frac{\nu_{\Sigma_0} - k\tau_{\Sigma_0} - \nu_{\Sigma_1}}{\tau_{\Sigma_1}} < Z < \frac{\nu_{\Sigma_0} + k\tau_{\Sigma_0} - \nu_{\Sigma_1}}{\tau_{\Sigma_1}}\right)\\ &= P\left(a < Z < b\right) = \Phi\left(b\right) - \Phi\left(a\right)\end{aligned}$$

dimana

$$a = \frac{\nu_{\Sigma_0} - \nu_{\Sigma_1} - k\tau_{\Sigma_0}}{\tau_{\Sigma_1}}$$

$$b = \frac{\nu_{\Sigma_0} - \nu_{\Sigma_1} + k\tau_{\Sigma_0}}{\tau_{\Sigma_1}}$$

Dibawah proses in control benar:

$$a = \frac{\nu_{\Sigma_0} - \nu_{\Sigma_0} - k\tau_{\Sigma_0}}{\tau_{\Sigma_0}} = -k$$

$$b = \frac{\nu_{\Sigma_0} - \nu_{\Sigma_0} + k\tau_{\Sigma_0}}{\tau_{\Sigma_0}} = k$$

Sehingga,

$$\beta = \Phi\left(k\right) - \Phi\left(-k\right)$$

Jadi,

$$\begin{aligned}ARL_0 &= \frac{1}{1 - \left[\Phi\left(k\right) - \Phi\left(-k\right)\right]}\\ &= \frac{1}{2\Phi\left(-k\right)}\\ &= \frac{1}{\alpha}\end{aligned}$$

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 73-86

BAYESIAN WEIBULL MIXTURE MODELS FOR DENGUE FEVER

Sri Astuti Thamrin, Andi Kresna Jaya and **La Podje Talangko**
Mathematics Department, Faculty of Mathematics and Natural Sciences,
Hasanuddin University, Jl Perintis Kemerdekaan Km 10, Tamalanrea,
Makassar, Indonesia, 90245. Email: tuti@unhas.ac.id

ABSTRACT

Dengue fever disease has become common problem in developing countries including Indonesia. Mixture models are usually used in modelling data consisting of several groups, where each group has different properties and characteristics of the one family but uses the same distribution. Weibull mixture models have received increasing attention in recent statistical research with applications in the field of survival analysis. The advances in the Bayesian paradigm have substantially expanded the methodology and application of Weibull mixture models. One problem of current interests is the analysis of survival times of patients. Dengue hemorrhagic fever data can be used to make inference about patient survival. In this study, we focus on the use of Bayesian Weibull mixture models for estimating survival. A simulation study that investigates the impact of censoring on these models is also described.

KEYWORDS

Bayesian, Censoring, Dengue hemorrhagic fever, Mixtures, Survival analysis, Weibull distribution.

1. INTRODUCTION

Indonesia is a tropical country with two seasons: rainy season and dry season. Tropical natural environment, poor sanitation, and lack of public awareness were the main reasons of the spread of dengue in Indonesia. Demographic and societal changes such as population growth, urbanization, and modern transportation appear to play an important role in the increased incidence and geographical spread of dengue virus (Gubler, 2002). Dengue Hemorrhagic Fever (DHF) in Indonesia is a disease caused by the bite of Aedes aegypti. If a person gets infected, this virus may result in her/his death (National Geographic Indonesia, 2012).

DHF usually affects children, but in recent decades it has been a trend of increases in the proportion of more mature age. Indonesia occupies the highest position in the case of DHF in Southeast Asia with 10,000 cases in 2011 (National Geographic Indonesia, 2012). Although, Makassar is categorized one of 10 regions ranked highest DHF (detikHealth, 2012), efforts to prevent the spread of this disease and to reduce mortality needs to be done. Government of Makassar sets 5 Sub districts of 14 districts in Makassar, as its prone areas to the spread of dengue disease (Viva news, 2013).

Preventive Efforts to reduce mortality should be done, such as increasing awareness and a healthy way of life in order to prevent the occurrence and spread of dengue fever (Kautler et al., 1997). On the other hand, the reduction of the mortality rate can be r curative by providing appropriate medical therapy. On the latter, the issue of patients' survival with DHF becomes very important to be studied.

Survival time of patients can be affected by many factors. In addition, the survival of patients with DHF is largely determined by the ability to develop and implement appropriate methods to a) identify factors that influence survival and b) get a model based on those factors. One of the popular methods is proportional hazard regression Nguyen and Rocke, 2002, Rosenwald et al., 2002, Kleinbaum dan Klein, 2005). This method only uses present data as a basis to estimate patient survival times and does not takes into account available prior information, other unknown features of the model or the model structure itself. In a Bayesian framework, such information can be informed through prior distributions and uncertainty in the model structure can be accommodated (Kaderali et al. 2006; Tachmazidou et al. 2008). Since the advances in computational and modeling technique, Bayesian methods are now becoming quite common for survival data (Kaderali et al., 2006; Lee dan Mallick, 2004; Thamrin et al., 2012, Thamrin et al., 2013, Thamrin, 2013).

Given the nature of DHF data to describe biological systems and outcomes of patients, and hence the potential of these covariates to produce more precise inferences about survival, the use of a single parametric distribution to describe survival time may not be adequate. DHF data may enable the description of several homogeneous subgroups of patients with respect to survival time. This research therefore used Bayesian Weibull mixture models for better estimation and prediction of this outcome, following the notion that the Weibull distribution is a popular parametric distribution for describing survival times (Dodson, 1994) and mixture models are commonly used in describing data consisting of several groups, where each group has different properties and features of the one family but uses the same distribution. These models provide a convenient and flexible mechanism to identify and estimate distributions, which are not well modelled by any standard parametric approaches (Stephens, 1997). There is developing literature on the application of Weibull mixture models in the field of survival and reliability analysis. The advances in EM algorithm Dempster et al. (1977), the Bayesian paradigm (Berger, 1985, Besag et al., 1995), and Markov chain Monte Carlo (MCMC) computational methods (Diebolt and Robert, 1994) have substantially expanded the methodology and application of Weibull mixture models. In the Bayesian context, Marin et al. (2005a) described methods to fit aWeibull mixture model with an unknown number of components. Farcomeni and Nardi (2010) proposed a two component mixture to describe survival times after an invasive treatment. Quiang (1994) also used a mixture of a Weibull component and a surviving fraction in the context of a lung cancer clinical trial. Tsionas (2002) considered a finite mixture of Weibull distributions with a larger number of components for capturing the form of a particular survival function. Based on previous studies, it is known that there are not many studies that analysed the survival time estimates of HDF, the factors that influence in Indonesia by using a Bayesian Weibull mixture and consider model choice issues.

The main aim of this paper is to develop and apply the Bayesian Weibull mixture approach to model DHF patients' survival. This aim is addressed through the estimation of a two component Weibull mixture model in a simulation study and an application to a dengue fever dataset. The number of components does not need to be confined into two components. Farcomeni and Nardi (2010) stated that while extending the model to the general case is straightforward, in their experience the two Weibull mixture is already sufficiently flexible. The paper is organised as follows. In Section 2, we define the Weibull mixture model and Bayesian computational approach for parameter estimation. We also provide the method of simulation and discuss the issue of label-switching and model evaluation in Section 2. In Section 3, we illustrate the model using simulated datasets and a DHF dataset. The results are discussed further in Section 4.

2. MATERIAL AND METHODS

2.1. Model Formulation

In this section, we define the Weibull mixture model for analysing survival data. We confine ourselves to survival times that are the difference between a nominated start time and a declared failure (uncensored data) or a nominated end time (censored time). Let T be a nonnegative random variable for a person's survival time and t be any specific value of interest as a realisation of the random variable T. Kleinbaum and Klein (2005) give some reasons for the occurrence of right censoring in survival studies, including termination of the study, drop outs, or loss to follow-up. For the censored observations, one could impute the missing survival times or assume that they are event-free. The former is often difficult, especially if the censoring proportion is large, and extreme imputation assumptions (such as all censored cases fail right after the time of censoring) may distort inferences (Leung and Elashoff, 1997, Stajduhar et al., 2009). In this study, we treat all censored cases as event-free regardless of observation time.

Initially, we assume that we observe survival time t on patients possibly from a heterogeneous population. The two parameter Weibull density function for survival time is given by

$$W(t|\alpha,\gamma) = \alpha\gamma t^{\alpha-1}\exp(-\gamma t^{\alpha})$$

for $\alpha > 0$ and $\gamma > 0$, where α is a shape parameter and is a scale parameter (Ibrahim et al., 2001b). A mixture of K Weibull densities (Marin et al., 2005a) is defined by

$$f(t|K,\pi,\alpha,\gamma) = \sum_{m=1}^{K}\pi_m W(t|\alpha_m,\gamma_m)$$

where $\alpha = \alpha_1,\alpha_2,\dots,\alpha_K$, $\gamma = \gamma_1,\gamma_2,\dots,\gamma_K$ are the parameters of each Weibull distribution and $w = w_1,w_2,\dots,w_K$ is a vector of nonnegative weights that sum to one.

The corresponding survival function S(t|K,π,α,γ) and hazard function S(t|K,π,α,γ) are as follows:

$$S(t|K,\pi,\alpha,\gamma) = \sum_{m=1}^{K}\pi_m \exp(-\gamma_m t_m^{\alpha})$$

$$h(t|K,\pi,\alpha\gamma) = \frac{f(t|K,\pi,\alpha,\gamma)}{S(t|K,\pi,\alpha,\gamma)}$$

Let xij be the jth covariate associated with patient i, for j = 1, 2, …, p. In our application, xij could indicate, for example, the thrombosis'. The covariates can be included in the model as follows (Farcomeni and Nardi, 2010)

$$\log(\gamma_m) = x_i'\beta_m = \lambda_m \quad (1)$$

where $x_i = x_{i1}, x_{i2}, \dots, x_{ip}$, $\gamma_m = \gamma_{1m}, \gamma_{2m}, \dots, \gamma_{pm}$ and $\beta_m = \beta_{1m}, \beta_{2m}, \dots, \beta_{pm}$ for i = 1, 2,…, n and m = 1, 2, …, K.

We now assume that we observe possibly right-censored data for n patients $y = y_1, y_2, \dots, y_n$ where y_i=(t_i,δ_i) and δ_i is an indicator function such that (Marin et al., 2005a) δ_i = 1, if the lifetime is uncensored, i.e., Ti = ti and δ_i = 0, if the lifetime is censored, i.e., Ti > ti.

Thus, the likelihood function becomes:

$$L(\pi, \alpha, \gamma | K, t_i, \delta_i, x) \propto \prod_{i=1}^{n} f(t_i | K, t_i, \delta_i, x)^{\delta_i} S(t_i | K, t_i, \delta_i, x)^{1-\delta_i}$$

Here, the incomplete information is modelled via the survivor function, which reects the probability that the ith patient was alive for duration greater than ti.

In WinBUGS (Lunn et al., 2000, Ntzoufras, 2009, Spiegelhalter et al., 2002), possibly right censored data can be modelled using a missing data approach via the command I(.,) as follows

t[i] ~ dweib(alpha[i]; gamma[i])I(cens.time[i];)

where cens:time[i] is either zero for uncensored outcome or the ith recorded survival time for censored outcomes. Hence, censored survival times are assumed to be drawn from a truncated Weibull distribution.

The following prior distributions were placed on the parameters π and α:

$$\pi | K \sim \text{Dirichlet}\ (\phi_1, \phi_2, \dots, \phi_K)\ ,\ \phi_m = \phi, \forall m = 1, 2, \dots, K$$

$$\alpha_m \sim \text{Gamma}(u_\alpha, v_\alpha), m = 1, 2, \dots, K$$

For a model without covariates, we employ the following prior for γ_m.

$$\gamma_m \sim \text{Gamma}(u_\gamma, v_\gamma), m = 1,2, \dots, K$$

We choose small positive values for $u_\alpha, v_\alpha, u_\gamma, v_\gamma$ to express vague prior knowledge about these parameters and we set ϕ = 1 (Marin et al., 2005a). For a model with covariates, in this paper, we employed an independent normal prior on each β_m, so that

$$\beta_m | K \sim N(0, \Sigma)$$

and we allow β_m to be diagonal with elements σ_j^2, j= 1, 2, …, p. We express a vaguely informative prior by setting σ_j^2=10. The diagonal matrices were used here but this changed recently (Bhadra and Mallick., 2013), so one may argue that a non-diagonal variance-covariance matrix may be more appropriate.

2.2. Computation Method

The model described in Section 2.1 can be fitted using MCMC sampling with latent values Zi to indicate component membership of the ith observation (Diebolt and Robert, 1994, Robert and Casella, 2000). Since $\pi m = Pr(Z_i = m)$, we can write $Z_i \sim M(\pi_1, \pi_2, \ldots, \pi_k)$. In this scheme, the Z_i is sampled by computing posterior probabilities of membership, and the other parameters are sampled from their full posterior distributions, conditional on the latent indicators. This was implemented in the WinBUGS software package (Lunn et al., 2000).

Label switching, caused by non-identifiability of the mixture components, was dealt with post-MCMC using the reordering algorithm of Marin et al. (2005b). The algorithm proceeded by selecting the permutation of components at each iteration that minimised the vector dot product with the so-called "pivot", a high density point from the posterior distribution. The MCMC output was then reordered according to each selected permutation. In this paper, the approximate maximum a posteriori (MAP) (i.e. The realization of parameters corresponding to the MCMC iterate that maximised the unnormalised posterior) was chosen as the pivot.

2.3. Model Evaluation and Comparison

The appropriateness of the Weibull model can be checked by applying goodness of fit measures which summarize the discrepancy between observed values and the values expected under the model in question (Gupta et al., 2008). The most commonly used assessments of model fit are in the form of information criteria, such as the Bayesian Information Criterion (BIC) (Schwarz, 1978a),

$$\mathrm{BIC} = -2\,\mathrm{logL}(t|\theta) + k\,\log(n),$$

and Akaike Information Criterion (AIC) (Akaike, 1973),

$$\mathrm{AIC} = -2\,\mathrm{E}\big(\log\big(\mathrm{L}(t|\theta)\big)\big) + 2k$$

For AIC, θ are unknown parameters of the model and k is the number of free parameters in the model. The term 2k in the AIC is also a complexity measure. Both the BIC and DIC can be calculated from the simulated values based on MCMC results; smaller values indicate a more suitable model in terms of goodness of fit and short-term predictions (McGrory and Titterington, 2007), (Spiegelhalter et al., 2002).

The selection of variables in regression problems has occupied the minds of many statisticians. Several Bayesian variable selection methods have been applied to gene expression and survival studies. For example, Volinsky and Raftery (2000) investigated the Bayesian Information Criterion (BIC) for variable selection in models for censored survival data and Ibrahim et al. (2008) developed Bayesian methodology and computational algorithms for variable subset selection in Cox proportional hazards model with missing covariate data. Other papers that deal with related aspects are Cai and Meyer (2011) and Gu et al. (2011). Cai and Meyer (2011) used conditional predictive ordinates and the DIC to compare the fit of hierarchical proportional hazards regression models based on mixtures of B-spline distributions of various degrees. Gupta et al. (2011) presented a novel Bayesian method for model comparison and assessment using survival data with a cured fraction.

In our analysis, for demonstration we compute the DIC and BIC for all possible subsets of variables and select these models with smallest DIC and BIC values (Burnham and Anderson, 2002). We also evaluate the model by applying posterior predictive checks based on the validation dataset.

2.4. Simulation Algorithm

Our interest in this study was to estimate the parameter of Bayesian Weibull mixture. The models developed here was the Weibull mixture model. We used the simulation algorithm for analyses. The probability models with five explanatory variable were used in simulations

For these study, data were simulated from two component Weibull mixture models with the following parameter configurations:

$$W(t|k = 2, \pi, \alpha, \gamma) = 0{,}5W(3{,}1) + 0{,}5W(2{,}2),$$

The censoring levels20% was applied to model and a sample size of n = 200 was used for all experiments. The following steps were applied to carry out the simulations.

1. Generate t_i, from the respective model, for i=1,2,…,n.
2. Generate censoring times by assuming that the largest C% survival times are right censored.
3. Generate each covariate $x_i = (x_{1i}, x_{2i}, \dots, x_{5i})$ from independent standard normal distributions, and then set γ_m using equation 1. For the purpose of the simulation study, we fixed the coefficient values relating to the covariates in each component to $\beta_1 = (1,1,1,1,1)$ and $\beta_2 = (2,2,2,2,2)$.
4. Fit the model based on the data yi = (ti; _i), with 100,000 iterations, discarding the first 10,000 iterations as burn-in.
5. Record posterior estimates of the model parameters, namely median and standard deviation.

3. RESULTS

3.1. Simulation

We simulated the generated data by running for Weibull mixture model with noninformative prior. The averaged values over the 100 simulations for median of posterior means and standard deviations of theWeibull mixture parameters are reported in Table 1 for 200 sample size. The table confirms the accuracy of the parameter estimates in the 20% of censoring.

Table 1
Posterior estimates of parameters (α, π, β_m) for simulation data model with 20% levels of censoring.

C	Parameter	True value	Posterior Median	Posterior Standard Deviation
20%	α_1, α_2	3, 2	(2.967,2.122) (0.352, 0.211)	(0.308, 0.225)(0.057, 0.024)
	β_{11}, β_{12}	1, 2	(1.038, 1.986)(0.055, 0.069)	(0.051, 0.074)(0.007, 0.012)
	β_{21}, β_{22}	1, 2	(0.979, 1.983)(0.049, 0.084)	(0.046, 0.072)(0.007, 0.011)
	β_{31}, β_{32}	1, 2	(0.994, 2.007)(0.050, 0.052)	(0.046, 0.074)(0.007, 0.010)
	β_{41}, β_{42}	1, 2	(1.024, 1.977)(0.047, 0.031)	(0.048, 0.078)(0.008, 0.012)
	β_{51}, β_{52}	1, 2	(0.991, 1.991)(0.041, 0.061)	(0.048, 0.072)(0.007, 0.019)
	w_1, w_2	0.5, 0.5	(0.502, 0.498)(0.015, 0.015)	(0.048,0.048)(0.003, 0.003)

3.2 Application to Dengue Hemorrhagic Fever

Here, we analyse a dataset of medical records of patients with DHF. These data were taken in Dr. Wahidin Sudirohusodo hospital, Makassar in 2005 - 2007. This dataset contains DHF patients, comprising 2091 patient. Patients with missing values for a particular DHF element were excluded from all analysis involving that element. The response variable used in this study is the length of stay, which is a long hospitalisation of patients with DHF until otherwise be discharged as the situation improved and within the limits of the study period, in days, with provisions: (a) If a patient's inpatient admission to otherwise allowed to go home because the situation improved in the care of Dr Wahidin Sudirohusodo hospital and within the limits of the study period, the survival time is categorised as not censored survival data; (b) If an inpatient either exceeded the limits of research or died, o moved hospital then it is classified as censored survival data. While the predictor variables were used: (a) The number of trombocyte (X_1) is the amount of trombocyte when the patient was first declared inpatient admission with a value of 1 for amount of trombocyte <50,000/µl, a value of 2 for the amount of trombocyte is 50000/µl-100000/ml, a value of 3 for the amount of trombocyte is 100000/µl-150000/µl, and a value of 4 for the trombocyte amount is more than 150,000/µl. (b) Hematocrit levels (X_2) is a hematocrit levels when the patient was first declared the inpatient admission. (c) Variable age (X_3) is the age of the patients at admission hospital. (d) Variable gender (X_4) with a value of 1 for female and the value 2 for male.

We fitted Weibull mixture models to the dataset using the prior distributions described in Section 2. As in the simulation study, we use the WinBUGS software (see Lunn et al., 2009) to fit the MCMC, where for each model, 100,000 samples were collected and after a burning period of 10,000. Summary statistics of the dataset are given in Table 2.

Table 2
The description of Dengue Hemorrhagic Fever

Variable	Mean	Standard Deviation	Median	Minimum	Maximum
Survival time	3,519	2,002	3,000	1,000	15,000
Age	3,247	0,812	3,000	1,000	6,000
Hematocrit	38,53	6,495	39,000	3,000	78,000
Trombocyte	85400	39849	86000	4000	150000

From Table 2, we can see the average of the number of trombocyte of patients with DHF in Wahidin Sudirohusodo hospital Makassar was 85400, with a minimum and maximum amount of trombocyte were 4000/µl and 150000/µl, respectively. The smaller the number of trombocyte the worse anyway disease dengue fever of a patient will be and vice versa. The normal amount of human trombocyte is min 100,000/µl. Thus, from Table 1 above there are patients whose condition is very unstable because they have a trombocyte count of 4000/µl. Moreover, the average level of hematocrit of DHF patients was 38.53% with the lowest and the highest levels were 3% and 78%, respectively. Different from the trombocyte, for the hematocrit levels, the greater level of the patient's hematocrit, then the patients' condition is likely to be more severe and vice versa.

An increase of hematocrit is usually preceded by a decrease in trombocyte. This increase reflects increasing capillary permeability and plasma leakages. It should be noted that the hematocrit value is affected by the replacement fluid or bleeding. Hematocrit levels will continue to rise if there is bleeding and will always decrease after the administration of fluids to patients.

Based on the characteristics of long hospitalization of patients, subsequently we calculated the survival function and the hazard function. The result can be seen in Table 3.

Table 3
The Survival Probability and the Hazard Rate of DHF Data

t	St(1)	St(2)	ht(1)	ht(2)	h(t)
1	0.6482	0.3186	0.0692	0.0175	0.0356
2	0.5109	0.2924	0.0280	0.0346	0.0733
3	0.2735	0.2516	0.0101	0.0466	0.1071
4	0.0833	0.2024	0.0023	0.0514	0.1345
5	0.0121	0.1518	0.0003	0.0493	0.1556
6	0.0007	0.1061	0.0000	0.0422	0.1721
7	0.0000	0.0689	0.0000	0.0325	0.1865
8	0.0000	0.0416	0.0000	0.0227	0.2012
9	0.0000	0.0233	0.0000	0.0145	0.2177

Table 3 shows that the probability of survival function is progressively increasing, but function the hazard is progressively decreasing. This means that the longer the patients stayed in the hospital, the lower the patients' survival probability will be, but contrast to this, the patient's hazards' rate will be higher. The survival function gives the probability of survival of patients survive for all time t, for example, the probability of patients'

survive on day 4 was 0.0833, meaning that the number of patients who would not recover on day 4 was 8.33%, and based on the hazard function, on day 4 patients hazard rate was 0.0023.

Table 4
Posterior Estimates of Parameters (α, β, π) for DHF Data

k	Parameter	Variable	Posterior Mean	SD	95% Credible Interval (CI)
1	α_1		3.104	0.157	(2.817, 3.44)
	β_{11}	trombocyte	-0.344	0.120	(-0.583, -0.118)
	β_{21}	hematocrit	-0.552	0.057	(-0.664, -0.439)
	β_{31}	age	-0.212	0.061	(-0.334, -0.097)
	β_{41}	gender	-0.201	0.083	(-0.356, -0.026)
	π_1		0.570	0.041	(0,491 , 0,649)
2	α_2		1.797	0.058	(1.682, 1.919)
	β_{12}	trombocyte	-0.076	0.075	**(-0.223, 0.070)**
	β_{22}	hematocrit	-0.508	0.042	(-0.594, -0.425)
	β_{32}	age	-0.150	0.043	(-0.230, -0.066)
	β_{42}	gender	-0.314	0.067	(-0.442, -0.183)
	π_2		0.430	0.041	(0.351, 0.509)

As can be seen from Table 4, the 95% credible intervals for $\beta_1, \beta_2, \beta_3, \beta_4$ respectively do not include 0 in the first component. This finding is interesting, since it indicates that four of these variables substantially contribute to patients' survival times, namely trombocyte, hematocrit, age and gender, with a negative effect on the predicted survival time. In the second component, the hematocrit, age and gender substantially described patients' survival times and had a negative effect on the predicted survival time. Based on Table 4, the Weibull mixture models for the first component (W1) and the second component (W2) is

$$W = \pi_1 W_1(t_i|\alpha_1, \lambda_1) + \pi_2 W_2(t_i|\alpha_2, \lambda_2),$$

Where

$$W_1(t_i|\alpha_1, \lambda_1) = 0.57\left[3{,}104 t_i^{2{,}104} \exp\left(\lambda_1 - \exp(\lambda_1) t_i^{3{,}104}\right)\right];$$
$$\lambda_1 = -0{,}344 x_{\text{trombosit}} - 0{,}552 x_{\text{hematokrit}} - 0{,}212 x_{\text{umur}} - 0{,}201 x_{\text{gender}};$$
$$W_2(t_i|\alpha_2, \lambda_2) = 0{,}43\left[1{,}797 t_i^{0{,}797} \exp\left(\lambda_2 - \exp(\lambda_2) t_i^{1{,}797}\right)\right]; \text{ and } \lambda_2 = -0{,}508 x_{\text{hematokrit}} - 0{,}150 x_{\text{umur}} - 0{,}314 x_{\text{gender}}$$

From Table 5, we can see that the smallest value of BIC and AIC for the Weibull mixture model without were 6939 and 6909, respectively. Based on these, we conclude that the model with three components is more appropriate for this data set.

Table 5
The Value of BIC and AIC for DHF Data

Number of component (k)	BIC	AIC
1	8288	8278
2	7233	7213
3	6939	6909

4. CONCLUSION AND DISCUSSION

This paper has presented the Bayesian Weibull mixture model with MCMC computational methods. The case study that we considered involved DHF survival, with covariates given by trombocyte, hematocrit, age and gender.

Based on two goodness of fit criteria, we showed that selected covariates are substantially associated with survival in this study. Computation of DIC and BIC estimates for all possible combinations of the covariates facilitated full consideration of competing models. For example, models that are not best in the sense of goodness of fit (based on these two criteria) may be interpretable with respect to their biological and medical implications.

When viewed from the age, the majority of patients with this disease aged 21-30 years. Our study also supports the work of Karyanti et al. (2014) who indicated that in those aged 15 years or over, DHF incidence increased.

The first mixture components that affect a cure DHF patient were age, gender, hematocrit, and trombocyte. The results show that compared to female DHF patients, male patients tend to recover faster by 0.818 times. The greater the patient's hematocrit by one unit, the longer the recovery of the patients, which is 1,733 times, and DHF patients with trombocyte amount of between 4000/µl and 150 000/µl tend to recover faster by 0.708 times compared to those with different amount of trombocyte.

The second mixture components affecting a cure DHF were age, gender and hematocrit levels. The results show that male patients tend to heal faster by 0.371 times than female ones and the patient's with greater hematocrit by one unit, tends to have longer recovery for 1,661 times, and DHF patients with ages between 21 and 30 years tend to heal faster by 0.731 times than those with different ages.

Based on two goodness-of-fit criteria, we showed that selected variables are substantially associated with survival in this study. Computing by using the BIC and AIC can estimate all possible combinations of the number of components required in determining the best model. For example, models that are not best in the sense of goodness of fit (based on these two criteria) may be interpretable with respect to their biological and medical implications.

Apart from accuracy and precision criteria used for the comparison study, the Bayesian approach coupled with MCMC enable us to estimate the parameters of Weibull survival models and probabilistic inferences about the prediction of survival times. This is a significant advantage of the proposed Bayesian approach. Furthermore, flexibility of Bayesian models, ease of extension to more complicated scenarios such as a cure mixture model, relief of analytic calculation of likelihood function, particularly for non-tractable

likelihood functions, and ease of coding with available packages should be considered as additional benefits of the proposed Bayesian approach to predict survival times.

REFERENCES

1. Akaike, H. Information theory and extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki. (1973). Second International Symposium on Information Theory (pp. 267-281).
2. Berger, J. (1985). Statistical Decision Theory and Bayesian Analysis. 2nd ed. Springer: Verlag.
3. Besag J, Green E, DHigdon, and Mengersen, K. (1995). Bayesian computation and stochastic systems. Statistical Science, 10(1), 3-41.
4. Bhadra, A and Mallick, B. K. (2013). Joint high dimensional Bayesian variable and covariance selection with an application to eQTL analysis. Biometrics, 69, 447-457.
5. Burnham, K and Anderson, N. (2004). Multimodel inference: Understanding AIC and BIC in model selection. Sociological Methods Research, 33, 261–304.
6. Cai, B., and Meyer, R. (2011). Bayesian semiparametric modelling of survival data based on mixtures of b-spline distributions. Computational Statistics and Data Analysis 55, 1260–1272.
7. Collet, D. (1994). Modelling Survival Data in Medical Research. 1st ed. Chapman and Hall.
8. Dempster, AP., Laird, NM. and Rubin, DM. (1977). Maximum likelihood from incomplete data via EM algorithm (with discussion). Journal of the Royal Statistical Society: Series B, 39:1–38.
9. DetikHealth Jumat, 15/06/2012. (2012). 10 Peringkat Daerah Tertinggi Demam Berdarah di Indonesia, accessed on 20 Desember 2014.
10. Diebolt, J and Robert, CP.(1994). Estimation of finite mixture distributions through Bayesian sampling. Journal of the Royal Statistical Society: Series B, 56, 363–375.
11. Dodson, B. (1994). Weibull Analysis. American Society for Quality, Milwaukee.
12. Farcomeni, A and Nardi, A. (2010). A two-component Weibull mixture to model early and late mortality in a Bayesian framework. Computational Statistics and Data Analysis, 54, 416–428.
13. Gubler DJ. (2002). Epidemic dengue/dengue haemorrhagic fever as a public health, social and economic problem in the 21st century. Trends Microbiology, 10, 100–103.
14. Gilks, WR., Richardson, S. and Spiegelhalter, D.G. (1996). Markov Chain Monte Carlo in Practice. Chapman and Hall.
15. Gubler, DJ. (1997). Epidemic dengue/dengue hemorrhagic fever as a public health, social and economic problem in the 21st century. TRENDS in Microbiology, 10(2), 100-103.
16. Gupta, A., Mukherjee, B., and Upadhyay, S. K. (2008). Weibull extension model: A Bayes study using Markov Chain Monte Carlo simulation. Reliability Engineering and System Safety, 93(10), 1434-1443.
17. Ibrahim, JG., Chen, MH and Sinha, D. (2001). Bayesian semiparametric models for survival data with a cure fraction. Biometrics, 57:383–388.
18. Ibrahim, JG., Chen, MH., and Kim, S. (2008). Bayesian variable selection for the Cox regression model with missing covariates. Life Time Data Analysis, 14, 496–520.

19. Kaderali L, Zander T, Faigle U, Wolf J, Schultze JL, Schrader R. (2006). CASPAR: a hierarchical bayesian approach to predict survival times in cancer from gene expression data. Bioinformatics, 22, 1495-1502.
20. Karyanti, MR., Uiterwaal, CSPM., Kusriastuti, R., Hadinegoro, SR., Rovers, MM., Heesterbeek, H., Hoes, AW., Bruijning-Verhagen, B. (2014). The changing incidence of Dengue Haemorrhagic Fever in Indonesia: a 45-year registry-based analysis. BMC Infectious Diseases, 14, 412.
21. Kautler, I., Robinson, MJ. And Kubnle, U. (2007). Dengue virus infection: Epidemiology, pathogenesis, clinical presentation, diagnosis and prevention. The Journal of Pediatrics, 131(4), 516-524.
22. Kleinbaum, DG and Klein, M. (2005). Survival analysis: A Self-learning Text. Springer.
23. Lee, KE and Mallick, BK. (2004). Bayesian methods for variable selection in survival models with application to DNA microarray data. The Indian Journal of statistics, 6(4), 756–778.
24. Leung, K. M., Elashoff, R. M., and Afifi, A. A. (1997). Censoring issues in survival analysis. Annual Review of Public Health, 18, 83-104.
25. Lunn, DJ., Thomas, A., Best, N and Spiegelhalter, D. (2000). WinBUGS – A Bayesian modelling framework: Concepts, structure, and extensibility. Statistics and Computing, 10, 325–337.
26. Marin, JM., Bernal, MR. and Wiper, MP. (2005a). Using Weibull mixture distributions to model hererogeneous survival. Communication in Statistics Simulation and Computation, 34(3), 673–684.
27. Marin, JM., Mengersen, K and Robert, CP. (2005b). Bayesian Modelling and Inference on Mixtures of Distributions. Handbook of Statistics 25, D. Dey and C.R. Rao (eds). Elsevier-Sciences.
28. McGrory, CA and Titterington, DM. (2007). Variational approximations in Bayesian model selection for finite mixture distributions. Computational Statistics and Data Analysis, 51(11), 5352–5367.
29. Mengersen, K., Robert,C.P., and Titterington, M. (2011). Mixture: Estimation and Applications. Wiley.
30. National Geographic Indonesia, Berita Kesehatan, Tuesday 15 Mei 2012.
31. Nguyen and D.M Rocke. (2002). Partial least squares proportional hazard regression for application to DNA microarray survival data. Bioinformatics, 18(12), 1625-1632.
32. Ntzoufras, I. (2009). Bayesian Modelling using WinBUGS. Wiley, New Jersey.
33. Quiang, J. (1994) A bayesian weibull survival model. Ph.D. thesis. Institute of Statistical and Decision Sciences, Duke University.
34. Robert, CP and Casella, G. (2000). Monte Carlo Statistical Methods. Springer, New York.
35. Rosenwald A, Wright G,Wiestner A, Chan WC, Connors JM, Campo E, Gascoyne RD, Grogan T, Muller HK, Smeland EB, Chiorazzi M, Giltnane JM, Hurt EM, Zhao H, Averett L, Henrickson S, Yang L, Powell J,Wilson WH, Jaffe ES, Simon R, Klausner RD, Montserrat E, Bosch F, Greiner TC,Weisenburger DD, SangerWG,Dave BJ,Lynch JC,Vose J, Armitage JO, Fisher RI, Miller TP, LeBlanc M, Ott G, Kvaloy S, Holte H, Delabie J and Staudt LM. (2002). The use of molecular

profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma. The New England Journal of Medicine 346(25), 1937–1947.
36. Schwarz, GE. (1978a). Estimating the dimension of a model. Annals of Statistics, 6(2), 461–464.
37. Spiegelhalter, N., Best, N., Carlin, B and vanderLinde, A. (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society:Series B, 64 (4), 583–639.
38. Stajduhar, I., Basic, B. D., and Bogunovic, N. (2009). Impact of censoring on learning Bayesian networks in survival modelling. Artificial Intelligence in Medicine, 47, 199-217.
39. Stephens, M. (1997). Bayesian methods for mixtures of normal distributions. PhD thesis, the University of Oxford.
40. Tachmazidou, I., Andrew, T., Vercilli, C., Johnson, M. and De Lorio, M. (2008). Bayesian survival analysis in genetic association studies. Bioinformatics, 24, 2030-2036.
41. Thamrin, SA., McGree, JM. and Mengersen, KL. (2012). Bayesian Weibull survival model for gene expression data. In C. L. Alston, K. L. Mengersen and A. N. Pettitt (ed.). Case Studies in Bayesian Statistical Modelling and Analysis (1st ed., pp. 171-185).
42. Thamrin, SA. (2013). Bayesian Survival Analysis Using Gene Expression. PhD thesis, Queensland University of Technology.
43. Thamrin, SA., McGree, JM. and Mengersen, KL. (2013). Modelling survival data to account for model uncertainty: a single model or model averaging? Springer Plus, 2, 665.
44. Tsionas, E. G. (2002). Bayesian analysis of _nite mixtures of Weibull distribu- tions. Communication in Statistics - Theory and Methods, 31, 37-48.
45. VIVAnews, Demam Berdarah. Accessed on Thursday, 19 Desember 2013 at 11:03am.
46. Volinsky, C.T. and Raftery, A.E. (2000). Bayesian information criterion for censored survival models. Biometrics, 56, 256–262.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 87-96

SPATIAL ANALYSIS FOR THE DISTRIBUTION OF HUMAN DEVELOPMENT INDEX AND REGIONAL GOVERNMENT BUDGET IN EAST JAVA PROVINCE

Vinna Rahmayanti Setyaning Nastiti[1], Muhammad Nur Aidi[2]
and **Farit Mochammad Afendi[3]**
Bendungan Riam Kanan No 4, Malang, East Java Province, Indonesia
Email: vinna.rahmayanti@gmail.com

ABSTRACT

There is a gap of Human Development Index (HDI) between regency and city in East Java Province. This is due to the faster development of urban areas be compared to rural areas. On the other hand, movement from the regency to city is one of the causes of these imbalances. The impact of migration from the regency to city is low quality of human resources in the regency. HDI figures of regency and cities in East Java are calculated from the components of life expectancy at birth, literacy rate, mean years of schooling, and gross income per capita. Each component of HDI is expected to have spatial influence. Calculating of HDI is used Local Indicator of Spatial Autocorrelation (LISA). HDI is not only influenced by the local effect, but there is a relationship between HDI and Regional Government Budget. To determine the relationship between HDI and Regional GovernmentBudget be used General Spatial Model (GSM). This results of this study indicate that all components of the HDI have a spatial dependency based on city/regency in East Java. GSM model formed in East Java showed R^2_{adj} of 80.12%. GSM modelling also shows that the estimation result of HDI data is close to the actual data by the difference of 1.9978. Thus, it can be said that GSM model is good to be implemented in East Java.

KEYWORDS

Human Development Index, LISA, Local Government Budget, General Spatial Model.

1. INTRODUCTION

Human development index (HDI) was first introduced in Indonesia in 1996. Based on human rights and socio-economic rights in the Constitution of the Republic of Indonesia, there are four components that affect the HDI. Those are life expectancy at birth, literacy rates, mean years of schooling, and gross income per capita. The rate of HDI figures in Indonesia has increased from year to year. According to the United National Development Program in 2012, Indonesian HDI figures in the period 2011-2012 have increased from 0.624 to 0.629. However, compared to the 5 countries of ASEAN, Indonesia HDI figures are still relatively low. The HDI figure in the countries such as Singapore, Brunei Darussalam, Malaysia, Thailand, and the Philippines are respectively 0.866, 0.838, 0.761, 0.682, and 0.644. In addition, Indonesia's HDI figures based on city/regency showing inequality between the western and eastern regions. Cities and

regencies that have a high HDI are mostly located in the western region of Indonesia, while cities and regencies have the low HDI level are located in the eastern region of Indonesia.

The provinces on the island of Java show inequality of HDI figures. Based on the Human Development Index 2012, DKI Jakarta Province was ranked 1 and Yogyakarta Special Region 4th rank, while the province of Central Java, West Java, and East Java consecutively ranked 14, 16 and 17. It is caused by the inequality of HDI in the area of city/regency. Urban areas generally have better infrastructure than the surrounding regency, causing urbanization from the regency to the city. The impact of migration is causing the regency to have lower human qualities. HDI figures regencies and cities in East Java are calculated based on its components. Each of the components of the HDI is thought to influence spatial. The emergence of the spatial influence was caused by the proximity factor between geographic regions. Therefore, analysis is required in order to know the influence of HDI figures in the city/regency toward the city/regency in the vicinity. In addition, the grouping and correlation analysis between neighboring regions are required. Analysis of correlation is based on the influence of the location called spatial correlation. Local Indicator of Spatial Autocorrelation (LISA) is a statistical method that aims to identify and describe the spatial correlation. LISA is being used to analyze the components of the HDI indicators spread across the city/regency in East Java.

HDI values are not only influenced by the effects of the location, but there are indications of its relationship with the Regional Government Budget. This can be seen in previous studies conducted by Nasution (2010) and Paramita (2012). Nasution (2010) examined the impact of the realization of the budget to the improvement of the human development index in Binjai. This study used a multiple linear regression method and the results of this study indicate that the budget for the education sector and the health sector brings a positive influence on the human development index. Paramita (2012) conducted a study on the relationship between the budget realization and the Human Development Index in Makassar from 2000 to 2009. This study used a linear logarithmic and the results of this study indicate that the realization of the budget realization gives effect on the human development index because local/city governments have more freedom in using its budget in accordance with the needs of the community. Both studies only analyzed the relationship between the HDI and the budget without looking at the influence of the location.Based on both of research, the influence of the location add in this studyby using the method of the General Spatial Model (GSM).The purposes of this study are (1) to analyze the spatial correlation, globally and locally for all constituent HDI 2012 in East Java; (2) Prepare the General Spatial Model HDI with budgets in the province of East Java

2. METHODOLOGY

The data used in this study are HDI data and budget data. Data Human Development Index in 2012 by city and regency in the province of East Java was obtained from the Badan Pusat Statistik. HDI data was obtained include component life expectancy at birth, literacy rates, mean years of schooling, gross income per capita in each city and regencies in East Java. Life expectancy at birth is based on the calculation of the average age of people who live in the area. This figure indicates the rate of a person's health. HDI

components the education sector is affected by the literacy rate and mean years of schooling. Mean years of schooling describes the amount of time which has spent to carry out formal education in the age of 15 years and over in the population, while the literacy rate is the percentage of the population aged 15 years and over who can read and write.Gross income per capita is the average costs to be spent by people in an area within a certain time. This figure illustrates the rate of well-being and explains the condition of the people who can meet the standards of living. Regional Goverment Budget in 2012 obtained from Directorat Genereal of Finance in the Ministry of Home Affairs. Regional Goverment Budget shows the allocation of expenditure to implement programs of the local government. That programs will be implemented to encourage regional economic growth, income distribution, and development in many aspects.

Procedureanalysis steps which were conducted in this study is as follows:

1. Created, calculated, and established contiguity weighting matrix for all locations across the regency / city of East Java Province. Weight calculation used the following equation.

$$w_{ij} = \frac{c_{ij}}{\sum_j c_{ij}} \tag{1}$$

2. Observed values of standardization as the basis to calculate the value of global and local moran index.

3. Identified spatial correlation with the index used the LISA method forming components of HDI. Moran global index equation is shown in equation.

$$I = \left[\frac{n}{\sum_i^n \sum_j^n W_{ij}}\right]\left[\frac{\sum_i^n \sum_j^n W_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i^n (x_j - \bar{x})^2}\right] \tag{2}$$

Furthermore, moran local index equation is shown in equation (3).

$$I_i = z_i \sum_j W_{ij} z_{ij} \tag{3}$$

4. Created spatial exploration using moran scatterplot and map of components of HDI.

5. Analyzed of the relationship between the HDI and the budget by using the General Spatial Model with the following equation.

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$
$$\mathbf{u} = \lambda\mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon} \tag{4}$$
$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

6. Testing the spatial effects that spatial dependency test with the Lagrange Multiplier method with the following equation.

$$LM_{GSM} = \frac{(\frac{\mathbf{e}'\mathbf{W}}{\mathbf{s}^2}\mathbf{e} - \frac{\mathbf{e}'\mathbf{W}}{\mathbf{s}^2})^2}{(nJ - T)} + \frac{(\frac{\mathbf{e}'\mathbf{W}}{\mathbf{s}^2})^2}{T} \tag{5}$$

7. Predicted parameters for spatial regression model equation.

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{B}'\boldsymbol{B}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{B}'\boldsymbol{B}\boldsymbol{A}\boldsymbol{y}$$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'(\boldsymbol{I} - \hat{\lambda}\boldsymbol{W_2})'(\boldsymbol{I} - \hat{\lambda}\boldsymbol{W_2})\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{I} - \hat{\lambda}\boldsymbol{W_2})'(\boldsymbol{I} - \hat{\lambda}\boldsymbol{W_2})(\boldsymbol{I} - \hat{\rho}\boldsymbol{W_1})\boldsymbol{y} \quad (6)$$

8. Determined the model by using the method of the General Spatial Model in East Java.

9. Compared HDI actual and prediction of HDI using graphic.

3. RESULT AND DISCUSSION

Local Indicator of Spatial Autocorrelation

East Java is one of provinces in Java. The total area of East Java reaches 47 963 km^2 which is divided by two parts. There are mainland of East Java of 42 521 km^2 and Madura island of 5422 km^2. East Java surrounded by Java Sea, the strait of Bali, Indian Ocean, and Central Java province.East Java is located at 111.00 to 114.40 east longitude and south latitude 7.120 to 8.480. Malang has the highest number of regencies. Having a large number of regencies does not automatically having a larger number of villages.Regency that hasthe largest number of villages is Lamongan, with 474 villages. Meanwhile, the region with the greatest land area is Banyuwangi with the total area of 3,606 km^2. Population of East Java in 2012 is 38.1 million inhabitants.

East Java Province HDI rateshows fluctuation when measured from before the crisis until 2012. In 1996 the HDI rate of East Java Province was 65.5 in 1999 it had decreased to 61.8. Then, in 2002 it increased to 62.64, while in 2005 it increased again to 65.89 which is almost the same position with the condition before the economic crisis. Furthermore, the HDI in 2008 reached 70.38 and in 2012 it reached the peak on 72.83. Improved HDI of East Java Province from 2002 to 2012 showed that the stability of the economic and human development has begun to show signs of improvement and of course cannot be separated from the contribution of the determining components, such as Life Expectancy at Birth, Literacy Rates, Mean Years of Schooling, and Gross Income per Capita. Descriptive statistics of the components of HDI are presented in Table 1.

Table 1
Descriptive Statistics of the Components of HDI in East Java on 2012

Statistic	HDI	Life Expectancy at Birth (Year)	Literacy Rates (Percent)	Mean Years of Schooling (Year)	Gross Income Per Capita (thousand rupiahs per capita)
Minimal	61.67	61.70	69.12	4.22	624.05
Maximal	78.43	72.80	98.35	10.87	660.38
Average	71.87	68.76	90.03	7.69	643.70

Table 1 shows the average HDI in East Java at 71.87 where the lowest HDI located in Sampang and the highest HDI located in Malang. The health dimension measured by Life Expectancy at Birth has an average of 68.76 years. The lowest of Life Expectancy at Birth is located in Probolinggo and the highest located in Blitar. The education components is measured by Literacy Rates and Mean Years of Schooling. There was an inequality in Literacy Rates which is the highest Literacy Rates is 98.35% located in

Surabaya and the lowest Literacy Rates is 69.12% located in Sampang. Mean Years of Schooling has an average of 7.69 years which the lowest Mean Years of Schooling is located in Sampang and the highest Mean Years of Schooling is located in Malang. The standard of living dimension is measured by gross income per capita. Gross Iincome per Capita has an average of 643 700 rupiahs per capita which the lowest Gross Income per Capita is located in Bojonegoro and the highest is located in Surabaya.Results of testing global spatial index in the province of East Java using Moran index can be seen in Table 2.

Table 2
Components of HDI Global Moran Index in East Java in 2012

HDI Components	Global Moran Index	*P-Value*
Life Expectancy at Birth	0.6709	0.0000
Literacy Rates	0.5630	0.0000
Mean Years of Schooling	0.3508	0.0022
Gross Income per Capita	0.3483	0.0024

Table 2 shows that all components of the HDI in East Java, including Life expectancy at Birth, Literacy Rates, Mean Years of Schooling, and Gross Income per Capita significant. The fourth component indicates that there is a spatial correlation between cities/regencieswith neighbouring spatial correlation either negative or positive in the province of East Java. Distribution of the four components of the HDI in East Java province with moran scatterplot is shown in Figure 1.

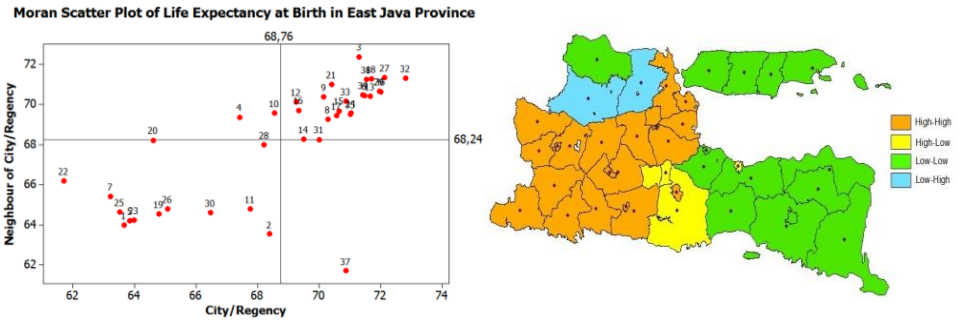


Figure 1: Moran Scatter Plot and Thematic Map ofLife Expectancy at Birth in East Java on 2012

Moran Scatter Plot of Life Expectancy at Birth in East Java shows clustered in first quadrant, which means that this component has a positive spatial correlation in the city/regency of East Java Province. Health dimension in the province of East Java indicated by high value of life expectancy at birth and the value in the surrounding area is also high. Areas that have 'high-high' criteria are Pacitan, Trenggalek, Tulungagung, Blitar, Kediri, and surrounding areas. In addition, significant regions also occur in third quadrant, which shows the region with'low-low' criteria, including Probolinggo, Jember,

Bangkalan, Pamekasan, Sampang, and Sumenep. In education dimension, HDI components indicated by Literacy Rates in Figure 2.

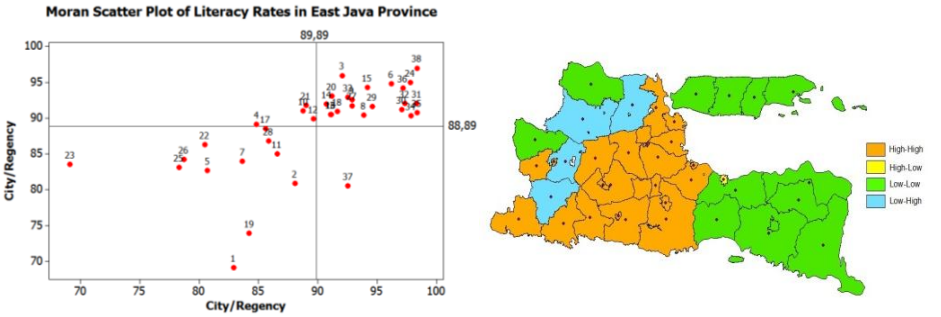


Figure 2: Moran Scatter Plot and Thematic Maps of literacy rate in East Java on 2012

Moran Scatter Plot Literacy Rate in East Java shows the pattern of clusters in first quadrant, which means that the component has a positive spatial correlation in the city/regency of East Java Province. Some significant regions in the 'high-high' quadrant on Literacy Rate components are Gresik, Mojokerto, Sidoarjo and Surabaya. This happens because the area is directly neighboring to the capital of the province of East Java that have an impact on the surrounding area. While the area around the island of Madura and Banyuwangi are on the 'low-low'criteria which means those areas have low Literacy Rate and are surrounded by other areas with low level of literacy as well. The education dimension as indicated by Mean Years of Schooling can be seen in Figure 3.

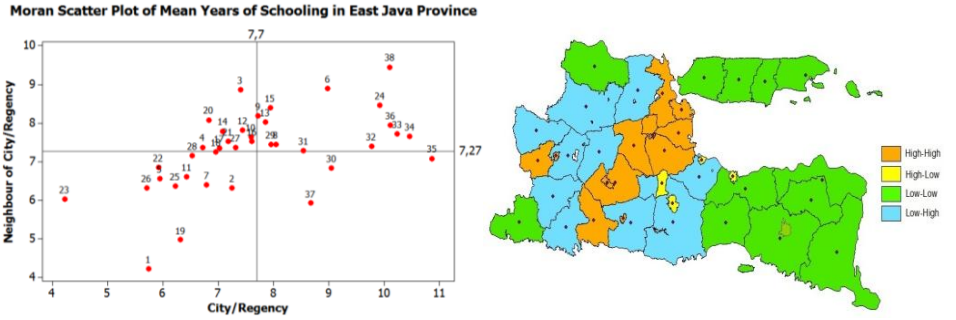


Figure 3: Moran Scatter Plot and Thematic Maps of Mean years of schooling in East Java on 2012

Figure 3 that shows the Mean Years of Schooling in East Java is clustered in fourth quadrant, indicating outliers pattern. The outlierspattern meansthat those area have short mean years of schooling, while the surrounding area have high duration. Regions which have these criteria are Lamongan, Bojonegoro, Madiun, Trenggalek, Blitar, and Malang. In addition, some regions showed lowmean years of schooling. These areas are Bangkalan,

Pamekasan, Probolinggo, Sampang, and Situbondo. On the economic dimension of the HDI which is shown by Gross Income per Capita can be seen in Figure 4.

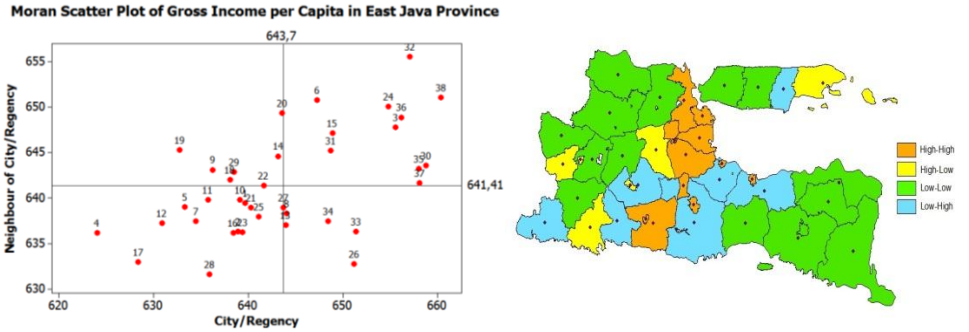


Figure 4: Moran Scatter Plot and Thematic Maps of Gross Income Per Capita in East Java on 2012

In this Gross Income per Capita, the figure indicates the cluster in third quadrant which is the 'low-low' quadrant. Significant regions in East Java Province are Bojonegoro, Jember, Madiun, Nganjuk, Ngawi, and Tuban. These areas are geographically so close together and it indicates a spatial effect. Overall, East Java has low rate in education, economics, and health. The cause of of low numbers HDI-forming components is due to the location of administrative area and numerous remote areas in East Java which are far away from the capital city of Indonesia.

General Spatial Model of East Java Province

The results of testing the assumption of spatial dependence in East Java using the Lagrange Multiplier (LM) test. The test shows that P-Value in East Java Province is smaller by 10% when compared to α. Based on the initial hypothesis of LM test, it can be concluded that LM has a significant coefficients. Thus, it can be concluded that there is a spatial dependence between between HDIamong one region to another neighboring region in the province of East Java. The presence of spatial dependence in East Java with GSM model can be formed by the following equation.

$$\hat{y}_i = 0.48099Wy + 35.3924 + 0.01888\,RGC + u \quad (7)$$

$$u = 0.47396Wu$$

Formed GSM model in equation (7) has R^2adj of 80.12%. The model explains that the diversity of the HDI variable is 80.12%, while 19.88% is explained by other variables outside the model. Predicted values are positive parameters independent variables. This means that one rupiah increase of the Regional Governement Budgete will increase by HDI by 0.0188 in the province of East Java. The ρ coefficient is significant with the value of 0.4809. This value indicates that an area that is surrounded by other n regions, then the influence of each region in East Java that surrounds it can be measured at 0.4809 multiplied by the average HDI variables in the surrounding region. The coefficient λ in the model is significant with a coefficient of 0.4736. This value indicates that if an area

surrounded by other regions of n, then the effect of each of the surrounding region can be measured at 0.4736 times the measurement error in the variables surrounding areas in East Java. Based on GSM models that form the HDI predictive value in East Java can be obtained. The results of the comparison between the actual value and the expected value of the HDI can be shown in Figure 5.

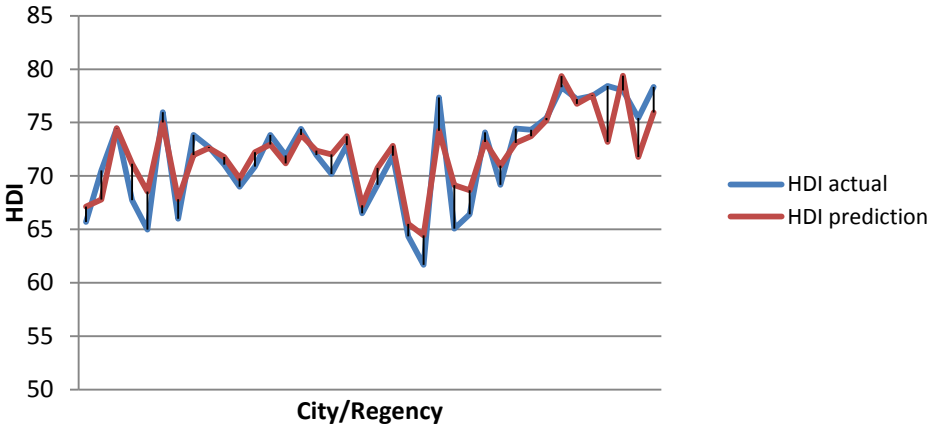


Figure 5: The Comparison of HDI Actual and HDI Prediciton in East Java

Figure 5 shows the comparison between the results of the estimation approach with actual HDI data in East Java. This is supported by the average of the difference with the estimation of actual data of 1.9978. GSM model in East Java involves the whole town and regency in East Java. The relation between the area and the distance between the regions neighboring to each other in the region of East Java caused the error formed is small enough. Hence, it can be said that the GSM model is the proper model for the province of East Java.

4. CONCLUSION

Distribution of components of the HDI is calculated based on Life Expectancy at Birth, Literacy Rates, Mean Years of Schooling, and gross income per capita. The fourth component inferred from spatial dependence. LISA analysis resulted in the spread of IPM components in East Java. Regencies around Bondowoso, Situbondo, Jember and Banyuwangi are at the 'low-low' quadrant, it means the area and surrounding areas in the region have a low HDI. It also occurs in regencys on the island of Madura, including Bangkalan, Pamekasan, Sampang, and Sumenep which haverelatively lower rate of Life Expectancy at Birth, Literacy Rates, and lower Mean Years of Schooling. High HDI component is dominated by the area around the city of Surabaya. This is due to the city of Surabaya is the capital of the province of East Java that has adequate facilities and infrastructure to develop the quality human resources. The relationship between the HDI with the Regional Government Budget was analyzed using the GSM model. GSM model shows the results of the estimation of HDI data approach the actual data to estimate the

average difference of 1.9978 so it can be said that the GSM model is appropriate to beimplemented in the province of East Java.

REFERENCES

1. Anselin L. (1988). *Spatial econometrics: Methods and Models.* New York (US): Kluwer Academic Publisher.
2. Anselin L. (1995). *Geographical Analysis. Local Indicator of Spatial Analysis-LISA*. 27,93-115.
3. Ardiansa D. (2010). *Analisis Spasial Untuk Sebaran Suara dan Perolehan Kursi Partai Politik pada Pemilu Legislatif 2009 di Wilayah DKI Jakarta dan Jawa Barat*. Bogor: Institut Pertanian Bogor.
4. Badan Pusat Statistik. (2013). *Indeks Pembangunan Manusia 2012*. Jakarta: Badan Pusat Statistik.
5. Nasution A. (2010). Analisis Dampak Realisasi APBD Terhadap Peningkatan Indeks *Pembangunan Manusia di Kota Binjai*. Binjai: Universitas Sumatera Utara.
6. Paramita A. (2012). *Analisis Dampak Realisasi APBD Terhadap Indeks Pembangunan Manusia di Kota Makassar Periode 2000-2009*. Makassar: Universitas Hassanudin.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 97-108

MOBILE LEARNING BASED FLASHLITE IN STATISTICS COURSE

Artanti Indrasetianingsih and **Permadina Kanah Arieska**
Statistics Department FMIPA Universitas PGRI Adi Buana Surabaya
Email: artanti.indra@gmail.com
permadina.kanah@gmail.com

ABSTRACT

The objectives of this research was to create a mobile learning (m-learning) as an alternative of learning in the subjects Statistics. M-learning programs by using mobile devices (mobile phone) will be able to be an exciting learning media, especially for subjects that still abstract and need explanation by simulation. Learning can be done anywhere and anytime. Platform was used to create m-learning applications was FlashLite 3.0. The samples were the students of Mathematics Education of PGRI Adi Buana University that took a Statistics Course at even semester 2013-2014 with a total of 43 students. The difficulties in Statistics Course often exist on Hypothesis Chapter. So this m-learning research discussed the hypothesis testing of the mean and hypothesis testing of two means. Students are given learning to m-learning and learning outcomes tested before and after utilizing m-learning applications. Most of the students stated that the applications could improve the understanding of the subjects and interesting to learn. The result of Wilcoxon test has p-value = 0.000. This shown that the average value of the students before and after using the m-learning was significantly different. It turns out that mobile learning can improve students learning outcomes to be better.

KEYWORDS

Flashlite, hypothesis testing, learning outcomes, wilcoxon tes.

INTRODUCTION

The development of Information and Communication Technology (ICT) very rapidly have affected various areas of human life, including in the field of education that is often referred to as e-learning. E-learning is learning that uses electronic circuits (LAN, WAN, or the Internet) to deliver learning content, interaction, or guidance. There also interpret the e-learning as a form of distance education is done via the Internet [4]. E-learning still has the disadvantage that requires users to deal with Personal Computer (PC) connected to the internet. As a solution of the shortage of e-learning is developed learning through mobile devices are referred to as mobile learning (m-learning).

M-learning is a learning method using mobile devices such as mobile phone/ smartphone, ipad or PDA. One mobile device applications for the development of m-learning is by using FlashLite 3.0. This application is an application with the Flash platform that is widely used because of the ease programming based WYSIWYG (What

You See Is What You Get). Therefore, the combination of mobile learning and web-based e-learning is expected to facilitate the learning process will be.

In the course of Statistics, the lecture usually uses the common method that sometimes makes students saturated. Mobile learning can be an alternative method of learning to the material in the course of Statistics attract more students. M-learning was developed with multimedia format by presenting text, images, audio, and animation. Mathematical concepts in statistics courses can be visualized with simulation and can be applied in mobile phone. The concept of m- learning is expected to promote the establishment of active learning, innovative, creative, effective and fun. Potential and prospects for the future development of mobile learning is open wide given the tendency of society increasingly dynamic and mobile as well as the demands of quality education and diverse.

Mobile learning is given, turned out to be a new medium for student learning. The objectives of this research was to create a mobile learning (m-learning) as an alternative of learning in the subjects Statistics, especially in hypothesis testing chapter.

METHODS

The samples are the students of Mathematics Education of PGRI Adi Buana University that take a Statistics Course at even semester 2013-2014 with a total of 43 students. The number of students are 23 from class A and 20 from class B.

M-learning applications are created using Flashlite 3.0 with the material in Hyphotesis Testing Chapter on the subject of Statistics, which is testing the hypothesis of the mean and hyphotesis testing of two means. The Students are given m-learning. Then the students were given a questionnaire about m-learning applications that have been created, which includes three aspects of the assessment, the software aspects (ergonomics, communicative, completeness), aspects of instructional design (attractive, easy to understand, benefits) and visual communication aspects (graphic visual, animation, color composition). Testing learning outcomes before and after using the m-learning is tested by using paired t test or wilcoxon test.

SYSTEM DESCRIPTION

System description m-learning of hypothesis testing chapter that used shown in Figure 3.1

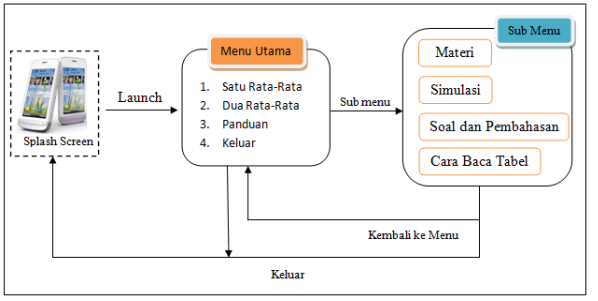


Figure 3.1: System design of m-learning application

Based on the above design, the first application on mobile phone will display a splash screen followed to the main menu. In this main menu the user can choose the option of learning consisting of Hypothesis testing of the mean, Hypothesis testing of two means, guides and exit. After that, the user can select the sub menu options consist of materials, simulations, and discussion about and how to read the table. From the sub- menu, the user can exit or return to the main menu.

SOUND DESIGN

In this application, the voice used for the application to be interesting and not boring. File used *.mp3 format. This sound file is used for background sounds and sound effects are also key to explaining tutorials on hypothesis testing. By using sound, the user can more easily understand the material provided.

INTERFACE SPLASH SCREEN AND MAIN MENU

The initial appearance before entering the main menu is the welcome splash screen. There are four choices in the main menu i.e. one mean, two means, guides and exit. Display the main menu are as follows:

Figure 3.2: The Main Menu Design

If the user selects the menu one mean, it will be associated with the following sub-menus:

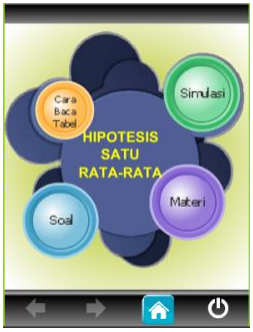


Figure 3.3: Design Sub Menu of Hypothesis Testing of one mean

Sub menu of the hypothesis testing of one mean has 4 buttons, namely material, simulation, problem and how to read table. In sub- menu contains explanations related to the hypothesis testing of one mean. Part of the sub menu materials are as follows:

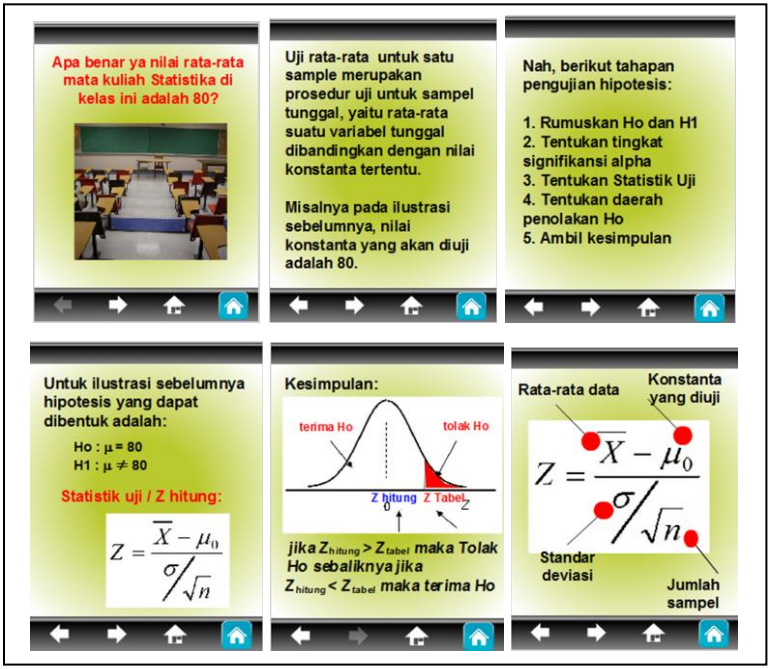


Figure 3.4: The contents of sub menu materials

Other keys are simulation button. In the simulation button is explained with examples of questions and then the solution is accompanied by animation normal table. In normal table animation, students will know the extent of the area on the normal curve. Simulation in reading normal table is also accompanied by the " play" then table will move itself in accordance with the desired value. If the observed direction of movement of the table, then the student will know the extent of the value of Z in question. Display Flashlite application if the user selects a button Simulation is as shown in the following figure.

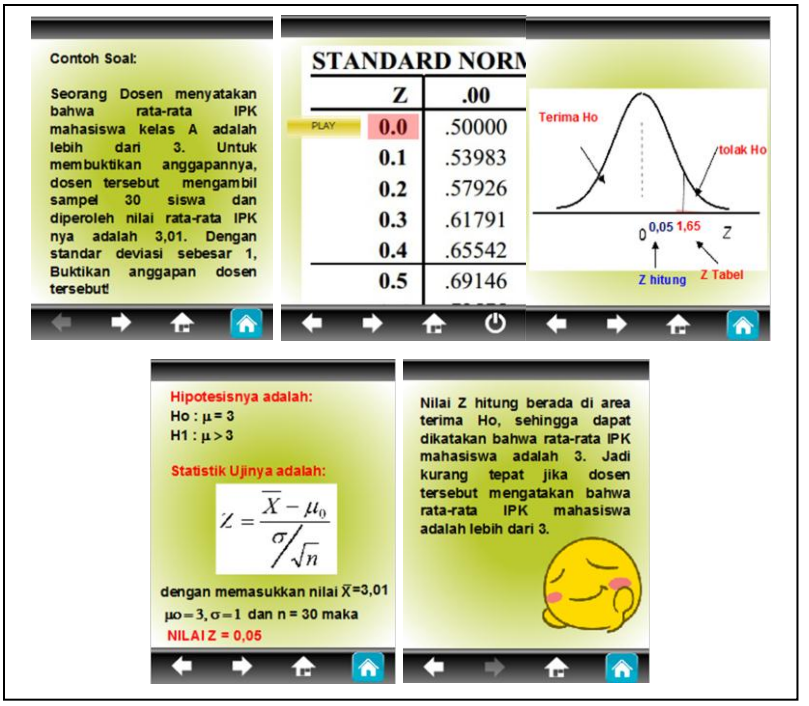


Figure 3.5: The contents of sub menu simulation

Meanwhile, if the user selects a button QUESTIONS, it will appear about as follows.

Figure 3.6: Start display of sub menu Questions

There are 5 questions given. Each was given a score of 1 so that if the user answered everything correctly, then the user will get a perfect score of 5. Here's one of questions in the sub menu.

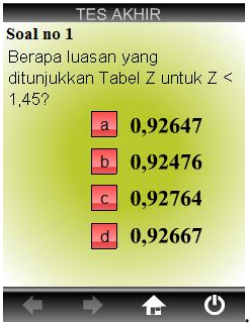


Figure 3.7: Display of sub menu end of question

Here are two possibilities animation of the user answered.

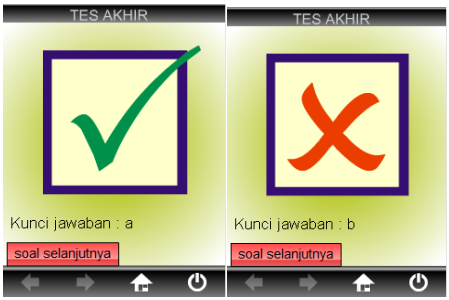


Figure 3.8: The display of two possibilities of the users answer true or false

After the user completing 5 questions given it will show the results of the overall response. There are 3 possible outcomes answer.

Table 3.1
The results of the overall response

NO	NUMBER QUESTION CORRECTLY	STATUS
1	0-3	LESS
2	4	MODERATE
3	5	PERFECT

If the result is less than the display will tell the user “you have to learn again”. If there are one wrong answer then the display will tell the user “learn again”. If all of the answers are correct then the display will tell the user “your answer is perfect”.

Figure 3.9: Display Comments of Final Results

The last sub menu is HOW TO READ TABLE. This is made because many students who cannot read Z Table. The display of this menu is in Figure 3.10.

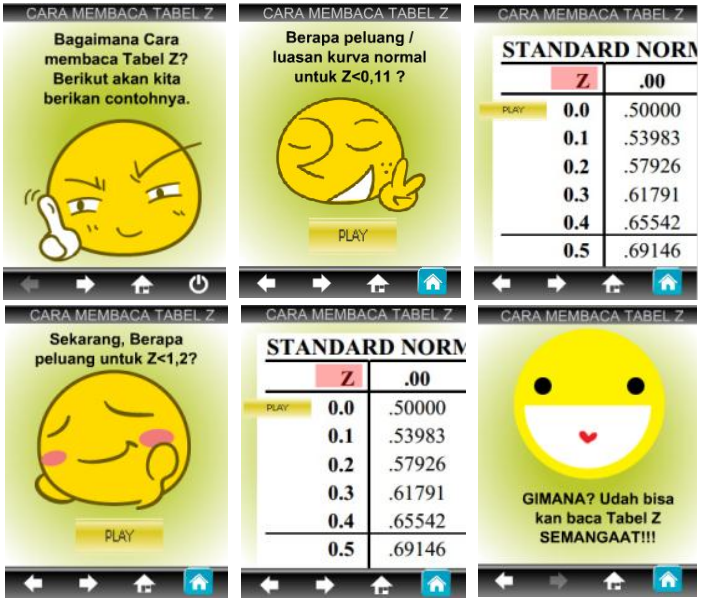


Figure 3.11: The Display of How to Read Table

The displays were adjusted according to the students so as not boring and close to the characteristics of the students. Therefore, the results of this application would be evaluated based on the advice and input of students as part of the user. Here are the results of student assessment to flashlite application of various aspects.

The questionare result in general, the students explained that the application was made is good. However, it should be improved, especially in terms of sound, in order to become more perfect. It also needs to be a more complete explanation of the material so that users clearly understand.

Table 3.2
Descriptive of Sofware Aspect, Design Learning Aspect and Communication Visual Aspect

Aspecs	Total	Percentage
Software Aspect		
1. Ergonomics aspect		
- Less easy	1	1.3
- Easy	30	69.8
- Very easy	12	27.9
2. Communicative aspect		
- Less communicative	2	4.7
- Communicative	34	79.1
- Very communicative	7	16.3
3. Completeness aspect		
- Less complete	14	32.6
- Complete	27	62.8
- Very complete	2	4.7
Design Learning Aspect		
1. Attractive aspect		
- Less attractive	1	2.3
- Attractive	25	58.1
- Very attractive	17	39.5
2. Easy to understand aspect		
- Less easy	3	7.0
- Easy	32	74.4
- Very easy	8	18.6
3. Benefits aspect		
- Usefull	23	53.5
- very usefull	20	46.5
Communication Visual Aspect		
1. Visual Graphics aspect		
- Less good	9	20.9
- Good	29	67.4
- Very good	5	11.6
2. Animation aspect		
- Less attractive	9	20.9
- Attractive	29	67.4
- Very attractive	5	11.6
3. Color composition aspect		
- Less good	8	18.6
- Good	32	74.4
- Very good	3	7.0

The test results of the application of learning material overall hypothesis testing are shown in Table 3.3. All test cases have status OK

Table 3.3
The result of testing the application m-learning

Test cases	Result	Status
Splash Screen	Animation goes well	OK
Main Menu	Animation goes well and the button works well	OK
Sub Menu: One mean	Button has been going well and right connected	OK
Sub Menu: Two mean	Button has been going well and right connected	OK
Simulation	Tutorial and button have been going well	OK
Material	Tutorial and button have been going well	OK
Test	Exercises successfully display and can check the user answers	OK
How to read table	Tutorial has been going well and button can be executed	OK
Exit	Exit button is functioning well	OK

LEARNING OUTCOMES TEST

Testing the normality of learning outcomes before and after using m- learning applications with Kolmogorov Smirnov Test. The result Normality test show that learning outcomes before and after both are not normally distributed, as a result both have p-value < 0.01, where the value is less than the value of $\alpha = 5$ %.

Therefore, the data are not normally distributed, the Wilcoxon test was used. The test result show that p – values= 0.000. This shows that the average learning outcomes of students before and after the use of mobile learning is different. The average value, appears that the learning outcomes of students after using m-learning applications is higher than before using m-learning.

WEBSITE DISPLAY

Flashlite application that has been created is uploaded on the web. This website aims to provide easy, especially for students so that students can access the course material with easy and effective. In addition, mobile learning applications using Flashlite can also be copied in the phone and downloaded for free. This site is also designed for lecturer to monitor the material that will be presented to students. It will be a two-way learning because students are also already know the material to be taught so that it can be used as a material consideration or evaluation transform the future of learning. Students and lecturer can open the web page at: http://statistika-unipasby.com

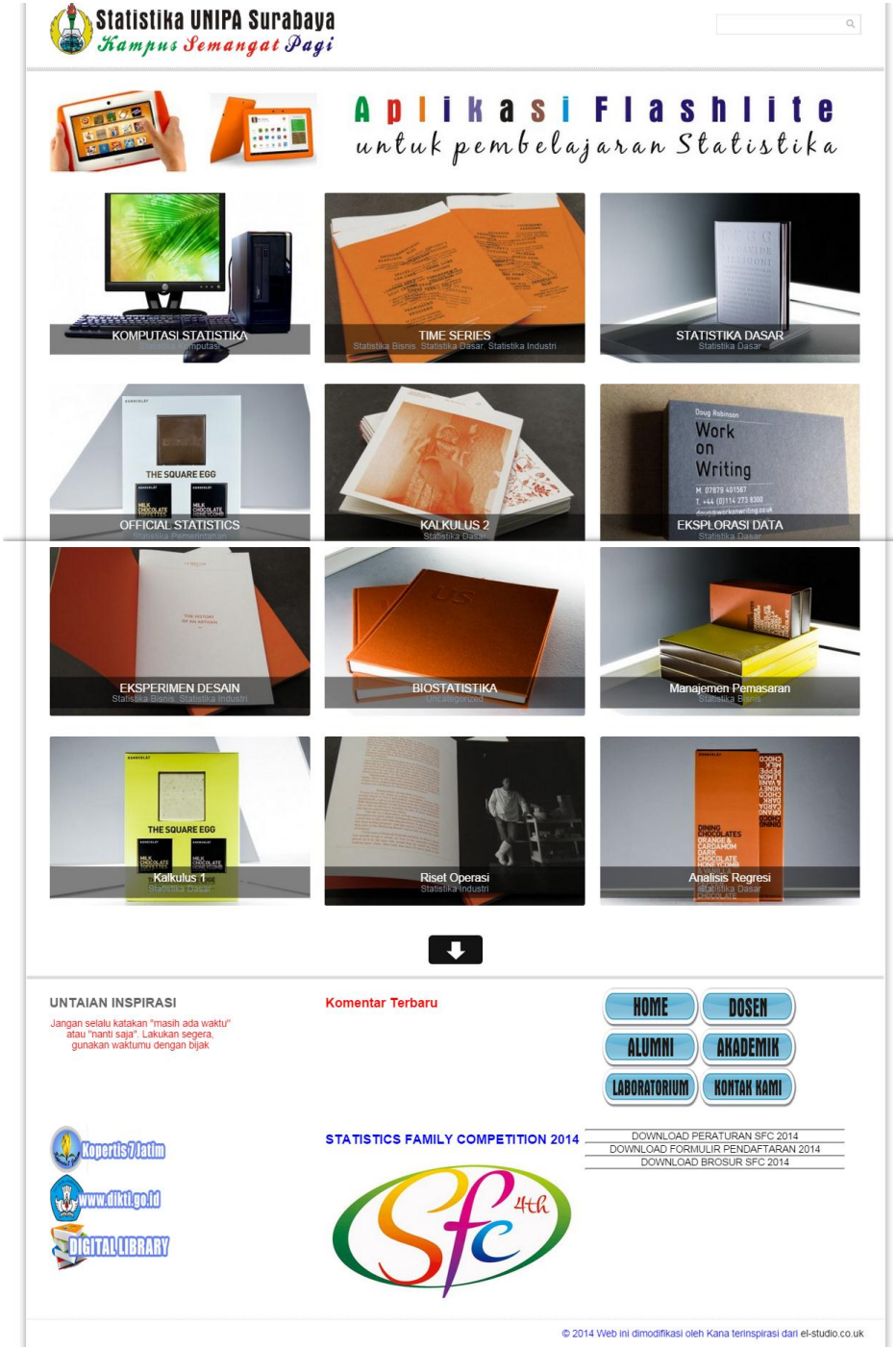


Figure 3.12: Display of Website

CONCLUSIONS

From the description of the results achieved in this study it can be concluded that m-learning applications based Flashlite can be used to facilitate students understand course material Statistics. Some advantages of m- learning applications are easy to use and mobile. This excess is used to describe materials that tend complicated subjects like Statistics in chapter hypothesis testing. Most of the students stated that applications are made to increase understanding and to make learning interesting.

The result of learning outcomes test can be concluded that the average learning outcomes of students before and after the use of mobile learning is different. And learning outcomes after utilizing mobile learning applications is higher than before using mobile learning

REFERENCES

1. Achmad Buchori dan Herry Agus Susanto (2012). Pengembangan Media Mobile Learning Berbasis Software Classpad Casio Pada Mata Kuliah Geometri Datar Di Perguruan Tinggi. *Edumatica* ISSN 2088-2157 Volume 02 Nomor 01 April 2012.
2. Aditya Sri N, Munir, dan Heri Sutarno (2011). *Pengembangan dan Implementasi Mobile Learning Berbasis J2ME Untuk Mata Pelajaran Keterampilan Komputer dan Pengelolaan Informasi*. Universitas Pendidikan Indonesia. Bandung.
3. Al-Zoubi1 A, Sabina J. and Olivier, P. (2010). *Mobile Learning in Engineering Education: The Jordan Example*. June 9th-11th, New York, NY, USA. The International Conference on E-Learning in the Workplace 2010. www.icelw.org.
4. A.Z. Fanani dan Arry Maulana S. (2010). *Mudah Membuat Mobile Application dengan FlashLite 3.0*. Penerbit Andi Yogyakarta.
5. Didin Wahyudin. (2011). *Aplikasi MECCA Sebagai Model Pembelajaran Pengenalan Komponen Elektronika Dengan Smartphone Android*. Universitas Pendidikan Indonesia. Bandung.
6. Gu, X and Laffey. (2011). *Designing a mobile system for lifelong learning on the move*. Blackwell Publishing Ltd Journal of Computer Assisted Learning, 27, 204-215.
7. Hussein M.O.M. and Cronje, J.C. (2010). Defining Mobile Learning in the Higher Education Landscape. *Educational Technology & Society*. 13(3), 12-21.Cape Peninsula University of Technology South Africa.
8. Kalinić, Z. and dan Arsovski, S. (2009). Mobile Learning – Quality Standards, Requirements and Constrains. Scientific Review Paper (1.02). *International Journal for Quality research UDK*- 378:004 Vol.3, No. 1.
9. Keskin. N.O. and Metcalf, D. (2011). *The Current Perspectives, Theories and Practices of Mobile Learning*. *TOJET: The Turkish Online Journal of Educational Technology* – April 2011, Volume 10 Issue 2.
10. Lukita Yuniati (2011). *Pengembangan Media Pembelajaran Mobile Learning Efek Doppler Sebagai Alat Bantu Dalam Pembelajaran Fisika Yang Menyenangkan*. JP2F Volume 2 Nomor 2 September 2011.
11. Priyanto H, Aldi Daswanto dan Sulistyo P. (2011). *Asyik Membuat Mobile Game Edukatif Dengan Flash.* Penerbit Informatika Bandung.

12. Sarrab, M., Elgamel, L. and Aldabbas, H. (2012). Mobile Learning (M-Learning) and Educational Environments. *International Journal of Distributed and Parallel Systems (IJDPS)* Vol.3, No.4.
13. Wahana Komputer (2012). *Beragam Desain Game edukasi Dengan Adobe Flash CS5*. Penerbit Andi Yogyakarta.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 109-118

COMPARISON OF STOCHASTIC SOYBEAN YIELD RESPONSE FUNCTIONS TO PHOSPHORUS FERTILIZER

Mohammad Masjkur
Department of Statistics, Faculty of Mathematics and Natural Sciences
Bogor Agricultural University, Bogor, Indonesia
Email: masjkur@gmail.com

ABSTRACT

Stochastic yield response function to fertilizer was known better than deterministic version to determine optimum doses of regional fertilizer recommendation. However, the selection of functional forms suitable for certain cropping condition was also critical. This study was intended to know the best model of stochastic soybean yield response function to phosphorus fertilizer. The research was conducted based on multi location experimental data of soybean yield response to phosphorus fertilizer. The fixed parameter models (M1) of Linear plateau, Spillman-Mitscherlich, Quadratic and Logistic were compared with the random parameter models containing either 1 (M2) or 2 random effects (M3) using -2 log-likelihood, Akaike information criterion, and Bayesian information criterion. Results showed that the AIC values of M1 fixed parameter models sequentially were Linear plateau <Spillmann-Mitscherlich = Logistic < Quadratic. Meanwhile, the AIC values of M2 random parameter models sequentially were Linear plateau < Logistic < Spillmann-Mitscherlich < Quadratic. The AIC values of M3 random parameter models sequentially were Spillmann-Mitscherlich< Logistic < Linear plateau. The best model for soybean yield response function to phosphorus fertilizer was the stochastic Spillmann-Mitscherlich model with location intercept and potential maximum yield random effects.

KEYWORDS

Response functions; fixed effect; random effects; regional fertilizer; recommendation; soybean.

1. INTRODUCTION

Building phosphorus fertilizer recommendation for soybean was based on generalized curved of fertilizer response for each soil test classes using quadratic regression with ordinary least squares method with assumption that the residuals were normally, independently, and have constant variances. The parameters of model were assumed to have a fixed value. However, the assumption that the parameters of model have a fixed value was not realistic for multilocation fertilizer trials data. The fixed parameter model approach ignore the variability that probably exist between location and the correlation between observation (Wallach, 1995; Makowski et al., 2002; Makowski and Lavielle, 2006).

The alternative model was to estimate parameter of fertilizer response model using mixed model approach of random parameter model. The mixed model approach was possible to consider the random effects that represent the variability between location, the heterogenous variance, and the correlation that probably exist between observation. Some studies showed that the random parameter model approach was better than the fixed parameter model approach for determining optimum doses of fertilizer recommendation (Makowski et al., 2001; Makowski et al., 2002; Tumusiime et al., 2011; Boyer et al., 2013).

Furthermore, the quadratic function commonly used for fitting fertilizer data was not always the best model. Tumusiime et al. (2011) and Park et al. (2012) showed that the stochastic linear plateau and Mitscherlich's exponential type functions were better than the quadratic function. Boyer et al. (2013) revealed that the stochastic linear plateau function was better than the stochastic quadratic plateau function for corn response to nitrogen fertilizer.

This study was intended to know the best model of stochastic soybean yield response function to phosphorus fertilizer.

2. METHODOLOGY

2.1 Data

The study using data of multilocation trials of phosphorus fertilizer on soybean in Java and Sumatera (Nursyamsi and Sutriadi, 2004; Nursyamsi et al., 2004). Each trial consists of five levels of phosphorus fertilizer treatment. The phosphorus fertilizer levels applied were as 0, 20, 40, 80 and 160 kg P/ha of SP36. The response measured was soybean grain dry weight (ton/ha). The experiment using a randomized complete block design with three replications. The soybean grain yield responses obtained with different phosphorus fertilizer treatments in fourteen experiments was shown in Figure 1.

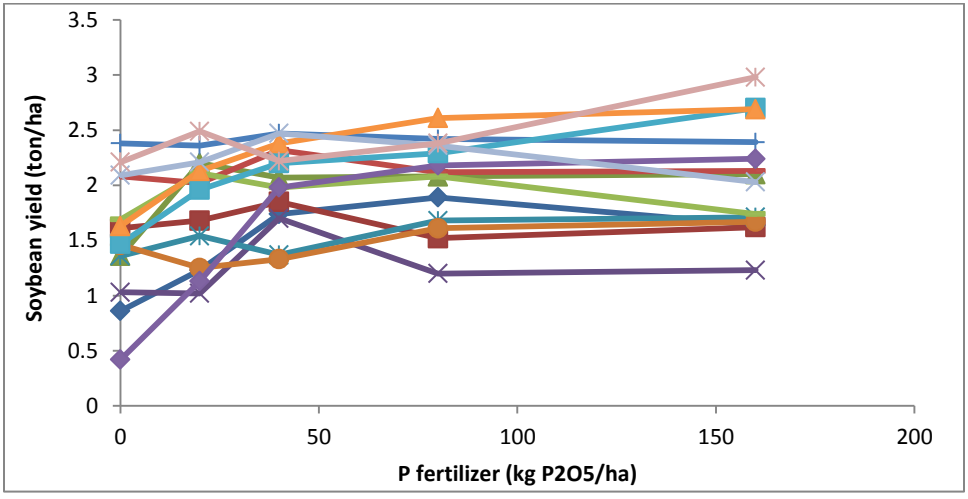


Fig. 1: The Soybean Yield Responses to Applied Phosphorus for Fourteen Locations

2.2 Methods

2.2.1 The Stochastic Response Model

The stochastic response model can be expressed as

$$Y_{ij} = f\left(\boldsymbol{\beta}_{\mathbf{ij},}, \mathbf{u}_{\mathbf{ij}}\right) + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma_{\varepsilon}^2) \quad [1]$$

where Y_{ij} is the i^{th} observation (soybean yield) on the j^{th} location (i=1, 2, ….., n_j); j = 1, 2, …,L); L is the total number of locations and n_j is the number of observations on the j^{th} location; f is the linear or nonlinear function relating soybean yield to phosphorus fertilizer and other possible covariates $\boldsymbol{u_{ij}}$ varying with location; $\boldsymbol{\beta}_{\mathbf{ij}}$ is a vector with the parameters of the linear or nonlinear function; ε_{ij} is the residual term; and σ_{ε}^2 is the variance for the residuals.

The $\boldsymbol{\beta}_{\mathbf{ij}}$ vector may be modeled in a second stage as the sum of 2 components: a fixed (population) component, $\boldsymbol{\beta}$, common to all location, and a random component, $\mathbf{b}$, specific to each location. Therefore:

$$\boldsymbol{\beta}_{\mathbf{ij}} = \mathbf{A}_{\mathbf{ij}}\boldsymbol{\beta} + \mathbf{B}_{\mathbf{ij}}\mathbf{b}_{\mathbf{j}}, \mathbf{b}_{\mathbf{j}} \sim N(0, \sigma_u^2), \quad [2]$$

where $\mathbf{A}_{\mathbf{ij}}$ and $\mathbf{B}_{\mathbf{ij}}$ are design matrices for the fixed and random effects, respectively; $\boldsymbol{\beta}$ is a p-dimensional vector of fixed population parameters; $\mathbf{b}_{\mathbf{j}}$ is a q-dimensional random effects vector associated with location (not varying with i), and σ_u^2 is the variance of the random effects. It is assumed that observations made on different location are independent, and ε_{ij} is independent of $\mathbf{b}_{\mathbf{j}}$ (Lindstrom and Bates, 1990).

The stochastic linear plateau response model was as follows,

$$Y_{ij} = \min\left(\alpha_1 + (\alpha_2 + u_{j3})X_{ij};\ \mu_p + u_{j2}\right) + u_{j1} + \varepsilon_{ij} \quad [3]$$

where Y_{ij} is the soybean yield in i^{th} plot and j^{th} location; X_{ij} is the phosphorus fertilizer level; α_1 is the intercept parameter; α_2 is the linear response coefficient; u_p is the plateau yield; u_{j1} is the (intercept) location random effects; u_{j2} is the plateau location random effects; u_{j3} is the slope random effects; and ε_{ij} is the random error term.

The stochastic Spillman-Mitscherlich response model was as follows,

$$Y_{ij} = \beta_1 - \left(\beta_2 + u_{j2}\right)\exp\left(-\beta_3 + u_{j3}\right)X_{ij} + u_{j1} + \varepsilon_{ij} \quad [4]$$

where β_1 is the maximum or potential yield attainable by applying phosphorus fertilizer in experimental condition; β_2 is the increase in yield by applying phosphorus fertilizer; β_3 is the ratio of successive increment in output β_1 to total output Y; u_{j1}; u_{j2}; u_{j3} are the random effects; and ε_{ij} is the random error term.

The stochastic quadratic response model was as follows,

$$Y_{ij} = \gamma_1 + (\gamma_2 + u_{j2})X_i + (\gamma_3 + u_{j3})X_i^2 + u_{j1} + \varepsilon_{ij} \quad [5]$$

where γ_1 is the intercept parameter whose position (values) could shift up or shift down from location to location by location intercept random effect u_{j1}; γ_2 is the linear response

coefficient with the random effect u_{j2}; γ_3 is the quadratic response coefficient with the random effect u_{j3}; and ε_{ij} is the random error term.

The stochastic logistic response model was as follows,

$$Y_{ij} = (\delta_3 + u_{j3})/[1 + \exp(\delta_1 - (\delta_2 + u_{j2})X_i) + u_{j1} + \varepsilon_{ij} \quad [6]$$

where δ_3 is the maximum yield, δ_1 is the intercept parameter; δ_2 is the response coefficient by applying phosphorus fertilizer; u_{j1}; u_{j2}; u_{j3} are the random effects; and ε_{ij} is the random error term (Tembo et al., 2003; Tumusiime et al., 2011; Brorsen, 2013).

If the model was non-stochastic, then the random effects u_{j1}, u_{j2} and u_{j3} would be zero. In the stochastic models the random effects were entered sequentially. The first was the model with one random effect u_{j1} (M2) (Park et al., 2012) and then the model with two random effects u_{j1} and u_{j2} (M3) and the model with three random effects u_{j1}, u_{j2} and u_{j3} (M4). However, the models with three random effects u_{j1}, u_{j2} and u_{j3} (M4) were not convergent. The random parameters u_{j1} and u_{j2} are assumed to have a mean of zero with a 2 x 2 unstructured covariance matrix.

2.2.2 Statistical Analysis

The response model was estimated using nonlinear mixed model procedure. Under normal assumption the selection of the best model using criteria of -2log-likelihood (-2LL), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and residual variance (σ_ε^2). The smaller values of -2LL, AIC, BIC and residual variance indicate a better model fitting to the data.

3. RESULTS AND DISCUSSION

3.1 Model parameter estimate

The parameter estimate of linear plateau model with fixed effect and mixed effects are presented in Table 1. In M2, the between location variability was modeled by varying location intercept coefficient. In M3, the between location variability was modeled by varying location intercept and plateau parameter.

There are differences of parameter estimate of intercept, the linear response coefficient and the plateau mean between the fixed model M1 with the random parameter model M2 and M3, although the all three parameter were significant, except for intercept parameter of the random parameter model M3. The SE for residual variance was lowest in M3 following by M2, with the M1 being the greatest. In M2, residual variance was separated into within location variation (σ_ε^2) and between location variation (σ_{u1}^2). The residual variance was reduced by about 75.63 percent in M2 compared with M1 indicating an improvement in the accuracy of model parameter estimation. The log-likelihood, AIC and BIC value of M2 smaller than M1 suggesting a better fit of M2 to the data.

Table 1
Parameter Estimate of Linear Plateau Model

Parameter	Fixed Model (M1)	Mixed Model (M2)	Mixed Model (M3)
	Estimate (SE)	Estimate (SE)	Estimate (SE)
α_1	1.56** (0.11)	1.61** (0.09)	0.10 (0.12)
α_2	0.01** (0.004)	0.01* (0.005)	0.02** (0.003)
u_p	2.05** (0.08)	2.10** (0.07)	0.62** (0.13)
σ_ε^2	0.002**(0.0003)	0.0007**(0.0001)	0.0003**(0.00006)
σ_{u1}^2	-	0.002**(0.0005)	0.01**(0.0017)
σ_{u2}^2	-	-	0.002**(0.0004)
σ_{u1u2}	-	-	$3\ 10^{-7}$
-2 LL	86.7	45.0	63.8
AIC	94.7	55.0	75.8
BIC	103.7	58.2	79.7

* significant at 5% level; ** significant at 1% level; σ_ε^2=residual variance; $\sigma_{u1}^2, \sigma_{u2}^2$= the individual variance within population; σ_{u1u2}= individual covariance between random effects;

M1 = no random effect, M2= 1 random effect, M3= 2 random effects,
SE= standard error

In M3, the between location variation was further divided into variance due to varying intercept parameter (σ_{u1}^2), variance due to varying the plateau mean (σ_{u2}^2), and the covariance among them (σ_{u1u2}). The residual variance of M3 was reduced by about 89.92 and 58.62 percent compared with M1 and M2, respectively. From M3, it appears that the large proportion of between location variation accounted for by location intercept variation (σ_{u1}^2) than the plateau variation (σ_{u2}^2), although the plateau variation also significant at 5 percent level. The covariance between the two random effects relatively small. Based on the model fitting criteria, the stochastic linear plateau M3 was the best model compare with the stochastic linear plateau M2 and the deterministic model M1.

The Spillman-Mitscherlich model parameter estimate with the fixed effect and the mixed effects are given in Table 2. In M2, the between location variability was modeled by varying location intercept. In M3, the between location variability was modeled by varying location intercept and potential maximum yield parameter.

Table 2
Parameter Estimate of Spillman-Mitscherlich Model

Parameter	Fixed model (M1)	Mixed model (M2)	Mixed model (M3)
	Estimate (SE)	Estimate (SE)	Estimate (SE)
β_1	2.06** (0.10)	2.04** (0.10)	2.15** (0.07)
β_2	0.52** (0.15)	0.52** (0.09)	0.46** (0.12)
β_3	0.04 (0.03)	0.04* (0.02)	0.035** (0.007)
σ_ε^2	0.002** (0.0003)	0.0007**(0.0001)	0.0003**(0.00008)
σ_{u1}^2	-	0.0016 (0.001)	0.0012**(0.0002)
σ_{u2}^2	-	-	0.0034*(0.001)
σ_{u1u2}	-	-	$2\ 10^{-8}$
-2 LL	86.8	46.1	29.6
AIC	94.8	56.1	41.6
BIC	103.8	59.3	45.4

* significant at 5% level, ** significant at 1% level

M1 = no random effect, M2= 1 random effect, M3= 2 random effects

The parameter estimate of β_2 and $\beta_3$forthe fixed parameter model M1 and the random parameter model M2 and M3 were similar, but the β_3 parameter of M1 was not significant (P-value=0.0601). The estimate of potential yield parameter β_1 was differ between the three models. The SE for residual variance was lowest in M3 following by M2, with the M1 being the greatest. In M2, the residual variance reduced by about 73.11 percent compared with M1 indicating the accuracy of model parameter estimation increases. The log-likelihood, AIC and BIC values of M2 smaller than M1 suggesting that M2 was better for fitting the data.

The residual variance of M3 reduced by about 78.15 and 18.75 percent compared with M1 and M2 respectively. In M3, the proportion of between location variability much more explained by location intercept variability (σ_{u1}^2) than the maximum yield variability (σ_{u2}^2). The maximum yield variability was not significant. The covariance between the two random effects relatively small. The fitting model criteria indicate that the stochastic Spillman-Mitscherlich model M2 was the best model compared with the stochastic Spillman-Mitscherlich model M3 and the fixed model M1.

The parameter estimate of quadratic model with the fixed effect and the mixed effects are given in Table 3. In M2, the between location variability was modeled by varying the location intercept. In M3, the between location variability was modeled by varying the location intercept and the location quadratic response parameter. However, the M3 model was not convergent or not applicable.

Table 3
Parameter Estimate of Quadratic Model

Parameter	Fixed Model (M1)	Mixed Model (M2)
	Estimate (SE)	Estimate (SE)
γ_1	1.60** (0.10)	1.60** (0.11)
γ_2	0.0099** (0.0036)	0.0099** (0.0022)
γ_3	-0.00004* (0.00002)	-0.00004** (0.00001)
σ_ε^2	0.002**(0.0003)	0.0007**(0.00014)
σ_{u1}^2	-	0.0013*(0.00055)
σ_{u2}^2	-	-
σ_{u1u2}	-	-
-2 LL	87.8	48.9
AIC	95.8	58.9
BIC	104.8	62.1

* significant at 5% level, ** significant at 1% level
M1 = no random effect, M2= 1 random effect, M3= 2 random effects

The parameter estimate of intercept, linear response and quadratic response of the fixed parameter model M1 and the random parameter model M2 were similar indicating that in the quadratic model the expected means of mixed effect model as the same as that of the fixed effect model. The SE for residual variance in the M2 was lower than the M1. In M2, the residual variance reduced by about 75.41 percent compared with M1 indicating an improvement in the accuracy of model parameter estimation. The log-likelihood, AIC and BIC values of M2 was smaller than M1 suggesting a better fit of M2 to the data.

The parameter estimate of logistic model with the fixed effect and the mixed effects were shown in Table 4. In M2 the between location variability was modeled by varying location intercept. In M3, the between location variability was modeled by varying location intercept and location response coefficient.

Table 4
Parameter Estimate of Logistic Model

Parameter	Fixed model (M1)	Mixed model (M2)	Mixed model (M3)
	Estimate (SE)	Estimate (SE)	Estimate (SE)
δ_3	2.06** (0.09)	2.04** (0.099)	2.06** (0.11)
δ_1	-1.087** (0.34)	-1.075** (0.21)	-1.08** (0.21)
δ_2	0.05 (0.03)	0.05* (0.02)	0.05* (0.018)
σ_ε^2	0.002**(0.0003)	0.0007**(0.00014)	0.0007**(0.00014)
σ_{u1}^2	-	0.0016 (0.0010)	0.0012 (0.0008)
σ_{u2}^2	-	-	0.00015 (0.0009)
σ_{u1u2}	-	-	$-6\ 10^{-7}$
-2 LL	86.8	46.0	46.1
AIC	94.8	56.0	58.1
BIC	103.8	59.2	61.9

* significant at 5% level, ** significant at 1% level
M1 = no random effect, M2= 1 random effect, M3= 2 random effects

There are differences of parameter estimate of maximum yield, the intercept coefficient and the fertilizer response between the fixed model M1 and the random parameter model M2 and M3. The SE for residual variance in M3 as same as in M2, with the M1 being the greatest. In M2, residual variance was reduced about 73.11 percent compared with the fixed model M1 indicating the accuracy of parameter model estimation improved. The log-likelihood, AIC and BIC value of mixed model M2 smaller than the fixed model M1 suggesting that the mixed model M2 was better for fitting the data.

The residual variance of the mixed model M3 reduced by about 74.79 and 6.25 percent compared with the fixed model M1 and the mixed model M2 respectively. In the mixed model M3, the large proportion of the between location variability explained by the location intercept variability than the maximum yield variability. The maximum yield variability was not significant. The covariance of the two random effects was small. The fitting model criteria show that the stochastic logistic M2 was the best model compared with the mixed model M3 and the fixed model M1.

3.2 The Model Comparison

The log-likelihood, AIC and BIC values of all the fixed model M1 and the mixed model M2 and M3 were shown in Table 5. In the fixed model M1 the log-likelihood, AIC and BIC values sequentially were linear plateau <Spillmann-Mitscherlich = logistic <quadratic indicating that the linear plateau model was the best fixed model. In the mixed model M2 thelog-likelihood, AIC and BIC values sequentially were linear plateau < logistic < Spillmann-Mitscherlich < quadratic indicating that the linear plateau model was also the best M2 mixed model. However, in the mixed model M3 the log-likelihood, AIC and BIC values sequentially were Spillmann-Mitscherlich < logistic < linear plateau. Therefore the Spillmann-Mitscherlich model was the best M3 mixed model.

Table 5
The Log-likelihood, AIC, and BIC Values of Fixed and Mixed Effects Model

Model	Fixed Model (M1)			Mixed model (M2)			Mixed model (M3)		
	-2LL	AIC	BIC	-2LL	AIC	BIC	-2LL	AIC	BIC
LP	86.7	94.7	103.7	45.0	55.0	58.2	63.8	75.8	79.7
SM	86.8	94.8	103.8	46.1	56.1	59.3	29.6	41.6	45.4
Q	87.8	95.8	104.8	48.9	58.9	62.1	-	-	-
LOG	86.8	94.8	103.8	46.0	56.0	59.2	46.1	58.1	61.9

LP = linear plateau; SM = Spillman-Mitscherlich; Q = quadratic; LOG = logistic

The log-likelihood, AIC, and BIC values of the mixed model M2 and M3 were smaller than the fixed model M1 in all response functions indicating that the mixed model was better for fitting the data than the fixed model version (Table 5). This phenomena were caused by the fact that in the mixed model decomposing of variance-covariance associating with the random effects make it possible to separate the between location variability from the within location variability. However, the mixed model with the larger number of random effects was not always the best model. Based on the fitting model criteria the stochastic Spillmann-Mitscherlich model with location intercept and potential maximum yield random effects (M3) was the best model for yield response function of soybean to phosphorus fertilizer.

4. CONCLUSIONS

The stochastic response model was more accurate than the deterministic version to estimate parameter of fertilizer response model. The best model for soybean yield response function to phosphorus fertilizer was the stochastic Spillmann-Mitscherlich model with location intercept and potential maximum yield random effects.

ACKNOWLEDGMENT

I would like to thank The Bogor Agricultural University for financial support and The Soil Research Institute for data availability.

REFERENCES

1. Boyer, C.N., Larson, J.A., Roberts, R.K., McClure, A.T., Tyler, D.D. and Zhou, V. (2013). Stochastic corn yield response functions to nitrogen for corn after corn, corn after cotton, and corn after soybeans. *Journal of Agricultural and Applied Economics*, 45(4), 669-681.
2. Lindstrom, M.L. and Bates, D.M. (1990). Nonlinear mixed effects models for repeated measure data. *Biometrics*, 46, 673-687.
3. Makowski, D. and Wallach, D. (2002). It pays to base parameter estimation on a realistic description of model errors. *Agronomie*, 22, 179-89.
4. Makowski, D., Wallach, D. and Meynard, J.M. (2001). Statistical methods for predicting the responses to applied N and for calculating optimal N rates. *Agronomy Journal*, 93, 531-539.

5. Makowski, D. and Lavielle, M. (2006). Using SAEM to estimate parameters of response to applied fertilizer. *Journal of Agricultural, Biological, and Environmental Statistics*, 11, 45-60.
6. Nursyamsi, D. and Sutriadi, M.T. (2004). Selection of Phosphorus Extraction Method for Inceptisols, Ultisols and Vertisols for Soybean. *Proceedings of National Seminar on Land Resources*. Center for Research and Development of Soil and Agroclimate. Bogor.
7. Nursyamsi, D., Sutriadi, M.T. and Kurnia, U. (2004). The Extraction Method and Fertilizer Requirement of Phosphorus for Soybean in Typic Kandiudox, Papanrejo Lampung. *Journal of Soils and Climate*, 22, 15-25.
8. Park, S.C., Brorsen, B.W., Stoecker, A.L. and Hattey, J.A. (2012). Forage response to swine effluent: a Cox nonnested test of alternative functional forms using a fast double bootstrap. *Journal of Agricultural and Applied Economics*, 44, 4, 593-606.
9. Tembo, G., Brorsen, B.W. and Epplin, F.M. (2003). Linear Response Stochastic Plateau Functions. *Selected Paper prepared for presentation at the Southern Agricultural Economics Association annual meetings*, Mobile, Alabama.
10. Tumusiime, E., Brorsen, B.W., Mosali, J., Johnson, J., Locke, J. and Biermacher, J.T. (2011). Determining optimal levels of nitrogen fertilizer using random parameter models. *Journal of Agricultural and Applied Economics* 43(4), 541-552.
11. Wallach, D. (1995). Regional optimization of fertilization using a hierarchical linear model. *Biometrics*, 51, 338-346.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 119-130

STATISTICAL ANALYSIS FOR NON-NORMAL AND CORRELATED OUTCOME IN PANEL DATA

Annisa Ghina Nafsi Rusdi[§], Asep Saefuddin and **Anang Kurnia**
Department of Statistics, Bogor Agricultural University, Indonesia
Email: [§]ghina.rusdi@gmail.com

ABSTRACT

There are many cases that cannot fulfill some assumptions in statistical analysis, such as normality and independence. In many practical problems, the normality as well independent assumption is not reasonable. For example, data that repeated over time tend to be correlated. If analysis ignores the non independent outcome the Standard Error (SE) on the parameter estimates tends to be too small. Generalized Linear Model (GLM), Generalized Estimating Equation (GEE), and Generalized Linear Mixed Model (GLMM) can be used for non-normal data using the link functions. GEE includes working correlation matrix to accommodate the correlation in the data. GLMM may overcome the repeated observation and allows individual have different baseline/intercept. The study is aim at comparing result based on those approaches. The data that are used in this study are from BPS and the outcome that are used is poverty proportion. Based on R^2_{MAR} the study shows that GEE approach is better than GLM for marginal model, and GLMM approach is better than GEE with dummy variable.

1. INTRODUCTION

Methods of statistical analysis depend on the measurement scales of the response and explanatory variables. In many practical problems, the normality assumption is not reasonable. In some cases the response variable can be transformed to improve linearity and homogeneity of variance. But this approach has some drawbacks such as response variable has changed (not original) and transformation must simultaneously improve linearity and homogeneity of variance.

Panel analysis combines the time series and cross-sectional data. Models that usually used in panel analysis are Pooled Model, Fixed Effects Model, and Random Effects Model. The outcome that are used in this study is not normal and correlated.

Data that repeated over time tend to be correlated or dependent each other. Individuals tend to be more similar to themselves over time than the other independent individuals. Group of individuals may have dependent outcome. The Standard Error (SE) of the estimate is very small in the case of ignore dependent outcome.

If the assumptions are violated, such as response variable is not normally distributed and correlated, the classic model like Panel Regression may produce missleading conclusion. There are some approaches for this kind of data (non-normal and correlated outcome). Pooled Model is approached by Generalized Linear Model (GLM) and Generalized Estimating Equation (GEE), Fixed Effects Model is approached by

Generalized Estimating Equation with dummy variable, and Random Effects Model is approached by Generalized Linear Mixed Model (GLMM).

GLM can solve the problem of non-normal data distributed. GLM introduced by Nelder and Wedderburn (1972) are standard method used to fit regression model for univariate data that are presumed to follow an exponential family distribution (Horton and Lipsitz 1999). It is possible to fit models of data from normal, inverse Gaussian, gamma, Poisson, binomial, geometric, and negative binomial by suitable choice of the link function g(.).

GEE can solve the problem of non-normal and correlated data. Liang and Zeger (1986) introduced GEE to take into account correlation between observations in GLM. GEE takes into account the dependency of observation by specifying a "working correlation matrix". The working correlation matrix is not usually known and must be estimated.

GLMM is a little bit different than GLM and GEE. It has random effect where you can have different intercept for every observation if you assume that the intercept of each observation is different.

GLM and GEE are marginal model, the parameter estimates the marginal population mean. In contrast, GLMM is a mixed effect modeling approach. Parameters in GLMM have a subject-specific interpretation.

The data that was used in this study are poverty proportion by province (POV) that was obtained from BPS (2008-2012) for all provinces in Indonesia. The explanatory variable that were used are Human Development Index by province (1-100) (HDI), Unemployment Rate by province (%) (UR), Gross Domestic Regional Product per Capita by province (Rp10.000) (GDRPC). The objective of the study is to determine the most appropriate model for non-normal and correlated outcome.

2. METHODS

2.1 Panel Data

Data structure (Dobson 2002):

$$y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix}, \boldsymbol{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{bmatrix},$$

so y has length $\sum_{i=1}^{N} n_i$, where N is the amount of the object and n_i is the observation in the object (n). x_i is a $p\ x\ 1$ vector of explanatory variable,

$$x_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}, \text{ so } x_i' = [x_{i1} \quad \dots \quad x_{ip}]$$

and β is the $p\ x\ 1$ vector of regression parameters $\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$. The vector x_i is the *i*th column of the design matrix **X**

$$\mathbf{X} = \begin{bmatrix} x_1^T \\ \vdots \\ x_N^T \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Np} \end{bmatrix}.$$

Panel regression that are usually used (Gujarati 2004):

- Pooled model: $y_{it} = x_i\beta + \varepsilon_{it}$
- Fixed Effects Model: $y = x_i\beta + \tau_i + \varepsilon_{it}$; where $\tau_i = 0$
- Random Effects Model: $y = x_i\beta + \tau_i + \varepsilon_{it}$; where $\tau_i \sim N(0, \sigma_\tau^2)$.

2.2 GLM

GLM model for independent data are characterized by:

$$g\big(E(Y_i)\big) = g(\mu_i) = x_i'\beta$$

where $\mu_i = E(Y_i)$, g is a link function. Some examples of the link function can be seen in Table 1.

Table 1
Example of link function

Outcome (Y)	Distribution	Link	Function	Var
Continuous	Normal	Identity	$g(\mu_i) = \mu_i$	$V(\mu_i) = 1$
Proportion	Binomial	Logit	$g(\mu_i) = \text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$	$V(\mu_i) = \mu_i(1-\mu_i)$
Count	Poisson	Log	$g(\mu_i) = \log(\mu_i)$	$V(\mu_i) = \mu_i$

Response variable which is assumed to share the same distribution from exponential family. The values of the β coefficients are obtained by maximum likelihood estimation. The maximum likelihood estimator of the $p\ x\ 1$ parameter vector β is obtained by solving estimating equation for β.

$$\sum_{i=1}^{m} \frac{\partial \mu_i'}{\partial \beta} v_i^{-1}\big(y_i - \mu_i(\beta)\big) = 0$$

with variance-covariance matrix (*V*) is

$$V = \begin{bmatrix} V_1 & O & \cdots & O \\ O & V_2 & \ldots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & V_N \end{bmatrix}$$

V_i is vector of variance, and assuming that responses for different subjects are independent, where **O** denotes a matrix of zero, (Dobson 2002).

Hence the GLM model for binomial distribution using logit link function was expressed as the following:

$$log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 HDI + \beta_2 UR + \beta_3 GDRPC.$$

2.3 GEE

GEE is a method of estimation of regression model parameters when dealing with dependent data and extension of Generalized Linear Model (GLM) for longitudinal data analysis using quasi-likelihood estimation. When data are collected on the same units across in time, these repeated observations are dependent over times. If this dependent is not taken into account then the Standard Error (SE) of the parameter estimates will not be valid and hypothesis testing result will be non-replicable.

GEE model for correlated outcome is the same as GLM:

$$g\big(E(Y_i)\big) = g(\mu_i) = x_i'\beta$$

GEE accommodated correlated outcome using Working Correlation Matrix. GEE includes a working correlation matrix in the SE calculation.

Working variance-covariance matrix for y_i equals:

$$\boldsymbol{V_i = \emptyset A_i^{1/2} R_i A_i^{1/2}}$$

Table 2
Common Working Correlation Matrix

Structure	Example	Explanation
Independence	$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$	No correlation between repeated observation
Exchangeable	$\begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$	Correlation between repeated observation is the same accross all individuals
Unstructured	$\begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{22} & 1 & \cdots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{pmatrix}$	Correlation varies with each group of observations
Auto-Regressive	$\begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \rho & \vdots \\ \vdots & \rho^2 & \rho & \ddots & \rho \\ \rho^{n-1} & \ldots & \rho^2 & \rho & 1 \end{pmatrix}$	Correlation is a function of time between measurements

(Horton and Lipsitz 1999)

where A_i is n x n diagonal matrix with elements var(y_{ik}), R_i is n x n "working" correlation matrix for y_i, Ø is a constant to allow for over dispersion.
GEE estimator of β is the solution of:

$$\boldsymbol{U} = \sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \beta} \boldsymbol{V_i^{-1}} \left(Y_i - \mu_i(\beta)\right) = 0$$

GEE model for binomial distribution using logit link function for this study was:

$$log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 HDI + \beta_2 UR + \beta_3 GDRPC.$$

2.4 GLMM

Subject-specific models assume that every region has its own intercept. To avoid correlation away regions, each province is assumed to have its own model. The GLMM of subject-specific models rope with this condition.

GLMM model for binomial distribution using logit link function with random intercept was expressed as the following:

$$g\left(E(Y_i)\right) = g(\mu_i) = x_i'\beta + v_i$$

where v_i is the random effect (one of each subject). These random effect represent the influence of subject i on repeated subjects that is not captured by observed covariates.
The mixed model equations are (Henderson 1984)

$$\begin{bmatrix} \boldsymbol{X'R^{-1}X} & \boldsymbol{X'R^{-1}Z} \\ \boldsymbol{Z'R^{-1}X} & \boldsymbol{Z'R^{-1}Z + G\left(\widehat{\theta}\right)^{-1}} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{v} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'R^{-1}}y \\ \boldsymbol{Z'R^{-1}}y \end{bmatrix}$$

GLMM model for this study is:

$$log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_1 HDI + \beta_2 UR + \beta_3 GDRPC + v_i$$

Random effects assume that $v_i \sim N\ (0, \sigma_v^2)$ and independence of the fixed effects and the error terms (ε).

2.5 Marginal R^2 (R^2_{MAR})

An extension of the R^2 measure is calculating using:

$$R^2_{MAR} = 1 - \frac{\sum_{i=1}^{n} \sum_{t=1}^{n_i} (Y_{it} - \hat{Y}_{it})^2}{\sum_{i=1}^{n} \sum_{t=1}^{n_i} (Y_{it} - \bar{Y})^2}$$

This measure is interpreted as the proportion of variance in the outcome that is explained by the model (Hardin and Hilbe 2003).

3. RESULT AND DISCUSSION

3.1 GLM Approach

The constructions of GLM are by deciding on response and explanatory variables of the data and choosing and appropriate link function and response probability distribution.

Because the response variable in the data is proportion so the probability distribution for the response variable is binomial with log $\left[\frac{\mu_i}{1-\mu_i}\right]$ as the link function.

The parameters estimate and Standard Errors (SEs) for GLM can be seen in Table 3.

Table 3
Parameters Estimate and Standard Errors (GLM Approach)

Parameter	Estimate	Standard Error
Intercept	3.7367	0.0030
HDI	-0.0708	0.0000
UR	-0.0394	0.0000
GDRPC	-0.0025	0.0000

R^2_{MAR} 31.1569%

The model estimation for GLM approach in this study is:

$$log\left[\frac{\mu_i}{1-\mu_i}\right] = 3.7367 - 0.0708\,HDI - 0.0394\,UR - 0.0025\,GDRPC$$

and R^2_{MAR} for GLM approach is 31.1569%. It is interpreted as the 31.1569% proportion of variance in the outcome that is explained by the model.

Scatter plot between prediction values and residual in Appendix 3 shows that residual scale are still large, that are from -0.12 until 0.23. It is bacause the approach that used is not enough to explain the data and it is in accordance with the R^2_{MAR} from GLM approach that is only 31.1569%.

3.2 GEE Approach

The construction of GEE is by deciding the working correlation matrix that we will use. In this study, working correlation matrix that was used is autoregressive (AR(1)). The reason for using AR working correlation matrix is measurement closer in time are likely to be more correlated than measurement further apart in time.
Working correlation matrix (AR(1)) for this approach is:

$$\begin{bmatrix} 1.0000 & 0.9999 & 0.9998 & 0.9997 & 0.9996 \\ 0.9999 & 1.0000 & 0.9999 & 0.9998 & 0.9997 \\ 0.9998 & 0.9999 & 1.0000 & 0.9999 & 0.9998 \\ 0.9997 & 0.9998 & 0.9999 & 1.0000 & 0.9999 \\ 0.9996 & 0.9997 & 0.9998 & 0.9999 & 1.0000 \end{bmatrix}$$

The parameters estimate and Standard Errors (SEs) for GEE approach can be seen in Table 4.

Table 4
Parameters Estimate and Standard Errors (GEE approach)

Parameter	Estimate	Standard Error
Intercept	9.3656	1.3063
HDI	-0.1616	0.0194
UR	0.0055	0.0083
GDRPC	0.0008	0.0009

R^2_{MAR} 37.8220%

The model estimation for GEE approach (using autoregressive working correlation matrix) in this study is:

$$log\left[\frac{\mu_i}{1-\mu_i}\right] = 9.3656 - 0.1616\,HDI - 0.0055\,UR - 0.0008\,GDRPC$$

and R^2_{MAR} for GEE approach is 37.8220%. It is interpreted as the 37.8220% proportion of variance in the outcome that is explained by the model. R^2_{MAR} for GEE is higher than GLM.

Scatter plot between prediction values and residual in Appendix 3 shows that residual scale are still large, that are from -0.14 until 0.16. It is in accordance with the R^2_{MAR} from GEE approach that is 37.8220%. GEE approach has better R^2_{MAR} than GLM even though it is not significant difference.

3.3 GLMM Approach

In the previous approaches, GLM and GEE are marginal model where you estimate model for population. As we can see in Figure 1, the province variance is very high. It can give missleading output if you assume that every province has same mean in poverty proportin (intercept). To solve that problem we can apply GLMM approach so that every province can have different intercept by assuming that intercept is random effect.

The parameters estimate and Standard Errors (SEs) for GLMM approach can be seen in Appendix 1. R^2_{MAR} for GLMM approach is 99.2349%. It is interpreted as the 99.2349% proportion of variance in the outcome that is explained by the model.

As we can see in Appendix 1, it is assumed that the intercept is random effect and different for every province. The variance of the random intercepts on the logit scale is estimated as $\sigma_v^2 = 0.0016$.

The model estimation for GLMM approach in this study is:

$$log\left[\frac{\mu_i}{1-\mu_i}\right] = 6.8584 - 0.1584\,HDI - 0.0053\,UR - 0.0008\,GDRPC + v_i$$

Scatter plot between prediction values and residual in Appendix 3 shows that residual scale is from -0.03 until 0.03. That is much better that the previous approaches. It is in accordance with the R^2_{MAR} from GLMM approach that is 99.2349%.

3.4 GEE with dummy variable approach

GEE approach doesn't have coefficient or variable that shows the subject specifict/effect as in GLMM. It is kind of not suitable to compare between GEE and GLMM. The model that uses GEE approach in this study take subject cepecifit, in this study is provinces, as dummy variable. Dummy variable put into the model in GEE approach. Dummy variable that used for every province can be seen in Appendix 2. This method had been done by Anwar (2012) in his bachelore thesis.

The difference between GLMM and GEE with dummy variable is, in GLMM we assume that there are random effects but in GEE with dummy variable we assume that all variable is fixed.

The working correlation matrix that used is Auto-Regressive. Working correlation matrix (AR(1)) for this approach is:

$$\begin{bmatrix} 1.0000 & 0.2514 & 0.0632 & 0.0159 & 0.0040 \\ 0.2514 & 1.0000 & 0.2514 & 0.0632 & 0.0159 \\ 0.0632 & 0.2514 & 1.0000 & 0.2514 & 0.0632 \\ 0.0159 & 0.0632 & 0.2514 & 1.0000 & 0.2514 \\ 0.0040 & 0.0159 & 0.0632 & 0.2514 & 1.0000 \end{bmatrix}$$

The parameters estimate and Standard Errors (SEs) for GLMM approach can be seen in Appendix 2. R^2_{MAR} for GLMM approach is 99.2227%. It is interpreted as the 99.2227% proportion of variance in the outcome that is explained by the model.

Scatter plot between prediction values and residual in Appendix 3 shows that residual scale is from -0.03 until 0.03. It is similar with GLMM approach. The R^2_{MAR} from GEE approach with dummy variable is 99.2227% and it is also similar with GLMM approach.

3.5 Model Comparison

The SEs of parameter estimate for GLM approach are very small which are 0.000 (under estimate), in contrast GEE, GLMM, and GEE with dummy variable approach have bigger SEs of parameter estimate. That problem occured because GLM ignore the corrrelation between outcomes.

The R^2_{MAR} for every approach in this study are GLM approach is 31.1569%, GEE is 37.8220%, GLMM is 99.2366%, and GEE with dummy variable is 99.2227%. For marginal model, GEE is better than GLM in R^2_{MAR} because GEE can overcomes the correlation problem. For subject-specifict model, GLMM has better R^2_{MAR} than GEE with dummy variable. GLMM approach has the best R^2_{MAR} which is means that it is better in explaine the proportion of variance in the outcome by the model.

Scatter plot between prediction values and residual from GLM and GEE approaches still haven't fulfil the exppected result as in the theory. Hence, it is possible to execute model emendation with another approach such as quasi likelihood. For GLMM and GEE with dummy variable approaches, the scatter plots between prediction values and residual are much better than the previous approaches.

4. COMMENTS AND CONCLUSION

For population mean model, GEE approach is better than GLM to overcome non-normal and correlated outcome in the case of this study which R^2_{MAR} for GEE is 37.8220%. For subject-specific model, GLMM approach is better than GEE with dummy variable to overcomes non-normal and correlated outcome in the case of poverty proportion in this study where the variance of subjects is very big. The R^2_{MAR} for GLMM approach in this study is 99.2366%.

REFERENCES

1. Anwarr, N. (2012). Pemodelan Tingkat Pengangguran di Lima Negara Anggota ASEAN dengan Regresi Data Panel dan *Generalized Estimating Equation.* Bogor Agricultural University.
2. Dobson, A.J. (2001). *An Introduction to Generalized Linear Models*. 2nd ed. New York: CRC Press.
3. Gujarati, D.N. (2004). *Basic Econometrics*. 4nd ed. Florida: Chapman & Hall.
4. Henderson, C.R. (1984). Best Linear Unbiased Prediction of Performance and Breeding Value.
5. Horton, N.J. and Lipsitz, S.R. (1999). Review of Software to Fit Generalized Estimating Equation Regression Models. *The American Statistics*, 53, 160-169.
6. McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear, and Mixed Model*. New York: Wiley.
7. Nelder, J.A., Wedderburn, R.W. (1972). Generalized Linear Models. *Royal Statistics Society*, 135(3), 370-384.
8. Zeger, S.L. and Liang, K.Y. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrics*, 73, 13-22.

APPENDIX

Appendix 1: Parameters Estimate and Standard Errors (GLMM Approach)

Parameter		Estimate	Standard Error
Intercept		6.8584	0.0531
Random effects	province 1	-0.2940	0.0024
	province 2	-0.7291	0.0039
	province 3	-0.9828	0.0038
	province 4	-0.8720	0.0053
	province 5	-1.1839	0.0053
	province 6	-1.1333	0.0034
	province 7	-0.5083	0.0031
	province 8	-1.4401	0.0035
	province 9	-0.2370	0.0036
	province 10	-0.4231	0.0026
	province 11	-1.8808	0.0087
	province 12	-0.9600	0.0028
	province 13	-1.6976	0.0022
	province 14	-0.4281	0.0030
	province 15	-0.0561	0.0051
	province 16	-0.6538	0.0023
	province 17	-1.8445	0.0030
	province 18	-0.9199	0.0035
	province 19	-0.5994	0.0030
	province 20	-1.4980	0.0019
	province 21	-1.1841	0.0044
	province 22	-2.0085	0.0019
	province 23	-1.2635	0.0063
	province 24	-0.7809	0.0054
	province 25	-0.3263	0.0027
	province 26	-0.5243	0.0024
	province 27	-1.0000	0.0025
	province 28	-0.7086	0.0022
	province 29	0.0092	0.0030
	province 30	-1.4221	0.0028
	province 31	-0.3158	0.0036
	province 32	-0.8956	0.0026
	province 33	0.0000	0.0531
HDI		-0.1584	0.0003
UR		0.0053	0.0002
GDRPC		0.0008	0.0000
R^2_{MAR}			99.2349%

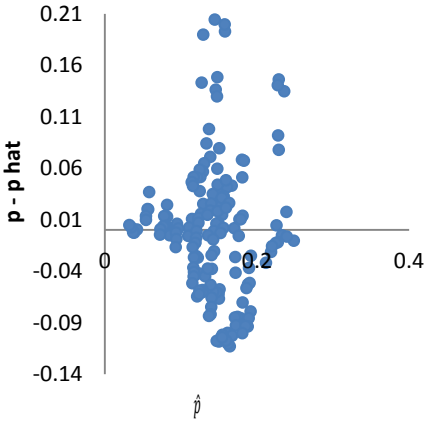
APPENDIX 2

Parameters Estimate and Standard Errors (GEE Approach with Dummy Variable)

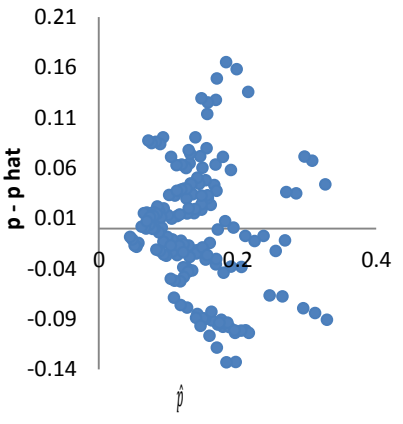
Parameter	Estimate	Standard Error
Intercept	10.0993	0.8494
HDI	-0.1588	0.0122
UR	0.0051	0.0078
GDRPC	0.0008	0.0007
Prov 1	-0.1766	0.0601
Prov 2	-0.4860	0.0777
Prov 3	-0.7605	0.0784
Prov 4	-0.5233	0.0696
Prov 5	-0.8644	0.0699
Prov 6	-0.9563	0.0701
Prov 7	-0.3213	0.0632
Prov 8	-1.2500	0.0497
Prov 9	-0.0515	0.0744
Prov 10	-0.3211	0.0641
Prov 11	-1.4056	0.1684
Prov 12	-0.8188	0.0798
Prov 13	-1.6325	0.0733
Prov 14	-0.2694	0.0720
Prov 15	0.2584	0.1002
Prov 16	-0.5294	0.0378
Prov 17	-1.6913	0.0515
Prov 18	-1.0987	0.0613
Prov 19	-0.6920	0.0712
Prov 20	-1.4869	0.0399
Prov 21	-0.9102	0.0682
Prov 22	-1.9612	0.0316
Prov 23	-0.9072	0.1119
Prov 24	-0.4576	0.1191
Prov 25	-0.2844	0.0711
Prov 26	-0.4282	0.0503
Prov 27	-0.8906	0.0645
Prov 28	-0.6702	0.0533
Prov 29	0.1013	0.0901
Prov 30	-1.4316	0.0640
Prov 31	-0.5136	0.0694
Prov 32	-0.8705	0.0602
R^2_{MAR}		99.2227%

APPENDIX 3

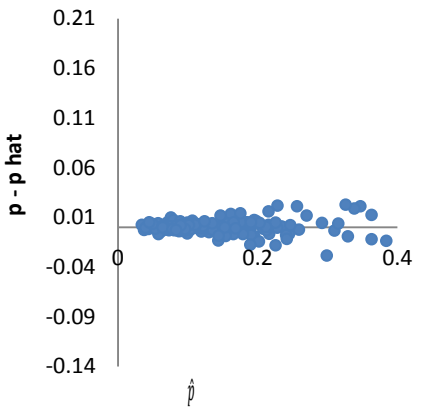
Scatter Plot between Prediction Values and Residual



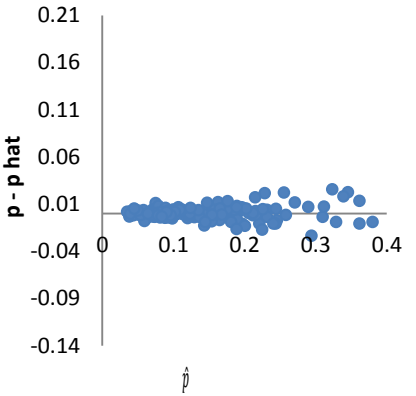
(c) GLM Approach



(b) GEE Approach



(c) GLMM Approach



(d) GEE with Dummy Variable Approach

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 131-138

MULTIDIMENSIONAL DEPRIVATION SPECTRUM: AN ALTERNATE ROUTE TO MEASURE POVERTY

Taseer Salahuddin
University of Central Punjab, Lahore, Pakistan
Email: salahuddin.taseer@gmail.com

ABSTRACT

The construct of poverty has been extensively explored within international literature and practice of development planning especially after its inclusion in Millennium Development Goals as the first and foremost goal. Yet there is a still no agreement on how to define and measure poverty. Two major paradigms currently existing are income based poverty and multidimensional poverty. Both these paradigms however, take individuals as units of analysis and on the basis of poverty lines and cutoffs classify them as poor and non-poor. Social stigma and labeling theory state that this label of poverty impacts negatively on self-esteem of people or develops paternalism in them. This paper suggests that poverty should be measured using dimensions of life as unit of analysis and in turn suggests a new measure of poverty measurement in the form of deprivation spectrum. It will help suggest community poverty alleviation strategies based on the improvement of these dimensions of life instead of individual poverty alleviation. The study offers practical implications for policy makers and governmental agencies working on poverty alleviation programs.

KEYWORDS

Poverty measurement; Multidimensional Deprivation; Poverty Alleviation.

1. INTRODUCTION

Since the advent of the construct of poverty it has been thought as lack of income (Chambers, 2006). This paradigm is based on rationality assumption which argues that as every human is rational, if income is provided to him he can take care of his needs himself. Xavier (2005)however, found contrary evidence from research of human life. Overtime it was realized that poverty is multidimensional in nature (Thorbecke, 2005). Income is a means to ends and not an end in itself therefore for measuring poverty, direct measurement of all dimensions of human life like education, health shelter, living standard, water & sanitation and much more should be done. Many efforts have been done so far as to capture this multidimensionality of poverty. Scope of this study does not include this debate as to which dimensions are to be added or not to poverty measurement. That is a different issue. This study focuses on in depth measurement of community poverty and its analysis.

Income based poverty paradigm is the oldest and still the most commonly understood and used poverty definition around the globe. This paradigm has developed both in

literature and operationalization over time. World Bank uses this paradigm to measure poverty and do international comparisons. The famous 1$ a day and 2$ a day definitions of poverty are based on income-poverty idea. It gave birth to many poverty measures like head-count ratio (Booth, 1889 and Rowntree, 1901), Income-gap ratio (Batchelder, 1971) is, poverty-gap ratio (Sen, 1976), FGT measure (Foster, Greer & Thorbecke, 1984) and many more. It is in use now for more than a century. Over time it has grown to its apex. Nevertheless it is not criticism free. Most prominent criticism faced by income-poverty comes from (Rawls, 1971) social justice theory which rejects income as the only measure of poverty of human life. Income and expenditures are not only subject to human preferences but also to some external factor beyond human control. Income based poverty measures fail to cover these external factors and the non-monetary deprivations (Bougorian, 2002).A practical failure of this aspect of income based measure is seen by recent improvement of developing countries in income poverty accompanied with paradoxical increase in deprivations in many aspects of basic human needs and their redistribution (Citro & Micheal, (1995), Besharov & Germanis, (2004) and Jencks, Mayer & Swingle, (2004a)). Furthermore, income poverty requires direct transfer of money to the beneficiaries, which has challenges of its own. These include true beneficiary selection, decision about amount of money transfer needed to alleviate poverty, training and awareness of beneficiaries in order to effectively use money allocated, monitoring and evaluation of such programs to ensure success. Even if all these conditions are fulfilled diversification of individual deprivations makes it impossible to achieve desirable results without health, education, empowerment and other dimensions provisions. Income paradigm fails to support such allocations. Last but not the least income measures are one-size-fit-all tools which do not cater for the differences in economic, social and other factors impacting incomes and expenditures.

In the light of above shortfalls of income paradigm, a need to develop alternate definitions and measures of poverty was felt. This resulted in the development of multidimensional poverty paradigm. This alternate paradigm was based on Sen's Capability Approach (Sen, 1975, 1977), which argues that deprivations should be measured on the basis of capabilities or endowments available to a person and not on the basis of choices he/she made. There have been practical advances by few researchers in this area. One of these measures Alkire & Foster (2007) had practical success by developing a multidimensional index of poverty capturing all these dimensions in real sense. It has been used by UNDP human development report (2011) to access human deprivation globally. There is a lot of room for improvement by multidimensional measures. These measures show flexibility to adjust to individual country needs, giving a much clearer picture of reality. Yet, there are limitations these measures face. Some of the major challenges faced by them include decision about dimensions to be included, operationalization of these dimensions to be added, complexity of measurement models, generalization of framework developed (Bougorian 2002).

Though multidimensional poverty measures when overcome above mentioned limitations, does help to depict a zoomed-in picture of poverty (Alkire& Seth,2007), still like income measures, these multidimensional poverty measures also focus on individuals as unit of analysis. It appears as if the main goal of all these measures is to classify humans into categories of poor and non-poor. According to social stigma theory and

labeling theory this label of 'poor' hurts self-esteem of some people and develops paternalism in others (Bos, A.E.R., Pryor, J.B., Reeder, G.D. & Stutterheim, S.E. (in press). Identifying people as 'poor' is equivalent to remedial measures at individual level. It will still require selection of beneficiaries and then meeting their requirements in terms of dimension deprivations. Even if they give an idea about overall deprivation in every dimension, it is an indirect way of measuring community deprivations.

Basic argument of the current study is that if a country knows its level of deprivation in different dimensions like education, health, living standard etc. and there is direct provision for attaining such endowments, deprivations will automatically eradicate at individual levels. For example if poverty is high in education dimension, limiting capabilities of the population, government should increase educational facilities and budget. This will be helpful in two ways. First it will save the cost in terms of money and time to trace beneficiaries and secondly it will help create facilities that will develop capabilities of the whole community.

Multidimensional deprivation spectrum is an extended form of Alkire-Foster Multidimensional poverty index. Just like its parent index it is a dual cut-off model where initially cut-offs are applied within dimensions to decide if an individual will be considered deprived or not. On second stage an overall cut-off is applied to see how many dimensions an individual has to be deprived in before he will be considered poor or overall deprived. There are two major differences between spectrum deprivation analysis and Alkire-Foster measure. The step forward in this analysis is that instead of just applying single cut-off or poverty line a spectrum of cut-offs are applied at both stages, resulting in not-deprived, borderline case, slightly deprived, moderately deprived, highly deprived and absolutely deprived cases depending upon the predefined deprivation cut-offs at both the cut-off levels. Secondly, it does not focus to pinpoint individuals as poor or non-poor rather it focuses on total number of deprivations in each spectrum band in each dimension so that policy makers can know the exact extent and cause of deprivations, on the basis of area, locality, gender, ethnic or any other class and then design policy targeted measures of deprivation alleviation accordingly. Just like Alkire-Foster Methodology, an aggregation measure M_0 is calculated. It is a product of head count ratio H and average deprivation share A. However, here, M is calculated twice, once at dimension level and then an average M is calculated at overall deprivation level.

2. METHODOLOGY

Mathematically it is a matrix approach where rows represent the individuals and columns represent dimensions. Let n represent the number of persons and N be the number of dimensions under consideration. Let y = [yij] denote the n * N matrix of achievements, where the typical entry yij> 0 is the achievement of individual i = 1,2,…, n in dimension j = 1,2,…, d. Each row vector yi lists individual i's achievements, while each column vector yj gives the distribution of dimension j achievements across the set of individuals. In what follows we assume that d is fixed and given, while n is the sample size of the survey being used. Let $z_{0\text{-}5} > 0$ denote the cutoffs below which a person is considered to be deprived at various spectrum levels in dimension j, and let z be the matrix of dimension specific cutoffs.

A methodology M for measuring multidimensional poverty consists of identification method and aggregation method. Former is represented in such a way that ρ (yj; z) = 5 if there is absolute deprivation in a particular dimension, ρ (yj; z) = 4 if there is high deprivation in a particular dimension, ρ (yj; z) = 3 if there is moderate deprivation in a particular dimension, ρ (yj; z) = 2 if there is slight deprivation in a particular dimension, ρ (yj; z) = 1 if it isa borderline case and ρ (yj; z) = 0 if there is no deprivation at all. Applying ρ to each dimension achievement vector in y yields the matrix Z {1,…, n} of dimensions deprivations at various levels in y given $z_{0\text{-}5}$. The aggregation step then takes ρs as given for each level of spectrum and associates them with the matrix y and the cutoff matrix z for an overall level M(y; z) of multidimensional deprivation in each dimension. The resulting functional relationship M (Adjusted Headcount) is called an index, or measure of multidimensional deprivation.

$$Md = M_0 + M_1 + M_2 + M_3 + M_4 + M_5$$

or

$$Md = (H_0 * A_0) + (H_1 * A_1) + (H_2 * A_2) + (H_3 * A_3) + (H_4 * A_4) + (H_5 * A_5)$$

for each dimension

Then Overall Adjusted Headcount M is calculated as an average of dimension wise adjusted headcounts,

$$M = (Md1 + Md2 + Md3 + \cdots + MdN)/N$$

where subscripts 0, 1, 2, 3, 4 and 5 represent not deprived, borderline case, slightly deprived, moderately deprived, highly deprived and absolutely deprived part of the spectrum respectively. H = frequency of deprivation at each spectrum level/ total sample size (n), and A= level of deprivation/ maximum deprivation level.

3. STEPWISE APPROACH

Step 1: Choose Dimensions. For example: Health, education, living standard etc.

Step 2: Choose Indicators. For example: immunization, birthplace for health or reading, writing, arithmetic skills and highest class level for education etc.

Step 3: Set Poverty Lines as per spectrum in two stages (for each indicator first apply AD, HD, MD, SD, BLC & ND and then do the scoring for each dimension). The first cut-offs step 1.

Table 1
Example for application of spectrum of cut-offs at dimensions level (Step 1)

	Health			Education				Living Standard		
	Immunization	Birth place	Score	Read	Write	Math	Score	House	Electricity	Score
Person 1	5	1	6	1	5	0	6	5	0	5
Person 2	4	1	5	2	4	1	7	4	0	4
Person 3	3	1	4	3	3	2	8	3	0	3
Person 4	2	2	4	1	5	3	9	2	1	3
Person 5	1	1	2	2	4	4	10	1	2	3
Person 6	1	0	1	3	4	2	9	1	3	4

Step 4: In second step of identification Apply Poverty Lines. AD, HD, MD, SD, BLC & ND with score range predefined for each value from 5 to 0 for the score of each dimension. For example: In health it is

Score Range	Cut-off	Spectrum Band
0-1	0	ND
2	1	BLC
3	2	SD
4	3	MD
5	4	HD
6	5	AD

Similarly for each dimension scores should be translated to spectrum bands.

Table 2
Example for application of spectrum band cutoffs at dimensions level (Step 2)

	Health		Education		Living Standard	
	Score	Spectrum Cut-off	Score	Spectrum Cut-off	Score	Spectrum Cut-off
Person 1	6	5	6	3	5	4
Person 2	5	4	7	4	4	3
Person 3	4	3	8	4	3	2
Person 4	4	3	9	5	3	2
Person 5	2	1	10	5	3	2
Person 6	1	0	9	5	4	3

Step 6: Count the Number of Deprivations in each spectrum band.

This means that for health Counts per spectrum bands are

Count	Spectrum Band
1	ND
1	BLC
0	SD
2	MD
1	HD
1	AD

Presenting that out of 6 people one is not deprived, one is borderline case, two are moderately deprived, one is highly deprived and one is absolutely deprived.

Step 7: Calculate the Headcounts, H_0 -H_5 for each dimension. Divide the number of deprivations in each band with total sample size n. For example here for health it is

Headcount Ratio	Spectrum Band
1/6 = 0.1667	ND
1/6=0.1667	BLC
0/6=0	SD
2/6=0.3333	MD
1/6=0.1667	HD
1/6=0.1667	AD

Table 3
Example for application of spectrum of cut-offs at overall level (Step 1)

	Health	Education	Living Standard
	Spectrum Cut-off	Spectrum Cut-off	Spectrum Cut-off
Person 1	5	3	4
Person 2	4	4	3
Person 3	3	4	2
Person 4	3	5	2
Person 5	1	5	2
Person 6	0	5	3
Scoring	16	26	16

Step 8: Calculate total score the deprivations in each dimension.

Step 9: Apply second cutoffs and redefine these score brackets into deprivation spectrum AD, HD, MD, SD, BLC & ND.

Score Range	Cut-off	Spectrum Band
0-1	0	ND
2	1	BLC
3	2	SD
4	3	MD
5	4	HD
6	5	AD

Step 10: Calculate the Average Deprivation Gap (A) for each band. A is the average number of deprivations suffered in each spectrum band. It is calculated by adding up proportions of total deprivation score in a dimension.

Average Deprivation Gap	Spectrum Band
0/30=0	ND
1/30=0.033	BLC
0/30=0	SD
6/30=0.2	MD
4/30=0.133	HD
5/30=0.083	AD

Here if all six persons were absolutely deprived then the score would be 30. So, maximum score in each band (obtained by multiplying band cutoff with number of deprived people in the band) when divided by 30 results in Average deprivation gap of that band of spectrum. This means that a person deprived in no dimensions at all is represented in not deprived band must have 0 % weight in analysis. This automatically removes from each dimension data of all non-deprived people. Whenever any person's deprivation band changes, it changes the average deprivation gaps of both the bands depicting the change.

Step 11: Calculate the Adjusted Headcounts for each dimension,

$$M_d = M_0 + M_1 + M_2 + M_3 + M_4 + M_5$$

which is calculated as a sum of the product of band-wise adjusted headcounts,

$$M = (H_0\ A_0) + (H_1\ A_1) + (H_2\ A_2) + (H_3\ A_3) + (H_4\ A_4) + (H_5\ A_5)$$

In above mentioned example for the dimension of health it would be as follows

$$M_h = (0.1667*0) + (0.1667*\ 0.033) + (0*0) + (0.3333*\ 0.2) + (0.1667*\ 0.133) + (0.1667*0.08)$$

$$M_h = 0 + 0.0055 + 0 + 0.066 + 0.022 + 0.0136$$

$$M_h = 0.107$$

This means that deprivation level in the dimension of health for this population is 10.7%. With 1.36% absolute deprivation, 2.2% high deprivation, 6.6% moderate deprivation and 0.5% borderline cases.

Step 12: Calculate the Overall Adjusted Headcount M as an average of dimension wise adjusted headcounts,

$$M = = \frac{Md1+Md2+Md3+\cdots+MdN}{N}$$

4. PROPERTIES OF MULTIDIMENSIONAL DEPRIVATION INDEX

Multidimensional poverty spectrum is a useful tool of analysis of poverty in more than one way worth sharing:

1. It is sensitive to any change in spectrum band for individuals, therefore portrays a realistic picture of poverty.
2. Like Alkire-Foster measure it also adjusts for the group size, resulting in international comparisons across various-sized countries.
3. It is a step-forward towards a closer look at poverty picture in terms of breakdown into different deprivation levels.
4. It can also depict a break-down of poverty in terms of regions, gender, ethnic group, rural urban and other classifications.
5. Its properties of breakdown at different classes into various deprivation bands will be a great help to policy makers and governments of countries for designing goal-targeted policies for poverty alleviation.

5. CONCLUSION

This study has argued that instead of classifying humans into poor and non-poor, poverty should be measured on the basis of its dimensions. This argument has been operationalized by developing a multidimensional deprivation spectrum. It is a step forward to Alkire-Foster multidimensional poverty index. Where methodology is similar being a dual cutoff method, builds a triple stage deprivation spectrums using dimensions as basic building block. It not only shows the level of deprivation within a dimension but also shows the degree of intensity and its spread across population. Further research into its application and international comparisons is underway.

6. BIBLIOGRAPHY

1. Alkire, S. and Seth, S. (2007). *Measuring multidimensional poverty in India- a new proposal' OPHI* working paper 15. Oxford University: Oxford Poverty & Human Development Initiative.
2. Alkire, S. and Foster, J. (2007). *Counting and Multidimensional Poverty Measures, OPHI* Working Paper 7 Oxford University: Oxford Poverty & Human Development Initiative.71
3. Batchelder, A.B. (1971). The Economics of Poverty.www.ebay.com/ctg/Economics-Poverty...Batchelder-1971...-/4594866
4. Besharov, Douglas J. and Peter Germanis. (2004). *Reconsidering the Federal Poverty Measure*" Project Description, University of Maryland.
5. Bos, A.E.R., Pryor, J.B., Reeder, G.D. and Stutterheim, S.E. (in press). Stigma: Advances in theory and research. Basic and Applied Social Psychology.
6. Chambers, R. (2006). What is poverty? Who asks? Who answers? Institute of Development Studies, Sussex, UK, International Poverty Centre Poverty in Focus. http://www.ipc-undp.org/pub/IPCPovertyInFocus9.pdf
7. Citro, Constance F. and Robert T. Michael. (1995). *Measuring Poverty: A New Approach*, eds. Washington, D.C.: National Academy Press.
8. Foster, J., Greer, J. and Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica,* 52, 761-765
   hdr.undp.org/en/media/Multidimensional_poverty_index.pdf
   http://booth.lse.ac.uk/static/a/4.html
9. Jencks, Christopher, Susan E. Mayer, and Joseph Swingle. (2004a). *Can We Fix the Federal Poverty Measure so it Provides Reliable Information about Changes in Children's Living Conditions*?" Working Paper. Harvard University.
10. Rawls, J. (1999). 'The theory of Justice' Harvard University Press. Edition b. Bourgorain, F. 2002. 'From income to endowments: a difficult task of expanding income poverty paradigm'. Paper prepared for the conference on "Conceptual challenges in poverty and inequality analysis", Cornell University, 16-17 April 2002. citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.201.8699 ….
11. Thorbecke, E. (2005). Multidimensional poverty: conceptual and measurement issues' paper prepared for The Many Dimensions of Poverty International Conference, UNDP International Poverty Centre, Brasilia, August 29-31, 2005
12. www.envoy.uk.net/jrf/poverty.html
13. www.worldbank.org/en/topic/poverty

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 139-144

AN EVALUATION OF THE PERFORMANCE OF LOCAL POLYNOMIAL ESTIMATES IN GENERALIZED POISSON REGRESSION MODEL

Erni Tri Astuti
Department of Applied Statistics, Sekolah Tinggi Ilmu Statistik
Jakarta, Indonesia
Email: erni@stis.ac.id

ABSTRACT

Local Polynomial estimates are known has advantages over other smoothing techniques in nonparametric regression modeling. It is mathematically intuitive and simple and also has the capability of making inference for the estimator of regression function by constructing a confidence interval. The performance of local polynomial estimates depends on the bandwidth and degree of polynomial used in the model. In this research we studied the performance of local polynomial estimates in Generalized Poisson regression Model with count response. Some simulation study are conducted, and to evaluate the performance of the estimator we used averaged square error (ASE) which is the average of squared difference between estimated regression curve and the true regression curve. Based on the simulation study, we found that that this estimator performed quite good and we need not used degree of polynomial larger than 2 to represent complex regression curve.

KEYWORDS

Local polynomial, nonparametric regression, count response, smoothing technique, Generalized Poisson distribution, averaged square error.

1. INTRODUCTION

In many cases, the relationship between response and covariates cannot describe by simply fitting some parametric function such as linear, exponential or polynomial function. In such case, nonparametric regression seems to be a reliable and reasonable choice. The aim of nonparametric regression is to minimize the assumption about regression function and let the data seeking for the function itself [Hardle, 1990]. In nonparametric regression, scatter plot smoothing is the simplest method to estimate regression function. There are several approaches for determined the regression function, such as kernel, spine and local polynomial technique. These techniques known as local fitting methods because the estimation of regression function is done locally around some interval of points.

Unlike it counterpart in parametric regression model, the development of nonparametric regression for count response with local fitting is moving slowly. There is not much research in this area, except [5], [6]. Local likelihood is a concept introduced by [7] and developed more intensively by [8]. This method extends the nonparametric

regression analysis to maximum likelihood based regression model which also known as likelihood-based smoother. In this model, the mean of response variables are assumed depends on covariates with some nonlinear link function. Although, there are no presumed function for the regression curve itself.

In this research we develop a nonparametric regression model for count response using local polynomial approach for the estimation of regression function. The count response is assumed to have generalized Poisson distribution. We called the estimator as local likelihood estimator because it is determined by local maximum likelihood method. Based on Taylor development of degree *p* and considering the generalized Poisson regression locally, in a neighborhood of some points of interest of the covariate. We present some simulation result to evaluate the performance of the estimator show the behavior of the local likelihood estimator as well as the confidence band of the regression function.

2. LOCAL POLYNOMIAL ESTIMATES

The major concern in nonparametric regression is to minimized the assumption of regression function and let the data found the form of the function itself (Hardle, 1990). If *n* data points $(\boldsymbol{x_i}, \boldsymbol{y_i}), \boldsymbol{i = 1, 2, \cdots, n}$, form an independent and identically distributed sample from a population $\boldsymbol{f(y|x)}$, with *y* is reponse variable and *x* as covariate. Of interest is to estimate the regression function $\boldsymbol{s(x) = E(Y|x)}$ given as

$$\boldsymbol{y_i = s(x_i) + \varepsilon_i} \qquad \textbf{(1)}$$

$s(\cdot)$ is unknown regression function that has a $(p+1)^{th}$ continuous derivative at the point x_0. The simplest method for determining the estimates of $s(\cdot)$ is by scatter plot smoothing. One of the smoothing technique that has been widely used is local polynomial technique. According to Fan and Gijbels (1996) local polynomial technique is applying polynomial regression locally in some neighborhood are resulting a separate line in a window around each x_0 value. The value of the estimated line at $x_0$is the estimates of the regression function at x_0. The size of the data in the local neighborhood is called bandwidth which plays as smoothing parameter in the model. The local polynomial technique has the advantages comparing to other smoothing techniques, not only that it is mathematically intuitive and simple but also the capability of making inference for the estimator of regression function by constructing a confidence interval.

Suppose that $(p+1)^{\text{th}}$ derivative of $s(x)$ at the point x_0 exist. We then approximate the unknown regression function $s(x)$ locally by a polynomial of order *p*. A Taylor expansion gives, for *x* in a neighborhood of x_0,

$$s(x) \approx s(x_0) + s^{(1)}(x_0)(x - x_0) + \cdots + \frac{s^{(p)}(x_0)(x-x_0)^p}{p!} \qquad (2)$$

If we set $\beta_j = \frac{s^{(j)}(x_0)}{j!}$, then we can rewrite equation (2) into,

$$s(x) \approx \sum_{j=0}^{p} \beta_j (x - x_0)^j \qquad (3)$$

This polynomial can be fitted locally by a weighted least square regression problem: minimize

$$\sum_{i=1}^{n}\{y_i - \sum_{j=0}^{p}\beta_j(x - x_0)^j\}^2 K_h(x_i - x_0)$$

where h is a bandwidth controlling the size of the local neighborhood and $K_h(\cdot) = K(\cdot/h)/h$ with K a kernel function assigning weights to each datum point.

There are several issues concerning this technique. First is the choice of the bandwidth parameter h which plays a crucial role. The model complexity is effectively controlled by the bandwidth h. As h increases from 0 to $+\infty$, the model runs from the most complex model (interpolation) to the simplest model (polynomial model). Fan and Gijbels (1997) stated that a too large bandwidth under parameterizes the regression function causing a large modeling bias, while too small bandwidth will over parameterizes the unknown faction and result in noisy estimates. Ideal or optimal model is lying between the two models, which can be obtained by different criteria. One of many methods for finding suitable bandwidth is cross validation technique. Another issue is the choice of the order of polynomial p. Since modeling bias is primarily controlled by the bandwidth, this issue is less crucial however. Unlike traditional polynomial parametric regression, local polynomial model just requires a small degree of polynomial for a complex pattern of data. In the case when p=0, the technique is called local constant and when p=1, it is known as local linear model etc. The choice of the kernel function is not a crucial issues, because the result is almost similar for any kind of kernel function including Epachnecnikov, Gaussian or Boxcar Kernels. In this study we investigate the behavior of the estimated regression function under different bandwidth and order of polynomial in some simulation with generated data set.

3. LOCAL GENERALIZED POISSON REGRESSION MODEL

In 1987, Tibshirani and Hastie extend the idea of local fitting in nonparametric regression analysis to the class of generalized linear model, also known as maximum likelihood-based regression models. They called this approach as local likelihood technique. Fan et al. (1998) present a framework for assessing the bias and the variance of the local likelihood estimator as well as a bandwidth selection procedure. In likelihood based smoother instead of using weighted least square method, the estimated regression function fit locally by maximized the weighted likelihood function,

$$L_{p,h}(\boldsymbol{\beta}, x_0) = \sum_{i=1}^{n} \ln\{f(y_i|x_i)\}\, K_h(x_i - x_0) \tag{4}$$

where $\mathbf{ln}\{\boldsymbol{f(y_i|x_i)}\}$ is a contribution of each datum point to local log-likelihood function. Finding an estimates of regression function $\boldsymbol{s(x)}$ is equivalent to search for maximum likelihood estimates of $\boldsymbol{\beta} = (\boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \cdots, \boldsymbol{\beta_p})'$ locally in each neighborhood of $\boldsymbol{x_0}$.

In case of count response regression modeling, Santos and Neves (2008) adopted this method to perform local Poisson regression, which is assumed Poisson distribution for count response and estimated the regression function using local polynomial technique with p=1 (local liner). In 2013, Astutiet all developed this model to local generalized Poisson regression to avoid over dispersion problem. In local generalized Poisson regression model, the count response y_i is assuming follows the generalized Poisson distribution, with the probability density function given by

$$\boldsymbol{f(y;\mu,\varphi)=\left(\frac{\mu}{1+\varphi\mu}\right)^{y}\frac{(1+\varphi y)^{y-1}}{y!}exp\left[-\frac{\mu(1+\varphi y)}{1+\varphi\mu}\right],y=0,1,\cdots}$$

with $E(Y)=\mu$ and $V(Y)=\mu(1+\varphi\mu)^2$. The parameter φ plays as dispersion parameter. When $\varphi=0$, it will reduce to Poisson probability density. When $\varphi<0$ this model is underdispersed, and when $\varphi>0$ it will overdispersed relative to Poisson distribution respectively (Famoye, 2000).

In local generalized Poisson model, $\boldsymbol{\mu(x)=E(Y|x)}$ are assumed depends on some covariates through a link function or regression function**,**

$$\boldsymbol{\mu_i(x)=\exp[s(x_i)]} \quad (5)$$

The weighted likelihood function in (4) then given as,

$$L_{p,h}(\boldsymbol{\beta},\varphi,x_0)=\sum_{i=1}^{n}\left\{y_i ln\left(\frac{\mu_i(x)}{1+\varphi\mu_i(x)}\right)+(y_i-1)\ln(1+\varphi y_i)-(1+\varphi y_i)\frac{\mu_i(x)}{1+\varphi\mu_i(x)}\right\}K_h(x_i-x_0) \quad (6)$$

If (3) can be stated in vector operation as,

$s(x)=\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$**,** with $\mathbf{x}_i=(1,(x_i-x_0),\cdots,(x_i-x_0)^p)^T$

Then the weighted maximum likelihood estimator for regression function, is the solution of (p+2) equation:

$$\frac{\partial L}{\partial\boldsymbol{\beta}}=\sum_{i=1}^{n}\frac{(y_i-\exp[\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}])}{(1+\varphi\exp[\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}])^2}\mathbf{x}_i K_h(x_i-x_0)=\mathbf{0} \quad (7)$$

$$\frac{\partial L}{\partial\varphi}=\sum_{i=1}^{n}\left\{-\frac{y_i\exp[\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}]}{1+\varphi\exp[\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}]}+\frac{y_i(y_i-1)}{1+\varphi y_i}-\frac{\exp[\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}](y_i-\exp[\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}])}{(1+\varphi\exp[\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}])^2}\right\}K_h(x_i-x_0)=0 \quad (8)$$

The solution of the system which is called local polynomial (maximum likelihood) estimator can be solved by iterative procedure such as Newton Raphson Methods.

The log-likelihood function in (6) depends on two quantities, the smoothing parameter (h) and the order of polynomial (p). As we mention above, the criterion for selecting suitable bandwidth parameter is an important step for finding the estimates of regression function. In this study we used maximum likelihood cross validation (MLCV) developed by Chauduri and Dewanji (199). MLCV is a function of bandwidth parameter h, given by

$$MLCV(h)=\sum_{i=1}^{n}\ln f\left(y_i,\hat{s}_{-i}(x_i)\right) \quad (9)$$

where $\hat{s}_{-i}(x_i)$ is the estimates of regression function in point $x_i$without involving the point (x_i,y_i) itself. This technique is included in some methods known as leave-one-out cross validation based on prediction error. The optimal bandwidth is the ones that gives the greatest value of MLCV.

4. SIMULATION STUDY

The objectives of this paper is to evaluate the performance of local polynomial estimator in generalized Poisson regression model. The estimator of the regression function according to log-likelihood function in (6) depends on two quantities, the smoothing parameter (h) and the order of polynomial (p). In this simulation study, we will evaluate how bandwidth or polynomial order impact the estimates of regression function. Suppose that 2 true regression function in (5) is given by

$$\mu_1(x) = \exp(2 - x^2) \tag{10}$$

$$\mu_2(x) = \exp\left(4sin^2(2x)\right) \tag{11}$$

with $s_1(x) = 2 - x^2$ and $s_2(x) = 4sin^2(2x)$ respectively. We generated 200 pseudo data $(x_1, y_1), \cdots, (x_{200}, y_{200})$. The value of covariate $\{x_i\}$ are from U(-2,2) for the first regression function and U(-1,1) for the second, while the value of response variables $\{y_i\}$ generated form generalized Poisson with mean (10) and (11) and dispersion parameter $\varphi = 0.1$.

5. RESULT

We generated 100 pseudo data (*x*1, *y*1),..., (*x*100, *y*100) from *yi*= μ(*xi*) + FIGURE 1 Best global fits (solid curves) with the true function (dotted curve) for the parametric (a) linear *m*1 and (b) cubic model *m*2.{εi} are independent with a uniform distribution on (−3, 3) and *N*(0, 0.22), respectively. We applied to this data set the local likelihood method described in Section 2. Figure 2 shows how the average squared error

ASE*(h)* = 1
100
_100
i=1
[*mˆh(xi) −μ(xi)*]2

for each model changes as *h* varies, where *mˆh* is the local likelihood estimator of μ. The ASE(*h*) is a natural approximation to

[*mˆh(x) −μ(x)*]2 d*F(x),*

where *F* is the covariate distribution.

The bandwidths which minimize ASE(*h*) are found to be 0.19 and 1.3 for the models *m*1 and *m*2, respectively. This is a typical situation which one may encounter with other data sets too: the case where the true regression function is remote from the parametric family leads to a choice of small *h* to overcome the model misspecification, while a large *h* is preferred in the case where the true function is near the parametric family. The two panels (a) and (b) in Figure 3 depict the local likelihood estimates where the optimal bandwidths were used.

The central question to be examined in this paper is how the degree of model misspecification, when it changes, alters the way in which bandwidth selection determines the performance.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 145-156

DETERMINANTS OF MATERNAL DELIVERY-CARE SERVICES: STUDY FROM RURAL AREA OF BANGLADESH

Md. Shah Jahan[1], **Ahm Shamsuzzoha**[2] and **Hasina Akhter Chowdhury**[3]
[1] Dept. of Public Health, Daffodil International University
Dhaka, Bangladesh. Email: drshahjahan@daffodil.bd
[2] Dept. of Production, University of Vaasa
PO Box 700, FI-65101, Finland. Email: ahsh@uwasa.fi
[3] Dept. of Biostatistics, Bangladesh University of Health Sciences
Dhaka, Bangladesh. Email: hachowdhury01@gmail.com

ABSTRACT

The maternal delivery-care services in Bangladesh are not optimally available and are often underutilized; especially in the rural areas, where high maternal mortality due to lack of access to professional delivery care. In this research study, an attempt is made to investigate the determinants associated with the use of maternal delivery-care services in a selected rural area of Bangladesh. In total 360 postnatal mothers who visited the Expanded Programme on Immunization (EPI) centers were randomly selected with the objective to obtain explicit information on various socio-demographics and practices on delivery-care services. Among the respondents it is estimated that about 76% were literate, 22% were from the poorest socioeconomic class, 62.5% were utilized ante natal care (ANC) service, 80.3% were delivered their babies at home and 66.4% were felt that home delivery was comfortable. A considerable percentage of maternal deliveries (50.6%) were attended by traditional birth attendants. The results from this study reveals that the determinants like educational level of the mothers, living children, access to mass media, wealth index, antenatal care use and number of ANC visit had a positive significant ($p<0.001$) effect on the place of maternal delivery-care. The results from binary logistic regression analyses also showed that education of the mothers ($p<0.05$), living children ($p<0.002$), access to mass media ($p<0.02$), wealth index ($p<0.05$), antenatal care use ($p<0.001$) and number of ANC visit ($p<0.001$) are the important correlates of maternal delivery-care. It is suggested from this research study that the policy makers and health planners need to target, plan, and improve utilization of maternal delivery care services through a combined effort of increasing awareness to avail skilled care and made the services more accessible with better quality to the rural women of Bangladesh.

KEYWORDS

Determinants; maternal delivery-care; health services, rural area, Bangladesh.

1. INTRODUCTION

Several studies were conducted both in developed and developing countries where there is growing concerns to the importance of need to birth delivery assistance from a

trained and well-equipped provider to reduce maternal mortality of babies (Tuncbilek. E, 1988; Boerma JT et al., 1992; Panis CWA et al., 1995). It is also noticed based on ecological and historical evidences that in European and Asian countries, health professions with midwifery skill were considered the key factors to decrease maternal mortality [Worku, AG, et al., 2013]. The Netherlands, Norway and Sweden are considered as the low mortality rates in the 20th century which was the result of an extensive collaboration between physicians and skilled midwives (Högberg U, 2004). Similarly, Malaysia, Thailand and Sri Lanka reduced their maternal mortality rates 50% within ten years by increasing the number of midwives in the 1950s and 1960s (Ronsmans C, Graham WJ, 2006).

We know that customer satisfaction is probably one of the most frequently measured marketing constructs. Safety of mother and new born even after the pregnancy period depends highly on delivery care and place of delivery during pregnancy. Insufficient care during delivery is largely responsible for maternal and infant deaths that occur just before or during delivery.

Although most pregnancies end with the birth of a live baby, on many occasions, childbirth is a time of pain, fear, suffering, and even death. Almost all maternal deaths occur in developing countries (Hill K et al., 2007), where the majority of women deliver at home without skilled birth attendant. These mothers are at increased risk from unpredictable obstetric complications, often ending in death either at home or after transfer to a health facility (Ronsmans C, Graham WJ, 2006; Thaddeus S, Maine D, 1994). Pregnancy and childbirth related complications are among the leading causes of maternal mortality in Bangladesh (BDHS2011). The health and family planning program of Bangladesh has made remarkable progress in the last two decades as evident from the decline in rates of maternal mortality-the death of women during pregnancy, childbirth, or in the 42 days after delivery, but the maternal mortality ratio (MMR) remain very high (194 maternal deaths per 100,000 live births) (BMMS, 2010).

The majority of maternal deaths and disabilities occur suddenly and unpredictably between the third trimester and the first week after the end of pregnancy due to hemorrhage, sepsis, and obstructed or prolonged labor [WHO, 2005]. Women and their families face socioeconomic and cultural barriers to seeking professional delivery care, such as high costs, long distances to health facilities, lack of knowledge about danger signs during pregnancy, and a tradition of using untrained local practitioners during delivery (Syed U, Khadka N, Khan A, Wall S, 2008).

Maternal mortality is one of the most important health challenges the world is facing today. More than 20 million women experience ill health as a result of pregnancy each year. The risk of a woman dying as a result of complication related to pregnancy in developing countries can be as much as 100 times that of women in Western Europe or North America (Estimation of Maternal mortality 2000).

In Bangladesh, high maternal mortality and morbidity rates are underpinned by the fact that 85 per cent of women give birth at home, most with unskilled attendants or relatives assisting. Bangladesh Maternal Mortality and Health Care Survey 2010 reveal that almost 2.4 million births take place at home annually, especially in rural areas. However, only 27% of all births in Bangladesh are assisted by skilled professionals and

only 23% of births take place in a health facility (Bangladesh Maternal Morality and Health Care Survey).

Although improvements in maternal health care have been priority in recent years, maternal morbidity and mortality still remains a major public health issue in most developing countries including Bangladesh. This paper seeks this gap through study and identifies the causal factors that lead to the improved use of healthcare services for women in Bangladesh. The rest of the paper is organized as follows: section 2 presents methodology as taken during this study, whereas section 3 demonstrates the outcomes from this study. Section 4 highlights the findings from this research in the form of discussion. Several managerial implications are provided in section 5, while section 6 concludes this paper.

2. MATERIALS AND METHODS

2.1 Study Design and Sampling

A community-based cross-sectional study, which was approved by the ethical review committee of the Bangladesh Medical Research Council (BMRC), was conducted in the Madhupur Upazila (a mid-level administrative unit) of Tangail district in Bangladesh between January and June 2012. This Upazila is located in 140 km from northwest of Dhaka with a population of 308,846 (Population and Housing Census 2011). Most of them (80 percent) live in rural area and economically dependent on agriculture. We collected information from the randomly selected postnatal mothers focusing on antenatal care (ANC) and utilization of delivery care of their recent deliveries. To draw our sample, first we prepared a sampling frame of postnatal mothers who visited different centers of the Expanded Program on Immunization (EPI) in the Madhupur Upazila. It should be noted that there are various EPI centers in this Upazila and each center maintains an EPI register to record all postnatal mothers during their visits. Briefly, all the EPI registers of the Madhupur upazila constituted the sampling frame. The sample size was then determined based on the information (particularly using the prevalence of postnatal mothers who visited healthcare facilities for PNC) of the Bangladesh Demographic and Health (BDHS) 2007. According to this report, about 30% of the postnatal mothers visited healthcare facilities for PNC services. After adjusting a non-response rate of 10%, the sample size became 360. From the sampling frame, we then randomly selected the required number of postnatal mothers for interviews.

2.2 Questionnaire and Data

A pre-tested structured questionnaire was used to collect information with special focus on maternal socio-economic factors and practices on delivery care. Only some of these factors are analyzed in this study, which are grouped (with categories in parentheses) as follows:

2.2.1 Maternal Socio-economic Factors

Maternal age (< 20, 21-25, 26-30, 30+ years), age at marriage (< 20, 21-25, 25 + years), maternal education in years (0/no formal education, 1-5/primary, 6-10/ secondary, 11-12/higher secondary+), maternal occupation (housewife, service / agriculture / else), total number of living children (1, 2-3, 4+), reading newspapers

(yes, no), listening to radio (yes, no), watching TV (yes, no), having mobile phone (yes, no), wealth index (poorest/poor, middle, rich/richest).

2.2.2 ANC and Delivery Factors

Receiving antenatal care (yes, no), number of ANC visit (0, 1, 2-3, 4+), place of delivery (home, hospital), decision maker in family (husband, mother/father in law, others), conduction of delivery (trained birth attendant, traditional birth attendant, neighbor/relatives), reason for home delivery (comfortable, family decision, financial problem, hospital is far, other).

2.2.3 Delivery Care

Delivery care was defined as a care that has taken place during delivery at health center or a hospital or in home facilitated by skilled birth attendants.

2.2.4 Place of Delivery

Place of delivery means where delivery of a baby takes place. In this study place of delivery was the main dependent variable while characteristics that determine the place of delivery service namely age, education, occupation, total number of living children, access to mass media (reading newspapers, watching TV and listening to radio), wealth index, receiving antenatal care and number of ANC visit were the independent variables.

2.3 Data Analyses

Data were analyzed using the SPSS software for Windows (version 17). Analysis of data on postnatal mother was done taking into consideration socioeconomic status and delivery care. Descriptive summary statistics, such as mean and standard deviation (SD), were computed for continuous variables and proportions for categorical characteristics of the mothers. The significance of the differences in patterns among values of the associated factors was tested using chi-square test at a 5% level of significance. Odds ratios with 95% confidence interval (CI) were calculated using the logistic regression model to control confounders and identify the factors associated with the practice of delivery care.

2.4 Limitations of the Study

There were several limitations during this research study. First of all, it is noticed that women faced difficulties to differentiate the type of skilled providers during interview session. To mitigate such problem, data collectors gave further clarifications to the interviewers with respect to the types of providers such as doctor, health officer, nurse, midwife, health extension worker or others. There were other errors like knowing the categories of different health professionals. However, during data collection the interviewers did not face serious difficulties in categorizing the service providers as skilled or non-skilled and also to identify various determinants to maternal delivery care. The potential limitations as experienced during the data collection process did not affect the interpretation of study results.

3. RESULTS

Out of 360 respondents, 145(40.3%) were between 21-25 years of age with mean age 24 (SD=±4.4) years. Most (95.6%) women were housewives and 272(75.5%) were literate. Regarding age of the marriage, highest percentage of the respondent's (81.9%) was up to 20 years. Twenty-two percent of the respondents were from the poorest

socioeconomic class, 20% were from the middle class, and 19% were from the richest class. In the case of accessibility to mass media, 83% had never read newspapers, 90% had never listened to radio, 40% had never watched television (TV), and 54.4% had no mobile phones (Table 1).

Among the respondents 62.5% took ante natal care, among them only 14.4% women had received ANC services at a health facility at least once during their last pregnancy. Majority (80.3%) of the delivery take place at home and 66.4% respondents felt it comfortable to deliver at home. In most cases (47.2%) decision were taken by husband. Half (50.6%) of the deliveries were conducted by traditional birth attendants (Table 2).

Regarding bivariate analyses, the socio-demographic characteristics, such as education of the respondent's ($p<0.001$), wealth index ($p<0.053$), total number of living children ($p<0.002$), watching TV ($p<0.026$) and listen to radio ($p<0.05$) were positively associated with place of delivery. On the other hand, significant associations were shown between place of delivery and utilization of antenatal care along with number of ANC visit ($p<0.001$) (Table 3).

Binary logistic regression analysis was conducted taking into consideration place of delivery as a dependent variable. The model suggested that education ($p<0.05$), living children ($p<0.001$), reading newspaper ($p<0.035$), watching TV (0.028), utilization of antenatal care ($p<0.001$) and number of ANC visit ($p<0.001$) were the important correlates of delivery care of the postnatal mothers (Table 4).

Table 1
Socio-Demographic Characteristics of Respondents (n=360)

Variable		**Number**	**Percentage**
Age-group (years)	Up to 20	105	29.2
	21 – 25	145	40.3
	26 – 30	89	24.7
	>30	21	5.8
Mean±SD		**24 ±4.4**	
Occupation of mother	Housewife	344	95.6
	Farming	2	.6
	Service	10	2.8
	Others	4	1.1
Number of total living children	1	151	41.9
	2-3	184	51.1
	4+	25	6.9
Mean±SD		***1.9±1.1***	
Age at marriage	Up to 20	295	81.9
	21 – 25	24	6.7
	>25	40	11.1
Education of mother	No formal education	88	24.5
	Primary	122	33.9
	Secondary	138	38.3
	Higher secondary	12	3.3

Variable		Number	Percentage
Wealth index	Poorest	80	22
	Poor	64	18
	Middle	72	20
	Rich	75	21
	Richest	69	19
Access to mass media			
Reading newspaper	Yes	63	17.5
	No	297	82.5
Listening to radio	Yes	35	10
	No	325	90
Television watching	Yes	217	60
	Never	143	40
Having mobile phone	Yes	164	45.6
	No	196	54.4

Results were expressed as n (%) and mean±SD

Table 2
ANC and Delivery Care of Respondents (n=360)

Variable		Number	Percentage
Receiving antenatal care	Yes	225	62.5
	No	135	37.5
No. of ANC visit	0	135	37.5
	1	52	14.4
	2-3	122	33.9
	4+	51	14.2
Place of delivery	Home	289	80.3
	Hospital	71	19.7
Decision maker in family	Husband	170	47.2
	Mother/Father in law	154	42.8
	Other family members	36	10.0
Conduction of delivery	Trained birth attendant	72	20.0
	Traditional birth attendant	182	50.6
	Neighbor/Relatives	106	29.4
Reason for home delivery	Comfortable	239	66.4
	Family decision	21	5.8
	Financial problem	45	12.5
	Hospital is far	12	3.3
	Other	43	11.9

Results were expressed as n (%)

Table 3
Association between Place of Delivery and other Related Demographic Variables

Variable	Home	Hospital	p value
Age			
Up to 20	105	29.2	
21 – 25	145	40.3	0.200
26 – 30	89	24.7	
>30	21	5.8	
Education			
No formal education	78	10	
Primary	107	15	**0.001**
Secondary	100	38	
Higher secondary	4	8	
Living children			
1	109	42	
2-3	95	20	**0.002**
4+	85	9	
Read newspapers			
Yes	39	24	**0.001**
No	250	47	
Listen to radio			
Yes	26	9	0.348
No	263	62	
Watching TV			
Yes	166	51	**0.026**
No	123	20	
Wealth index			
Poorest	69	11	
Lowest	53	11	
Middle	60	12	**0.053**
Upper middle	60	15	
Richest	47	22	
Utilization of antenatal care			
Yes	165	70	**0.001**
No	124	11	
No. of ANC visit			
0	124	11	
1-3	133	41	**0.001**
4+	32	19	

The χ^2-test was conducted. The level of significance at $p<0.05$

Table 4
Binary Logistic Regression Analysis Considering Place of Delivery as Dependent Variables

Independent variable	β	p value	OR
Age	0.051	0.277	1.052
Education			
No formal education	Reference		
Primary	0.089	0.837	1.093
Secondary	1.087	0.005	2.964
Higher secondary	2.747	0.001	15.600
Living children			
1	Reference		
2-3	0.764	0.005	0.466
4+	2.224	0.032	0.108
Wealth index			
Poorest	Reference		
Poor	0.089	0.837	0.435
Middle	1.087	0.005	0.621
Upper middle	2.747	0.001	0.448
Richest	1.077	0.009	1.936
Read newspapers			
No	Reference		
Yes	0.759	0.035	2.137
Watch TV			
No	Reference		
Yes	0.636	0.028	1.228
Receiving antenatal care			
No	Reference		
Yes	1.411	0.001	0.244
No. of ANC visit			
0	Reference		
1-3	1.246	0.001	3.475
4+	1.901	0.001	6.693
Constant	-0.246	0.825	0.759

The level of significance at $p<0.05$

4. DISCUSSIONS

The majority of the literature on the delivery care in low-income countries focuses on the barriers encountered by women (Gabrysch S, Campbell OM, 2009). However, the role of the family, socio-cultural factors primarily influence decision-making on whether to seek care, rather than affecting whether women reach a facility.

This study was designed to identify the determinants of delivery care in a selected rural area of Bangladesh. The study revealed that literacy rate of the respondents was about 75.5% of whom 122(33.9%) had primary level of education. This is much higher

than that of previous national literacy rate of female which was 48.8% (Ekele BA, Tunau KA, 2007).

The findings of this study showed that among the respondents 62.5% took ANC during pregnancy which is quite similar (89.9%) with previous study done in Dhamrai upazila in Dhaka district of Bangladesh [19]. Utilization of antenatal care services by a considerable proportion of respondents with only 14.4% of women attained at least once antenatal care visits during their last pregnancy. This finding is almost similar to the percentage (16.4%) reported by the Bangladesh Demographic Health Survey [20]. Regarding the place of delivery it was evident that the practice of home deliveries was higher than that of hospital deliveries. In a similar study conducted among the urban women of Nepal showed that planned home deliveries were 58.3%, which was much less than present study (Mekonnen Y, 2003). Majority of the respondents (66.4%) felt that home delivery was comfortable. In a similar study conducted among the urban women of Nepal shows that 25.7% home delivery was conducted due to comfort (Meherunnessa Begum, et al., 2013). The result of the study showed that a considerable percentage of deliveries (50.6%) were attended by traditional birth attendants. Trained birth attendant was involved only in 20.0% cases which is almost same (30.1%) as revealed by another study conducted in a union of Mirsarai, Chittagong (Bangladesh Demographic and Health Survey). Prominent role in decision making was of husband in 47.2%. Similarly, in another study, husband's advice was the dominant feature for home or institutional delivery of 48.8% respondents (Bolam A et al., 1998).

Maternal education is strongly associated with delivery care. It is obviously that higher educated mother are more conscious than illiterate mother and they are more likely to receive delivery care services during delivery. In binary logistic analyses mother's education appeared as important predictor to determine practice of delivery care ($p<0.05$). Mothers with primary education and secondary education had received hospital facilities as delivery care 0.39 and 0.67 times more than mothers having no education. Many previous studies conducted in developing countries have found education of mothers to be among the most important determinants of skilled delivery care utilization (Chandrashekhar T, Hari SJ, Binu VS, Sabitri G, 2006; Parveen AK, Quaiyum MA, Islam A, Ahmed S, 2000). There are a number of explanations that speculate as to why education is a key determinant of skilled delivery care demand. For example education is likely to enhance female autonomy so that mothers develop greater confidence and capabilities to make decision regarding their own health, as well as their children. It is also more likely that educated women demand higher quality service and pay more attention to their health in order to insure better health for themselves. Moreover, educated women are more likely to be aware of difficulties during pregnancy and as a result, they are more likely to use maternal health care services (Chandrashekhar T, Hari SJ, Binu VS, Sabitri G, 2006). Our study showed significant association between numbers of children with place of delivery. Similar results were seen in a study from Nigeria where more women of high parity delivered at home ($p<0.05$) (Beun MH, Wood SK, 2003). In a Bangladeshi study found that TV sets and wealth index appears positive effect on receiving delivery care (Rahman KMM, 2009). Our study also showed similar findings.

In this study utilization of antenatal care and number of ANC visit showed positive association with delivery care. Another study was found that mother who received adequate antenatal care is 3.67 times more likely to receive satisfactory delivery care than who didn't receive any antenatal care (Rahman KMM, 2009).

5. MANAGERIAL IMPLICATIONS

The results of this study have critical effect on maternal healthcare for the developing countries like Bangladesh, where both mothers and newborn babies health's are an important factors for social integrity and prosperous. This important social issue can be improved substantially through ensuring awareness and utilization of skilled maternal service throughout the country. Such awareness among women may influence through getting proper education and knowledge on the importance of skilled maternal care in different ways such as previous exposure to skilled maternal services, community based health educations, through community media or due to their better educational status.

Child birth is a time of transition and social celebration in many societies. Women's progression from birth to child bearing is influenced by economy, religion, kinship system and the complexity of communications and medical technology (Lauderdale J et al., 1999). Women are most in need of skilled care during delivery and immediate postpartum period. Traditional birth attendants, whether trained or untrained can neither predict nor cope with serious complications. Thus, this study was designed to identify the determinants of delivery care in a selected rural area of Bangladesh.

From this extensive study, it is obvious that several determinants such as education, cost, male dominated society, remote area, non-skilled midwives, etc., directly influence the overall maternal health care in Bangladesh. It is identified that home delivery was one of the major determinants of low quality antenatal care instead of trained urban health centers. From this study, it is highly recommended to initiate preventive measures in well ahead for remedial of new born babies during delivery phase. At the same time it is suggested to provide proper training and technologies in rural areas of Bangladesh and promote social awareness to offer safe and sound delivery care in the village level health care centers.

6. CONCLUSIONS

In order to improve the maternal delivery care in any country or region it is critical to identify and assess the potential determinants of antenatal care. Such determinants play a significant role in determining the utilization of skilled antenatal care. For instance, health facility characteristics, including quality and cost of service were important determinants for use of skilled antenatal care than maternity services. Improvement of community health care services, community awareness and perception on skilled care providers by targeting women are very important. Safe and sound motherhood education through the available communication networks in the rural communities can be adopted as the strategies to improve the intended awareness and perceptions of rural women.

In this study, various factors are assessed by health extension and community health workers with the objective to improve antenatal care use. Various critical determinants for maternal delivery care services are identified. The results of this study revealed that

education of the mothers, living children, mass media and member of a NGO are the important determinants of delivery care during pregnancy. Thus the results of the study suggest that antenatal care advice can contribute to increase the practice of delivery care. The result of the study also revealed that a considerable percentage of deliveries (50.6%) were attended by traditional birth attendants. About 47.2% husbands of rural women intervene the decision making process to arrange home delivery for their wives. With such circumstances, it is highly recommended to initiate different aspects of health education program need to be strengthened to improve maternal health care in the rural area of Bangladesh.

In the future study, more districts or regions of rural Bangladesh are planned to be brought under this study in order to bring such study more authenticate and reliable in nature. In addition to this studied determinants other determinants like daily working schedule (life style), information exchange, family size, job pattern, social status, etc., are also can be studied for better understanding the antenatal healthcare for rural Bangladeshi women. These additional determinants will contribute to organize and plan better antenatal care within rural women in Bangladesh.

REFERENCES

1. Boerma J.T. and Bicego, G.T. (1992). Preceding birth intervals and child survival: searching for pathways of influence. *Stud Fam Plann.*, 23 (4), 243-56.
2. Bhatia J.C. and Cleland, J. (1995). Determinants of maternal care in a region of South India. *Health Transit Rev.*, 5(2), 127-42.
3. Bangladesh Maternal Morality and Health Care Survey (BMMS). Summary of key findings and implications. Available from: <http://www.dghs.gov.bd/dmdocuments/BMMS_2010.pdf.
4. Bangladesh Demographic and Health Survey (2011). Preliminary Report, National Institute of Population Research and Training, Dhaka, Bangladesh.
5. Bolam, A., Manandhar, D.S., Shrestha, P., Ellis, M., Malla, K. and Costello, A.M. (1998). Factors affecting home delivery in the Kathmandu valley, Nepal. *Health policy Plan.*, 13, 152-58.
6. Beun, M.H. and Wood, S.K. (2003). Acceptability and use of clean home delivery kits in Nepal: a quantitative study. *J Health Popul Nutri.*, 21, 367-73.
7. Chandrashekhar, T., Hari, S.J., Binu, V.S. and Sabitri, G. (2006). Home delivery and new born care practices among urban women in western Nepal: a questionnaire Survey. *BMC Pregnancy Childbirth.* 6, 27.
8. Ekele, B.A. and Tunau, K.A. (2007). Place of delivery among women who had antenatal care in a teaching hospital. *Acta Obstet Gynecol Scand.*, 86(5), 627-30.
9. WHO (2003). *Estimation of Maternal mortality in 2000*, Annual Report 2003, Geneva.
10. Gabrysch, S. and Campbell, O.M. (2009). Still too far to walk: Literature review of the determinants of delivery service use. *BMC Pregnancy Childbirth*, 9, 34.
11. Högberg, U. (2004). The Decline in Maternal Mortality in Sweden. *Am J Public Health*, 94(8), 1312-20.

12. Hill, K., Thomas, K., Abou Zahr, C., Walker, N., Say, L., Inoue, M. and Suzuki, E. (2007). Estimates of maternal mortality worldwide between 1990 and 2005: An assessment of available data. *Lance*,. 370(9595), 1311-19.
13. Lauderdale, J. (1999). Childbearing and transcultural nursing care issues. In: Andrews M, Boyle J, editors. *Transcultural concepts in nursing care, 3rd ed. Philadelphia: Lippincott.* 81-106.
14. Mekonnen Y. (2003). Pattern of maternity care utilization in southern Ethiopia: evidence from community and family survey. *Ethiopian J Health Dev.*, 17(1), 27-33.
15. Meherunnessa Begum, Khondoker Bulbul Sarwar, Nasreen Akther, Rokshana Sabnom, Asma Begum, Kawser Ahmed Chowdhury (2013). Socio Demographic Determinants of Delivery Practice in Rural Women of Bangladesh. *Delta Med Col J*., 1(2), 42-45.
16. National Institute of Population Research and Training (2011). Bangladesh Demographic and Health Survey 2011.
17. Organization, W.H. (2005). The World Health Report: 2005: Make Every Mother And Child Count.
18. Panis, C.W.A. and Lillard, L.A. (1995). Child mortality in Malaysia: explaining ethnic differences and the recent decline. *Popul Stud (Camb)*, 49(3), 463-79.
19. Population and Housing Census (2011). Bangladesh Bureau of Statistics, Community Report Tangail Zila June 2012, Statistics and Informatics Division, Ministry of Planning.
20. Parveen, A.K., Quaiyum, M.A., Islam, A. and Ahmed, S. (2000). Complications of pregnancy and childbirth: Knowledge and practices of women in Rural Bangladesh, [WP131, 2000]. *ICDDR, B: Centre for Health and Population Research.* 131, 1-22.
21. Rahman, K.M.M. (2009). Determinants of Maternal Health Care Utilization in Bangladesh. *Research Journal of Applied Sciences*, 4(3), 113-19.
22. Ronsmans, C. and Graham, W.J. (2006). Maternal mortality: who, when, where, and why. *Lancet.* 368(9542) 1189-1200.
23. Ronsmans, C. and Graham, W.J. (2006). Maternal mortality: who, when, where, and why. *Lancet*., 368(9542), 1189-1200.
24. Syed, U., Khadka, N., Khan, A. and Wall, S. (2008). Care-seeking practices in South Asia: Using formative research to design program interventions to save newborn lives. *J Perinatol*, 28(2), S9-S13.
25. Statistical Pocket Book of Bangladesh (2010). Sample vital Registration System (SVRS) 2007, 370.
26. Thaddeus, S. and Maine, D. (1994). Too far to walk: maternal mortality in context. *Soc Sci Med*., 38(8), 1091-1110.
27. Tuncbilek, E. (1988). Infant Mortality in Turkey: Basic Factors. Semih of Set Matbaacilik, Ankara.
28. Worku, A.G., Yalew, A.W. and Afework, M.F. (2013). Factors affecting utilization of skilled maternal care in Northwest Ethiopia: a multilevel analysis. International Health and Human Rights. *BMC Int. Health Hum Rights*, 13:20.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 157-176

SPATIAL REGRESSION ANALYSIS TO EVALUATE INDONESIAN PRESIDENTIAL GENERAL ELECTION

Alan Duta Dinasty, Asep Saefuddin and **Yenni Angraini**
Statistics Department, Bogor Agricultural University, Indonesia
Email: alandutadinasty@gmail.com

ABSTRACT

Indonesia has held two presidential elections in 2004 and 2009. The result of these elections shows that number of voter abstention is very high. On the first term of 2004 presidential election, about 21.77% of official voters were abstain. On the second term, it increased to 23.37%. Five years later, the number of voter abstention was still high which increased to 27.19%. This research aims to identify spatial pattern and spatial relationship of 33 provinces that produce voter abstention number in the 2009 presidential elections and to identify voter characteristics. The results of this research show that there is positive spatial autocorrelation for voter abstention data, which means that there are similar proportion of provinces to their neighbors in Indonesia. Provinces that are significant to spatial autocorrelation are Riau, Riau Islands, West Sumatera, Central Kalimantan, and Gorontalo. According to significant provinces, there are not province identified as hotspot or cold spot observation. Determining factors that are significantly affected to voter abstention using Spatial Error Model (SEM) is better than Spatial Autoregressive Model (SAR) and Multiple Linier Regression. There are six significant explanatory variables, i.e. percentage of poor people, monthly average of wage/ salary/ income of employee, mean years of schooling population 15 years of age and over, human development index, school enrollment ratio 16-18 years of age, and life expectancy at birth.

KEYWORDS

Election; k-nearest neighbor; spatial autocorrelation; spatial error model; voter abstention.

1. INTRODUCTION

General election is part of democracy that has been improving and developing in Indonesia. Since 1955, Indonesia has implemented ten general elections for legislative and presidential election. Presidential technical election that is previously chosen by People's Consultative Assembly of the Republic of Indonesia (MPR) has been replaced and agreed as a direct election by all of Indonesian peoples. This new rule election has been implemented for a twice, i.e. in 2004 and 2009.

The policy to change the authority of vote from MPR to peoples has raised some problems. One of the problems is voter abstention, which has been actually happening since the first legislative election in 1955. The result of the last two elections shows that

number of voter abstention is very high. In the first term of 2004 presidential election, 21.77% of official voters were abstain. In the second term, it increased to 23.37%. Five years later, it was still in a high level, which increased to 27.19%. This condition will be feared to be increased for the next election.

According to Arianto (2010), abstention is a behavior that refuse to give a vote as a participation in election based on some factors and reasons. In the survey of social and political science from Eriyanto (2007), there are three reasons why peoples become abstain in election, i.e. administrative reason, individual or technical reason, and political reason. Political reason became the lowest choice than other reasons based on that survey. Moreover, three categories of voter abstention are identified by DeSipio *et al.* (2006). The first is Registered Not Voted, that is citizens who have a vote and actually registered but they do not use their vote. Second is Citizen Not Registered, that is citizens who have a vote but does not registered so that they unable to get a vote. The last is Non Citizen, that is peoples who do not have a vote because they are not citizen on the region which hold an election. This research use those first two categories of abstention produced by provinces in Indonesia to explore some clusters formed by neighboring provinces that is indicated spatial effect. Thus, this research desire to observe the pattern of voter abstention in Indonesia.

A research about voter abstention had been done by Nyarwi (2009) using a correlation. The result explained that voter abstention was influenced by several factors using some approaches, that were model approach of sociology, psychology and rational choice. Mujani (2007) did a survey and research about voter abstention behavior which almost produce a similar influenced categories that was compared with Nyarwi (2009). In an aspect of sociology, factor of income, occupation, education, urbanism, age, religion, and gender have a significant correlation to voter abstention behavior. In an aspect of psychology, significant correlation coming from factor of political information and partisanship. The previous research about voting behavior are regardless to consider spatial effect between neighboring province. According to Anselin (1988) in W Tobler's First Law about Geography, observation in one region is affected by other observation in another region, everything is related each other, but something close has more effect than one far. Determination hotspot provinces and factors that affect significantly to voter abstention is important to be evaluated for next election. Thus, this research aims to identify factors that affect significantly to voter abstention behavior in Indonesia with enhancing factor of spatial in order to be useful for related stakeholders for improving the better general election in Indonesia.

2. METHODOLOGY

The data used are secondary data from General Election Commission of Indonesia (KPU) for an official result of 2009 presidential election and Central Bureau Statistics of Indonesia (BPS) for characteristics of provinces. The dependent variable used is voter abstention of 2009 presidential general election (Y), while the explanatory variables are percentage of poor people (X1), monthly average of wage/salary/income of employee (X2), open unemployment rate (X3), mean years of schooling population 15 years of age and over (X4), human development index (X5), school enrollment ratio 16-18 years of age (X6), "A" number of accredited school (X7), and life expectancy at birth (X8).

Methods

Spatial Regression Model

Spatial regression is one of analysis of statistics that evaluates the relationship between variable and other variables by considering spatial effect. General model of spatial regression is:

$$\boldsymbol{y} = \rho\boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \quad (1)$$
$$\boldsymbol{u} = \lambda\boldsymbol{W}\boldsymbol{u} + \boldsymbol{\varepsilon}\,,\ \boldsymbol{\varepsilon} \sim N(0, \sigma^2\boldsymbol{I})$$

where $\boldsymbol{y}$ is a dependent variable $(n \times 1)$, ρ is a spatial autoregressive model coefficient, $\boldsymbol{W}$ is a spatial weight matrix $(n \times n)$ with zero diagonal of element, $\boldsymbol{X}$ is an independent variables $(n \times (p+1))$, $\boldsymbol{\beta}$ is vector of regression coefficient $((p+1) \times 1)$, $\boldsymbol{u}$ is a vector of regression errors that has assumed to have random effects and a spatial autocorrelation, λ is a spatial error model coefficient and $\boldsymbol{\varepsilon}$ is an error autocorrelation vector $(n \times 1)$ (Anselin 1988).

In the general model of spatial regression, if $\rho \neq 0$ and $\lambda = 0$ then the model is called Spatial Autoregressive Model (SAR). This model is a linier regression with the response variable having spatial autocorrelation. General model of SAR is:

$$\boldsymbol{y} = \rho\boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,,\ \boldsymbol{\varepsilon} \sim N(0, \sigma^2\boldsymbol{I}) \quad (2)$$

The estimator of $\boldsymbol{\beta}$ is:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{I} - \rho\boldsymbol{W})\boldsymbol{y} \quad (3)$$

and the estimator of ρ and σ^2 are obtained by log likelihood function of SAR, those equations are:

$$\hat{\rho} = (\boldsymbol{y}'\boldsymbol{W}'\boldsymbol{W}\boldsymbol{y})^{-1}\boldsymbol{y}'\boldsymbol{W}'\boldsymbol{y} \quad (4)$$

$$\hat{\sigma}^2 = \frac{(\boldsymbol{y} - \rho\boldsymbol{W}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \rho\boldsymbol{W}\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})}{n} \quad (5)$$

(Anselin 1988, referred from Arisanti 2011).

If $\rho = 0$ and $\lambda \neq 0$, the model is called Spatial Error Model (SEM). This model is a linier regression with the error variable having spatial autocorrelation. General model of SEM is:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \quad (6)$$
$$\boldsymbol{u} = \lambda\boldsymbol{W}\boldsymbol{u} + \boldsymbol{\varepsilon}\,,\ \boldsymbol{\varepsilon} \sim N(0, \sigma^2\boldsymbol{I})$$

The estimator of $\boldsymbol{\beta}$ is:

$$\widehat{\boldsymbol{\beta}} = [(\boldsymbol{X} - \lambda\boldsymbol{W}\boldsymbol{X})'(\boldsymbol{X} - \lambda\boldsymbol{W}\boldsymbol{X})]^{-1}(\boldsymbol{X} - \lambda\boldsymbol{W}\boldsymbol{X})'(\boldsymbol{y} - \lambda\boldsymbol{W}\boldsymbol{y}) \quad (7)$$

$$\hat{\sigma}^2 = \frac{[(\boldsymbol{I} - \lambda\boldsymbol{W})(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})]'[(\boldsymbol{I} - \lambda\boldsymbol{W})(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})]}{n} \quad (8)$$

Obtaining parameter of λ which maximizing log likelihood function of SEM needs a numerical iteration (Anselin 1988, referred from Arisanti 2011).

Methodologies of this research are summarized as follows:

1. Exploring the data using descriptive statistics.
2. Predicting and testing parameters that are significant to voter abstention using multiple linier regression analysis.
3. Defining spatial weight matrix.

   Spatial regression analysis needs spatial weight matrix ($\boldsymbol{W}$) to construct the model. It is generated to visualize the proximity between location. Initially, the matrix is filled by c_{ij} that is only value of 1 if the *i* location is adjacent with *j* location and the other is filled by 0. There are several approaches constructing the matrix, one of them is *k*-Nearest Neighbor (*k*-NN) which is used in this research. The steps of defining spatial weight matrix as follows:

   a. Determining the centroid of each province. This research uses centroid of the capital city of each province.

   b. Calculating the distance between centroid using euclidean distance d_{ij}. To get the distance of province i located on coordinate (u_i, v_i) and province j located on coordinate (u_j, v_j), the equation used is :

   $$d_{ij} = \sqrt{(u_i - u_j)^2 - (v_i - v_j)^2} \tag{9}$$

   c. Setting for k, the closest neighbors for each province. $k = 3, k = 4, k = 5$, $k = 4$ with population density, and $k = 5$ with population density. Population Density (PD) is a measurement of population per unit area or unit volume used as an additional requisite of province claimed as adjacent unit. Data of PD is obtained from (BPS 2010) and categorized into four classes which are not dense (less than 50 people per km^2), less dense (51-250 people per km^2), dense enough (251-400 people per km^2) and very dense (greater than 400 people per km^2). The categorization follows the Government Regulation in Lieu of Law number 56 of 1960.

   d. Ranking the centroid distance from each province i to all unit $j \neq i$ as follows $d_{ij(1)} \leq d_{ij(2)} \leq \cdots \leq d_{ij(n-1)}$. Then for each $k = 1, \ldots, n-1$, the set $N_k(i) = \{(j(1), j(2), \ldots j(k))\}$ contains the k closest provinces to i. For each given *k*, the *k*-NN weight matrix ($\boldsymbol{W}$), then has spatial weights of the form :

   $$c_{ij} = \begin{cases} 1 \,;\, j \in N_k(i) \; or \; i \in N_k(j) \\ 0 \,;\, otherwise \end{cases} \tag{10}$$

   e. Constructing weight matrix based on Threshold Distance (TD) as a comparation. The equation is :

   $$c_{ij} = \begin{cases} 1 \,;\, d_{ij} \leq d \\ 0 \,;\, d_{ij} > d \end{cases} \tag{11}$$

   f. Calculating the element of weight matrix w_{ij} becomes row standardized by equation as follows :

   $$w_{ij} = \frac{c_{ij}}{\sum_{j=1}^{n} c_{ij}} \tag{12}$$

4. Testing spatial autocorrelation for voter abstention data using Global Moran Index.

   The hypothesis of Global Moran Index (Moran's I) is:
   H_0 : I = 0 (there is no spatial autocorrelation)
   H_1 : I > 0 (there is positive spatial autocorrelation)
   I < 0 (there is negative spatial autocorrelation)

   The equation of Moran's I is :

$$I = \left[\frac{n}{\sum_i^n \sum_j^n w_{ij}}\right]\left[\frac{\sum_i^n \sum_j^n w_{ij}(y_i-\bar{y})(y_j-\bar{y})}{\sum_i^n (y_i-\bar{y})^2}\right] \quad (13)$$

   with n is number of observation, y_i is value for location *i*, y_j is value for location *j* and $\bar{y}$ is the mean of y_i from n locations. The test of statistics of Moran's I is:

$$z(I) = \frac{I-E(I)}{\sigma(I)}, E(I) \approx -\frac{1}{n-1} \quad (14)$$

   where I is Moran Index, $z(I)$ is the value of test of statistics from Moran Index, $E(I)$ is expected value from Moran Index, $\sigma(I)$ is standard deviation from Moran Index and n is number of area (Ward & Gleditsch 2007).

5. Testing spatial autocorrelation for voter abstention data using Local Moran Index.

   Local Moran is defined as:

$$I_i = \frac{(y_i-\bar{y})}{(\sum_i y_i-\bar{y})^2}\sum_j w_{ij}(y_j - \bar{y}) \quad (15)$$

$$L_i = f(y_i, y_{j_i}) \quad (16)$$

   where y_i is value of observation in location *i*, y_j is value of observation in location *j*, $\bar{y}$ is the mean from attributes of observation, and W_{ij} is value of spatial weighted matrix for region *i* and *j*, f is a function from (y_i, y_{j_i}), y_i is value of observation in province *i*, and y_{j_i} is other value of observation in province *j* in area or cluster *i* (Anselin 1995).

6. Testing spatial dependence effect and spatial homogeneity effect using Lagrange Multiplier and Breusch-Pagan Test.
   a. Lagrange Multiplier Test
      1. Spatial Autoregressive Model (SAR)
         Hypothesis : H0 : $\rho = 0$ and H1 : $\rho \neq 0$

         Test of Statistics :

$$LM_{lag} = \frac{\left[\frac{\boldsymbol{\varepsilon}'\boldsymbol{W}\boldsymbol{y}}{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/N}\right]^2}{D} \quad (17)$$

         with

$$D = \left[\frac{(\boldsymbol{W}\boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')(\boldsymbol{W}\boldsymbol{X}\hat{\boldsymbol{\beta}})}{\hat{\sigma}^2}\right] + tr(\boldsymbol{W}'\boldsymbol{W} + \boldsymbol{W}\boldsymbol{W}) \quad (18)$$

         and $\boldsymbol{\varepsilon}$ is a vector of error $(n \times 1)$ from Ordinary Least Square (OLS), $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ are determined from OLS, and *tr* is an operator (Anselin 1999).

Statistics of LM_{lag} follows chi-square distribution $X^2_{(1)}$. If LM_{lag} greater than $X^2_{(1)}$, the H_0 is rejected.

2. Spatial Error Model (SEM)

   Hypothesis : H0 : $\lambda = 0$ and H1 : $\lambda \neq 0$

   Test of Statistics :

$$LM_{error} = \frac{\left[\frac{\boldsymbol{\varepsilon' W \varepsilon}}{\boldsymbol{\varepsilon' \varepsilon / N}}\right]^2}{tr(\boldsymbol{W'W + WW})} \quad (19)$$

   and $\boldsymbol{\varepsilon}$ is a vector of error $(nx1)$ from Ordinary Least Square (OLS) and *tr* is an operator (Anselin 1999). Statistics of LM_{error} follows chi-square distribution $X^2_{(1)}$. If LM_{error} greater than $X^2_{(1)}$, the H_0 is rejected.

b. Breusch – Pagan Test

   Hypothesis: H0: $\sigma_i^2 = \alpha_i$ and H1: $\sigma_i^2 \neq \alpha_i$

   Test of Statistics:

$$BP = \frac{1}{2}(\sum_{i=1}^{n} x_i f_i)(\sum_{i=1}^{n} x_i x_i')(\sum_{i=1}^{n} x_i f_i) \quad (20)$$

   with $f_i = \left(\frac{\varepsilon_i}{\hat{\sigma}} - 1\right)$, $\varepsilon_i = \left(y_i - \widehat{\boldsymbol{\beta}}' \boldsymbol{x_i}\right)$, and $\hat{\sigma}^2 = \sum_{i-1}^{n} \varepsilon_i^2$. Statistics of BP follows chi-square distribution $X^2_{(k-1)}$. If BP greater than $X^2_{(k-1)}$, the H_0 is rejected (Arbia 2006).

7. Predicting and testing parameters that are significant to voter abstention using spatial regression analysis.
   a. Determining spatial kind of model according to Lagrange Multiplier in step 6.
   b. If $\rho \neq 0$ and $\lambda = 0$ then the model is called Spatial Autoregressive Model (SAR).
   c. If $\rho = 0$ and $\lambda \neq 0$ then the model is called Spatial Error Model (SEM).
   d. Checking and testing for the assumption of spatial regression model.

8. Comparing the goodness of fit of spatial regression model based on value of AIC and Log Likelihood. The lower AIC value is, the better model is likely to choose. The equation to calculate AIC is:

$$AIC = -2l + 2p \quad (21)$$

   where l is log likelihood and p number of parameters. Ward & Gleditsch (2007) identified log likelihood for classical regression, spatial autoregressive model, and spatial error model. Log likelihood l for classical regression (17), SAR (18), and SEM (19) are:

$$\ln L(\beta, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(2\sigma^2) - \frac{(y - \boldsymbol{X}\beta)'(y - \boldsymbol{X}\beta)'}{2\sigma} \quad (22)$$

$$\ln L(\beta, \sigma^2, \rho) = \ln|\boldsymbol{I} - \rho\boldsymbol{W}| - \frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(2\sigma^2) - \frac{(y - \rho\boldsymbol{W} - \boldsymbol{X}\beta)'(y - \rho\boldsymbol{W} - \boldsymbol{X}\beta)'}{2\sigma^2} \quad (23)$$

$$\ln L(\beta,\sigma^2,\lambda) = \ln|\boldsymbol{I}-\lambda\boldsymbol{W}| - \frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(2\sigma^2) - \frac{(y-\lambda\boldsymbol{W}-\boldsymbol{X}\beta+\lambda\boldsymbol{W}\boldsymbol{X}\beta)'(y-\lambda\boldsymbol{W}-\boldsymbol{X}\beta+\lambda\boldsymbol{W}\boldsymbol{X}\beta)'}{2\sigma^2} \quad (24)$$

with β is parameter estimates and $\boldsymbol{I}$ is value of Moran Index.

9. Interpretating significant variables in the best model.
10. Conclusion.

3. RESULT AND DISCUSSION

Data Exploration

According to General Election Commission of Indonesia (KPU), the final voter list of 2009 presidential election is 175.263.774 peoples. About 27.19% or 47.648.099 of them did not give a vote. This percentage is obtained by comparing number of voter abstention and final voter list in each provinces. The distribution of voter abstention number of each provinces is shown on Figure 1. Riau Islands is recorded as the highest province's number of voter abstention with 37.91% and followed by North Sumatera with 35.57%. Nevertheless, it is different with the provinces located in the East part of Indonesia. Papua and Gorontalo come up to the lowest number with 14.05% and 18.03%. Based on Table 2, mean and median of voter abstention are 26.52% and 25.31%. A quite similar mean and median indicates that data is normally distributed.

In Figure 2, the map shows geographically the distribution of voter abstention for all of provinces. There are some typical patterns created by neighboring provinces. For example, the darkest orange huge cluster formed consisted of North Sumatera, Riau, Riau Islands, and Bangka-Belitung in Sumatera and Central and East Kalimantan in Kalimantan Island that have similar high number of abstention with range from 30.5% until 37.91%. Other cluster formed are West Sumatera and Jambi in Sumatera and Central and East Java in Java Island for brighter orange with range of 25.69% until 29.63%. The second brighter orange consisted of South Sumatera, Lampung, and Banten

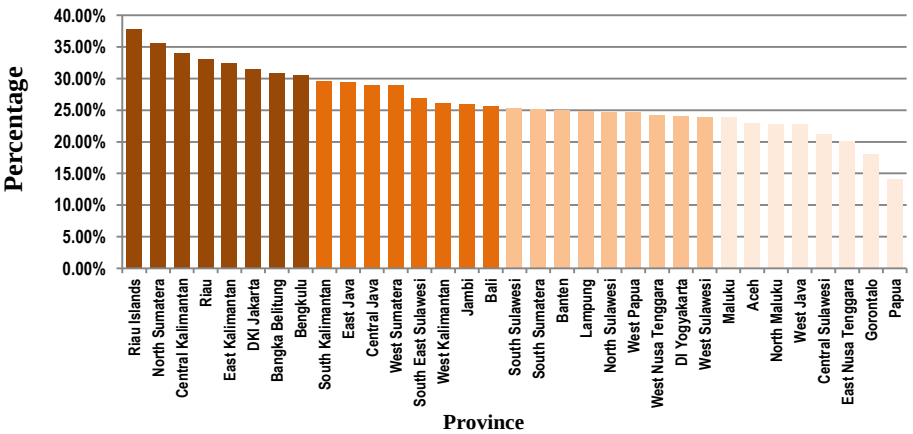


Fig. 1: The Distribution of Voter abstention for All Provinces

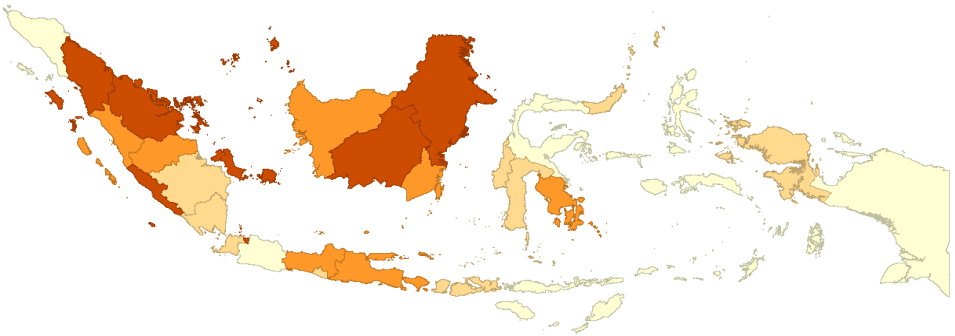


Fig. 2: Quantile Map of Voter Abstention

Table 2
Descriptive Statistic of Voter Abstention

Statistic	Voter Abstention
Mean	26.52%
Median	25.31%
Maximum	37.91%
Minimum	14.05%
Std. Error	0.90%
Std. Deviation	5.00%

with 23.92% until 25.31% of range. For the brightest orange - range between 14.05% until 23.91% - are dominated by provinces in East part of Indonesia, that are Gorontalo, Central Sulawesi, Maluku, North Maluku, East Nusa Tenggara, and Papua. This pattern could be indicated that there is factor of spatial correlation for voter abstention in neighboring provinces.

The indication of spatial correlation for neighboring provinces does not focused in each islands. It is also considered to analyse the correlation between province in other islands. As the reason of national scope observation for an Archipelagic Indonesia, this research uses distance approach to accommodate a thorough spatial correlation. *k*-Nearest Neighbors is part of distance approach used in this research to construct spatial weight matrix.

Multiple Linier Regression

Generating multiple linier regression model is aimed as a beginning step of spatial regression to compare each result then choose the best model for the data. This modeling produce some significant explanatory variables for $\alpha = 10$ %, that are percentage of poor people (X1), human development index (X5), and life expectancy at birth (X8). The R-squared of the model is 63.8% which means that model resulted by multiple linier regression can tell the variance of data for about 63.8%, while 36.2% is explained by other explanatory variables outside the model. Testing is also performed the following assumption:

1. Normality
   Normality test is performed by Kolmogorov-Smirnov test with its value of 0.120 and p-value greater than 0.1 which means that normality assumption for Ordinary Least Square (OLS) residual is completed.
2. Homoscedasticity
   Breusch-Pagan Test is used for checking this assumption. BP-value resulted from the data is 5.526 and p-value is 0.700. Based on it, p-value of BP test is greater than 0.1 which means that this assumption is also completed.
3. Independent error
   The assumption of independent error can be analyzed by considering the graph on Appendix 2. The graph shows that there are not certain patterns formed. Therefore, this assumption is not violated.
4. Multicolinierity
   According to Table 3, there are not VIF value which are greater than 10. It is indicated that there is no multicolinierity among explanatory variables.

Table 3
Parameter Estimates and p-values of Eight Explanatory Variables

Predictor	Coefficient	p-Value	VIF
Intercept	-0.037	0.895	
X1	-0.260	0.016*	1.767
X2	0.004	0.119	1.711
X3	-0.003	0.446	2.253
X4	-0.024	0.143	5.063
X5	0.016	0.003*	5.873
X6	0.002	0.127	1.874
X7	0.101	0.463	1.563
X8	-0.011	0.058*	3.356

*significant at $\alpha = 10\ \%$

Spatial Autocorrelation

Spatial autocorrelation is one of the most important effect in spatial analysis. It refers to a relationship between a value of an attribute at one location and other value of attribute in other locations (Fotheringham & Rogerson 1993). According to Figure 2, provinces which have such similar number of voter abstention that they tend to construct some clusters. These clusters distinguished by color sign that there is a positive spatial autocorrelation. It means that there is a similar value owned by neighboring locations (Silk 1979).

In 1995, Anselin introduce a thorough technique for spatial association that is Local Indicator of Spatial Association (LISA). It is a tool of data exploration area that can detect hotspot and cold spot in spatial analysis and find a local pattern of spatial autocorrelation (Local Moran Index) with testing each area and it's influence to the global aspects (Global Moran Index).

Global and Local Moran Index

Moran Index is divided into global and local correlation. Global Moran Index or Statistics Moran's I is a measurement of correlation between observation in one location and other locations which is contiguous each other (Anselin 1995).

Number of voter abstention in Indonesia affected by neighboring provinces can be detected by measuring Global Moran's I. The result of Global Moran's I for voter abstention is 0.244 with p-value 0.016 (significant at α = 10 %). This result means that there is a positive correlation so that neighboring provinces establish some clusters which have similar characteristics inside of each cluster. This condition is supported by previous explanation about spatial autocorrelation.

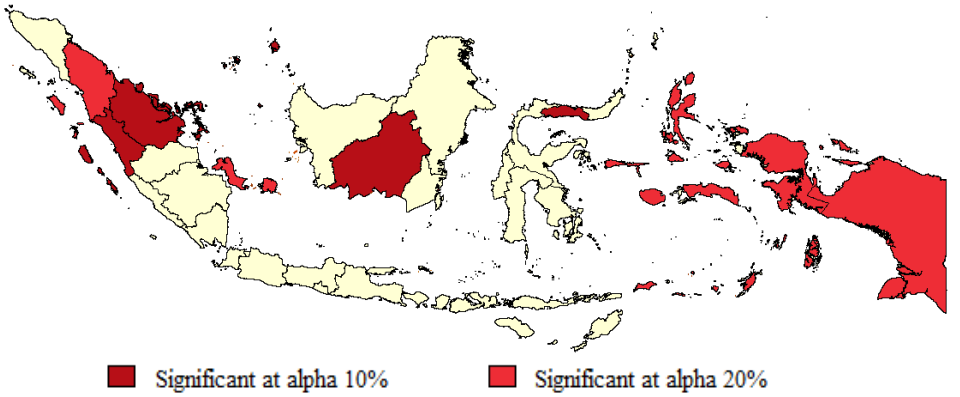


Fig. 3 Significance Map of Local Moran Index

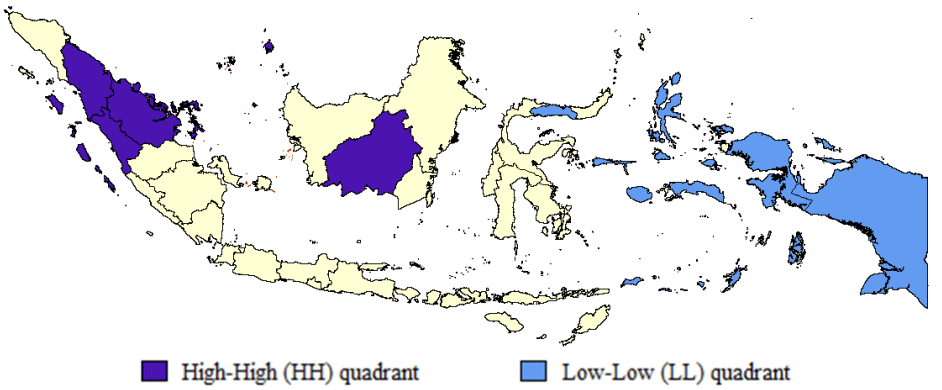


Fig. 4 Cluster Map of Local Moran Index

Global Moran's I explained that there is positive spatial correlation for case of voter abstention in Indonesia, but it did not tell what province affected significantly to spatial clustering around its observations. Anselin (1995) explained about Local Moran's I that the indication level of significant spatial clustering of similar values around observation

could be given by LISA for each observation. There are five significant provinces for α = 10 %, that are Riau, Riau Islands, West Sumatera, Gorontalo, and Central Kalimantan and also six for α = 20 % that are North Sumatera, Bangka Belitung, Maluku, North Maluku, Papua, and West Papua. In Figure 3, these 10% of alpha significant provinces are shown with darker red and brighter red for 20% of alpha. Thus, according to equation 16, this result of significant province means that there is a correlation between the value observed in the neighborhood of one province in the area of another significant province.

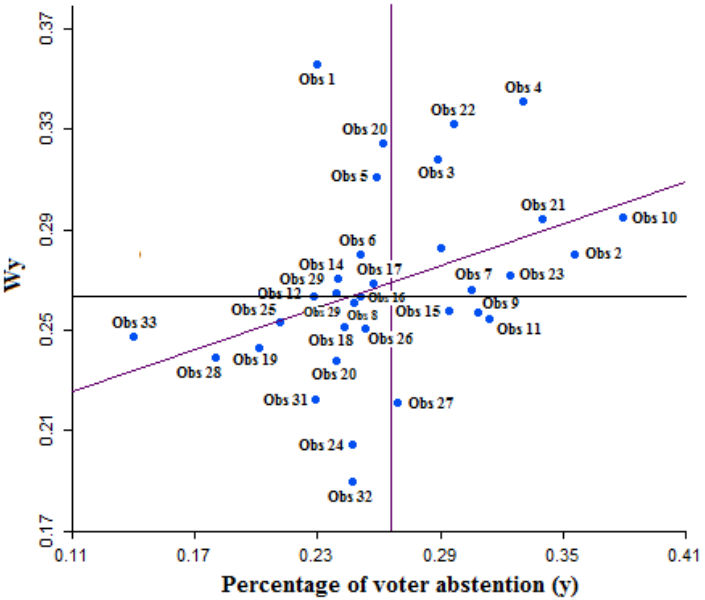


Fig. 5 Moran Scatter Plot

In Figure 5 as a Moran Scatter Plot, all of significant provinces are identified to several quadrants, that are High-High (HH) and Low-Low (LL) located in the upper right and lower left of quadrants. HH and LL indicate spatial clustering of similar high value (that is, more than the mean) and similar low value. Riau Islands, Riau, North and West Sumatera, and Central Kalimantan included in the high number of voter abstention are identified to HH provinces. The low number of abstention's provinces that are Gorontalo, Maluku, North Maluku and Papua step into LL province. According to Cluster Map in Figure 4, the darker blue signs to HH provinces and brighter blue for LL provinces. Other kinds of quadrants are High-Low (HL) and Low-High (LH) located in the lower right and upper left indicate high values surrounded by low neighboring values (HL) and low values surrounded by high values (LH). These two quadrants in this research do not filled by any significant provinces.

These results above show typically that province which has cluster or neighboring similar value of voter abstention tends to be affected significantly to its neighbors. Beside that, the significant provinces which are then plotted to quadrants have a positive

association where high number of voter abstention's provinces are identified to spatial clustering of similar high value (HH) and another one indicates otherwise.

Spatial Regression Model

Spatial Weight Matrix

In spatial modeling, one of the most important part that should not be separated is constructing spatial weight matrix (Getis & Aldstadt 2003). The common spatial weight such as contiguity approach does not appropriate for this research because of geographical condition of the observation. There are seven observations which can be called province and island at once, i.e. Riau Islands, Bangka Belitung, Bali, West and East Nusa Tenggara, Maluku, and North Maluku. It is problem to construct the matrix by using contiguity because it will result spatial weight matrix of certain singularities (rows and columns composed entirely of zeros). Direct contiguity can be too restrictive as islands typically have some contact with non-contiguous units. Another approach is using *k*-NN for social science cases as an alternative way to construct spatial weight matrix (Ward & Gleditsch 2007).

The result of *k*-NN using diverse number of *k* produces the best matrix evaluated by value of AIC and log likelihood (see Table 4). Based on p-value of Lagrange Multiplier (LM), there are not significant spatial dependences by SAR, while SEM gives all of significant spatial dependence, except k=3. According to that significance of SEM, *k*-NN with k=5 obtains the lowest AIC value with -122.60 that is lower than AIC of OLS from Multiple Linier Regression with -118.14. It also has the biggest log likelihood value with 72.29 which is higher than OLS with 69.07. The modification of *k* for this case gives different result. For lower *k* (k<5), it tends to make some no neighbor provinces so that excluding province supposed to be included will be happened. While, for bigger *k* (k>5), there are unnecessary provinces included to the matrix.

Making a comparation between *k*-NN and other kind of spatial weight matrices is considered important. TD weight matrix is used as a consideration of distance as an important criterion of spatial influence, while *k*-NN emphasizes basically to *k*-closest units. The threshold *(d=6.765)* used is based on GeoDa Software threshold default which is obtained by iterating all of possibility units in order to ensure that each province has at least one neighbor. TD matrix produce more neighboring provinces than *k*-NN. It is because high default threshold used can include inappropriate provinces to the matrix, although each province has at least one neighbor. For example, Lampung has 12 neighbors according to TD, where D.I. Yogyakarta as one of it's neighbor. Figure 7 shows that TD gives 25 province which have neighbor greater than 5. Contrastingly, too small *d* could be produce an islands so it will create some no neighbor provinces. *k*-NN gives more stable result of neighboring province than TD if it is compared with real condition. According to Table 4, it is strengthen the previous explanation that *k*-NN is better than TD. AIC value of *k*-NN with -122.60 is smaller than TD with -120.56, while log likelihood value of *k*-NN with 72.29 is greater than TD with 71.28. As those consideration, *k*-NN spatial weight matrix is used in this research for further analysis.

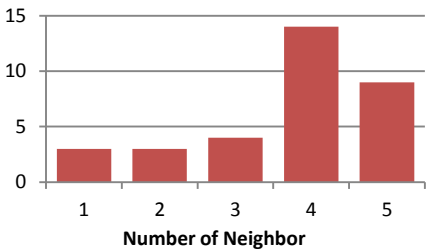


Fig. 6: Connectivity Histogram of *k*-NN *(k=5)*

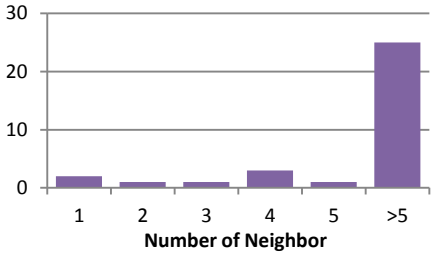


Fig. 7: Connectivity Histogram of Threshold Distance

Table 4
Result of diverse approaches of weight matrix

	***k*-Nearest Neighbor**					**Threshold Distance**
	***k*=3**	***k*=4**	***k*=4 + PD**	***k*=5**	***k*=5 + PD**	
Moran's I	0.21	0.28	0.22	**0.24**	0.17	0.19
p-Value LMSAR	0.20	0.66	0.20	**0.38**	0.10	0.51
p-Value LMSEM	0.10	0.08	0.08	**0.03**	0.05	0.09
AIC SAR	-117.64	-116.33	-117.97	**-116.90**	-119.12	-116.49
AIC SEM	-119.81	-120.19	-120.58	**-122.60**	-122.14	-120.56
AIC OLS	-118.14	-118.14	-118.14	**-118.14**	-118.14	-118.14
Log likelihood SAR	69.81	69.16	69.98	**69.44**	70.56	69.24
Log likelihood SEM	70.90	71.09	71.29	**72.29**	72.07	71.28
Log likelihood OLS	69.07	69.07	69.07	**69.07**	69.07	69.07

Spatial Effect Test

Spatial effect test is divided into two effects, there are spatial dependence effect and spatial heterogeneity effect. Test of spatial dependence effect using Lagrange Multiplier Test and using Breush-Pagan Test for spatial heterogeneity effect.

Table 5
Value of Lagrange Multiplier

Model	LM Value	P-value
SAR	0.760	0.383
SEM	4.440	0.035*

*significant at $\alpha = 10\ \%$

The result of multiple linier regression explained that all of the assumptions are completed, but exploring data from map indicates that there is spatial correlation between observations. Therefore, it seems necessary to check spatial dependence by Lagrange

Multiplier. According to Table 4, p-value of SEM is 0.035 that is significant at α = 10 % with LM value about 3.440. It is indicates that there is spatial dependence of error for case of voter abstention in Indonesia so that Spatial Error Model (SEM) is the suitable spatial model to consider spatial effect which is not determined in Multiple Linier Regression.

Spatial effect test for heterogeneity effect using Breusch-Pagan is actually same with test assumption in Multiple Linier Regression. Breusch-Pagan (BP) test summary for BP-value and p-value are 5.526 and 0.700 indicates that the residual of variance from the data is homogeny.

Spatial Error Model (SEM)

In the general model of spatial regression, if $\rho = 0$ and $\lambda \neq 0$ then the model is called Spatial Error Model (SEM). The table and equation resulted by SEM are:

Table 6
Result of Spatial Error Model (SEM)

Predictor	Coefficient	p-Value
Intercept	0.018	0.938
Lambda	-0.612	0.011*
X1	-0.319	0.000*
X2	0.006	0.002*
X3	-0.002	0.519
X4	-0.035	0.006*
X5	0.016	0.000*
X6	0.002	0.000*
X7	0.120	0.240
X8	-0.010	0.021*

*significant at α = 10 %

$$y = -0.319\,X1 + 0.006\,X2 - 0.035\,X4 + 0.016\,X5 + 0.002\,X6 - 0.010\,X8 - 0.612\,\boldsymbol{W}u$$

According to Table 6, there are six significant explanatory variables, that are percentage of poor people (X1), monthly average of wage/salary/income of employee (X2), mean years of schooling population 15 years of age and over (X4), human development index (X5), school enrollment ratio 16-18 years of age (X6), and life expectancy at birth (X8). SEM gave more significant explanatory variables than Multiple Linier Regression. Five significant variables on full model still on the track, it was then added by three more significant variables in SEM model. Beside that, SEM also obtain Lambda (λ) that is significant with p-value = 0.000 and coefficient of 0.612. It means that province which is surrounded by n-province(s) so that the effect from each surrounding

province is about 0.612 multiplied by mean of residual. This model has a better AIC value than full model. AIC value of SEM is -122.6 lower than full model with -118.14.

The significant explanatory variables divided into two parts, i.e. positive and negative significances. There are three positive and three negative significant variables. Coefficient of X1 is -0.319, it means, every 1% increase of poor people's percentage, it will reduce percentage of mean of voter abstention for about 0.319 in each province. X2 with coefficient of 0.006 means, every Rp 1 increase of monthly average of employee's wage/salary/income, it will increase percentage of mean of voter abstention for about 0.006 in each province. Coefficient of X4 is 0.035, means that mean of voter abstention's percentage will be reduced 0.035 if there is 1 point increment of mean years of schooling population 15 years of age and over. X5 with coefficient of 0.016 means, every 1 point increase of human development index, it will increase percentage of mean of voter abstention for about 0.016. School enrollment ratio (X6) has 0.002 of coefficient which means, if there is 1 point increase of it, for about 0.002 as an enhancement of mean of voter abstention's percentage in each province. Coefficient of X8 is 0.010, it means, every 1 point increase of life expectancy at birth, it will enhance 0.010 of mean of percentage of voter abstention. Each variable considers the assumption that the change of one variable will be built if other variables are constant.

Constructing an interpretation of case study of social sciences is more likely complex than other case. In Brown (2009), some studies in social science show that considering people's socioeconomic status, a measurement which factor in person's education, occupation, and income, will directly affect their behavior, including person's voting behavior. Significant variables in this research related to education are mean years of schooling population 15 years of age and over (X4) and school enrollment ratio 16-18 years of age (X6). X4 is affected negatively to voter abstention, while X6 is positive. This distinction of sign's significances of education status should have taken into account. According to Sigelman *et al.* (1985) about considering level of education to voting behavior, the more years of formal education one has, the greater the probability that one will vote. It is correspond to the result with including spatial factor that mean years of schooling population 15 years of age and over is negatively significant to voter abstention. The higher province's years of schooling is, the more its citizen is likely to vote. In contrast, the school enrollment ratio 16-18 years of age variable is different to others. Regardless spatial factor, Mujani (2007) implied that voting behavior in Indonesia is not contrary with voter abstention. The higher person's education level is, the more they less likely to vote. People who is being aphatetic and septic toward politics comes from higher educational groups. Indeed, this pattern is different from U.S presidential election. Logically, education is a powerful gun for every problem. School enrollment ratio 16-18 years of age can be equated as senior high school level that they are as an beginner voter with less of political awareness. Print *et al.* (2009) emphasizing to strengthen a civic education for youth political participators. The principle objective of civic education is to teach civic literacy, which can be a knowledge for understanding of the basic principles of government and enhancing skills to face democracies.

Beside education, economic status is also being noted. Variables related are percentage of poor people (X1) and monthly average of wage/salary/income of employee (X2). Both of them have different sign of significance to voter abstention. While X1 is

negative, X2 is positive. The SEM result show that, poor people is more likely to vote than the rich. This result is correspond with Figure 1 and 2 on data exploration. Provinces which have low number of voter abstention are most located in East part of Indonesia, such as Papua, Gorontalo and Maluku. Data of percentage of poor peoples used in this research then tabulated with voter abstention and the result is that most of provinces which have high number of percentage of poor people are having low abstain (see Figure 8). Mujani (2007) again implied that election in Indonesia is like a party for poor people. Characteristics of country may indicate the result. In some other countries, Sigelman *et al.* (1985) explained about economic status that poor people is being less likely to vote. However, Gupta (2004) implied in the research of election in India which has the same characteristics of developing country as Indonesia that the poor are more likely to vote than the rich. In one side, it is a good condition when people give a vote on election, but in other side it is susceptible to election violation such as money politics and voter suppression.

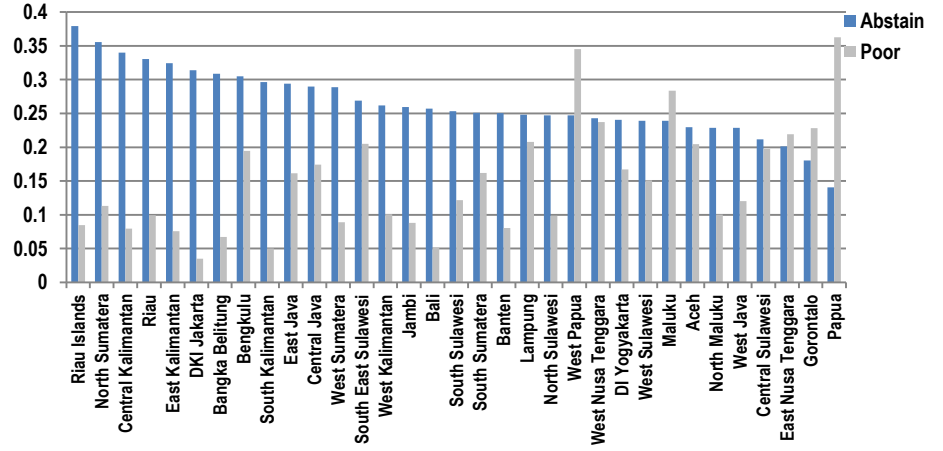


Fig. 8 Distribution of voter abstention and percentage of poor people

The result for monthly average of wage/salary/income of employee gives a significant and positive value to voter abstention. It indicates that the higher people's income is, the more they less likely to vote. According to Figure 9, high number of voter abstention's province also have high number of income. However, there are four provinces which have high number of income, while those have low abstention, such as Papua, West Papua, Maluku, and North Maluku. Those provinces have high income of employee because the life's cost are higher than other provinces, so it is assumed that those are identified as a low income of employees. In Eriyanto (2007), most abstain voter in Indonesia come from peoples who have monthly salary/income. Other past research show contrastingly, such as Nevitte *et al.* (2000) and Sigelman *et al.* (1985), are from developed countries. It may indicates that characteristics of each countries affect the result.

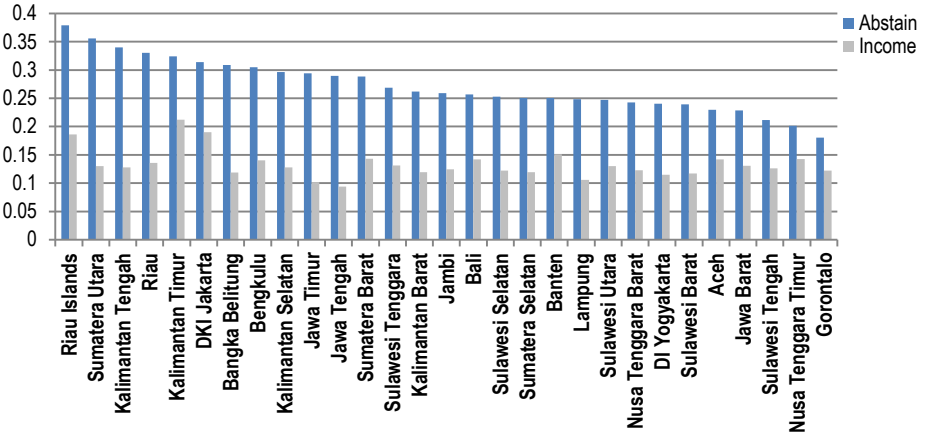


Fig. 9 Distribution of Voter Abstention and Monthly Average of wage/salary/income of employee

The interpretation of the relationship between education, economic status and voter abstention, regarding spatial factor, have different pattern from other country explained by other related research. Human Development Index (HDI) which have utility to represents country identified as poor, developing, or developed one, then can be observed. HDI (X5) gives positive and significant with voter abstention. Gupta (2004) again gives an example of the research about voting behavior in India which have similar characteristics of the result to Indonesia's case. According to UNDP report, Indonesia and India are identified as a medium development by HDI value. Contrastingly, characteristics of region or countries with very high and high value of HDI have a negative association to voter abstention. According to Figure 10, provinces with high number of voter abstention are also have high rate of HDI. Then, the brighter colour show that 17 from 33 provinces are identified as provinces with each HDI value below to the average of national value. Those provinces having low number of abstention located mostly in East part of Indonesia. HDI figures a comprehensive overview of the level of achievement of human development as a result of development activities done by government. By that condition, there is some disparities between provinces with distinguish value of HDI. Activities of development in Java and Sumatera are so more intensive than the East part that the province's citizen with low HDI tend to be provacated easily to do some election violation such as money politics or voter suppression.

Then, life expectancy at birth (X8) has negative association with voter abstention. This kind of variable related to people's health and prosperity will reduce voter abstention if it is increased because healthy people will easily to go to voting stand to give their vote.

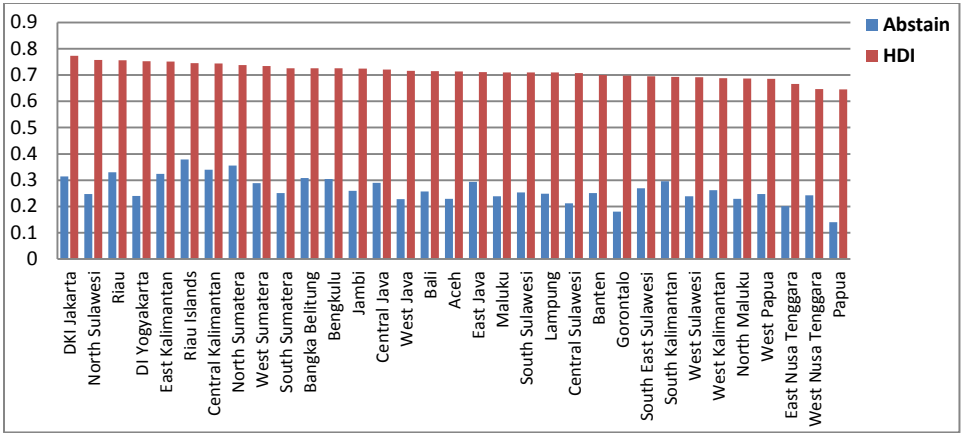


Fig. 10 Distribution of Voter Abstention and Human Development Index

6. COMMENTS AND CONCLUSION

There is positive spatial autocorrelation for voter abstention data, which means that there are similar values owned by neighboring provinces in Indonesia. Provinces that are significant to spatial autocorrelation are Riau, Riau Islands, West Sumatera, Central Kalimantan, and Gorontalo. According to significant provinces, there are not province that identified as hotspot or cold spot observation.

Determining factors that are significantly affected to voter abstention using Spatial Error Model (SEM) is better than Spatial Autoregressive Model (SAR) and Multiple Linier Regression. There are six significant explanatory variables, that are percentage of poor people (X1), monthly average of wage/salary/income of employee (X2), mean years of schooling population 15 years of age and over (X4), human development index (X5), school enrollment ratio 16-18 years of age (X6), and life expectancy at birth (X8).

Mahalanobis Distance for case of social science could be used as a comparation between Euclidean for calculating centroid distance in designing spatial weight matrix.

REFERENCES

1. Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht (NLD): Academic Publisher.
2. Anselin, L. (1995). Local Indicators of Spatial Association. *Research Paper 9331 Regional Research*. Virginia (US): Institute West Virginia.
3. Anselin, L. (1999). Spatial Econometrics. Dallas (US): Bruton Center.
4. Arbia, G. (2006). *Spatial Econometrics: Statistical Foundation and Application to Regional Convergence*. Berlin: Springer-Verlag.
5. Arianto, B. (2010). Analisis Penyebab Masyarakat Tidak Memilih dalam Pemilu. Tanjungpinang (ID): Fakultas ISPOL, Universitas Maritim Raja Ali-Haji. 1(1).
6. Arisanti, R. (2011). Performance Spatial Regression Models for Detecting Factors of Poverty in East Java Province [Thesis]. Bogor: Faculty of Mathematics and Natural Statistics, Bogor Agricultural University.

7. Badan Pusat Statistik [BPS] (2010). *Trends of the Selected Socio-Economic Indicators of Indonesia*. Jakarta: Badan Pusat Statistik.
8. Brown, C.R. (2009). Issue Brief: Voting Behavior based On Socioeconomic Status.
9. Chatterjee, S. and Hadi, A.S. (2006). *Regression Analysis by Example*. New Jersey (US): John Wiley & Sons, Inc.
10. DeSipio, L., Masuoka, N. and Stout, C. (2006). The Changing Non-Voter: What Differentiates Non-Voters and Voters in Asia, American and Latino Communities? *CSD Working Paper*. California (US): University of California.
11. Eriyanto, (2007). Golput dalam Pilkada. Kajian Bulanan Lingkaran Survei Indonesia. [Internet]. [Edisi 05 September 2007]. Jakarta (ID).
12. Fotherningham, A.S. and Rogerson, P.A. (1993). GIS and Spatial Analytical Problems. New York (US): University of Buffalo.
13. Getis, A. and Aldstadt, J. (2003). Constructing the Spatial Weight Matrix Using a Local Statistics. *Journal of Geographical Analysis*, Vol. 36, No. 2. Ohio (US): The Ohio State University.
14. Gupta, D. (2004). Analysis of Election in India. Sydney (AUS).
15. Komisi Pemilihan Umum [KPU] (2009). Sertifikat Rekapitulasi Perhitungan Suara Pemilu Presiden dan Wakil Presiden Tingkat Nasional. Jakarta: Komisi Pemilihan Umum.
16. Mujani, S. (2007). Voting Behavior Kasus Indonesia. Bahan Kuliah Program Pasca Sarjana Master Manajemen Komunikasi. Jakarta (ID): FISIP, Universitas Indonesia.
17. Nevitte, N., Blais, A., Gidengil, E. and Nadeau, R. (2000). Socio-Economic Status and Non-Voting - A Cross – National Comparative Analysis. *Prepared for presentation at the XVIIIth World Congress of the International Political Science Association*. Quebec.
18. Nyarwi, A. (2009). Golput Pasca Orde Baru : Merekonstruksi Ulang Dua Perspektif. *Jurnal Ilmu Sosial dan Ilmu Politik*. 12(2), 257-390.
19. Print, M. and Milner, H. (2009). Civic Education and Youth Political Participation. Rotterdam (NLD): Sense Publishers.
20. Sigelman, L., Roeder, P., Jewell, M. and Baer, M. (1985). Voting and Nonvoting: A Multi-Election Perspective. *American Journal of Political Science*, 29(4). 749-765.
21. Silk, J. (1979). Statistical Concept in Geography. London (UK): George Allen & Unwin.
22. Ward, M. and Gleditsch, K. (2007). An Introduction to Spatial Regression Models in the Social Sciences.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 177-184

COMPARISON OF BINARY, UNIFORM AND KERNEL GAUSSIAN WEIGHT MATRIX IN SPATIAL ERROR MODEL (SEM) IN PANEL DATA ANALYSIS

M Nur Aidi, Tuti Purwaningsih, Erfiani and **Anik Djuraidah**
Department of Statistics, Bogor Agricultural University, Indonesia
Email: purwaningsiht@yahoo.com; nuraidi@yahoo.com

ABSTRACT

Spatial Analysis is grow up nowadays. One of field in statistics like econometrics is discussing about spatial influence in many economics data. This research try to analyze spatial in panel data model. Panel data is combining cross-section data and time series data. If the cross-section is locations, need to check the correlation between locations. λ is parameter in spatial error model to cover effect of data correlation between location. Value of λ will influence the goodness of fit model, so it is important to make parameter estimation. The effect of another location is covered by make contiguity matrix until get spatial weighted matrix (W). There are some type of W, it is Binary W, Uniform W, Kernel Gaussian W and some W from real case of economics condition or transportation condition from locations. This study is aim to compare Binary W, Uniform W and Kernel Gaussian W in spatial error model in panel data analysis (SEM panel data) using RMSE value. The result of analysis showed that Kernel Gaussian Weight has RMSE value less than Binary and Uniform Weight in SEM panel data.

KEYWORDS

Binary Weight, Uniform Weight, Kernel Gaussian Weight, Spatial Error Model, Panel Data.

1. INTRODUCTION

Panel data analysis is combining cross-section data and time series data, in sampling when the data is taken from different location, it's commonly found that the observation value at the location depend on observation value in another location. In the other name, there is spatial correlation between the observation, it is spatial dependence. Spatial dependence in this study is covered by Generalized spatial model which is focussed on dependence between locations and error [1]. If there is spatial influence but not involved in model so error assumption that between observation must be independent will not fulfilled. So the model will be in bad condition, for that need a model that involve spatial influence in the analysis panel data that will be mentioned as Spatial Panel Data Model.

Some recent literature of Spatial cross-section data is Spatial Ordinal Logistic Regression by Aidi and Purwaningsih [2], Geographically Weighted Regression [3]. Some of the recent literature of Spatial Panel Data is forecasting with spatial panel data [3] and spatial panel models[4]. For accomodate spatial dependence in the model, there is Spatial weighted matrix (W) that is important component to calculate the spatial

correlation between location. Spatial parameter in spatial error model in panel data analysis, known as ρ. There are some type of W, it is Uniform W, Binary W, Invers distance W and some W from real cases of economics condition or transportation condition from the area. This research is aim to compare Binary W, Uniform W and Kernel Gaussian W in spatial error model (SEM) in panel data analysis using RMSE value which is obtained from simulation.

2. LITERATURE REVIEW

2.1. Data Panel Analysis

Data used in the panel data model is a combination of crosssection and time-series data. crossection data is data collected at one time of many units of observation, then time-series data is data collected over time to an observation. If each unit has a number of observations across individuals in the same period of time series, it is called a balanced panel data. Conversely, if each individual unit has a number of observations across different period of time series, it is called an unbalanced panel data (unbalanced panel data).

In general, panel data regression model is expressed as follows:

$$y_{it} = \alpha + \mathbf{x}'_{it}\boldsymbol{\beta} + u_{it} \qquad (1)$$
$$i = 1,2, \ldots, N \; ; \; t = 1,2, \ldots, T$$

with i is an index for crossection data and t is index of time series. α is a constant value, β is a vector of size K × 1, with K specifies the number of explanatory variables. Then y_{it} is the response to the individual cross-i for all time periods t and x_{it} are sized K × 1 vector for observation i-th individual cross and all time periods t and u_{it} is the residual / error[5].

Residual components of the direction of the regression model in equation (1) can be defined as follows:

$$u_{it} = \mu_i + \varepsilon_{it} \qquad (2)$$

where μ_i is an individual-specific effect that is not observed, and ε_{it} is a remnant of crossection-i and time series-t [5].

2.2 Spatial Weighted Matrix (W)

Spatial weighted matrix is basically a matrix that describes the relationship between regions and obtained by distance or neighbourhood information. Diagonal of the matrix is generally filled with zero value. Since the weighting matrix shows the relationship between the overall observation, the dimension of this matrix is NxN [6]. There are several approaches that can be done to show the spatial relationship between the location, including the concept of intersection (Contiguity). There are three types of intersection, namely Rook Contiguity, Bishinop Contiguity and Queen Contiguity.[6]

After determining the spatial weighting matrix to be used, further normalization in the spatial weighting matrix. In general, the matrix used for normalization normalization row (row - normalize). This means that the matrix is transformed so that the sum of each row of the matrix becomes equal to one. There are other alternatives in the normalization of this matrix is to normalize the columns of the matrix so that the sum of each column in the weighting matrix be equal to one. Also, it can also perform normalization by dividing

the elements of the weighting matrix with the largest characteristic root of the matrix ([6]; [7]).

There are several types of Spatial Weight (W): binary W, uniform W, invers distance W (non uniform weight) and and some W from real case of economics condition or transportation condition from the area. Binary weight matrix has values 0 and 1 in off-diagonal entries; uniform weight is determined by the number of sites surrounding a certain site in ℓ-th spatial order; and non-uniform weight gives unequal weight for different sites. The element of the uniform weight matrix is formulated as,

$$W_{ij} = \begin{cases} \frac{1}{n_i^{(l)}}, \text{j is neighbor of i in l} - \text{th order} \\ 0, \text{others} \end{cases} \quad (3)$$

$n_i^{(l)}$ is the number of neighbor locations with site-i in ℓ-th order. The non-uniform weight may become uniform weight when some conditions are met. One method in building non-uniform weight is based on inverse distance. The weight matrix of spatial lag k is based on the inverse weights $1/(1 + \text{d\_ij})$ for sites i and j whose Euclidean distance dij lies within a fixed distance range, and otherwise is weight zero. Kernel Gaussian Weight follow this formulla :

$$w_j(i) = \exp\left[-\frac{1}{2}\left(d_{ij}/b\right)^2\right] \quad (4)$$

with d is distance between location i and j, then b is bandwith which is a parameter for smoothing function.

2.3 Spatial Error Model in Panel Data Analysis (SEM-Panel Data)

Spatial error model in panel data analysis expressed in the following equation:

$$y_{it} = \mathbf{x'}_{it}\boldsymbol{\beta} + \mu_i + \phi_{it}; \quad (5)$$

$$\phi_{it} = \lambda \sum_{j=1}^{N} w_{ij}\phi_{it} + \varepsilon_{it}; \quad (6)$$

where δ is the spatial autoregressive coefficient and w_{ij} is elements of the spatial weighted matrix which has been normalized (W). Estimation of parameters in this model use Maximum Likelihood Estimator. [7]

3. METHODOLOGY

3.1 Data

Data used in this study was gotten from simulation using SEM panel data model as equation (5) with initiation of some parameter. Simulation was done use R program. The following step in methods is used to generate the spatial data panel which is consist of index n and t. Index n indicates the number of locations and index t indicates the number of period in each locations, the result can be look at Table 1.

3.2 Methods

1. Determine the number of locations to be simulated is N = 3, N = 9 and N = 25
2. Makes 3 types of Map Location on step 1

3. Creating a Binary Spatial weighted matrix based on the concept of Queen Contiguity of each type of map locations. In this step, to map the 3 locations it will form a 3x3 matrix, 9 locations will form a 9x9 matrix and 25 locations form a 25x25 matrix.
4. Creating Spatial Uniform weighted matrix based on the concept of Queen Contiguity of each type of map locations.
5. Making weighted matrix kernel gaussian based on the concept of distance. To make this matrix, previouly researchers randomize the centroid points of each location. After setting centroid points, then measure the distance between centroids and used it as a reference to build Kernel Gaussian W.

   Gaussian kernel W as follows:

   $$w_j(i) = \exp\left[-\frac{1}{2}\left({}^{d_{ij}}/_{b}\right)^2\right] \text{ [3]}$$

6. Specifies the number of time periods to be simulated is T = 3, T = 6, T = 12 and T = 24
7. Generating the data Y and X based on generalized spatial panel data models follows equation (5) and (6).
8. Cronecker multiplication between matrix Identtity of time periods and W, then get new matrix named IW.
9. Multiply matrix IW and Y to obtain vector WY.
10. Build a spatial panel data models and get the value of RMSE
11. Repeat steps 7-9 until 1000 replications for each combination on types of W, N, T, ρ and λ.

    Description:
    Types of W: W Binary, W Uniform and Gaussian kernel W
    Types of N: 3, 9 and 25 locations
    Types of T: 3, 6, 12 and 36 Series
    Types of ρ = 0.3, 0.5, 0.8 and λ = 0.3, 0.5, 0.8
12. Get the RMSE value for all of 1000 replications oh each combination between W, N, T and ρ.
13. Determine the best W based on the smallest RMSE for all combinations.

4. RESULTS AND DISCUSSION

Simulation generate data for vector Y as dependent variable and X matrix as independent variable. Y and X is generate with parameter initiation. After doing simulation, we can get RMSE for each combinations and proccessing it, then we can calculate RMSE for each W, N, T and λ. Here is the result. With the result in Table 1 then continued to figure it into graphs in order to look the comparison easily.

Table 1
Value of RMSE resulted from Simulation for all the combinations (W, N, T, ρ)

W Types	Location Types	Periods Types	RMSE	Average RMSE	Average RMSE
Binary W	N=3	T=3	2.891	3.776	3.109
		T=6	3.738		
		T=12	4.090		
		T=36	4.383		
		Average	3.776		
	N=9	T=3	3.535	3.657	
		T=6	3.668		
		T=12	3.689		
		T=36	3.735		
		Average	3.657		
	N=25	T=3	1.865	1.895	
		T=6	1.884		
		T=12	1.903		
		T=36	1.927		
		Average	1.895		
Uniform W	N=3	T=3	2.132	2.707	2.172
		T=6	2.669		
		T=12	2.938		
		T=36	3.090		
		Average	2.707		
	N=9	T=3	1.800	1.988	
		T=6	1.964		
		T=12	2.051		
		T=36	2.138		
		Average	1.988		
	N=25	T=3	1.757	1.822	
		T=6	1.816		
		T=12	1.847		
		T=36	1.867		
		Average	1,822		
Kernel Gaussian W	N=3	T=3	1.398	1.841	1.790
		T=6	1.854		
		T=12	2.009		
		T=36	2.101		
		Average	1.841		
	N=9	T=3	1.353	1.431	
		T=6	1.425		
		T=12	1.461		
		T=36	1.484		
		Average	1.431		
	N=25	T=3	1.922	2.099	
		T=6	2.076		
		T=12	2.166		
		T=36	2.232		
		Average	2.099		

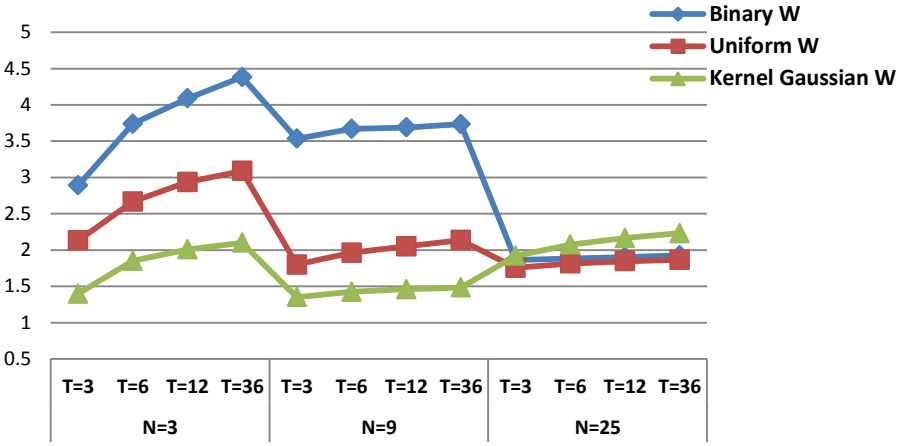


Fig.1 RMSE between Binary, Uniform and Kernel Gaussian Weight for Combinations N and T

Based on figure 1 can be said that Kernel Gaussian W has smaller RMSE than Binary and Uniform W for almost combinations of N types and T types. If we look the level of stabilization, Kernel Gaussian W is better than Binary and Uniform W. We can look at the graph in green line as Kernel Gaussian W, it has value only in range 1 until 2.5 then Uniform W has range from 1.5-3.5 and Binary from 1.5-4.5. So can be concluded that Kernel Gaussian W is better than Binary and Uniform W in SEM panel data model.

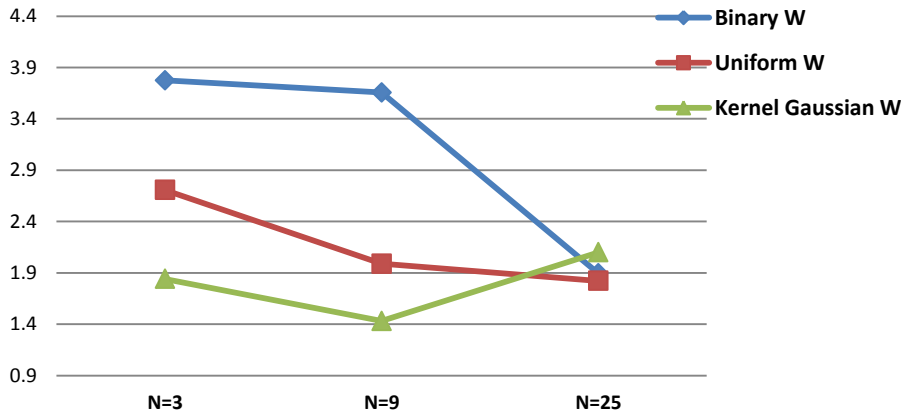


Fig. 2 Comparison RMSE of W based on N types

Figure 2 try to analyze diferencies between the W based on N types (3 locations, 9 locations and 25 locations). With graph above can be concluded that Kernel Gaussian W has smallest RMSE in all N types.

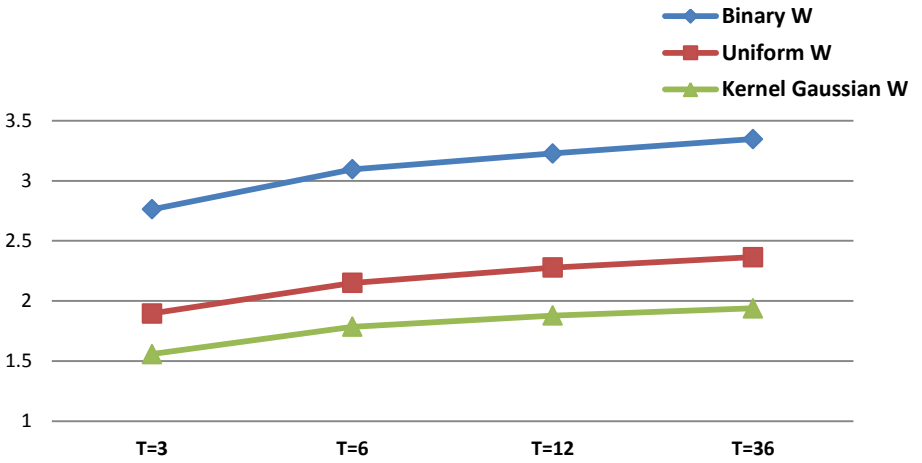


Fig. 3 Comparison RMSE of W based on T types

Figure 3 try to analyze diferencies between the W based on T types (3 periods, 6 periods, 12 periods and 36 periods). With graph above can be concluded that Kernel Gaussian W has smallest RMSE in all types of T.

5. CONCLUSION

Based on simulations result and after explorating the RMSE, can be concluded that Kernel Gaussian W is the best W in SEM panel data model.

ACKNOWLEDGEMENTS

The first, authors would like to thankful to Allah SWT, my parents, lecturer and all of friends.

REFERENCES

1. Anselin, L., Julie, G. and Hubbert, J. (2008). *The Econometrics of Panel Data*. Berlin: Springer.
2. Aidi, M.N. and Purwaningsih, T. (2012). Modelling Spatial Ordinal Logistic Regression and the Principal Component to Predict Poverty Status of Districts in Java Island. *International Journal of Statistics and Application.* DOI: 10.5923/j.statistics.20130301.01
3. Fotheringham, A.S., Brunsdon, C. and Chartlon, M. (2002). *Geographically Weighted Regression, the analysis of spatially varying relationships*. John Wiley and Sons, LTD.
4. Elhorst. (2011). Spatial panel models. *Regional Science and Urban Econometric*.

5. Baltagi, B.H. (2005). *Econometrics Analysis of Panel Data*. Ed ke-3. England: John Wiley and Sons Ltd.
6. Dubin, R. (2009). *Spatial Weights. Fotheringham AS, PA Rogerson*, editor, Handbook of Spatial Analysis. London: Sage Publications.
7. Elhorst, J.P. (2010). *Spatial Panel Data Models. Fischer MM, A Getis, editor, Handbook of Applied Spatial Analysis*. New York: Springer.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 185-200

FAST: A WEB-BASED STATISTICAL ANALYSIS FORUM

Dwiyana Siti Meilany Dalimunthe[1], Debi Tomika[1], Eka Miftakhul Rahmawati[1], Muchriana Burhan[1], Yayan Fauzi[1], Muchammad Romzy[2], I Made Arcana[1], Imam Machdi[2], Usman Bustaman[2], Widyo Pura Buana[2] and **Setia Pramana[1§]**
[1] Sekolah Tinggi Ilmu Statistik, Jl. Otista 64C, Jakarta, Indonesia
[2] Badan Pusat Statistik, Jl. Dr. Sutomo 6-8, Jakarta, Indonesia
[§] Corresponding Author Email: setia.pramana@stis.ac.id

ABSTRACT

Performing standard and advanced statistical analysis using statistical packages and interpreting the results or outputs is still a hurdle faced by beginners. Knowledge related to the use of the statistical packages are often derived from tutorials or books and the internet, which is very broad and dispersed. Furthermore, there was no good knowledge management on statistical analysis that can be used as a reference by the beginners. The results of the analysis still become personal knowledge, since no media can spread them. To overcome the limitation, a web-based forum named FAST is built to provide a discussion board that make the dissemination of knowledge to be faster. In addition, this forum also provides a web-based statistical applications built with the shiny framework as the user interface and R as back-end-program that can be used to analyze the data that has been provided or uploaded by the user. FAST currently provides several statistical analysis such as regression (linear, logistic, ridge, and tobit regression), ARIMA, survival data analysis, as well as cluster analysis. Users can place the results of the analysis conducted on the application to the discussion forum and gallery of analysis to have feedback and discussion from other users. In addition, users can also create reports of the results of the analysis conducted.

KEYWORDS

Internet forum, web-based statistical analysis, shiny framework, R

1. INTRODUCTION

Statistical science has been developed since the 17th century to the present. It was originally used to collect data. But this time, the statistical science is not only about data collecting, but also processing and interpreting. Collected data in the field have to be processed by using statistical analysis tools to produce the meaningful information. Statistical science that was originally synonymous with mathematics, is increasingly widespread, in which statistical science is now often applied to other sciences such as economics, social, health, and so forth. With the increasement in application of statistical science, of course, the used of statistical analysis tools is growing in accordance with the different conditions and circumstances.

The development of statistical science which followed by the development of technology led to the emergence of new analytical tools as well as a modification of the previous analysis tools. Most of these analysis tools are complex analysis tools that will take long time for manual counting. Therefore, we needtools to perform calculations in the form of applications that can streamline the process of statistical analysis.

Statistical application consists of two types when viewed by the way of acquiring it, is a paid statistical applications such as SPSS, STATA, and SAS, as well as freestatistical applications such as R, zaitun time series, etc. While terms of the technology used, statistical applications are divided into desktop-based applications and web-based applications. Where a desktop-based application requires the installation with certain specifications that cannot be installed on a device that has lower specification, but they do not require an internet connection because the applications typically run offline, although for some purposes the internet connection is required. Most desktop applications are paid applications, so many users simply find it difficult to acquire[1]. While the web-based application, users can access the application using a browser that is available in every device, as well as the platforms are not affected. Most of the web-based applications are not paid because the application can be widely accessible. But of course, because this application is web-based, so an Internet connection is necessary.

Related to the use of applications that exist, of course, users will find difficulty in understanding the theory of statistics, and learn how to use the analysis tools available within the application. The procedure to use the features that are available are often not provided clear and examples that support, so users often have difficulty in doing the analysis. Starting from simple things like how to load the data, process the data with a variety of analysis tools available, or how to interpret the results of the analysis of the application. The big differences in the design of the application were often confusing for users, especially the differences in the terms and procedures which must be carried out between the applications with each other.

Until now, when a user finds a difficulty in theory, of course, users will find the solutions or answers from books or related literature, and the most often thing that users do when they find difficulties in practice is seeking the answer in the form of tutorials or guidelines over the internet. The Internet users must know about googleforum[2] and stackoverflow[3] site, which are often used as a place to look for information when users encounter difficulties in applied statistics. These sites accommodate the needs of the user in finding information such as the procedures that must be done to solve a problem, as well as the theoretical discussion of the existing analysis tools. These sites use the discussion board concept that known as Internet forums, where everyone who has registered can write the correct solution, in accordance with the existing problems and questions. With the technique of two way communication, information spread more easily, anyone who knows the answer of a question can directly answer it. In addition, the information can also be spread faster. It is called a good knowledge management. The

[1]http://www.ncss.com/price-list/
[2]https://groups.google.com/forum/
[3]stackoverflow.com

documentation of the knowledge that has been done by others carried out systematically so that the excavation of the related knowledge becomes easier.

However, the problem that arises is, what if the version of the application used was different? Or what if users use a different platform? Obviously these differences that cause the unexpectedly problems. This is due to a separate forum for discussion of the application used. When someone uses a version of the application then appears the problem, of course, the resulting solution may not work if it is used to solve problems that arise in the same application but with a different version.

The use of internet forums for knowledge management must be one of the right solution to be used, but by looking at the constraints of the application version, platform, etc, so it would be better if the internet forum is combined with statistical application that can be directly used to analyze the data. FAST is an internet forum that is integrated with statistical applications. Everything that has been done in the application can be deployed to bereferences or just information for other users. Merger between internet forums and statistical applications certainly make easiness for the user to analyze the data, and then distribute to other users, so that will be created good knowledge management.

2. DESIGN APPROACH AND PHILOSOPHY

FAST is an internet forum that is integrated with a web-based statistical applications and built in order to facilitate statistical analysis of the data while creating a good knowledge management. Generally, FAST architecture is as follows:

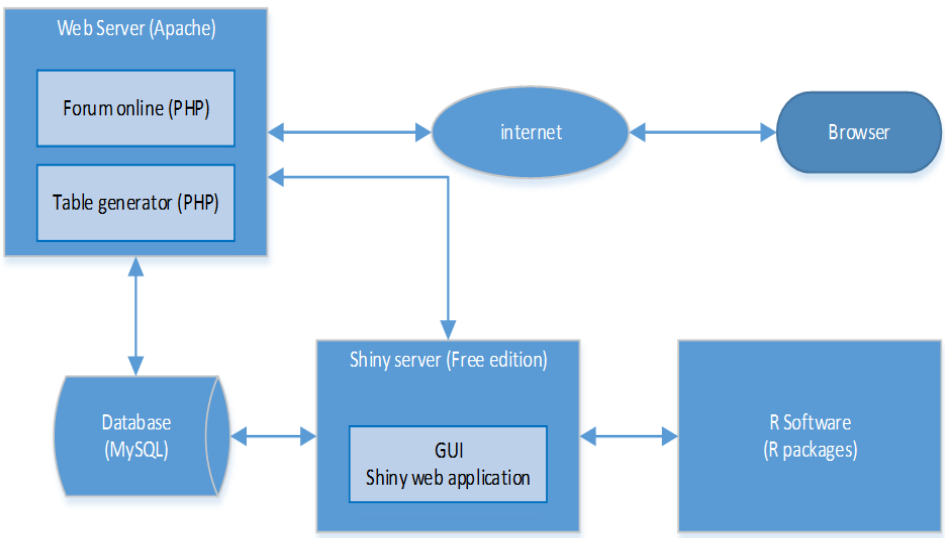


Figure 1: FAST Architecture

Internet forum and table generator are built using PHP, while application is built using R as well as user interface utilizing shiny framework. Users can access the forum and table generator through a browser. Database on FAST consists of two, forum database using Apache web server, and database of statistical application using free edition of shiny server.

This statistical application is built using R programming language, and utilizing packages available in R with some necessary modifications, also with the help of R-software version 3.12 and R-Studio software version 0.98. Then for the user interface of the statistical application is used shiny framework, namely the R package that can be used to build web applications. The usage of shiny for user interface development is because the shiny framework can be directly integrated with those methods are written in R, and also it provides some highly variable widgets, especially for displaying tables, plots and graphs, and can be written with or without javascript. Shiny framework also allows development of reactive applications, without the need to reload the page, when one of the parameters in the application is changed, then the application will react directly. Menus are provided in FAST including:

Home

Home is a start screen that displays key information comprising a link to the results of the analysis into the most popular thread, but it also contained a link to the latest analysis and general information about FAST itself.Home can only be accessed using an internet connection, this menu requires access to the database to display the results of the analysis especially for the latest and most popular categories. Without access to the database, this menu can still be accessed, with the limitations of some parts that do not appear.

Forum

Forum is a menu to get into the media discussion. It provides several categories according to the analysis tools are available on the statistical application and other categories based on commonly used statistical programs such as SPSS, SAS, and STATA. Moreover, it provides a feature for users who are logged in to write a thread or article based on the results of the analysis conducted on the statistical application available. Then after a user wrote a thread, also given the facility to comment on the thread for other registered users. Therefore, it can create two-way communication that speeds up the process of knowledge dissemination. Opening this menu require access to the forum database. Without access to the database, it can only be seen up to a category level, users can not see the threads that have been posted.

Table Generator

Menu table generator is provided to accommodate the needs of the data, particularly for strategic indicators from data provided by the Badan Pusat Statistik (BPS), such as employment, demographics, and so forth (Fitriani, 2011). Users who want to analyze the available data in the table generator can download it directly by adjusting dimensions such as year and region according to their individual needs. The data have downloaded will be in Microsoft Excel format, so as to further purposes of the user, they can change it easily.To access the table generator, users need to access the database that require special configuration, needed usage of php with version 5.3.2 for this table generator database.

Without access to the database as well as access to databases that do not meet the specific settings, menu generator table is not accessible.

Analysis Gallery

This menu accommodates a collection of the results of the analysis conducted by the users. Analysis gallery can be a reference as well as information regarding the analysis that can be done to the possibilities of data. Analysis gallery is also expected to bring innovation to the analysis tools used to enrich the knowledge and to provide whatever information analysis tools that can be used as an alternative to analyze the data. Of course, the data can be analyzed using one or more statistical analysis tools are available, therefore, the gallery analysis is designed to enable users to search for inspiration as well as help the beginners to choose the analysis tools can be used to analyze the data. Access to analysisgallery requires access to the database.Without access to the database, this menu can still be opened, but it will not appear on the analysis that has been done by the user.

Login

This facilitates the login menu for users to be able to go in and write a thread in the forum, because only registered users who can write and give comment in the forum. This course is designed with security considerations, built to identifying users and their constraints. In addition, with this policy, administrators can control the correctness of the user-written article, as well as give rank on users who active in writing thread.Access to the login menu does not require access to the database. However, if the user wants to log in, then the connection to the database is a must.

Analysis

This is a menu to enter into statistical application. Statistical application in FAST accomodates several analysis tools ranging from descriptive statistics, linear regression (OLS), ridge regression, logit, probit, and tobit regression, forecasting, cluster, as well as survival. Each of these analysis tools can be used to analyze data in accordance with the conditions and benefits of each analysis tool.Access to the application can be done with or without a login. And can be opened with or without access to the database. When there is no access to the database, then the share feature does not work, but other features can be implemented as intended. Here's an explanation regarding some of the analysis tools available in statistical applications:

a. **Descriptive statistics**

   Descriptive statistics are used simply to describe the sample you are concerned with. They are used in the first instance to get a feel for the data, in the second for use in the statistical tests themselves, and in the third to indicate error associated with results and graphical output. Two general types used to describe the data are

   Measures of central tendency, these are ways of describing the central position of a frequency distribution for a group of data. In this case, the frequency distribution is simply the distribution and pattern of marks scored by the 100 students from the lowest to the highest. We can describe this central position using a number of statistics, including the mode, median, and mean.

Measures of spread: these are ways of summarizing a group of data by describing how spread out the scores are. For example, the mean score of our 100 students may be 65 out of 100. However, not all students will have scored 65 marks. Rather, their scores will be spread out. Some will be lower and others higher. Measures of spread help us to summarize how spread out these scores are. To describe this spread, a number of statistics are available to us, including the range, quartiles, absolute deviation, variance, standard deviation, skewness, and kurtosis.

The descriptive statistics module is built using stats package available in R, and some of additional packages such asggplot2[4], which help in displaying graphs and plots.

b. Linear Regression (OLS)

Linear regression is a regression in which the dependent variable *(Y)* linked or described in one or more than one variable, maybe two, three and so on independent variables (x, x_1, x_2……..x_n) but still showed a linear relationship diagram. The addition of independent variables is expected to better explain the characteristics of the relationships that exist even though there is still some variables are neglected. OLS consider some of classic assumption as shown at Gujarati (2004).

The OLS module is built in the statistical application using MASS[5] package available in R as well as shiny user interface framework. MASS package accommodates almost all of the functions related to the regression, of course with some necessary modifications.

c. Ridge Regression

Ridge regression is one method used to reduce the effects of multicollinearity by modifying the method of least squares or Ordinary Least Square (OLS). Modifications on Ridge Regression method in question is the value of the independent variable. Ridge transformed first by centering and scaling procedures. Then on the main diagonal correlation matrix of independent variables added constant bias (k) or also called lambda (λ). In general, the value of λ is between 0 and 1. When k = 0, Ridge Regression estimators will be equal to the OLS. When λ > 0, Ridge Regression estimators will be biased, but more stable than the OLS (Neter, 1989). Although biased, the estimates produced by Ridge Regression tend to have smaller MSE than the OLS estimates that more precision.

According to Neter et al. (1989), if a sufficient estimator has a small bias, but more precise than an unbiased estimator, the estimator is likely to be chosen because it has a greater likelihood of large to approach the parameters. Figure 2 illustrates the situation. The estimator b is not biased but less precision, while b^R closer than the actual value estimator b.

[4]cran.r-project.org/package=ggplot2
[5]cran.r-project.org/package=MASS

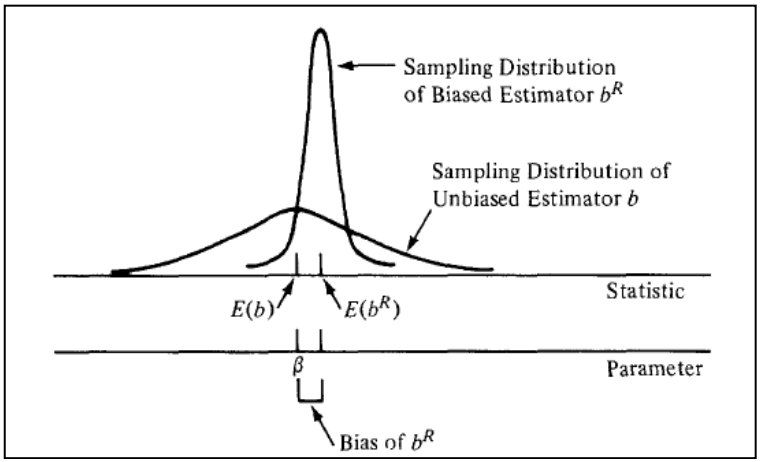


Figure 2: Comparison of estimates with OLS Estimates Ridge (Neter, 1989)

Ridge regression module built using MASS, ridge[6], and car package which are available in R with some necessary modifications. Ridge package accommodates almost all functions to find the parameters of ridge regression, besides the results of the calculation of the functions available in this package are similar with the results obtained when the parameters calculated manually.

d. Logistic Regression

In social science research, categorical data are often collected through surveys. Categorical data consist of nominal and ordinal variables, they take only a view values that do not have a metric. Logistic regression analysis is a method that used to examine the relationship between the response variable which is categorical data with explanatory variables consisting of categorical and numerical data. Logistic regression analysis consists of a binary logistic regression and multinomial logistic regression. Binary logistic regression is used to find the influence of the independent variables with the response variable is binary or dichotomous (Hosmer dan Lemeshow, 2000). The output of the response variable consists of two categories: success and failure that are denoted by y=1 (success) dan y=0 (failed). Whereas, multinomial logistic regression involving the response variable that have more than two categories.

Hosmer and Lemeshow (2000) formulate a common form of probability binary logistic regression models formulated as follows:

$$\Pi(x) = \frac{exp(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)}{1+exp(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)} \quad (1)$$

where:

$\pi(x)$: Probabilty of successful events that when Y = 1.

[6]cran.r-project.org/package=ridge

β_j: Logistic regression coefficient / parameter values with j=1,2,…,p

p: The number of the parameters in the model

For multinomial logistic regression, a response variable with j categories will form logit equation as much j-1 where each of these equations form a binary logistic regression comparing a group of categories against the reference category. In general, the steps are performed in a logistic regression analysis are:

1. Testing parameters simultaneously to determine the suitability of the model.
2. Testing parameters partially to determine the independent variables that influence in the model.
3. Interpretation of the value of the odds ratio.

Logistic regression model built using stats package (glm) function) and net package (multinom) function) which are available in R with some necessary modification.

e. Tobit Regression

The tobit model that was first proposed by Tobin (1958) is for metric response variable and when it is "limited" in the sense we observe it only if it is above or below some cut off level. The tobit model also called censored regression model. Censoring can be from below or from above, also called left and right censoring. The tobit models are defined as follows (Fair, 1977).

$$\boldsymbol{y_i} = \begin{cases} \boldsymbol{y_i}^*, \boldsymbol{y_i}^* > \mathbf{0} \\ \mathbf{0}\ ,\ \boldsymbol{y_i}^* \leq \mathbf{0} \end{cases} \quad (2)$$

where i = 1,2,3,..., n and $\boldsymbol{y_i}^*$ is response varaiable.

Reasoning behind using tobit model are:

1. If we include the censored observations as y = 0, the censored observations on the left will pull down the end of the line, resulting in underestimates of the intercept and over estimates of the slope.
2. If we exclude the censored observations and just use the observations for which y>0 (that is, truncating the sample), it will overestimate the intercept and underestimate the slope.
3. The degree of bias in both will increase as the number of observation that take on the value of zero increase. (see Figure 3)

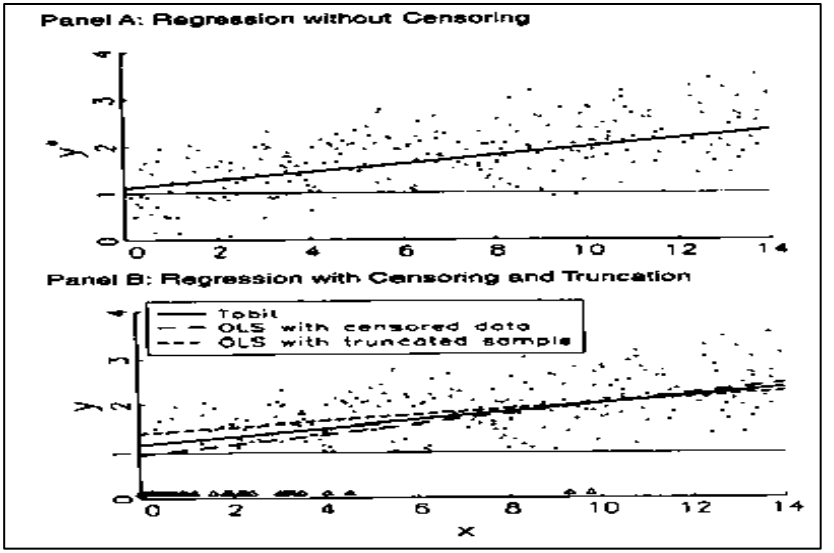


Fig. 3. Linear Regression Model and Without Censoring and Truncation (J.S. Long, 1997)

Tobit regression assumes a normal distribution and homochedasticity of error. If the assumption is violated, the resulting estimates are not consistent (Long, 1997). The steps are performed in the tobit regression analysis are testing parameters simultaneously, testing parameters partially, interpretation of model and testing assumptions. Tobit regression modul built using censReg[7] package which is available in R with some necessary modification.

f. Forecast

ARIMA modelcan be interpretedas a combinationof twomodels, namely theautoregressive model(AR) andmoving average(MA). This modeldoes not haveadifferent variableas theindependent variable, butuse the information inthe sameseriesin theshapemodel (Nachrowi andUsman, 2006). Box-Jenkins methodis used to selectthe appropriateARIMA modelto thetime series dataareused. Makridakis, et.al(1998) saysthat this procedureincludes threestages, identification, forecasting and testing, and application.

g. Cluster

Cluster analysis is a technique used for combining observations into groups or clusters such that each group or cluster is homogenous or compact with respect to certain characteristics. That is, observations in each group are similar to each other. Moreover, each group should be different from other groups with respect to the same characteristics. That is, observations of one group should be different from the observations of other groups(Sharma, 1996).

[7]cran.r-project.org/package=censReg

There are two techniques of clustering namely hierarchical and non-hierarchical (k-means). hierarchical technique is a technique that begins with n clusters, where each observation is considered as a cluster. The distance between clusters is measured in order to obtain a distance matrix of size nxn. Then the two objects of observation with the shortest distance merged into one new cluster. The distance between the new cluster with other clusters recalculated. Repeat these steps as many (n-1) times so that all the objects of observation are members of a cluster. Meanwhile, non-hierarchical technique, known as k-means is a division of observations into subsets (clusters) that do not overlap so that each observation is exactly included in one cluster. This technique begins by determining in advance the k initial cluster centroids randomly. Furthermore, the iterative process is done for grouping observations into clusters that minimizes the sum square error (SSE).

Cluster module built in this statistical applications accommodates the hierarchical method as well as non-hierarchical, utilizing the help of the cluster[8] package available in R, with some modifications in certain parts. Hierarchical methods include agglomerative and divisive, while for non-hierarchical or partitional methods include k-means and pillar k-means (Barakbah, 2009) which is a modification of the k-means. This module can be used to analyze the data you want to see the tendency of cluster will be conducted based on a combination of variables that exist. For example poverty cluster based on expenditure, number of family members, and so forth.

h. Survival

Survival Analysis is a statistical method for data analysis where the outcome variable of interest is the time to the occurrence of an event (Kleinbaum, 1996). Hence, survival analysis is also referred to as "time-to-event" analysis. The presence of censored observation provide complexities led to the development of a new field of statistical methodology. In survival data, only a few individuals experience the event and others do not experience the event until the end of the study. This is the concept of censoring. Censoring occurs when information about survival time available.

There are three methods in survival analysis: non-parametric, semi-parametric, and parametric method. One if the oldest and most straightforward non-parametric methods for analyzing survival data is to compute the life table, which was proposed by Berkson and Gage [3]. Another important development in non-parametric analysis methods was obtained by Kaplan and Meier[9], named Kaplan Meier Estimate (KME). While non-parametric methods work well for homogeneous samples, they do not determine whether or not certain variables are related to the survival time. So, this needs leads to the application of regression methods for analyzing survival data. But, the standard multiple linear regression model is not well suited to survival data for several reasons. Firstly, survival time are rarely normally distributed. Secondly, censored data result in missing values for the dependent variable (survival time) [10].

[8] cran.r-project.org/package=cluster

The Cox Proportional Hazards model is now the most widely used for the analysis of survival data in the presence of covariates or prognostic factors. This is because of its simplicity, and not being based on any assumptions about the survival distribution. The model assumes that the underlying hazard rate is a function of the independent covariates, but no assumptions are made about the shape of the hazard function[9].The Cox PH is known as the semi-parametric method in survival analysis.

The Cox PH model may not be appropriate in many situations such as when the proportional hazard assumption is not tenable and other modification such as stratified Cox model or Cox model with time-dependent variables. In return, the parametric can be alternative methods for the analysis of the survival data. There are two model in parametric methods: parametric Proportional Hazard (PH) model and Accelerated Failure Time (AFT) model.

PH model is known as the parametric version of the Cox PH model.The key difference between the two kinds of models is that the baseline hazard function is assumed to follow a specific distribution when a fully parametric PH model is fitted to the data, whereas the Cox model has no such constraint. The coefficients are estimated by partial likelihood in Cox model but maximum likelihood in parametric PH model.The commonly applied models are exponential, Weibull, or Gompertz models.

Although parametric PH models are very applicable to analyze survival data, there are relatively few probability distribution for the survival time that can be used with these models. In these situation, the AFT model is an alternative to the PH model for the analysis of survival time data[4]. Under AFT models, the direct effect of the explanatory variables based on the survival time instead of hazard. This characteristic allows for an easier interpretation of the resultsbecause the parameters measure the effect of the correspondent covariate on the meansurvival time. Unfortunately, recently the AFT model is not commonly used for the analysis of clinical trial data, although it is fairly common in the field of manufacturing. Similar to the PHmodel, the AFT model describes the relationship between survival probabilities and a setof covariates.Under an AFT model, the covariate effects are assumed to be constant and multiplicative on the time scale, that is, the covariate impacts on survival by aconstant factor (acceleration factor).AFT models are fitted using the maximum likelihood method.Commonly applied models are exponential, weibull, log-logistic, log-normal, or gamma models. Summary of commonly applied parametric models can be seen in Figure 4.

[9]Hazard function gives the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t.

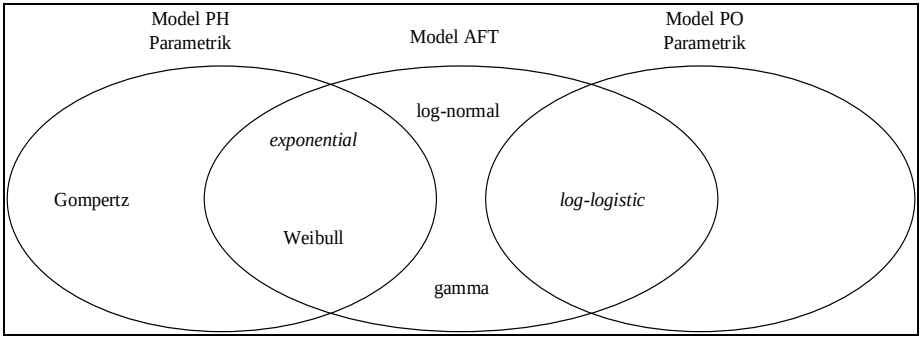


Fig. 4. Summary of Parametric Models

Survival module built in this statistical applications provide the parametric methods to analyze survival data, by utilize the survival[10] package in R. This module focus on parametric AFT model, include model checking by using statistical criteria to compare all these AFT models. This module is using the Akaike information criterion (AIC) instead of likelihood ratio (LR) test, sincethe non-nested models cannot be compared using LR test. Non-nested model in the commonly applied in parametric survival is gamma model. It is because the exponential, Weibull, and log-normal model are nested within gamma model. The smaller AIC is the better model.

3. IMPLEMENTATION

To access the application, users need to access the main page of FAST[11]. This page can be accessed with or without logging in, because all users can view the content of this page, which are general informations that need to be known. After that the user can access the available menus. To access the application, users need to select the analysis menu. In this application there are also a few menu items such as data, and some menus that accommodate the tools provided such as regression analysis, cluster, forecast, and survival. Data menu is an initial view when the application menu is accessed, which can be seen in figure 6. As default view will be provided the initial data, which is an example data called diamond with records of three thousand which areabout prices, carats, and other details of the diamond. This data can also be used for experimental analysis or to simply learn the types of data that can be used for analysis. In some menu tabs available data relating to the processing of the data itself, as the beginning and the default view when first time access to data is the manage tab. This tab provides facilities for data uploads in rdaand csv type, and paste feature of the data supplied from the microsoft excel worksheet. This feature is provided to cope manually data edit feature which is not provided in this application. In addition, also provided example data that can be loaded by the user.

[10]cran.r-project.org/package=survival
[11]http://rb.bps.go.id/fast

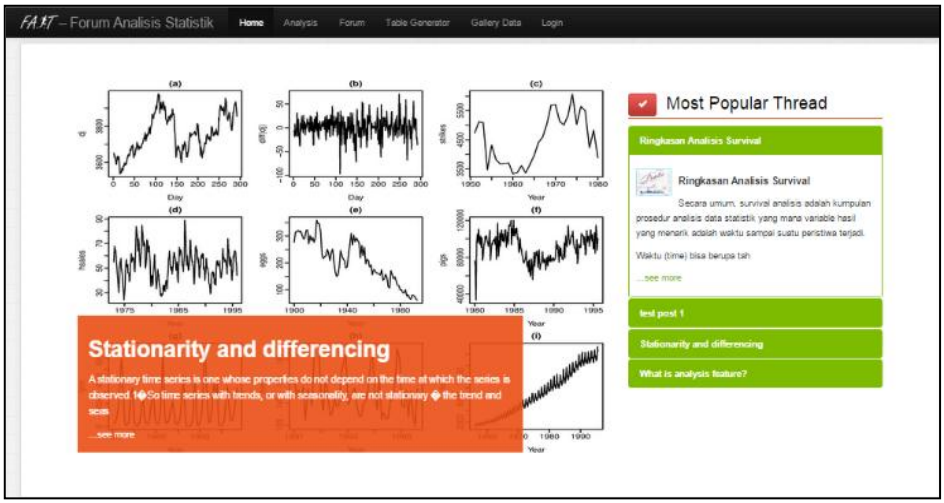


Fig. 5. FAST Main Page

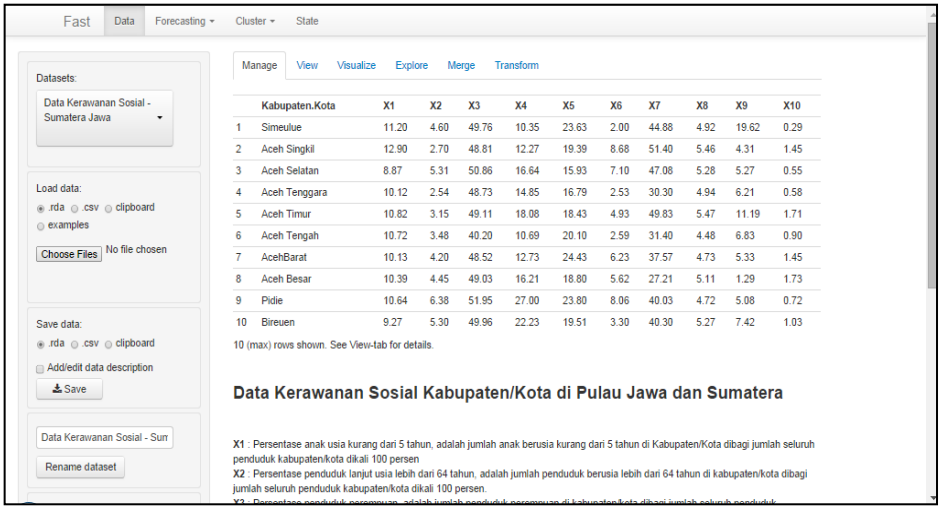


Fig. 6. Statistical Application of FAST

In addition to loading the data, the user can also save the data from the application to the rda or csv type. Or users can choose to copy the data to microsoft excelworksheet. Before storing the available data, users can add or edit the description of the data which can contain information of the data,apart from the contents. Then this data can be saved with rda-type, because this type can accommodate other variables in the form of a separate statement of core data.

In the manage tab, users can only view the top ten records of the data. If the user wants to see the complete data, the user can select the viewtab. In this tab, users can perform sorting of data based on certain variables, determine the number of records you

want to display, or conduct a search of the desired record. Furthermore, there is a visualize tab which present the data as the chart such as line chart or scatter plot which can be used for the initial analysis of the data. Explore tab accomodatesdata grouping based on certain variables. Merge tab accommodates the merging of data that has the same column names. Lastlytransform tab accomodates data transformation with change that have been provided or specified by the user.

Furthermore, when the predetermined data has beenchoosen, the user can continue the analysis by selecting the desired analysis tool. For example, an explanation will be given for cluster analysis.

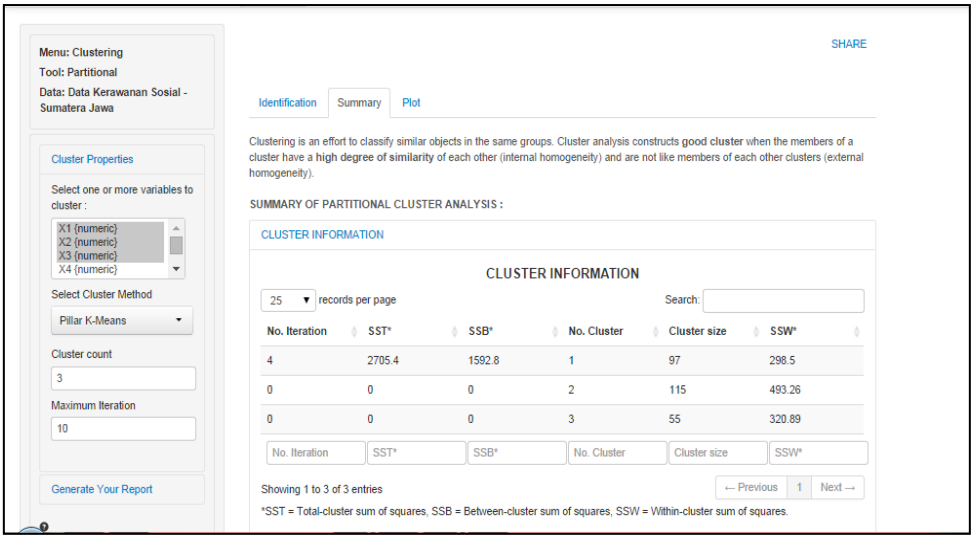


Figure 7. Cluster Analysis in FAST

The cluster menu consists of hierarchical and partitional sub-tabs, at this time will be explained about partitional sub-tab. It provides three tabs which can be used to analyze the data. First is the identification tab, namely identification of facilities to accommodate the data, in this case is to determine the optimum number of clusters can be formed, as well as the identification of outliers that can affect the results of clustering. Furthermore, the second tab is the summary which accommodates the results of cluster analysis. Note that the left panel contains the parameters supplied in the form of variables, the method used, the number of clusters, and the maximum number of iterations, which can be changed and give the results in the right panel, which is reactive when there is a change of one of the available parameters.

On the summary tab, there are three parts are shown, namely the cluster information that contains information such as the number of iterations of the cluster, the value of total cluster sum of squares (SST), the value of the between cluster sum of squares (SSB), the value of the within cluster sum of squares, and the size of the cluster. Then the cluster centers serving middle value of each cluster of each variable. Last is cluster membership serving where the membership of each observation.

Furthermore, the last tab is the plot tab that displays the results of the analysis in the form of images. Cluster plots show clustering results in the form of a scatter plot with different colors for each cluster formed. Then also described the midpoint for each of the clusters formed.

In each menu analysis tools are provided, the help facility is also available that can help the user to analyze data. Help facility are presented with examples of cases and how to use analysis tools. To simplify the user, in this application provided a feature to generate reports automatically in ms. word, pdf, and ms. Excel format. Then there is a facilitiy that connect these applications to the gallery of analysis, namely share. This facility accommodates the spread of knowledge from the results of the analysis have been done by users. When users want to deploy hail the analysis, the user can choose which parts they want to spread, and provide an explanation of what they had done.

Switch from the menu analysis, users can select the analysis gallery to see the results of the analysis that have been conducted and disseminated. Then if users want to write a thread in the forum, they can use the results of previously generated reports into a ms. Wordformat. Users can upload an attachment in the form of images or explanation about results of the analysis that have been done. To write or leave a comment in the forum, the user must log in first.

4. FUTURE EXTENSIONS

Some features have not accommodated yet are manually edit data, and the addition of the analysis tools provided in statistical application. So far the analysis tools available are still limited, it is expected in the future, this application also includes computational tools that can be used for soft computing such as artificial neural networks, genetic algorithms, etc.

5. COMMENTS AND CONCLUSION

FAST is an internet forum that is integrated with the statistical applications that can be used to analyze the data, as well as disseminate the resultsto the forumto share knowledge. Availableapplication is a web-based application that is reactive to changes in parameters occurred. Although there are still some shortcomings in this application, but the development of a better direction is being done in order to achieve optimal results.

REFERENCES

1. Agresti, Alan. (1990). *Categorical Data Analysis*. New York: John Wiley.
2. Barakbah, A.R. and Kiyoki, Y. (2009). A Fast Algorithm for K-Means Optimization by Distance Maximization for Initial Centroid Designation. *International Journal of Information and Communication Engineering*. 6, 85-88.
3. Berkson, J. and Gage, R.P. (1950).Calculation of survival rates for cancer. *Proceedings of the Staff Meetings of the Mayo Clinic.* 25, 270-286.
4. Cox, C. (1988). Delta method. P. Armitage and T. Colton, Eds. *In Encyclopedia of Biostatistics*. 2, 1125-1127.
5. Damodar, G. (2003). *Basic Econometrics.* 4th Ed. New York: Mc. Graw Hill.

6. Fair, R.C. (1977). A Note on the Computation of the Tobit Estimator. *Journal Econometrica*. 45, 1723-1727.
7. Henningsen, A. (2013). *CensReg: Censored Regression (Tobit) Models*. R package version 0.5,http://CRAN.R-project.org/package=censReg.
8. Hosmer, D.W. and Stanley, L. (2000). *Applied Logistic Regression*. 2nd ed. New York, USA: John Wiley & Sons.
9. Kaplan, E. and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *J. Amer. Statist. Assoc.*, 53, 457-481.
10. Kleinbaum, D.G. (1996). *Survival Analysis: A Self-Learning Text*. New York: Springer.
11. Long, J.S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. United States of Amerika: Sage Publications.
12. Makridakis, W. and Hyndman (1998). Forecasting: Method and Application. 3rd ed. New York, USA: John Wiley.
13. Nachrowi, N.D. and Usman, H. (2006). *Ekonometrika untuk analisis ekonomi dan keuangan.* Jakarta: Fakultas Ekonomi UI.
14. Neter, J. Wasserman, W. and Kutner, M.H. (1989). *Applied Linear Regression Models.* 2nd Ed. Boston: IRWIN.
15. Ripley, B. (2014). *nnet: Feed-forward Neural Networks and Multinomial Log-Linear Models.* R package version 7.3, http://CRAN.R-project.org/package=nnet.
16. Sharma, S. (1996). *Applied Multivariate Techniques.* Canada: John Wiley & Son, Inc.
17. Tobin, J. (1958). Estimation of Relationship for Limited Dependent Variable. *Econometrica,* 26(1), 24-36.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 201-212

WIRES: A USER FRIENDLY SPATIAL ANALYSIS SOFTWARE

Meidiana Pairuz[1], Nur'aidah[1], Hanik Devianingrum[1], Hergias Widityasari[1], Zumrotul Ilmiyah[1], Isna Rahayu[2], Arinda Ria[2], Diah Daniaty[2], Wahyu Hardi Puspiaji[2], Erma Purnatika Dewi[2], Pudji Ismartini[2], Robert Kurniawan[1], Sodikin Baidowi[2], Muchammad Romzi[2], Munawar Asikin[3], Karmaji[2] and **Setia Pramana[1]**

[1] Sekolah Tinggi Ilmu Statistik, Jl. Otista 64C, Jakarta, Indonesia
[2] Badan Pusat Statistik, Jl. Dr. Sutomo 6-8, Jakarta, Indonesia
[3] Pusdiklat BPS, Jl. Raya Jagakarsa No 70, Lenteng Agung Jakarta, Indonesia
Email: setia.pramana@stis.ac.id

ABSTRACT

Spatial analysis is gaining more popularity since Anselin in 1989 defined the newly born field as "The collection of techniques concerning the peculiarities caused by space in the statistical analysis of models on regional sciences" (Billé and Arbia, 2013). Many statistical methods and techniques have been developed for the analysis, however there are still few software available, such as GeoDa, GWR, ArcGIS, and several packages of R. Those software still have some limitation. GeoDa, GWR, and ArcGIS are limited to some spatial methods, while R covers more methods but still uses command line interface, which will be difficult for non-expert users. To overcome this limitation, WIRES, a GUI spatial analytical tool, is developed using C# and R as back-end program. WIRES covers more features, spatial weight matrix generator, exploratory spatial data analysis, spatial regression, explanatory spatial data analysis, spatial logistic regression, ordinary and simple Kriging, and other descriptive tools. WIRES also covers methods which have not been covered by other spatial analytical tools, i.e. spatial cluster, spatial regional inequality (using Gini and Theil method), and spatial shift-share.

KEYWORDS

Spatial Analysis, Graphical User Interface, C#, R.

1. INTRODUCTION

The use of statistics has been applied almost in all field including business, economics, biology, physics, astronomy, meteorology, chemistry, medicine, sociology, psychology and so on. For years, conventional statistical theory based on assumption that among observations are independent. But since Tobler (1970) came with his publication, the researchers finally realize that there are some situations where observations are not independent anymore but correlate in space, for example. This spatial correlation can't be overcome by conventional statistics method because it can lead to inconsistent estimator (Case in Bhat and Sener, 2009). Because of the

conventional statistics' limitation in inferencing spatial data, the spatial analysis exists to overcome the limitation.

Spatial statistics has triggered many new methods which are the extension of conventional statistics to appear. Despite that fact, recent technologies and global positioning system also have led to a rapid expansion of the number and size of available spatially dataset (Rosenberg and Anderson, 2011). To enable and facilitate spatial analysis on those spatial data, several spatial analysis tools were built.

Today some spatial analytical tools are now available to carry out a wide array of spatial data analytical techniques, they have been created as modules for programming languages, statistical platforms, geographic information systems (GIS) or as standalone software. For example, pySAL is a library of Python intended to aid the development of applications for spatial analysis, R is a statistical languages that can be used to write spatial analysis packages (such as gstat, spgwr, ade4, and so on), ArcGIS is popular for spatial data storage, visualization, and offers a handful of commonly used analysis functions in Spatial Statistics toolbox, PASSage 2 is a software for pattern exploration and spatial description, GeoDa is a free software that focuses in mapping and exploring spatial data, and so on (Rosenberg and Anderson, 2011).

However, the available spatial software still have some limitations. For example, ArcGIS is powerful software but is not free of charge and only cover some statistical modeling. Open GeoDa and PASSage 2 only focus in mapping and exploring spatial data, and not cover any spatial inference method. R covers more statistical techniques and modeling, but based on command line interface, which forces the user to understand its syntax before using it, of course it will be difficult for non-expert user who used to use graphical user interface (GUI)-based software.

All those limitations have delivered to the need of free, user friendly, and more complete spatial statistical software in describing and inferencing/modeling spatial data. That's why we build WIRES, a GUI-based spatial analytical software. WIRES is intended to help analyzing spatial data not only for expert user but also for non-expert user with its easy to use graphical interface. To describe more about WIRES, in section 2 we will explain about some spatial analysis covered by WIRES, in section 3 we will explain about WIRES, including its architecture and how we implement it into some interfaces, and in section 4 we will give case study to show how to use WIRES to analyze spatial data.

2.1 Spatial Statistics

Conventional statistics has been built on assumption that among observations are independent. This can be in contradiction with the fact that there are some situations were among observations are not independent but correlate in space, such as the low income-region tends to clustered with another low income-regions (Miller, 2004), the auto supplier plants prefer to locate in areas close to other suppliers (Klier and McMillen, 2008), and so on. To analyze the observations that are correlate in space, we can't use standard statistical methods because it can lead to inconsistent estimator (Case in Bhat and Sener, 2009). Anselin (1996) also stated that spatial autocorrelation in spatial data makes the data does not fulfill the assumption in conventional statistics which assumed the observations are independent. The consequence, the inference using conventional

statistics is not efficient anymore. Because of this limitation, the spatial analysis exists to overcome the limitation.

Spatial statistics is the collection of techniques and models in which spatial referencing of each data play an explicit role in the analysis of data (Goodchild and Haining, 2004). Spatial statistics can be used to describe spatial patterns (exploration) (Anselin, 2004), test hypothesis about patterns (inference) (Sherman, 2010), and predicting value in unobserved location (Fischer and Getis, 2010). There are many spatial statistics methods has been developed, in this paper and software we built we would cover some basic and important methods, including exploratory spatial data analysis, spatial weight, spatial clustering, spatial regression, spatial logistic regression, kriging, inequality region, and shift share.

2.2 Exploratory Spatial Data Analysis

The calculation of spatial autocorrelation is an important step in spatial analysis to detect the spatial structure of the data to be analyzed. Moran's I (Moran,1948) and Geary's C (Geary, 1954) can be used to detect overall/global spatial autocorrelation in the data that consists of one or two variables. Moran's I behaves like Pearson correlation coefficient, in case its value is between -1 and 1. Geary's C is similar to a distance-type structure function and is within the range 0 to 2. The difference of those two statistics is Moran's I is based on cross-products to measure value association, while Geary's C employs squared difference (Kalkhan, 2011).

Both Moran's I and Geary's C are statistics used to calculate overall/global spatial autocorrelation, but cannot be used to identify spatial autocorrelation in smaller range. For that purpose, Moran Scatterplot is used to check local spatial instability and Local Indicators of Spatial Association (LISA). According to Anselin (1995), LISA's value of each observations indicate how significant is spatial clustering among observation, and the sum of LISA's value of each observations is equal to indicator of global spatial association.

Moran, Geary and LISA are intended to measure spatial autocorrelation of data with one or two variable. For more than two variables, we can use MULTISPATI (Multivariate spatial analysis based on Moran's I) method which the idea is found by Wartenberg (2008) then is developed by Dray, Saïd, and Debias (2008).

2.3 Spatial Weight Matrix

Spatial weight matrix is an important aspect in spatial analysis. This matrix defines spatial interaction among observation (Anselin in Getis and Aldstadt, 2010). In practices, weight matrix which frequently used is a standardized matrix, therefore if we sum all the element in its row is must be one. There are some spatial weight matrix which is well known, including contiguity matrix, distance based-matrix, and k nearest neighbor matrix (Getis and Aldstadt, 2010).

2.4 Spatial Clustering

The collaboration of exploratory spatial data analysis and data mining results many new powerful methods that can enhance the analysis of data itself, one of them is spatial clustering. Spatial clustering is a method to cluster object by considering its spatial attribute and identify the closeness of observations in space (Cao, et al. 2013). Carvalho,

et al. (2009) showed how important spatial clustering is by stated that *"Specifically concerning geographical data, spatial clustering is a powerful technique that can be adapt to the most varied cases and thus has developed quickly and become increasingly popular"*. In his research, Carvalho, et al. (2009) modify hierarchical clustering algorithm by considering the interaction among observations in space, this resulted new method called Spatial Hierarchical Clustering.

2.5 Spatial Regression

Regression is an important statistical method to examine the relationship between a variable of interest (dependent variable) and one or more explanatory variables (predictors). One of the regression assumptions is independence of observations. If this doesn't hold, we obtain inaccurate estimates of the coefficients, and the error term. Spatial dependency is one of reason that can this problem. If spatial dependency occurs, the observations are not independent anymore but correlate in space. To overcome this problem, spatial regression method is developed such as:

- **Spatial lag or spatial autoregressive (SAR) model**. This model includes spatial lag dependent variabel (WY) used as dependent variable.
- **Spatial error model.** This model includes spatial autocorrelation occurs in random error ($W\varepsilon$) used as independent variable.

2.6 Spatial Logistic Regression

Spatial logistic regression is an analysis to model the association between explanatory variables and discrete response variable by considering spatial autocorrelation in characteristics of observations. Spatial autocorrelation which gained a serious attention is spatial autocorrelation in latent response variable(spatial lag) which shows that logit value of an observations is influenced by logit value of neighbor observations, because it can't be handled by classical logistic regression which assumed that among observations are independent and some ways to solve it needs many times iterations and inversion of nxn matrix to get the estimated coefficients, which means it only can be used for small sample (Klier and McMillen, 2008).To overcome this situation, Klier and McMillen (2008) developed spatial logistic regression using linearized GMM. By using linearized GMM, estimated regression coefficients can be calculated in short time and result in more accurate statistics than other method such as non-linearized GMM developed by Pinkse and Slade (1998).

2.7 Regional Inequality

Income inequality is a condition where income is not equally distributed to all society in a region/country (Glaeser, 2006). It is also an effect caused by relative poverty (BPS, 2008). Regional inequality analysis is an analysis used to decide whether government succeed to decrease poverty rate in his country. There are some regional inequality methods based on average regional income relative to national income, but they are fail to explain the distribution of national individual income or the dispersion of income in a region (Metwally and Jensen, 1973). To overcome this, Akita (2000) develops regional inequality using decomposed Theil Index, which can explain regional inequality completely. Allinson (1978) also develops regional inequality analysis using Gini Index. Gini Index is a good index, because of its sensitive to transfer, that means the observations with income around average income will have higher effect in inequality calculations than the observations which is not.

2.8 Spatial Shift-Share

Shift-share analysis is an analysis used to compare growing rate in many kind of sectors between regency and province or between province and national (Tarigan, 2012). The purpose of this analysis is giving information about economy and man power in 3 sectors, those are: regional economy growth, industry-mix/proportional-shift, and regional-shift/differential shift.

In classical shift-share, the regions are considered to be independent. In fact, this situation could be really wrong because the regions are not isolated and a region could really dependent economically with another region. This idea is then used by Nazara and Hewings (2004) to modified classical shift share developed by Dunn (1960) for spatial shift-share. This modification is needed to consider the positive/negative interaction between one region and another regions.

2.9 Kriging

Sometimes we find that not all region in unit we want to observe has value of a variable. Whereas complete information of that variable is needed to perform the analysis. To get the missing value, we can use interpolation techniques. Interpolation is the process of finding unknown values from known values (Sunitha et. al., 2013).

Kriging is one of interpolation method to estimate a missing or unobserved value in a target region based on region around it. To estimate a missing value, Kriging considers spatial and non-spatial factors which can increase the accuracy such as number of sample, coordinate of sample, distance between sample and target point, and spatial continuity of involved variables. Compared to other method such as Invers Distance Weighted (IDW), triangulation, polygon, an others, Kriging estimation can be accounted statistically because it gives BLUE (Best Linear Unbiased Estimator) (Bohling, 2005).

3 WIRES

In 2012, the development of a spatial statistical software to analyze spatial data named WIRES was started by five students of Sekolah Tinggi Ilmu Statistik (STIS), Indonesia. At that time WIRES provides tools for creating weight, and analysis including spatial regression, spatial shift-share, classical shift share, exploratory spatial data (univariate Moran's, bivariate moran's, Geary's C and LISA), and inequality region. The purpose of WIRES is to make a user friendly GUI-based software.The technology behind WIRES is a combination of Java, C# language and R-software. WIRES uses R to call package of analysis that have been covered in R output (like spatial regression, spatial weight, and exploratory spatial data) and present it in a new and easy interpreted, and also to calculate analysis that have not been covered in R and other statistical software, like inequality region using Theil index and spatial shift-share.

Recently, several new analysis such as spatial logistic regression, multivariate spatial analysis using Moran's I, spatial hierarchical clustering, inequality region using Gini index, kriging, and geographically weighted regression were added to extend WIRES capabilities by other STIS' students. Several descriptive tools like quantile/percentile map, discrete map, equal map, and box map have been included as well.

3.1 Architecture of WIRES

Based on figure 1, architecture of WIRES consists of two main components, user interface component and analysis component.

1) User interface component

   For map visualization, we use DotSpatial 1.7 from MapWindow because it is the new project of MapWindow GIS that has combined all the best component in open souce application in .NET GIS like SharpMap, Proj.NET, and MapWindow 6, so it is more complete and easier in executing the available functions.

   For graph visualization, we use ZedGraph. ZedGraph is a class library, user control, and web control for .net, written in C#, for drawing 2D Line, Bar, and Pie Charts. It features full, detailed customization capabilities, but most options have defaults for ease of use.

   While for importing excel library, we use NPOI library

2) Analysis component

   All the analysis use R as back-end program, by using R we can invoke function that has been covered by R and also compute analysis that has not been covered by R with complete array and matrix operations in R. All these simplicity makes us can focus in designing the new and easier interpreted output for each analysis. Table 1 show our primary analysis and R packages that we use for those analysis.

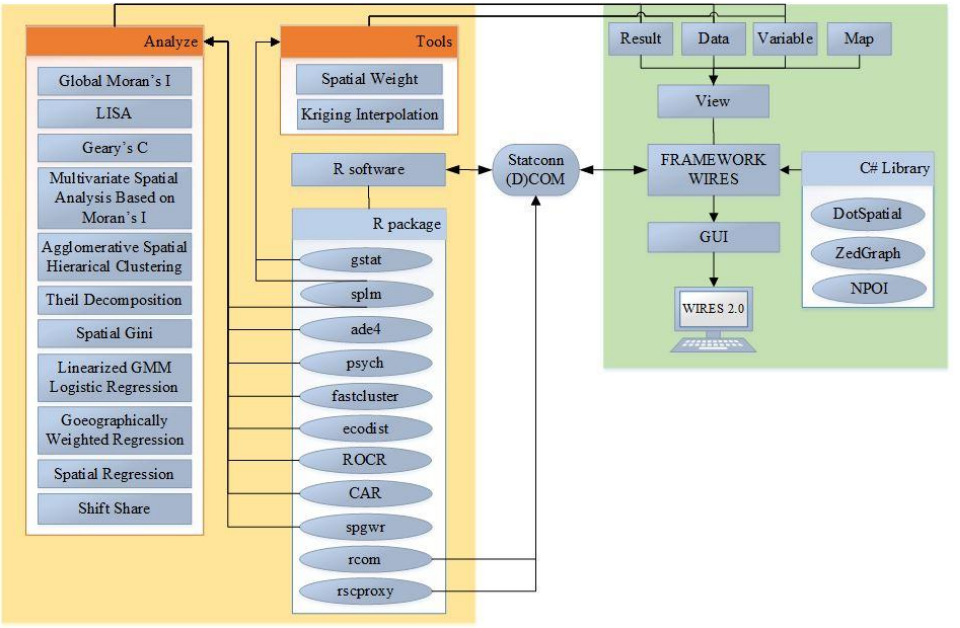


Fig. 1: Architecture of WIRES

Table 1:
WIRES's Primary Analysis and R Packages Used

Analysis	R packages
Explaratory spatial data analysis (Global Moran's I, LISA, and Geary's C)**	splm
Multivariate spatial analysis based on Moran's I (MULTISPATI)**	ade4, psych
Agglomerative spatial hierarchical clustering*	fastcluster
Inequality region using theil index*	-
Inequality region using Gini index*	-
Linearized GMM logistic regression ***	CAR, ROCR
Spatial regression*	-
Shift share*	-
Kriging**	Gstat

* Has not been covered by any other spatial software
** Has been covered by R, enhanced in output
*** Has been covered by R, enhanced in analysis and output

To connect user interface component and analysis component we use Statconn (D)COM developed by Baier dan Neuwirth (2010) and some packages in R including rconn and rscproxy. Statconn (D)COM is a middleware component that integrates R software to application by providing DCOM component for R.

3.2 Implementation of WIRES

In this part, we will explain some important WIRES' dialogs/screens that explain how we implement our designed architecture into desktop application. The first screen is a screen that appear when user RUN WIRES, it consist of three main functions, create new WIRES project, open WIRES project, or open recent WIRES project like in Figure 2.

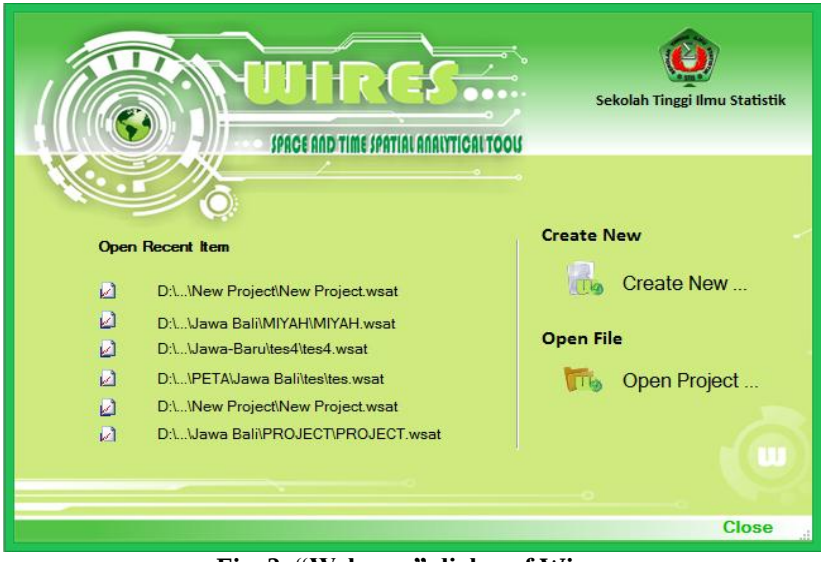


Fig. 2: "Welcome" dialog of Wires

When user choose to create new project, user will be asked to choose the shapefile with extension *.shp and variable that become identity of data that will be analyzed (Figure 3).

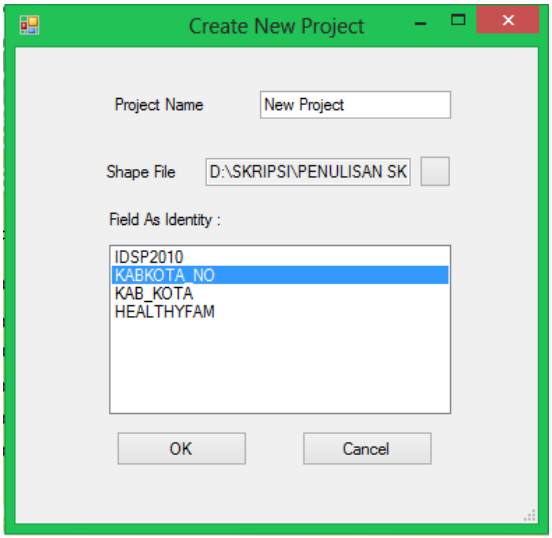
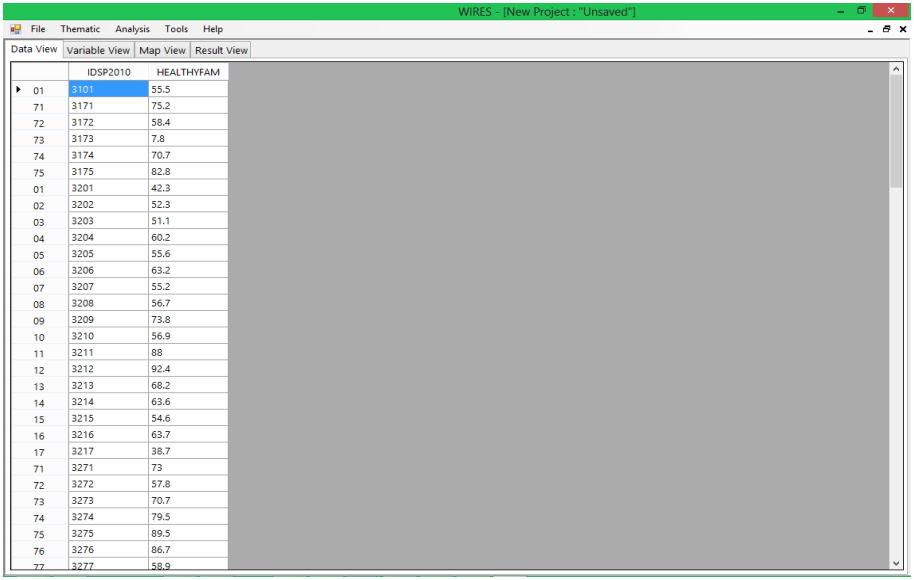


Fig. 3: Dialog to create new project

After choosing the shapefile, the structure of the chosen data will be shown in main dialog consists of menu bar and tab view (Figure 4). Tab view consists of 4 views: 1) Data view to see and manipulate the data in database file that correlate with shapefile, 2) Variabe view to delete, create, and rename the variable in database file that correlate with shapefile, 3) Map view to see and explore map/shapefile, 4) Result view to see the result of analysis that have been done.

Menu bar consists of 5 main menus, including: 1) File menu consists of submenus for creating new project, opening project, opening recent project, saving project, saving as project, importing variable from excel file, and exiting application, 2) Thematic menu to do some descriptive analysis using map, like see the distribution of variable in a map based on its percentile values, 3) Analysis menu consists of main analysis which have been explained in section 2, except Kriging and creating spatial weight matrix 4) Tools menu consists of Kriging and creating weight submenus, and 5) Help menu to help user getting further and detail explanation about how to use WIRES.



	IDSP2010	HEALTHYFAM
01	3101	55.5
71	3171	75.2
72	3172	58.4
73	3173	7.8
74	3174	70.7
75	3175	82.8
01	3201	42.3
02	3202	52.3
03	3203	51.1
04	3204	60.2
05	3205	55.6
06	3206	63.2
07	3207	55.2
08	3208	56.7
09	3209	73.8
10	3210	56.9
11	3211	88
12	3212	92.4
13	3213	68.2
14	3214	63.6
15	3215	54.6
16	3216	63.7
17	3217	38.7
71	3271	73
72	3272	57.8
73	3273	70.7
74	3274	79.5
75	3275	89.5
76	3276	86.7
77	3277	58.9

Fig. 4: Main dialog of WIRES

4. CALCULATING MORAN'S I USING WIRES

In this section we will show how to calculate Moran's I which reflects spatial autocorrelation of a variable using WIRES. The data we used here are drawn fromIndonesia Health Statistics in 2011, you can download the data including shapefile and database file in https://www.dropbox.com/s/3i4qqjwv20berb8/Jawa%20Bali.zip?dl=0. The observations are all regencies in Jawa and Bali island. Here we want to explore whether there is spatial autocorrelation of percentage of healthy life style-family in observed regencies.

The first step is creating the weight matrix that contains list of neighborhood between one observations to another. To create weight matrix we choose Tools menu, then choose Create Weight, and choose what kind of spatial matrix we want to create. Here we will choose rook contiguity matrix. After that the application will inform whether the weight matrix is successfully created and if succeed it willgive the result tab that consists of 1) map analysis that shows the neighborhood of each observations by connecting the observations with its neighbors with lines, 2) weight matrix that shows the structure of spatial weight matrix, and 3) weight graph that shows graph of observations and their number of neighbours.

After creating the spatial weight matrix, we can calculate Moran's I by select Analysis menu, then choose ESDA, choose Univariate Moran's I, and choose the variable we want to analyze, in this case is percentage of healthy life style-family, then choose the weight we have created. The application will result tab that consist of Moran's I statistics and its scatter plot. Figure 5 and 6 show the result tab for Moran's I.

	Variable(s)	Moran's I	E(I)	Var(I)	p-value	Standard Deviation
▶	HEALTHYFAM	0.1904	-0.0080	0.0045	0.0030	2.9675

Fig. 5: Statistics of Moran's I for Percentage of Healthy Life Style-Family Variable

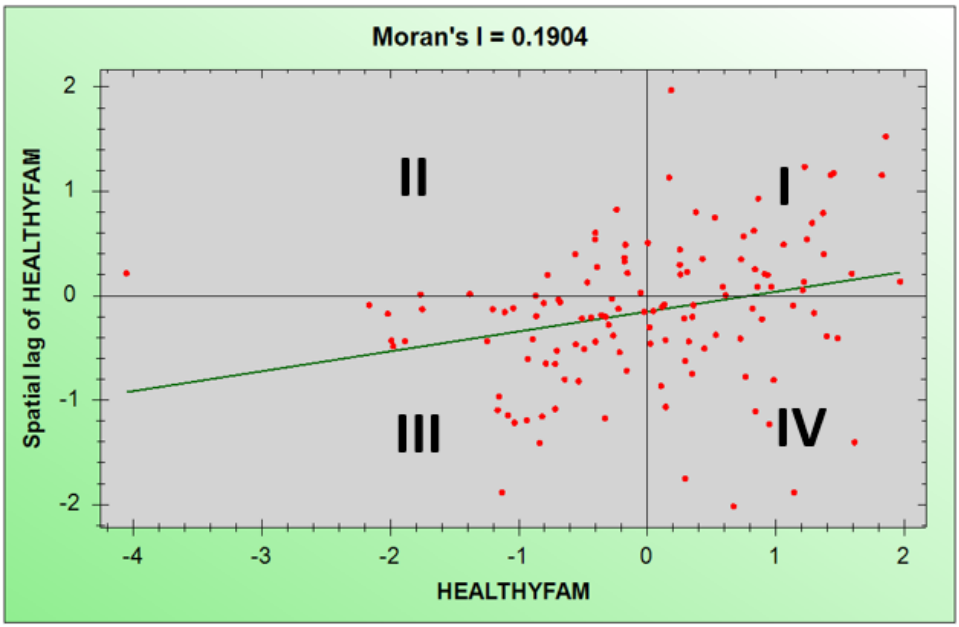


Fig. 6: Scatter Plot of Moran's I for Percentage of Healthy Life Style-Family Variable

From Figure 5 we get the Moran's I for percentage of healthy life style-family variable is 0, 1904 and its p-value is 0.0030. It means that with confidence rate (for example) 95%, we believe that there is positive correlation of percentage of healthy life style-family between regencies. To see the distribution of spatial autocorrelation in each regency, we can see Moran's I scatterplot in Figure 6. Figure 6 shows that many observations lies in quadrant I and III, it means that many regencies with high percentage of healthy life style-family have neighbors that also have high percentage of healthy life style-family, and vice versa.

CONCLUSION

The popularity of spatial analysis and huge number of available spatially dataset have encouraged the need of spatial analysis tools. But the available spatial analysis tools still have some limitations, like not free of charge, only cover descriptive functions, and based on command line interface. That's why we build WIRES.

We build WIRES using C# and powerful R as back-end program, we also use some C# library that enhance the output of WIRES. WIRES covered some tools and analysis to

help expert and non-expert user to do basic and complex spatial analysis. By building WIRES, we have create a free and user friendly application which combine the exploratory functions and inference functions altogether. We also have covered some analysis which have not been covered in any other spatial tools.

REFERENCES

1. Akita, T. (2000). *Decomposing Regional Income Inequality using Two-Stage Nested Theil Decomposition Method* (No. EMS_2000_02).
2. Allinson, P.D. (1978). Measures of Inequality. *American Sosiological Review*, 43(6), 865-880.
3. Anselin, L. (1995). Local Indicators of Spatial Association—LISA. *Geographical Analysis*, 27, 93-115.
4. Anselin, L. (1996). *The Moran Scatterplot as An ESDA Tool to Asses Local Instability in Spatial Association*. In spatial Analytical Perspective on GISA edited by M. Fischer, H. Scholten, and D. Unwin. London: Taylor and Francis.
5. Anselin, L. (1999). *Interactive techniques and exploratory spatial data analysis. Geographical Information Systems: Principles, Techniques, Management and Applications,* eds., P. Longley, M. Goodchild, D. Maguire, and D. Rhind. Cambridge: Geoinformation Int.
6. Anselin, L. (2004). Exploring spatial data with GeoDaTM: a workbook. *Urbana*, *51*, 61801.
7. Anselin, L. (2005). *Spatial Regression Analysis in R–A Workbook*. Urbana, 51, 61801.
8. Baier, T. and Eric, N. (2010). Statconn DCOM, http://rcom.univie.ac.at/.
9. Bhat, C.R. and Ipek N.S. (2009). *A* Copula-Based Closed-Form Binary Logit Choice Model for Accommodating Spatial Correlation Across Observational Units. *Journal of Geographical Systems*, 11(3), 243-272.
10. Billé, A.G. and Arbia, G. (2013). Spatial Discrete Choice and Spatial Limited Dependent Variable Models: A Review with an Emphasis on the Use in Regional Health Economics. *arXiv preprint arXiv*:1302.2267.
11. Bohling, G. (2005). Kriging. *Kansas Geological Survey, Tech. Rep*.
12. Cao, Z., Wang, S., Forestier, G., Puissant, A. and Eick, C.F. (2013). Analyzing the composition of cities using spatial clustering. *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing Article* No, 14, ACM: New York.
13. Carvalho, A.X.Y., Albuquerque, P.H.M., Almeida Junior, G.R. and Guimarães, R.D. (2009). Spatial Hierarchical Clustering. Brasil: Institute for Applied Economics Research (IPEA).
14. Dray, S., Sonia, S., and François, D. (2008). Spatial Ordination of Vegetation Data using a Generalization of Wartenberg's Multivariate Spatial Correlation. *Journal of Vegetation Scene*, Paris, 19, 45-56.
15. Dunn, E.S. (1960). A Statistical and Analytical Technique for Regional Analysis, *Papers of the Regional Science Association*, 6. Hal. 97-112.
16. Geary, R.C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 115-146.
17. Getis, A. and Jared, A. (2010). Constructing the spatial weights matrix using a local statistic. *Perspectives on Spatial Data Analysis* (pp.147-163). Berlin: Springer.

18. Glaeser, E.L. (2005). Inequality. *Harvard Institute of Economic Research, Discussion Paper* No. 2078. Cambrige: Harvard University.
19. Goodchild, M.F. and Haining, R.P. (2004). GIS and spatial data analysis: Converging perspectives. *Papers in Regional Science*, 83(1), 363-385.
20. Fischer, M.M. and Getis, A. (eds.) (2010): *Handbook of Applied Spatial Analysis*. Springer, Berlin, Heidelberg and New York (812 pp.).
21. Haining, R. (1989). Geography and spatial statistics: current positions, future developments. In B. MacMillan (ed.), *Remodeling Geography*, pp. 191-203. Oxford, Basil Blackwell.
22. Kalkhan, M.A. (2011). *Spatial statistics: geospatial information modeling and thematic mapping*. CRC Press.
23. Klier, T. and Daniel P. McMillen. (2008). Clustering of Auto Supplier Plants in the United States: Generalized Method of Moments Spatial Logit for Large Samples. *Journal of Business & Economics Statistics*, 26(4), 460-471.
24. Metwally, M.M. and Jensen, R.C. (1973). A Note on the Measurement of Regional Income Dispersion, Economic Development and Cultural Change, *Halaman* 135-136. Miller, H.J. (2004). Tobler's First Law and Spatial Analysis. *Annals of the Association of American Geographers*, 94, 284-289.
25. Moran, P.A. (1948). The interpretation of statistical maps. *J. Roy. Statist. Soc. Series B (Methodological)*, 10(2), 243-251.
26. Nazara, S. and Hewings, G.J.D. (2003). *Towards Regional Growth Decomposition with Neighbor's Effect: A New Perspective on Shift-Share Analysis, Regional Economics Application Laboratory (REAL)*, University of Illinois at Urbana-Champaign.
27. Pinkse, J. and Slade, M.E. (1998). Contracting in Space: An Application of Spatial Statistics to Discrete-Choice Models. *Journal of Econometrics*, 85(1), 125-154.
28. Rey, S.J. and Richard, J.S. (2012). A Spatial Decomposition of the Gini Coefficient. Cyber GIS Software Integration for Sustained Geospatial Innovation. *Lett. Spat. Resour. Sci.*, 6, 55-70.
29. Rosenberg, M.S. and Anderson, C.D. (2011). PASSaGE: pattern analysis, spatial statistics and geographic exegesis. Version 2. *Methods in Ecology and Evolution*, 2(3), 229-232.
30. Sherman, M. (2010). Spatial Models and Statistical Inference. *Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties*, 71-86
31. Sunitha, L., BalRaju, M. and Sasikiran, J. (2013). Data Mining: Estimation of Missing Values Using Lagrange Interpolation Technique. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, *2*(4), pp-1579.
32. Tarigan, Robinson. (2012). *Ekonomi Regional; Teori dan Aplikasi (Edisi Revisi).* Jakarta: Bumi Aksara.
33. Tobler, W. (1970). *A computer movie simulating urban growth in the Detroit region*. Economic Geography, 46(2), 234-240.
34. Wartenberg, D. (1985). Multivariate Spatial Correlation: A Method for Exploratory Geographical Analysis. Dalam Geographical Analysis, 17(4), 263-283.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 213-232

EDUCATION MAPPING IN INDONESIA USING GEOGRAPHICALLY WEIGHTED REGRESSION (GWR) AND GEOGRAPHIC INFORMATION SYSTEMS (GIS)

Robert Kurniawan and **Marwan Wahyudin**
Department of Computational Statistics, Institute of Statistic (STIS)
Jakarta, Indonesia,
Email: robertk@stis.ac.id

ABSTRACT

Indonesia is a unitary state which has many islands with 33 Provinces. There are 497 cities and districts in Indonesia are different geographically and socially. The number of regions in Indonesia led to unequal education in Indonesia. It is seen from the city / districts that the implementation of the 9-year basic education program has not been completed and visible from the high school level enrollment rates are low. Therefore, one way to deal with this phenomenon, the research conducted using Geographically Weighted Regression (GWR) to determine the factors that affect education in Indonesia spatially. These factors are also factors that influence both globally or locally in every city and county in Indonesia. Based on the results of testing with multiple linear regression and obtained a residual assumption 7 (seven) predictor variables that can be used in this research. Based on GWR method and integrated with GIS obtained 5 (five) group mapping influential education or linearly aligned with the independent variable.

KEYWORDS

GWR, Geographically Weighted Regression, Spatial, Education, Multiple Linear Regression, GIS.

1. INTRODUCTION

In everyday life, we tend to face many problems that are related to two or more variables in a form of some kind of relationship that is stated with a mathematical equation. That form of relationship between variables is called regression. (Sudjana, 2005).

The equation that connects between response variable and predictor variable is a form of a classic regression analytical result. Classical regression analysis assumes that approximated value of regression parameter will be the same for every observation location or in other words, affects globally. The classical regression analysis itself has a few assumptions as requirements that need to be fulfilled at first. Those assumptions are named classic assumption which consists of: normality, linearity, the absence of auto-correlation and multi-colinearity, and also homoscedasticity (Farhan, 2013).

Marwan (2014) stated that in data modeling that is affected by spatial aspect or observation location's geographical condition, there are few assumptions those are fairly

difficult to fulfill in that classical regression analysis implementation. Based on the data variety owned by spatial data, then the linearity assumption in classic linear regression model will be difficult to be fulfilled. Besides, another obstruction encountered is the assumption when the spatial data is modeled in classic linear regression model is the residual assumption that must be identical or homoscedasticity. So that if these assumptions are not yet fulfilled, then it will make spatial heterogeneity to occur.

Spatial data analysis is an analysis that concerns more about location. The difference between one location with another location, based on the geographical, cultural, and other conditions, becomes a cause for the occurrence of spatial hetorogenity. The impact of spatial heterogeneity is the spatially varying of the regression parameter. Regression method which can be used to analyze that is *Geographically Weighted Regression* (GWR), a method that uses geographical factors as a response variable that can affect predictor variables.

According to Fortheringham *et al.* (2002), GWR is a developed model of classic regression that is used to analyze spatial heterogeneity. Heterogenity that is mentioned is a condition when *Measurement of Relationship* between different variables between one location and another. In GWR, approximated parameter values produced are local, so that every observation location has different regression coefficient value. Parameter approximation in GWR is done by adding location weighting. The selection of the weighting function is the one of many factors that determines the GWR model analytical result.

LeSage (1998) wrote GWR modeling algorithm on Matlab software. In the development, softwares those were used in GWR calculations grew up fast, one of them was Spatial Analysis Software WIRES (Dekha, dkk, 2014), and GWR modeling specialized for Indonesian education sphere by using Matlab (Marwan, 2014).

In the applications, GWR tends to be used in social-economy researches related to spatial stuffs. As an example, research that was conducted by Hasbi (2011) by applicating GWR to observe crime rate in Colombus affected by income and housing factor. Suan-Pheng, dkk (2005) used GWR to observe poverty in Bangladesh. Whereas Ribut (2013) used GWR to observe poverty rate in Papua Province, Indonesia. Nathan (2007) also used GWR to observe students' attendance rate with education quality in Maynas, Peru. And the research related to education mapping in Indonesia that is nobody have been researching up until now.

Therefore, we are trying to propose this research in order to observe education mapping in Indonesia and how educational factors affect it based on the spatial condition by using GWR method. In this research, we were using the software that was developed by (2014) for GWR calculations that is directly integrated with GIS. We are expecting that this research will be useful for Indonesian government especially and for development of knowledge, especially in education sector.

2. METHODOLOGY

Geographically Weighted Regression (GWR)

Classic regression is used as a basic model former from data approximation or depiction of data variety. In using classic regression, it has to fulfill one of some assumptions; every observation has to be free in relationship meaning with other observations, or with other words, has to be independent. If in one observation data territorial-refereed data or geographical-related data and has an active relationship is found, it will cause a spatial heterogeneity.

According to Anselin (1988), *spatial heterogeneity* in classical regression analysis is a thing that should get special attention. The spatial heterogeneity may be caused by spatial units' condition in one research area those are basically not homogeneous. For example, in territorial income rate may be different.

GWR model is a technique that assumes that regression parameter is varying spatially. By using GWR, spatial variety in parameter estimation value will be discovered, so that different and valuable interpretations can be obtained for every location researched.

GWR model is a developed global linear regression model where the idea is basically referred from nonparametric regression. This model is a local linear regression model that produces local model parameter approximation for every location depending where the data is collected, so that every geographical location has its own regression parameter value. GWR model can be written as the following (Fotheringham *et al*., 2002):

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \quad (1)$$

where

I = 1,2, …, p

y_i = response variable observed value for the i-th location

(u_i, v_i) = geographical positioning coordinate (*longitude, latitude*) for the *i-th* observation location

$\beta_k(u_i, v_i)$ = *k*-th predictor variable regression coefficient for the *i-th* observation location

x_{ik} = *k*-th predictor variable observed value for the *i-th* observation location

ε_i = the *i*-th observation error that is assumed as identical, independent, and normally distributed with zero mean and constant variance σ^2.

GWR Model Parameter Estimation

In GWR model, it is assumed that observation data that is adjacent to the *i*-th point has a greater impact to the $\beta_k(u_i, v_i)$ estimation than the further one. That equation is used to measure model relationship for every *i*-th points. $\beta_k(u_i, v_i)$ regression coefficient is estimated with *Weighted Least Squares* (WLS), by giving different weightings for every location where the data are observed. According to (2000), the weightings those are given is in accordance with the Tobler's First Law: "*Everything is related to everything else, but near things are more related than distant things"*. For example, the weighting

for every (u_i, v_i) locations is $w_j(u_i, v_i), j = 1,2, \dots, n$ so the parameter in the (u_i, v_i) parameter location is estimated by adding $w_j(u_i, v_i)$ weighting elements in (2.1) equation and then minimizing residual sum of square, as the following:

$$\sum_{j=1}^{n} w_j(u_i, v_i)\varepsilon_j{}^2 = \sum_{j=1}^{n} w_j(u_i, v_i)\left[y_j - \beta_0(u_i, v_i) - \sum_{k=1}^{p} \beta_k(u_i, v_i)x_{jk}\right] \quad (2)$$

Or in the residual sum of square matrix form is

$$\boldsymbol{\varepsilon}^T \boldsymbol{W}(u_i, v_i)\boldsymbol{\varepsilon} = \boldsymbol{y}^T \boldsymbol{W}(u_i, v_i)\boldsymbol{y} - 2\boldsymbol{\beta}^T(u_i, v_i)\boldsymbol{X}^T\boldsymbol{W}(u_i, v_i)\boldsymbol{y} + \boldsymbol{\beta}^T(u_i, v_i)\boldsymbol{X}^T\boldsymbol{W}(u_i, v_i)\boldsymbol{X}\boldsymbol{\beta}(u_i, v_i) \quad (3)$$

GWR Model Weighting

To determine the size of the weighting for each own locations in GWR model, *kernel function* can be used to do so.

Kernel function is used to estimate parameter in GWR model if the (w_j) range function is a continuous and downward monotone function with (u_i, v_i)(Chasco *et al.*, 2007). The weighting that is formed by using this kernel function is *Gaussian Distance Function*, *Exponential function*, *Bi-square* function and kernel *Tricube* function. Each weighting functions can be written as following:

a. Gaussian

$$w_j(u_i, v_i) = \phi\left(\frac{d_{ij}}{\sigma h}\right) \quad (4)$$

where ϕ is standard normal density and σ shows standard deviation for d_{ij} distance vector.

b. *Exponential*

$$w_j(u_i, v_i) = \sqrt{exp\left(-\frac{d_{ij}{}^2}{h^2}\right)} \quad (5)$$

c. *Bi-square*

$$w_j(u_i, v_i) = \begin{cases} \left(1 - \left(\frac{d_{ij}}{h}\right)^2\right)^2 & \text{, for } dij \le h \\ 0 & \text{, otherwise} \end{cases} \quad (6)$$

d. *Tricube*

$$w_j(u_i, v_i) = \begin{cases} \left(1 - \left(\frac{d_{ij}}{h}\right)^3\right)^3 & \text{, for } dij \le h \\ 0 & \text{, otherwise} \end{cases} \quad (7)$$

with $d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2}$ *eucliden* distance between (u_i, v_i) coordinates as geographical positioning coordinates (*longitude*, *latitude*) for *i*-th observation location to (u_j, v_j) location and *h* is a non-negative parameter that is known as *bandwidth*.

Bandwidth

GWR model requires observation coordinate points as data. Those coordinates are used to get the distance between observation locations. GWR has two distance systems : projected distance system and coordinates distance system using *latitude-longitude* degrees. In measuring distance, projected coordinates system uses *phytagoras* rule. Coordinates in common maps are the example of projected coordinates system. Unlike projected coordinates system, in measuring distance between observation locations, *longitude-latitude* degrees coordinates system uses *great-circle distance* rules. *Great circle distance* is the shortest distance between two points on earth. Distance measurement on earth based on *World Geodetic System 1984* rules.

In parameter estimations using GWR, *bandwidth* has a major role. *Gaussian* function requires a *bandwidth* value to produce weighting matrix. *Bandwidth* can be analogized as a radius of a circle, so that a point that is inside a circle is still considered to have influence.

In establishing GWR model, *bandwidth* plays major role. The optimal *bandwidth* value shows that how many observations that create significant effect to GWR model establishment. Optimal *bandwidth* value can be obtained with *Cross Validation* (CV). According to Menis (2006), *crossvalidation* formula is as following:

$$CV(h) = \sum_{i=1}^{n}(y_i - \hat{y}_{i\neq 1}(h))^2 \quad (8)$$

With $\hat{y}_{i\neq 1}(h)$ is approximated value of y_i where observation in (u_i, v_i) location is removed from estimation process. Optimal *bandwidth* can be obtained if the minimum CV value has been obtained at first.

Hypothesis Testing on GWR model

Hypothesis testing on GWR model consists of GWR model conformity test and model parameter test. GWR model conformity test (*goodness of fit*) is done with hypothesis as following:

H_0: $\beta_k(u_i, v_i) = \beta_k$

(There is no significant difference between regression model and GWR)

H_1: At least there is one $\beta_k(u_i, v_i) \neq \beta_k$

where

$k = 0,1,2, \dots, p$

$i = 1,2, \dots, n$

The determination of test statistic based on *Residual Sum of Square*/RRS which is obtained for each H_0 and H_1. Under the condition of H_0, by using OLS method, RSS value can be obtained as following:

$$\begin{aligned} \mathrm{RSS}(H_0) &= \hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}} \\ &= (\boldsymbol{y} - \hat{\boldsymbol{y}})^T(\boldsymbol{y} - \hat{\boldsymbol{y}}) \\ &= \boldsymbol{y}^T(\mathbf{I} - \mathbf{H})\mathbf{y} \end{aligned} \quad (9)$$

with $\boldsymbol{H} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ idempotent.

Under the condition of H_1, the spatially varying regression coefficient in (2.1) equation defined with GWR method, so that RSS value can be obtained as following:

$$\begin{aligned} \text{RSS}(\text{H}_0) &= \hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}} \\ &= (\boldsymbol{y} - \hat{\boldsymbol{y}})^T (\boldsymbol{y} - \hat{\boldsymbol{y}}) \\ &= \boldsymbol{y}^T (\mathbf{I} - \mathbf{L})^T (\mathbf{I} - \mathbf{L}) \mathbf{y} \end{aligned} \tag{10}$$

By using the difference between residual sum of squares under H_0 and under H_1:

$$\begin{aligned} F &= \frac{\frac{RSS(\text{H}_0) - RSS(\text{H}_1)}{\tau_1}}{\frac{RSS(\text{H}_1)}{\delta_1}} \\ &= \frac{\boldsymbol{y}^T [(\mathbf{I}-\mathbf{H}) - (\mathbf{I}-\mathbf{L})^T (\mathbf{I}-\mathbf{L})] \mathbf{y} / \tau_1}{\boldsymbol{y}^T (\mathbf{I}-\mathbf{L})^T (\mathbf{I}-\mathbf{L}) \mathbf{y} / \delta_1} \end{aligned} \tag{11}$$

under H_0, F will approach F-distribution with degree of freedom

$$df_1 = \frac{\tau_1}{\tau_2} \tag{12}$$

and

$$df_2 = \left(\frac{\delta_1}{\delta_2}\right), \tag{13}$$

with

$$\tau_1 = tr\left([(\mathbf{I} - \mathbf{H}) - (\mathbf{I} - \mathbf{L})^T (\mathbf{I} - \mathbf{L})]^i\right) \tag{14}$$

where

$$i = 1,2.$$

If α significancy level is taken, then reject $\mathbf{H_0}$ if $F \geq F_{\alpha,df1,df2}$.

3. DATA AND DEFINITION

Data which was used in this research was education and educational infrastructures data that is referred from national social-economy survey (SUSENAS) on year 2011 and village potentials survey (PODES) on year 2011 which are done by Badan Pusat Statistik (BPS). One of them is Human Development Index (IPM) data per regencies and cities equals to 497 regencies/cities all over Indonesia as response variable (Y). While for the predictor table consists of 7 variables : literacy rate, percentage of population of junior high school graduates with 15 years old age over 15 years old population size, number of elementary schools, number of junior high schools, 7-12 years school participation rate, 13-15 years school participation rate, and 16-18 years school participation rate. Attached to Table 1, about response variable and predictor variable.

Table 1
Variables Used in the Research

Variables	Explanation	Measurement Scale
(1)	(2)	(3)
Y	Human Development Index	Ratio
X_1	Literacy Rate	Ratio
X_2	Percentage of population of junior high school graduates with 15 years old age over 15 years old population size	Ratio
X_3	Number of elementary schools	Ratio
X_4	Number of junior high schools	Ratio
X_5	School Participation Rate for 7 – 12 years old	Ratio
X_6	School Participation Rate for 13 – 15 years old	Ratio
X_7	School Participation Rate for 16 – 18 years	Ratio

United Nation Development Programme / UNDP (1990) stated that human development index is the formation of human abilities that is derived from health improvement, expertise, and knowledge. So that IPM formulation is calculated from infant mortality index for every 1000 live births, the average of population's lifespan and ability to read and write (literacy rate) and also per capita income.

According to Harjasujana (Mumuh, 2003), if it is related to the educational programs in schools, reading plays major role. Reading abiliy is the main factor that determines the learning achievements. WHO emphasize that illiteracy eradication must be an integrated part of economic reformation. Illiteracy can cause negative impacts on second generations, because illiterate mothers tend to not have sufficient knowledge for their children needs in early stage, which this stage is considered as *golden age*, so it affects their health, emotional, social, and intellectual development. (Kusnadi, 2005).

School participation rate is a percentage of a population that is still in school at certain age over a population on that age. This rate is one of many indicators that depicts how many people that have a chance to get educated. The increase of school participation rate means there is an accomplishment in educational sector, especially that is related to the effort to spread the range educational services. Or with more details that the increase of school participation rate connote that there is increasing number of people who are able to get educated. (Marwan, 2014).

According to E. Mulyasa (2004), educational infrastructures are tools and supplies those are directly used to support educational process, especially learning and teaching process such as : buildings, classrooms, tables and chairs, and other equipments and medias. Educational infrastructures are supporting infrastructures for learning and teaching process. Marwan (2014) concludes that educational infrastructures are all facilities, directly supporting educational process, especially learning and teaching

process, either mobile or static, so that educational goal accomplishment can run smoothly, regularly, effective, and efficient.

From predictors variables those are collected, then processed and mapped with the help of GIS method. Where GIS is an information system that is designed to work with spatially-referenced or geographically-coordinated data, or in other words, GIS is a data-based system with a special ability to handle spatial data along with sets of operations (Barus and Wiradisastra, 2000).

While according to (Anon, 2001) GIS is an information system that can integrate spatial data, text, and objects those are related geographically on earth (*georeference*). Besides, GIS can also combine, manage, and analyze data which will finally generate an output that can be used as a reference for decision-making factor in some cases of problems those are related to geography.

The main goal of the utilization of GIS is to simplify the process of obtaining a processed information and saved as a location's or object's attributes. The main characteristics of the data that can be utilized in GIS is the data that is already bound with locations and is a basic data that is not yet specified (Dulbahri, 1993).

4. RESULT AND DISCUSSION

Characteristics those are seen in Indonesia indicate the existence of influential variable differences in every regencies and cities in Indonesia. So it can be said that there is an uneven distribution of education in Indonesia.

As shown on Table 2 where the school participation rate significant differences for junior high school and senior high school level can be found, for junior high school level, the average of school participation rate is 87.76%, it shows that Indonesian people who are currently at 13-15 years old that participate in school is just equals to 87.76%. While for the senior high school level, the school participation rate decreases to 60.77%. And for the average of 15 years old or more population that have graduated from junior high schools in Indonesia is equals to 47.95% if compared to the average of literacy rate that reached 92.07%.

Table 2
Educational Variables' Totals, Averages, and Standard Deviations

Variable	Y1	X1	X2	X3	X4	X5	X6	X7
Totals	35371.2	45759.02	23833.02	168029	46074	47862.9	43616.07	30204.61
Averages	71.16	92.07	47.95	338.08	92.70	96.30	87.76	60.77
Std Dev.	5.27	11.82	15.43	313.71	87.55	8.45	9.30	12.83

Gaussian

Sumber Error	Jumlah Kuadrat	df	MS	F
Improvement	9.993e+02	1.072e+02	9.320e+00	4.080e+00
GWR Residu	8.722e+02	3.818e+02	2.285e+00	
Residu Global	1.871e+03	489		

Picture 1: GWR Model *Gaussian* Function Parameter Testing Result

Goodness of fit from GWR model or the conformity test of GWR model is used to see whether the GWR model is better than global regression model. Based on the picture above, the F-count value from GWR model with *Gaussian* weighting function is equals to 4.080 that has a greater value than the one obtained from the F-Table ($F_{1-\alpha,df1,df2}$) equals 1.277. It is shown that there are enough evidences to reject H0 where the GWR modeling with *Gaussian* weighting function with (α) significancy level equals 5%, and it can be concluded that GWR model is significantly different if compared to global linear regression.

Exponential

Sumber Error	Jumlah Kuadrat	df	MS	F
Improvement	1.020e+03	1.072e+02	9.517e+00	4.269e+00
GWR Residu	8.511e+02	3.818e+02	2.229e+00	
Residu Global	1.871e+03	489		

Picture 2: GWR Model *Exponential* Function Parameter Testing Result

GWR model with *exponential* weighting function has F-count value equals to 4.269. Based on the Picture 2 which the F-count of GWR model with *exponential* weighting function is greater than the one with *Gaussian* weighting function, that makes the value of the F-count greater than the one that is listed in the F-Table ($F_{1-\alpha,df1,df2}$) ,which equals to 1.277. H0 will be rejected if the F-count value is greater than the one that listed in the F-Table so that GWR modeling with *exponential* weighting function with (α) significancy level equals to 5%, it can be concluded that GWR model is significantly different with the global linear regression model.

Tricube

Sumber Error	Jumlah Kuadrat	df	MS	F
Improvement	1.800e+03	1.072e+02	1.679e+01	8.999e+01
GWR Residu	7.123e+01	3.818e+02	1.866e-01	
Residu Global	1.871e+03	489		

Picture 3: GWR Model *Tricube* Function Parameter Testing Result

In Picture 3 GWR model with *tricube* weighting function has F-count value equals to 8.999e+01. That value is greater than *Gaussian* weighting function and *exponential* weighting function so that the F-count value is greater than the one that is listed in the F-Table. The starting hypothesis will be rejected if the F-count value is greater than the

one that is listed in F-Table ($F_{1-\alpha,df1,df2}$), which equals to 1.277. According to Picture 3, GWR modeling with *tricube* weighting function has (α) significancy level equals to 5%, in conclusion, GWR model is significantly different with global linear regression.

Bi-Square

```
************************************************************
Sumber Error     Jumlah Kuadrat    df          MS          F
Improvement      1.805e+03       1.072e+02   1.684e+01   9.710e+01
GWR Residu       6.620e+01       3.818e+02   1.734e-01
Residu Global    1.871e+03             489 ************************
```

Picture 4: GWR Model *Bi-square* Function Parameter Testing Result

The starting hypothesis stated that there are no differences between GWR model and global regression model and will be rejected if the F-count value is greater than the one that is listed in the F-Table. The F-count value in Picture 4 for GWR model with *bi-square* weighting function is greater than the one that is listed in F-Table ($F_{1-\alpha,df1,df2}$) , which equals to 1.277, and the F-count value does not differ much with *tricube* weighting function, that is equals to 9.71e+01. Based on Picture 4, GWR modeling with *bisquare* weighting function with (α) significancy level equals to 5%, it can be concluded that GWR model is significantly different with global linear regression.

Based on Table 3, if the F-count values from all weighting functions compared, then it can be concluded that the *bi-square* function has the greatest R^2 that is most likely to approach 1. So it can be said that the best GWR model that is able to depict between response variable (human development index) and predictor variables (literacy rate, percentage of population of junior high school graduates with 15 years old age over 15 years old population size, number of elementary schools, number of junior high schools, 7-12 years school participation rate, 13-15 years school participation rate, and 16-18 years school participation rate) is by using *bi-square* weighting function.

Table 3
The F-count and R^2 Values for GWR Model on Different Weighting Functions

Weighting Functions	Gaussian	Exponential	Tricube	Bi-square
(1)	(2)	(3)	(4)	(5)
F-count Values	4.80	4.269	8.999e+01	9.71e+01
R^2 Values	0.9369	0.9385	0.9948	0.9952

The conformity test of GWR model for the starting hypothesis is there are no differences between GWR model and global linear regression model and will be rejected if the F-count value is greater than the one that is listed in F-table. The F-count value in Table 3 for GWR model with *bi-square* weighting function is greater than the one that is listed in F-table ($F_{1-\alpha,df1,df2}$) , which equals to 1.277, and the F-count value equals to 9.71e+01. The GWR modeling with *bi square* weighting function with (α) significancy level equals to 5% can be concluded that GWR model is significantly different with global linear regression model.

Based on Picture 5, the estimation coefficient total that is significant on *Intercept* reached 314 regencies/cities. While for the maximum value of the *Intercept* lies on 81.6 and the minimum value is -55,9 with coefficient total that has a positive sign (+) is as many as 352.

Variabel	Beta Mak lokal	Beta Min lokal	Jml Beta(+)	Jml Beta(sig)
variabel 1	8.16e+02	-5.59e+02	352	314
variabel 2	4.48e+00	-4.96e+00	371	335
variabel 3	1.07e+00	-7.48e-01	429	407
variabel 4	1.99e-01	-2.53e-01	259	209
variabel 5	9.81e-01	-6.82e-01	261	220
variabel 6	6.65e+00	-5.26e+00	247	215
variabel 7	2.71e+00	-2.10e+00	245	194
variabel 8	9.98e-01	-6.60e-01	213	173

****

Picture 5: GWR Model with *Bi-Square* Function Parameter Testing Result

On variable 2 and 3 GWR model *bi-square* weighting function are only as many as 335 and 407 regencies/cities. On variable 4, 5, 6, and 7 that is lineary parallel with the human development index variable have the regencies/cites totals for each variables : 209, 220, 215, 194. Variable 8 becomes a lineary parallel variable with human development index variable with the least observation location as many as 173. Overall, the predictor variable and response variable in GWR model with *bisquare* weighting function has the most models.

GWR Model for Education Mapping in Indonesia

GWR model mapping for education data will use the best GWR model, that is the GWR model with *bi-square* weighting function. In the mapping, the observation ranges between areas which have lineary parallel predictor variable with response variable. The greater the value of an area (represented by colors), the stronger the predictor variable's influence it will be, either it is positively (+) or negatively (-). The mapping divided into 5 large groups :

1. Group 1 (blue) ranges between minimum value – 20-th percentile.
2. Group 2 (yellow) ranges between 20-th percentile – 40th percentile.
3. Group 3 (light blue) ranges between 40th percentile – 60-th percentile.
4. Group 4 (green) ranges between 60-the percentile – 80-th percentile.
5. Group 5 (grey) ranges between 80-percentile – maximum value.

Picture 6: GWR Model Mapping with Human Development Index Parameter Estimation

In picture 6 can be seen that Sumatra Island seems to be varying from group 1 to group 5. In Aceh, North Sumatra, West Sumatra, Riau, Jambi, Bengkulu, South Sumatra, and Lampung Province, they have all color groups. While for Riau Archipelago and Bangka Belitung are dominantly being in group 2 and group 5.

In Java Island is can also be seen fairly varying from Banten, West Java, Central Java, Yogyakarta, and East Java Province. Group 5 are seen to be dominant in Banten, West Java, and East Java. Bali Province seems to be unicolored in group 3, while for Nusa Tenggara Archipelago varies between group 1 and group 4, but note that in the East Nusa Tenggara, group 5 is more dominant.

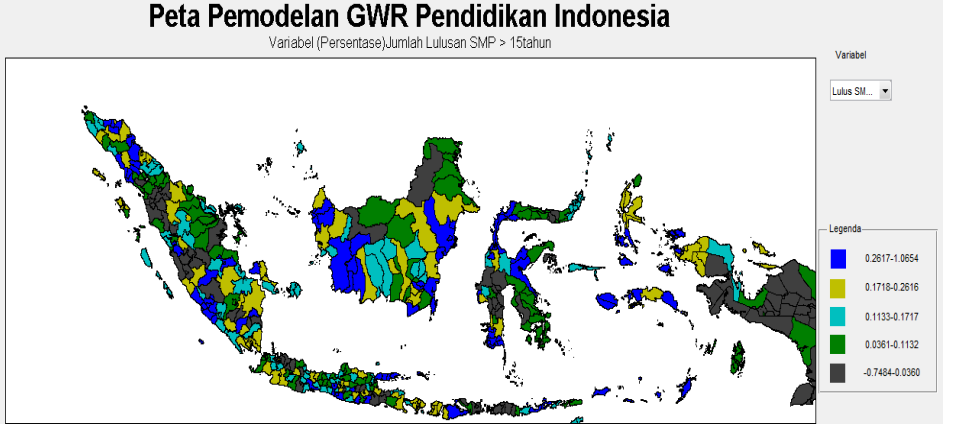
In Kalimantan Island is more dominated by group 4. Sulawesi Island and Maluku Archipelago are seem to be varying. In Papua and some part of West Papua are dominated by group 3. Human development index influence against other variables in group 5 is pretty strong and negative, which means, if the predictor variable increased by 1 measurement unit, then the human development index variable estimation will be decreased by amount of the *Intercept* value.

Picture 7: GWR Model Mapping with AMH Variable Parameter Estimation

Picture 7 shows that AMH variable for Sumatra Island is very variative, Riau Province, Riau Archipelago, and Bangka Belitung Province are dominated by group 5. In Java Island, Banten and DKI Jakarta Province are dominated by group 1 and for Central Java, Yogyakarta, and East Java are pretty varying, ranging from group 2 to group 4. From Bali Province to the eastern end of Nusa Tenggara Province lies between group 4, and then group 3, and finally group 2.

In Kalimantan Island, group 1 is only found in some part of East Kalimantan and Central Kalimantan Province. West Kalimantan is more likely to be dominated by group 2, Central Kalimantan is dominated by group 3, and East Kalimantan is dominated by group 4 and group 5. In northern parts of Sulawesi Island are dominated by group 5 and the Central parts are dominated by group 4. While for the southern parts are pretty varying those are dominated by group 1 and group 2. In Maluku Cluster Islands are dominated by group 1 and group 4. While in West Papua Province, group 1 is still existed, but the rest are dominated by group 2 and group 3.

Banten Province is dominated by group 1, which means, AMH variable influence against human development index variable is fairly strong in that area and if the AMH variable value increased by 1 measurement unit then the human development index variable will also be increased as well.



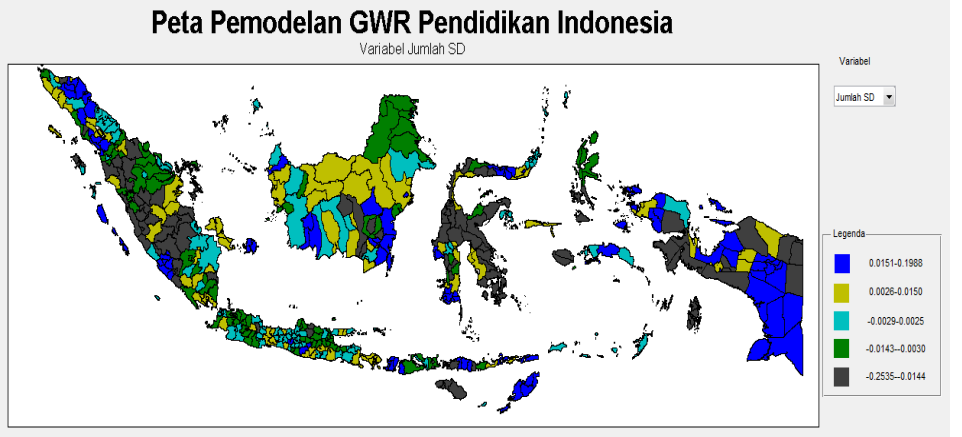
Picture 8: GWR Model Mapping with Persentage of Junior High School Graduates for Age 15+ Variablen Parameter Estimation

Percentage of junior high school for age 15+ has different variation for each area if we estimate the parameter using GWR model. In map 8, overall, as can be seen that Sumatera, Java, Bali, Nusa Tenggara, Kalimantan, and Sulawesi have no dominant group.

In Sumatra, especially for Aceh, there is no group 5, but there are some variation of junior high school graduation by group 1 to group 4. North Sumatera, Riau, and Riau Islands even don't have group 1, yet it is still have vary graduation by group 2 to group 5. While another province in Sumatera have a big variation in their graduation. Percentage of junior high school graduation for age 15+ have a big variation for each area. It is clearly to see that in map 8, Java has all criteria group. As well as in Kalimantan,

Sulawesi, Bali, and Nusa Tenggara. In Java and Kalimantan, there are only some part of them that have group 5.

Group 1 and group 2 are more dominant than another groups in Maluku Islands. While group 5 is dominant group in West Papua and Papua.

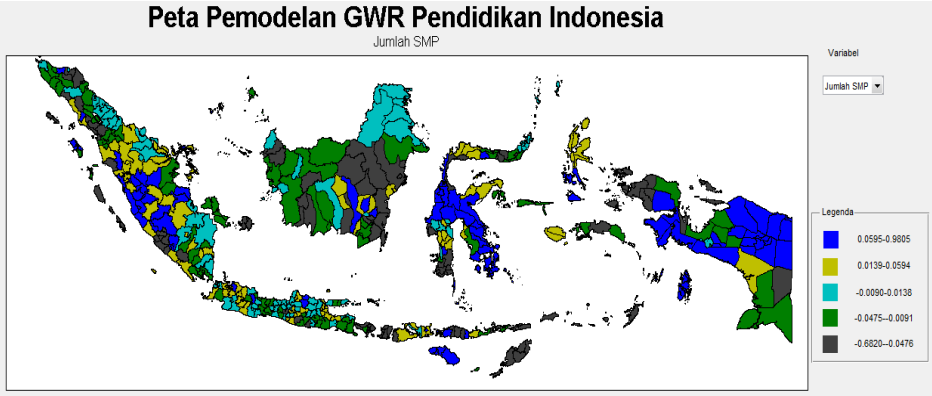


Picture 9: GWR Model Mapping with Numbers of Elementary Schools Parameter Estimation

Based on map 9, we can see that group 5 have its crawl on Sumatera and Sulawesi. In Aceh and North Sumatera have varying graduations. In Riau, Riau Islands, West Sumatera, Bengkulu, and Jambi, group 5 is very dominant. Group 1 and 2 are dominant in Bangka Belitung, while group 2 and 4 are dominant in Lampung.

Based on map 9, provinces in Java such as, Banten, DKI Jakarta, West Java, Central Java, and East Java are dominated by group 2 to group 5. Only in some regencies/cities in Central Java are dominated by group 1. Bali only has a group that dominated there. It is group 2. West Nusa Tenggara is dominated by group 1 and group 4, furthermore East Nusa Tenggara is dominated by group 1 and 5. Kalimantan looks varying in group 5. North areas of Kalimantan are dominated by group 4, they are different than West Kalimantan that is dominated by group 2.

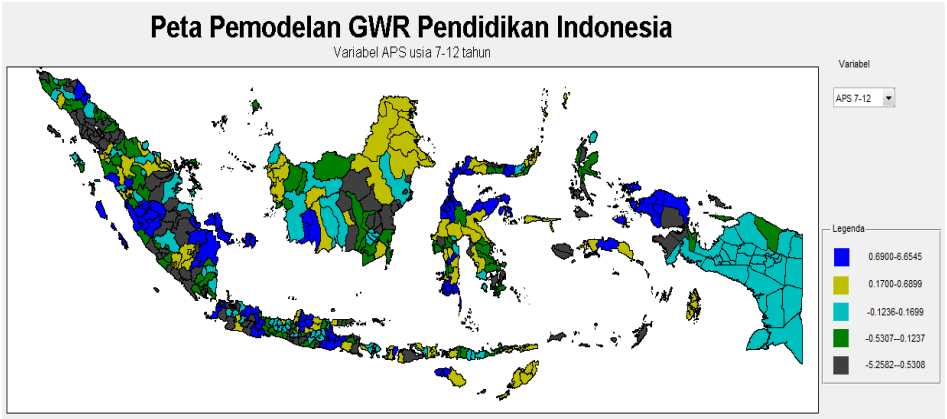
Percentage of junior high school in north and south area in Sulawesi are very varying. But, there are some provinces that dominated by group 5. They are Central Sulawesi, West Sulawesi, and Southeast Sulawesi. Group 4 is dominating in North Maluku, it is different than Maluku that have varying graduation. Some cities/regencies in West Papua have varying graduation, group 5 is dominating, though. In Papua, group 4 and 5 are more dominant than another province.



Picture 10: Number of Junior High Schools Variable Parameter Estimation GWR Model Mapping

Based on Picture 10, we can see that number of junior high school variable has different effect for each area. In Sumatera we can see that Aceh and North Sumatera have quite varied effect from group 3 until group 5. Sumbar, Riau, Kepri, Bengkulu, and Jambi has been dominated by group 1 and group 2, though it is still quite varied from Banten until East Java. We can take West Java region as an example. In West Java, group 2 and group 3 are dominating, in Central Java group 3 and group 4 are dominating, in Bali and NTT group 5 is dominating. While in NTB, group 1, 2, and 5 have a varied number.

In west area of Kalimantan, group 4 is dominating. Central Kalimantan has more varied number than East Kalimantan that very dominated by group 3 and group 5. In north and south area of Sulawesi, have varied number, as well. They are dominated by group 2 until group 5. But, in central and southeast area of Sulawesi is dominated by group 1. In North Maluku, there are 2 groups that dominate the number of junior high school. The groups are group 1 and group 2. The number of junior high school in West Papua and Papua is varied, group 1 is dominating, though.

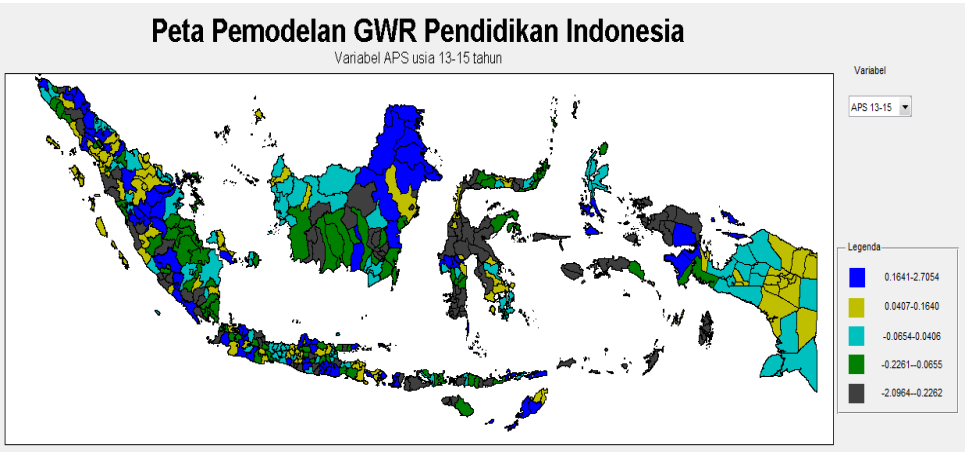


Picture 11: 7-12 Years Old School Participation Rate Variable Parameter Estimation GWR Model Mapping

School participation rate for age 7-12 can show us the effect of school participation rate for age 7-12 to human development index variable. In Picture 11, we can see that Sumatera is dominated by group 5. Aceh, West Sumatera, and Riau have varied school participation rate. North Sumatera is dominated by group 5. While Bangka Belitung is dominated by group 1.

School participation rate in Java varied from group 1 to group 5. Group 5 is dominating in Banten and Yogyakarta. West Java and East Java are dominated by group 1, while Central Java is dominated by group 3 an group 4. Bali is dominated by group 3. In NTT, school participation rate tend to dominated by group 2. In Kalimantan, especially East Kalimantan, we can see clearly that group 2 is dominating. But in West Kalimantan, even though it has vary school participation rate, we can't find group 5 there. We can find group 5 in Central Kalimantan and South Kalimantan.

In Southeast Sulawesi, group 1 is dominating. In other Province in Sulawesi, school participation rate is quite varied. Overall, Sulawesi is dominated by group 1 and group 2. Maluku Cluster Islands has varied school participation rate. In general, Maluku Cluster Islands are dominated by group 5. West Papua is dominated by group 1 and Papua is dominated by group 3.



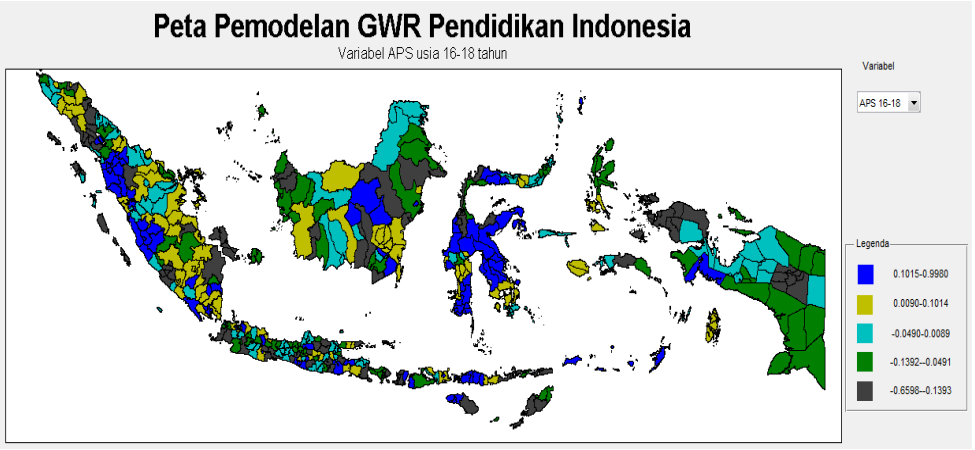
Picture 12: 13-15 Years Old School Participation Rate Variable Parameter Estimation GWR Model Mapping

Based on Picture 12, the effect of school participation rate for age 13-15 to IPM variable have different dominant area, especially for each Island. Sumatera has varied school participation rate for age 13-15. We can see that Aceh, North Sumatera, West Sumatera, Bengkulu, and Lampung have different dominating group. In Riau, South Sumatera, and Bangka Belitung, there is no group 5.

Java has varies school participation rate for age 13-15. Central Java is dominated by group 3. School participation rate for age 13-15 in other Province such as Banten, West Java, DKI Jakarta, and East Java are vary, as well as in Bali, NTB and NTT. West

Kalimantan is dominated by group 3. East Kalimantan is dominated by group 1. Central Kalimantan is dominated by group 4. While, South Kalimantan quite varies in group 3 to group 5.

It's clearly to see that Sulawesi is dominated by group 5 in central area. Even, in north area and southeast area Sulawesi varies, but group 5 is dominant in south area of Sulawesi, as well. North area of Maluku Islands is dominated by group 3 and in south area is dominated by group 5. West Papua is dominated by group 5, even though there is group 1 in some area. It is rather different than West Papua, in Papua, group 2 and group 3 are dominating.



Picture 13:1 16-19 Years Old School Participation Rate Variable Parameter Estimation GWR Model Mapping

The variety of school participation rate for\ age 16-19 variable to Human Development Index variable in some area can be seen in map 13. Aceh is dominated by group 2. In North Sumatera and Bangka Belitung looks varying, even though group 5 is more dominant. West Sumatera and Bengkulu is dominated by group 1. South Sumatera and Riau is dominated by group 2.

In Java, it is clearly to see that the effect of school participation rate for age 16-19 to Index Development Human variable is vary. Group 1 to group 5 are randomly spread in each Province in Java. There are not many differences than Java, in Bali, even though it varies, but group 3 is still dominant. Kalimantan looks varying in each Province. In West Kalimantan, group 3 is the most. In south Kalimantan, group 2 is more dominant than the other. East Kalimantan is dominated by group 3, 4, and 5. Sulawesi is dominated by group 1 in central area, south area, and southeast area. North area of Sulawesi is dominated by group 1 to group 5. It can be seen that group 2 and group 4 are dominant in North Maluku. Papua is dominated by group 5. While in Papua, there are 3 groups that dominate. They are group 3, 4, and 5.

5. COMMENTS AND CONCLUSION

Based on GWR Model Fit Test, we can see that GWR model is significantly different, and it will be better if we compare it with Global Linear Regression Model/Classic Regression Model. In educational mapping based on lineary parallel variable with response variable, we can see that there are some regencies and cities that have the same predictor variable.

The test show us that there are some disadvantages in GWR if we combine it with GIS, it is caused by the area's expansion processes every year in Indonesia. Because of that, it is impossible to compare the data from year to year. Consequently, GWR in Indonesia can only be used in specific time. And if we want to compare it from time to time, we must have different X and Y coordinates each year, and different area map. Therefore, it needs to make some more studies to solve this problem.

Not all predictor variables that affect Human Development Index in Education sector in this research are used. Therefore, in the next research the other predictor variable in education sector must be added, such as what UNESCO is currently doing.

REFERENCES

1. Anselin, L. (1988). *Spatial Econometrics: Methods and Models.* Dordrecht: Kluwer Academic.
2. Anselin, L. (1999). *Spatial Econometrics.* Richardson: Bruton Center, School of Social Sciences, University of Texas.
3. Ardhanacitri, D. and Ratnasari, V. (2013). *Pemodelan dan Pemetaan Pendidikan di Provinsi Jawa Timur Menggunakan Geographically Weighted Regression* [Skripsi]. Surabaya: Jurusan Statistika, Fakultas MIPA, Institut Teknologi Sepuluh November.
4. Bafadal, I. (2003). *Manajemen Peningkatan Mutu Sekolah Dasar: Dari Sentralisasi Menuju Desentralisasi.* Jakarta: Bumi Aksara.
5. Barus, B., and Wiradisastra, U.S. (2000). *Sistem Informasi Geografi: Sarana Manajemen Sumberdaya.* Bogor: Laboratorium Penginderaan Jauh dan Kartografi, Jurusan Tanah, Fakultas Pertanian IPB.
6. BPS (2012). Statistik Pendidikan 2012. Jakarta: Badan Pusat Statistik.
7. Chasco, C., Gracia, I. and Vicens, J. (2007). Modelling Spatial Variations in Household Disposible Income with Geographically Weighted Regression. *Munich Personal RePEc Arkhive (MPRA).* Paper No. 1682. 1 April 2014. http://mpra.ub.uni-muenchen.de/1682/1/MPRA_paper_1682.pdf
8. Dekha, DKK. (2014). *WIRES: A User Friendly Spatial Analysis Software.* Jakarta: STIS
9. Dulbahri (1993). *Sistem Informasi Geografi*. Yogyakarta: BAKOSURTANAL PUSPIC-UGM.
10. Engler, N.J. (2007). *Spatial Analysis of the Effect of Absenteeism on Education Quality in Maynas, Peru* [Tesis].Ontario: University of Waterloo.
11. Farhan, M.Q. (2013). *Analisis Regresi Terapan "Teori, Contoh Kasus, dan Aplikasi dengan SPSS"*. Yogyakarta: Penerbit Andi.
12. Fotheringham, A.S., Brundsdon, C. and Charlton, M. (2002). *Geographically Weighted Regression.* Chichester: Jhon Wiley & Sons.

13. Hasbi, Y. (2011). *Pemilihan Variabel pada Model Geographically Weighted Regression.* Media Statistik: 4(2), 63-72.
14. Harris, R. DKK. (2010). Grid-enabling Geographically Weeighted Regression: A Case Study of Participation in Higher Education in England. *Transaction in GIS.* 14(1), 43-61. 26 Maret 2014.
15. Huda, Ni'matul. (2008). *UUD 1945 dan Gagasan Amandemen Ulang*. Jakarta: Rajawali Pers.
16. Kusnadi, DKK. (2005). *Pendidikan Keaksaraan Filosofi, Strategi, Implementasi.* Jakarta: Departement Pendidikan Nasional.
17. LeSage, J.P. (1999). *The Theory and Practice of Spatial Econometrics.* Departement of Economics, University of Toledo. Ohio. USA.
18. Marwan, W. (2014). *Implemantasi Pemetaan Pendidikan di Indonesia Tahun 2011 dengan Model Geographycally Weighted Regression*. Jakarta: STIS.
19. Mennis, J. (2006). Mapping the Result of Geographically Weighted Regression. *The Cartographic Journal.* 43(2), 171-179. http://astro.temple.edu/~jmennis/pubs/mennis_cj06.pdf
20. Mumuh (2003). *Model Pelatihan Membaca Cepat: Penelitian Tindakan Kelas Terhadap. Kecepatan membaca Siswa SMU Negeri 1 Cisaat Kabupaten Sukabumi Tahun Pelajaran 2002/2003.* Bandung: Universitas Pendidikan Indonesia.
21. Mulyasa, E. (2004). *Manajemen Berbasis Sekolah.* Bandung: PT Remaja Rosdakarya.
22. Nathan, J.E. (2007). *Spatial Analysis of the Effect of Absenteeism on Education Quality in Maynas, Peru* [Thesis]. Canada: Master of Environmental Studies, University of Waterloo.
23. Pemerintah Republik Indonesia 2003. *Undang-undang Republik Indonesia No. 20 Tahun 2003 tentang Sistem Pendidikan Nasional*. Jakarta.
24. Pemerintah Republik Indonesia. (2005). *Undang-undang Republik Indonesia No. 14 Tahun 2005 tentang Guru dan Dosen*. Jakarta.
25. Ribut, N.T. (2013). *Faktor-faktor yang Berkorelasi dengan Kemiskinan di Provinsi Papua: Analisis Spatial Heterogeneity* [Thesis]. Jakarta: Magister Ekonomi, Universitas Indonesia.
26. Sudjana (2005). *Metode Statistika*. Bandung: TARSITO.
27. Widianantari (2008). Kebutuhan dan Jangkauan Pelayanan Pendidikan di Kecamatan Bandongan Kabupaten Magelang [Tesis]. Semarang: Program Pasca Sarjana Magister Teknik Pembangunan Wilayah dan Kota Universitas Diponegoro.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 233-244

ROW-COLUMN INTERACTION MODELS FOR ZERO-INFLATED POISSON COUNT DATA IN AGRICULTURAL TRIAL

Alfian F. Hadi[1] and **Halimatus Sa'diyah**[2]
[1] Department of Mathematics, The University of Jember, Indonesia. Email: afhadi@unej. ac.id
[2] Department of Agronomy, The University of Jember, Indonesia

ABSTRACT

Many zero observations makes some difficulties and fatal consequence in Poisson modeling and its interpretation. We consider to facilitates the analysis of two-way tables of count with many zero observations in agricultural trial. For example, in counting the pest or disease in plants. Plants that have no sign of attack, can occur because of two things, it could be resistant, or simply there is no spore disease (no endemics) or no pest attack. This is the difference between inevitable structural zero or sampling zero that is occurring according to a random process.

This paper describes a statistical framework and software for fitting row-column interaction models (RCIMs) to two-way table of count with some Zero observations. RCIMs apply some link function to the mean of a cell equaling a row effect plus a column effect plus an interaction term is modeled as a reduced-rank regression with rank of 2, then will be visualized by biplot. Therefore its potentially to be develop become AMMI models that accommodate ZIP count.

KEYWORDS

ZIP, AMMI Models, Row-Column Interaction Models, SVD Reparameteri-zation.

1. INTRODUCTION

The Poisson distribution is widely used in quality studies for count related data. Poisson regression models are basically modelling for counts. There are two strong assumptions for Poisson model to be checked: one is that events occur independently over of time or exposure period, the other is that the conditional mean and variance are equal. In practice, the Poisson with a large numbers of count, usually have greater variance than the mean are described as overdispersion. Poisson with smal value of mean, it also have small value of the variance, in this case, count data encounter with value of zero problems. This indicates that Poisson regression is not adequate. There are two common causes that can lead to overdispersion are additional variation to the mean or heterogeneity, an Negative Binomial model is often used and other cause counts with excess zeros or zero-infated. Poisson counts, since the excess zeros will give smaller conditional mean than the true value, this can be modeled by using zero-inflated Poisson (ZIP).

The proper model is needed to present a valid conclusion from the data count by the zero-inflation. Various applications related to this. For example, in calculating the pest or disease in plants. Plants that have no sign of attack, can occur because of two things, it could be as resistant to disease, or simply because there is no disease spores (no endemics) or no pest attack there. This is the difference between a structural zero, the inevitable, and zero sampling that is occurring according to a random process.

2. ZERO-INFLATED AND ITS CONSEQUENCES

Ignoring zero-inflation, especially when a sizeable proportion of the data is zero, implies that the underlying distributional assumptions will not be met. This will more than likely affect the results of an analysis, and hence lead to incorrect conclusions concerning the data. In addition to accounting for zero-inflation, we also need to consider the possibility of over-dispersion, which is variation larger than would be expected under the distributional assumptions. This is a commonly occurring phenomenon with Poisson models, and if ignored, can lead to underestimated standard errors and hence misleading inference about regression parameters (Hinde & Demetrio, [5]). Both zero-inflation and over-dispersion can occur simultaneously in a data set.

3. HANDLING ZERO INFLATED ON AN ADDITIVE MODEL

Suppose that yi is the number of occurrences of an attribute or event and xi ϵ R is a vector of covariates, both recorded for each of i = 1, . . . ,m sites. The simplest approach to modeling the relationship between yi and xi is an ordinary least squares fit of the transformed response, such as $\sqrt{yi}$ or log(yi). However, such transformations are not helpful when the data contain many zeros because the zeros are unchanged under a square-root transformation and are undefined under a logarithmic transformation. Additionally, the underlying distributional assumptions of linearity, homoscedasticity and Gaussianity do not hold so this approach is not suitable in this context.

A better approach is to fit a Poisson generalised linear model (GLM) with Log - link function. We can model over-dispersion by a proper variance function (McCullagh & Nelder [10]). However, the overall fit can be poor because there are typically many more zeros in the data than expected under a Poisson model which allows for over-dispersion.

Another approach to modeling zero-inflated count data concerns the classification of the zero observations into two different groups. Distributions which classify their zero counts in this way have been referred to as zero-modified distributions, distributions with added zeros, zero-inflated distributions or mixture distributions. In these distributions, the zeros are classified into two groups: one group, which along with the positive counts are modeled by a discrete distribution such as the Poisson or negative binomial distribution, occur with probability $1 - \omega$; and the other group, which represent the 'extra' zeros, occur with probability ω. Dietz & Bohning [3] discuss estimation of the parameter in zero-modified Poisson distributions and for illustration, they analyse counts recorded in a dental epidemiological study. Lambert [7], Welsh et al [15], and Bohning [1], also discuss applications of this approach to modeling zero-inflated count data.

4. INTRODUCING ROW-COLUMN INTERACTION MODEL FOR ZERO INFLATED

Zero-inflated models often found in additive models, but less in the model of interaction. In the study of pest/diseases, analyses concern with the interaction between genotype and environmental influences (GEI). Crops that have no sign of attack, can occur because of two things, it could be resistant, or simply there is no spore disease (no endemics) or no pest attack. This is the difference between inevitable structural zero or sampling zero that is occurring according to a random process.

In this research area of the GEI, AMMI model said to be most powerful one to analyzed the GEI, by main effects plus multiplicative interaction terms. Nowadays, AMMI model has been developed to be more generalized, named GAMMI. It can handle Poisson count as well. AMMI model is basically presents interaction through dimension reduction techniques. Here, it is very important to introduce a statistical methodology handling problems with inflation-zero count. It will be possible to model the zero count as an expression of resistance and not because chance did not affected.

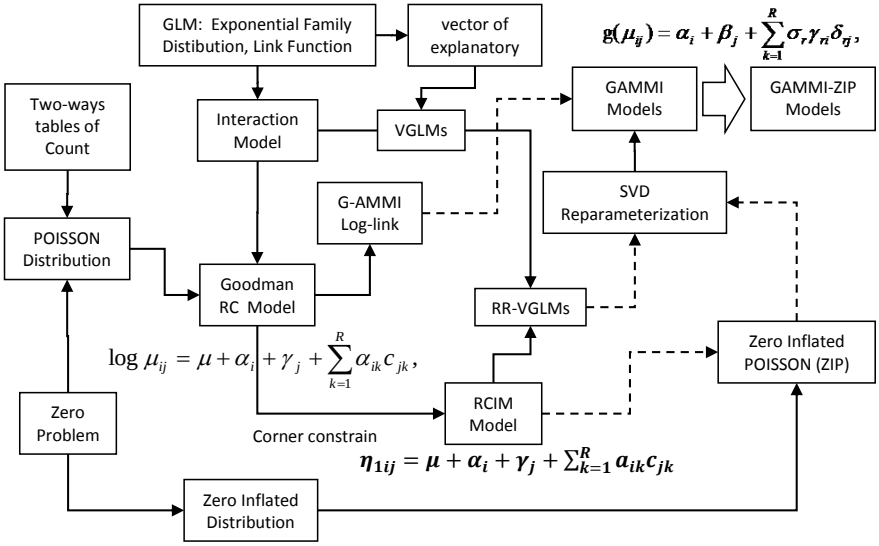


Fig. 1. TheFramework

This section will discuss the framework (Figure 1) by something's related to the development of zero-inflated in the multiplicative model. Beginning with the concept of mixture distribution of zero-inflated Poisson, and the regression framework terampat by reduction of dimensions or Reduced Rank Vector Generalized Linear Models (RR-VGLMs) which was introduced by Yee and Hastie [17]. Then the Row-Column Interaction Model (RCIM) Yee & Hadi [16].

4.1 The Zero-Inflated Poisson Distribution

A (discrete) random variable Yi is said to have a zero-inflated distribution if it has value 0 with probability ω, otherwise it has some other distribution with P(Y = 0) > 0. Hence P(Y = 0) comes from two sources, and the ω source can sometimes be thought of as a structural zero. The most famous of Zero-inflated distribution is the distribution of zero-inflated Poisson (ZIP) (Yee, [19]).

Zero-infated poisson (ZIP) model, well described by Lambert [7] is a simple mixture model for count data with excess zeros. The model is a combination of a Poisson distribution and a degenerate distribution at zero. Specifically if Yi are independent random variables having a zero-infated Poisson distribution, the zeros are assumed to arise in to ways corresponding to distinct underlyingstates. The first state occurs with probability zero ω_i and produces only zeros, whilethe other state occurs with probability $1-\omega_i$ and leads to a standard Poisson countwith mean λ and hence a chance of further zeros. In general, the zeros from the first state are called structural zeros and those from the Poisson distribution are called sampling zeros (Jansakul and Hinde, [12]). This two-state process gives a simple two-component mixture distribution with p.m.f

$$f(Y_i = y_i) = \begin{cases} \frac{(1-\omega_i)e^{-\theta_i}\theta_i{}^{y_i}}{y!}\ ;\ y = 1,2,3,\ldots \\ \omega_i + (1-\omega_i)e^{-\theta_i};\ y = 0 \end{cases}$$

$$\text{with } 0 \le \omega_i < 1 \qquad (1)$$

The mean and variance are:

$$E(Y_i) = \sum_{y_i=0}^{n} y_i[\mathrm{I}_{(y_i=0)}.\left(\omega_i + (1-\omega_i)e^{-\theta_i}\right) + \mathrm{I}_{(y_i>0)}.(1-\omega_i)e^{-\theta_i}\theta_i{}^{y_i}/y_i!]$$

$$E(Y_i) = \sum_{y_i=1}^{n} y_i\,(1-\omega_i)e^{-\theta_i}\theta_i{}^{y_i}/y_i! = \cdots = \theta_i(1-\omega_i) = \mu_i \qquad (2)$$

$$E(Y_i^2) = \sum_{y_i=0}^{n} y_i{}^2[\mathrm{I}_{(y_i=0)}.\left(\omega_i + (1-\omega_i)e^{-\theta_i}\right) + \mathrm{I}_{(y_i>0)}.(1-\omega_i)e^{-\theta_i}\theta_i{}^{y_i}/y_i!]$$

$$E(Y_i^2) = \sum_{y_i=0}^{n} y_i{}^2((1-\omega_i)e^{-\theta_i}\theta_i{}^{y_i})/y_i! = \cdots = (1-\omega_i)\left(\theta_i + \theta_i{}^2\right)$$

$$Var(Y_i) = E(Y_i^2) - \left(E(Y_i)\right)^2 = (1-\omega_i)(\theta_i + \theta_i{}^2) - \left(\theta_i(1-\omega_i)\right)^2.$$

$$Var(Y_i) = \cdots = \theta_i(1-\omega_i)(1+\omega_i\theta_i) = \cdots = \mu_i + \left(\frac{\omega_i}{1-\omega_i}\right)\mu_i^2 \qquad (3)$$

By writing $\mathrm{E}(\mathrm{Y_i}) = \mu_i$ dan $\mathrm{Var}(\mathrm{Y_i}) = \mu_i + \left(\frac{\omega_i}{1-\omega_i}\right)\mu_i^2$seen that the distribution of conditional $\mathrm{Y_i}$shows a phenomenon of over-dispersion, if $\omega_i > 0$. It is clear that this reduces to the standard Poisson model when $\omega_i = 0$. For a random sample of observations $y_1, y_2 y_3, \ldots, y_n$; the log-likelihood function is given by

$$\ell = \ell(\lambda, \omega; \mathbf{y}) = \sum_i \{ I_{(y_i=0)} \ln[\omega_i + (1-\omega_i)e^{-\lambda_i}] + I_{(y_i>0)}[\ln(1-\omega_i) - \lambda_i + y_i \ln\lambda_i - \ln(y_i!)]\},$$

where I(.) is the indicator function for the specified event, i.e. equal to 1 if the event is true and 0 otherwise. To apply the zero-infated Poisson model in practicalmodelling situations, Lambert [7] suggested the following joint models for
ω_i and λ_i as follows:

$$ln\left(\frac{\boldsymbol{\omega}}{1-\boldsymbol{\omega}}\right) = \boldsymbol{G\gamma} \text{ dan } ln(\boldsymbol{\lambda}) = \boldsymbol{XB}$$

4.2 Reduce – Rank Vector Generalized Linear Models

Suppose our data comprises(x_i, y_i) for $i = 1, \cdots, n$ where x_i denotes the vector of explanatory variables for the ith observation, and y_i is the response (possibly a vector). The first value of x_i is unity for an intercept term.VGLMs are similar to ordinary GLMs but allow for multiple linear predictors. VGLMSs handle M linear predictors (the dimension M depends on the model to be fitted) where the jth one is

$$n_j = n_j(x) = \beta_j^T x = \sum_{k=1}^{p} \beta_{(j)k}\, x_k \ , j = 1, \cdots, M \tag{4}$$

The ηj of VGLMs may be applied directly to parameters of a distribution, θj, rather than just to mean μ = E(Y) as for GLMSs, in general, $\eta_j = g_j(\theta_j)$ for some parameter link function g_j and parameter θ_j.

$$\eta = \eta(x) = \begin{bmatrix} \eta_1(x) \\ \vdots \\ \eta M(x) \end{bmatrix} = \mathrm{B}^{\mathrm{T}} x = \begin{bmatrix} \mathrm{B}_1^{\mathrm{T}} x \\ \vdots \\ \mathrm{B}_{\mathrm{M}}^{\mathrm{T}} x \end{bmatrix} \tag{5}$$

where B is a p × M matrix of (sometimes too many) regression coefficients. In many situations the regressions coefficients are related to each other. For example, some of the $\beta_{(j)k}$ may be equal, set to zero, or add up to a certain quantity. These situations may be dealt with by use of constraint matrices. VGLMs in general have

$$n_j(x) = \sum_{k=1}^{p} \mathbf{B}_k \beta^*_{(k)} x_k, \ \ j = 1,\ldots,M \tag{6}$$

where H1, H2 ,…, Hp are known full-column rank constraint matrices and $\beta^*_{(k)}$ are vectors of unknown coefficients. With no one constraint at all H1 = H2 = … =Hp = IM. then, for VGLMs,

$$\mathbf{B}^T = \left(H_1\beta^*_{(1)} \quad H_2\beta^*_{(2)} \ldots H_p\beta^*_{(p)}\right) \tag{7}$$

Partition x into $\left(x_1^T, x_2^T\right)^T$ (of dimension $p_1 + p_2 = p$) and B = $\left(\mathbf{B}_1^T, \mathbf{B}_2^T\right)^T$ if $\mathbf{B}_2$ has too many regression coefficients then we can reduce its number dramatically by a reduced-rank regression. RRVGLMs then have:

$$\eta = \mathbf{B}_1^T x_1 + \mathbf{B}_2^T x_2 \tag{8}$$

where we approximate $\mathbf{B}_2$ by a reduced-rank regression

$$\mathbf{B}_2 = \mathbf{C}\,\mathbf{A}^T. \tag{9}$$

Here, $\mathbf{C}$ and $\mathbf{A}$ are $p_2 \times R$ and $M \times R$ respectively, and they are 'thin' because the rank R is low, e.g., R = 1 or 2. Thus

$$\eta = \mathbf{B}_1^T x_1 + \mathbf{A}\nu \tag{10}$$

where $\nu = \mathbf{C}^T x_2$ is a vector of R latent variables.

To make the parameters unique, it is common to enforce corner constraint on $\mathbf{A}$. By default, the top $R \times R$ submatrix is fixed to be $\mathbf{I_R}$ and the remainder of $\mathbf{A}$ is estimated.

4.3 Row-Column Interaction Model for Data count in the RR-VGLM

We use Goodman's RC association model (Goodman, [3]) to explain what a RCIM is. For more background see Yee and Hastie [17]. How does Goodman's RC association model fit within the VGLMs framework? Suppose Y = [(yij)] be a n × M matrix of counts. Goodman's model fits a reduced-rank type model to Y by firstly assuming that Yij has a Poisson distribution, and that

$$\eta_{ij} = \log(\mu_{ij}) = \mu + \alpha_i + \gamma_j + \sum_{k=1}^{R} a_{ik} c_{jk} \tag{11}$$

where $\mu_{ij} = \mathrm{E}(Y_{ij})$ is the mean of the i-j cell. Identifiability constraint are needed in (11) for the row and column effects $\alpha_i$dan γ_j; we use corner constraints $\alpha_i = \gamma_j = 0$ in this article. The parameters a_{ik} and c_{jk} also need constraints, e.g., we use $k = 1, \ldots, R$ for $a_{ik} = c_{jk} = 0$. We can write (11) as $\log(\mu_{ij}) = \mu + \alpha_i + \gamma_j + \delta_{ij}$. Where the $n \times M$ matrix $\Delta = [(\delta_{ij})]$ of interaction terms is approximated by the reduced ran quantit y $\sum_{k=1}^{R} a_{ik} c_{jk}$. Goodman's RCassociation model fits within the VGLMs framework by letting $\eta_i = \log \mu_i$. Where $\mu_i = \mathrm{E}(Y_i)$ is the mean of ith row of Y. Then the matrix $(\eta_1, \ldots, \eta_n)^T$ fits into RR-VGLM framework as follows. From last section, we obtain

$$\mathbf{B}_1^T x_{1j} = \begin{pmatrix} \mu 1_M & \alpha_2 1_M & \cdots & \alpha_n 1_M \end{pmatrix} \begin{pmatrix} 1 \\ e_{(-1)i} \\ 1_{M-1} \end{pmatrix}$$

when a subscript "(-1)" means the first element or row is removed from the vector or matrix. This shows, for example,that the intercept and row score variables have 1M as their constraint matrices. Similiarly, because B2 is approximated by $\boldsymbol{CA}^T$, the i-th row of Δ will be approximated by $\boldsymbol{x_{2i}^T C A}^T$, or equivalently, Δ is approximated by

$$\begin{pmatrix} \boldsymbol{x_{21}} \\ \vdots \\ \boldsymbol{x_{2n}} \end{pmatrix} \boldsymbol{CA}^T.$$

The desired reduced-rank approximation of Δ can be obtained if $\boldsymbol{x}_{2\mathrm{i}} = \boldsymbol{e_i}$ so that$\boldsymbol{I}_{\mathrm{p2}} \boldsymbol{CA}^T = \boldsymbol{CA}^T$. Note that

$$\boldsymbol{\Delta} = \begin{pmatrix} \mathbf{0} & \mathbf{0}^T \\ \mathbf{0} & \tilde{\boldsymbol{\Delta}} \end{pmatrix} \approx \boldsymbol{CA}^T = \begin{pmatrix} \mathbf{0}^T \\ \boldsymbol{C}_{(-1)} \end{pmatrix} \begin{bmatrix} \mathbf{0} & \left(\boldsymbol{A}_{(-1)}\right)^T \end{bmatrix},$$

That is, the first row of matrix of A consist of structural zeros which are 'omitted' from the reduced rank regression of $\boldsymbol{\Delta}$.

One could define RCIMs as a RR-VGLM with

$$\boldsymbol{\eta_{1ij}} = \boldsymbol{\mu} + \boldsymbol{\alpha_i} + \boldsymbol{\gamma_j} + \sum_{k=1}^{R} \boldsymbol{a_{ik} c_{jk}} \tag{12}$$

Note that (12) applies to the first linear/additive predictor; for models with M >1 one can leave $\eta_2, \ldots, \eta_M$ unchanged. Of course, choosing $\eta_1$for (12) is only for convenience. The software chooses $g_1^{-1}(\hat{\eta}_{ij})$ as the fitted values of the model and the result should be the same dimensions as the two-way table.

4.4 Zero-inflated Poisson model in the RR-VGLM

ZIP model is very powerful in dealing with count data with excess zeros than the usual Poisson distribution, partly it is because the ZIP model also handles over-dispers. Distribution as ZIP equation (1) can be mentioned that the event (Y = 0) come from two sources and in this model it with the RR-VGLM η_1 and η_2 as

$$\eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} logit\psi \\ log\mu \end{pmatrix} \tag{13}$$

There are two processes of how the data occurs, the first data is zero and the Poisson count data. Both processes are modeled respectively by η_1 and η_2.

Liu and Chan [8] gives examples of research involving surveys in which the spatio-temporal aggregation of fish catches show a positive opportunity similar to a monotonic function of the middle value.

Liu and Chan [8] proposed several new methodologies that allow ZIP to handle the linkage between the two processes. They call it COZIGAM, which is constrained zero-inflated generalized additive models. Which in the fact now, can be seen simply that this is a dimension reduction regression models ZIP or reduced-rank zero-inflated Poisson model (RR-ZIP). RR-ZIP is given by

$$logit\psi = \eta_1 = \beta_{(1)1} + \alpha_{11}.\eta_2 \tag{14}$$

$$log\mu = \eta_2 = \beta_2^T \mathcal{X} \tag{15}$$

with β(1)1 and a(1)1 is coeficient who want predictable.

Actually, because ηj here is the linear predictor as in equation (4), then equation (14) and (15) should be called COZIGLM. Connectedness can be seen, for example, if μ increases, η2 increases, and then η1 and increasing ψ if $a_{(1)1}$appreciating positive.

Equation (14) and (15) is a model RR-ZIP was rank 1 with $H_1 = I_2$ and $H_2 = \ldots = H_p = (a_{(1)1})^T$. There is a trivial complication that the constraint angle (can use other constraints) imposed on parameters that are used instead of the first two. This can be simplified if the order parameter exchanged.

4.5 SVD-Reparameterization for the GAMMI Model

An overlapping methodology is the generalized additive main effects and multiplicative interaction models, or GAMMI models, of Turner and Firth [13]. These also comprise the row and column main effects plus one or more components of the multiplicative interaction. The singular value corresponding to each multiplicative

component is often factored out, as a measure of the strength of interaction between the row and column scores, indicating the importance of the component, or axis. For cell means μ_{ij}a GAMMI-R model has the form

$$g(\mu_{ij}) = \alpha_i + \beta_j + \sum_{k=1}^{R} \sigma_k \gamma_{ki} \delta_{kj}. \quad (13)$$

Based on (13) GAMMI models appear to be identical to RCIMs. Here they apply a SVD to our AC^T. While our interaction term uses corner constraints, their SVD parameterization is quite interpretable and is related to some of the other parameterizations described in Yee and Hastie [17]. The advantage of RCIMs is that it should work for any VGAM family function, thus the family size is much bigger. It is easy to perform some post-transformations such as applying svd() to the VGAM output to obtain the SVD parameterization for GAMMI models.

5. APPLICATION: LEAF RUST DISEASE ATTACKS ON MUNG BEAN

The data comes from the Indonesian Legumes and Tuber Crops Research Institute (ILETR) Malang. This trial involved 10 genotypes and two green beans varieties which planted in 5 different environments at Probolinggo, Jombang, Jember, Rasanae, and Bolo. Experiments conducted on plots of size 4x $5m^2$ with a spacing of 40 cmx 10cm, two seeds per hole. The design in each environment is completely randomized, with 3 replications. One of the researchers' attention is on resistance to *leaf rust* disease. This disease is a not major disease. Observations done on trials field without inoculated. Theatrically it allows to be happened what is called by the term "escape", event with no attack. Statistically it is such of the zero-inflated phenomenon. Table 1 shows the amount average from three replications.

Table 1:
Count of *Leaf Rust* Disease Attacks on Mung Bean

Genotype		Locations (Environments)				
		Proboliggo	**Jember**	**Jombang**	**Bolo**	**Rasanae**
MLG	1002	0	167	100	150	150
MLG	1004	0	217	250	233	250
MLG	1021	0	200	217	183	217
MMC	74d Kp1	0	133	200	183	133
MMC	71d Kp2	0	200	200	233	367
MMC	157d Kp1	0	133	150	167	150
MMC	203d Kp5	0	50	100	67	83
MMC	205e	0	50	67	100	67
MMC	100f Kp1	0	50	83	83	83
MMC	87d Kp5	0	83	117	133	83
MURAI		0	0	50	33	33
PER KUTUT		0	67	133	117	117

Data were analyzed using the VGAM package with SVD reparameterization in the RCIM model follows what is done by Turner & Fifth[13] on the Poisson distribution with the GAMMI model proposed by VanEeuwijk [14].

Zero inflated Poisson (ZIP) model give us two results, the logit and the log-part. The logit-pat give us the probability of being zero at random, from a intercept only model. The log-part is our main attentions since it give us more information. The model used is the RCIM model with rank = 0 and SVD-reparaterization on working residuals to get the interaction with rank =2. The data were number of crops attacked by leaf-rust, so that genotype with large numbers indicate that it is vulnerable. Genotype with average count in almost allocations, said to be *stable* (in fact, it is vulnerable) and will be located close to the origin point on biplot. That is, the zero origin biplot point does not always describe the resistance of the genotype. Probolinggo are drawn close to the zero point, because of in general, the overall genotypes have the same number of zeros. However, the zeros on the observation in Probolinggo sometimes called "escape" observation, where all the columns on this row is zero. The ZIP model relies on the assumption that "zero" are as structural zero and random zero. As random zero, ZIP model will provide us the probability of cell to be zero, and the fitted value for Poisson count, as well.

Table 1
The Log-Likelihood Ratio Test

Model	DF	Log-Likelihood	G	DF-Chisq	p-value	
FullModel (Rank=4)	58	-150.2266				
GAMMI Rank=3	50	-150.2266	0	8	1.00E+00	ok
GAMMI Rank=2	40	-160.1218	19.7904	18	3.45E-01	ok
GAMMI Rank=1	28	-182.6969	64.9406	30	2.23E-04	bad
Main Effects(Rank=0)	14	-249.6574	198.8616	44	1.44E-21	bad

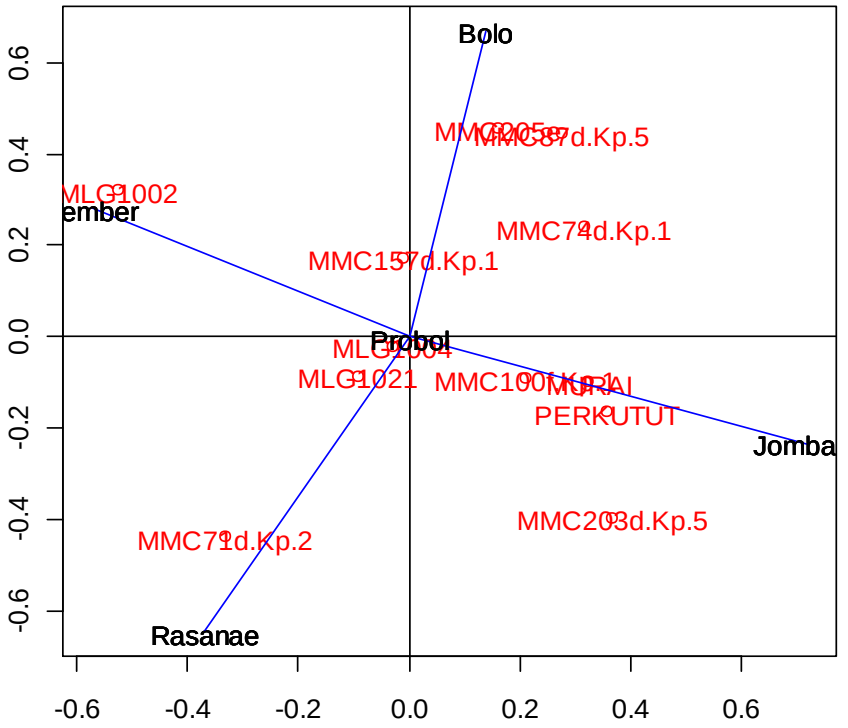


Figure 2: Biplot of the Interaction Effect on Log-Scale of Zero-Inflated Poisson

Determining the Rank 2 model, we test the existence the interaction term using Likelihood Ratio (LR) between the null model with the saturated model. In this case the LR value between Rank = 2 and Rank = 1 is equal to 19.7904 with a p-value of on Chi-square distribution (degrees of freedom = 18) which very small that is equal to 3.45e-01. With the ZIP model Rank = 2 of GAMMI-ZIP model, the biplot of it is presented in Figure 2. The variability shown by the eigen value of matrix interaction, the five root traits in a row are: 1.056, 0.807, 0.696, 0.000, 0.000. The first two eigen, explaining the total variability Biplot, hat is 72.78%.

ACKNOWLEDGEMENT

Special thanks to Thomas W. Yee, Department of Statistics, The University of Auckland for the VGAM package and for kindness during my previous visiting research. Thanks to Rudi Iswanto, Indonesian Legumes and Tuber Crops Research Institute, Malang, Indonesia for the data of agricultural trial.

REFERENCES

1. Andrews, D., Bickel, P.J., Hampel, F., Huber, P., Rogers, W. and Tukey, J.W. (1972). *Robust Estimates of Location: Survey and Advances*. Princeton, N.J.: Princeton University Press, 83-98.
2. Beaton, A.E. and Tukey, J.W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16, 147-186.
3. Brich, J.B. and Myers, R.H. (1982). Robust analysis of covariance. *Biometrics*, 38, 699-713.
4. David, H.A. (1981). *Order Statistics*. 2nd ed. New York: Wiley.
5. BoÈhning, D. (1998). Zero-Inflated Poisson Models and C.A.MAN: A Tutorial Collection of Evidence. *Biometrical Journal* 40- 7, 833-843.
6. Lord, D., Washingtonb, S.P. and Ivanc, J.N. (2005) Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis and Prevention* 37, 35-46.
7. Dietz, E. and BoÈhning, D. (2000). On estimation of the Poisson parameter in zero-modified Poisson models. *Computational Statistics and Data Analysis* 34, 441-459.
8. Goodman, L.A. (1981). Interaction models and canonical correlation in the analysis of cross-classifications having ordered categories. *Journal of the American Statistical Interaction*, 76, 320-334.
9. Hadi, A.F., Mattjik, A.A. and Sumertajaa, I.M. (2010). Generalized AMMI models for assessing the endurance of soybean to leaf pest. *Jurnal. Ilmu. Dasar*., 11(2), 151-159.
10. Hinde, J. and Demetrio, C.G.B. (1998). Over dispersion: models and estimation. *Computational Statistics and Data Analysis*, 27, 151-170.
11. Kaminski, R.J. and Thomas, D. (2009). Stucky Reassessing Political Explanations for Murders of Police. *Homicide Studies*, 13(1), 3-20.
12. Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34, 1-14.
13. Liu, H. and Chan, K.S. (2010). Introducing COZIGAM: An R package for unconstrained and constrained zero-inflated generalized additive model analysis. *Journal of Statistical Software*, 35:1-26. http://www.jstatsoft.org/v35/i11.
14. McCullagh, P. and Nelder, J.A. (1989). Generalized linear models, 2nd edition. London: Chapman & Hall.
15. Motta, M.R., Gianola, D., Heringstad, B. Rosa, G.J.M. and Chang, Y.M. (2007). A Zero-Inflated Poisson Model for Genetic Analysis of the Number of Mastitis Cases in Norwegian Red Cows. *J. Dairy Sci*., 90, 5306-5315.
16. Jansakul, N. and Hinde, J.P. (2002). Score tests for zero-inflated Poisson models. *Computational Statistics and Data Analysis*, 40, 75-96.
17. Turner, H. and Firth, D. (2009). Generalized nonlinear models in R: An overview of the GNM package. URL http://CRAN.R-project.org/package=gnm. R package version 0.10-0. UK: University of Warwick.
18. Van Eeuwijk, F.A. (1995). Multiplicative interaction in generalized linear models. *Biometrics*, 51, 1017-1032.
19. Welsh, A.H., Cunningham, R.B. and Chambers, R.L. (2000). Methodology for estimating the abundance of rare animals: seabird nesting on North East Herald Cay. *Biometrics*, 56, 22-30.

20. Yee, T.W. and Hadi, A.F. (2014). Row Column Interaction Models, with an R implementation. *Computational Statistics.* Online first. http://link.springer.com/article/ 10.1007/s00180-014-0499-9
21. Yee, T.W. and Hastie, T.J. (2003). Reduced-rank vector generalized linear models. *Statistical Modeling*, 3, 15-41.
22. Yee, T.W. (2010). The VGAM package for categorical data analysis. *Journal of Statistical Software*, 32, 1-34.
23. Yee, T.W. (2008). VGAM Family Functions for Positive, Zero-altered and Zero-Inflated Discrete Distributions. http://www.stat.auckland.ac.nz/

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 245-262

LAPAROSCOPIC VS OPEN REPAIR FOR INCISIONAL HERNIA META-ANALYSIS AND SYSTEMATIC REVIEW OF LAPAROSCOPIC VERSUS OPEN MESH REPAIR FOR ELECTIVE INCISIONAL HERNIA

Aiman Awaiz[1], Foyzur Rahman[2], Md Belal Hossain[2], Rossita Mohamad Yunus[3] Shahjahan Khan[4] Breda Memon[5] and **Muhammed Ashraf Memon[§]**

[1] Jinnah Sindh Medical University & Dow University of Health Sciences Karachi, Pakistan. Email: aiman.awaiz@gmail.com
[2] Department of Statistics, Biostatistics & Informatics, Dhaka University Dhaka, Bangladesh. Email: frshuweb@gmail.com; bjoardar2003@yahoo.com
[3] Institute of Mathematical Sciences, University of Malaya Kuala Lumpur, Malaysia. Email: rossita@um.edu.my
[4] School of Agricultural, Computing and Environmental Sciences, International Centre for Applied Climate Science, University of Southern Queensland, Toowoomba, Queensland, Australia. Email: shahjahan.khan@usq.edu.au
[5] Sunnybank Obesity Centre, Suite 9, McCullough Centre, 259 McCullough Street, Sunnybank, Queensland, Australia. Email: bmemon@yahoo.com
[6] Mayne Medical School, School of Medicine, University of Queensland, Brisbane, Queensland, Australia.
[7] Faculty of Health Sciences and Medicine, Bond University, Queensland, Australia
[8] Faculty of Health and Social Science, Bolton University, Bolton, Lancashire, UK
[§] Corresponding author Email: mmemon@yahoo.com

ABSTRACT

OBJECTIVES: The aim was to conduct a meta-analysis of RCTs investigating the surgical and postsurgical outcomes of elective incisional hernia by open versus laparoscopic method.

MATERIAL AND METHODS: A search of PubMed, Medline, Embase, Science Citation Index, Current Contents, and the Cochrane Central Register of Controlled Trials published between January 1993 and September 2013 identified all the prospective RCTs comparing surgical treatment of only incisional hernia (and not primary ventral hernias) using open and laparoscopic methods were selected. The outcome variables analyzed included (a) hernia diameter; (b) operative time; (c) length of hospital stay; (d) overall complication rate; (e) bowel complications; (f) reoperation; (g) wound infection; (h) wound hematoma or seroma; (i) time to oral intake; (j) back to work; (k) recurrence rate; and (l) post-operative neuralgia. The quality of RCTs was assessed using Jadad's scoring system. Random effects model was used to calculate the effect size of both binary and continuous data. Heterogeneity amongst the outcome variables of these trials was determined by the Cochran Q statistic and I^2 index. The meta-analysis was prepared in accordance with PRISMA guidelines.

RESULTS: Six RCTs were considered suitable for meta-analysis. A total of 378 patients underwent open mesh repair and 373 had laparoscopic repair. Statistically significant reduction in bowel complications was noted with open surgery compared to

the laparoscopic repair in five studies (OR 2.56, 95% CI 1.15, 5.72, p=0.02). Comparable effects were noted for other variables which include hernia diameter (SMD -0.27, 95% CI -0.77, 0.23, p=0.29), operative time (SMD -0.08, 95% CI -4.46, 4.30, p=0.97), overall complications (OR -1.07, 95% CI -0.33, 3.42, p=0.91), wound infection (OR 0.49, 95% CI 0.09, 2.67, p=0.41), wound hematoma or seroma (OR 1.54, 95% CI 0.58, 4.09, p=0.38), reoperation rate (OR -0.32, 95% CI 0.07, 1.43, p=0.14), time to oral intake (SMD -0.16, 95% CI -1.97, 2.28, p=0.89), length of hospital stay (SMD -0.83, 95% CI -2.22, 0.56, p=0.24), back to work (SMD -3.14, 95% CI -8.92, 2.64, p=0.29), recurrence rate (OR 1.41, 95% CI 0.81, 2.46, p=0.23), and postoperative neuralgia (OR 0.48, 95% CI 0.16, 1.46, p=0.20).

CONCLUSIONS: On the basis of our meta-analysis, we conclude that laparoscopic and open repair of incisional hernia is comparable. A larger randomized controlled multicenter trial with strict inclusion and exclusion criteria and standardized techniques for both repairs is required to demonstrate the superiority of one technique over the other.

KEYWORDS

Hernia; Incisional; Abdomen; Abdominal Wall; Abdominal wall surgery; Hernia surgery; Randomized controlled trials; Open methods; Laparoscopic methods.

INTRODUCTION

Every surgical procedure that requires access through the abdominal wall carries a risk of development of incisional hernia. Incisional hernias are mostly related to failure of the fascia to heal and involve technical and biological factors. They may cause pain, increase in size over time, and also result in severe complications such as bowel incarceration and strangulation. A vast majority of open surgical repair of incisional hernias are achieved using a prosthetic mesh which is still associated with early or late complications such as mesh complications and the recurrence rate of approximately 32% over a 10-year follow up period [Burger et al., 2004, Teserteli et al., 2008]. LeBlanc et al. in 1993 [LeBlanc & Both 1993] reported the first case of laparoscopic incisional hernia repair using a synthetic mesh to improve upon the open method. Since the introduction of this technique, a number of randomized control trials (RCTs) comparing laparoscopic and open methods have been published analyzing various aspects of these approaches. The objective of this meta-analysis was to determine the clinical outcomes, safety and effectiveness of laparoscopic repair compared with open repair for elective surgical treatment of incisional hernia only.

MATERIALS AND METHODS

Search Strategy and Data Collection

RCTs were identified by conducting comprehensive search of electronic databases, PubMed, Medline, Embase, Science Citation Index, Current Contents and the Cochrane Central Register of Controlled Trials published between January 1993 and September 2013 using medical subject headings (MESH); “hernia,” “incisional,” “abdominal,” “randomized/randomised controlled trial,” “abdominal wall hernia,” “laparoscopic repair,” and “open repair”; “Human”; and “English”. We further searched the reference lists of all included primary studies and existing meta-analysis by hand for additional

citations. Data extraction, critical appraisal and quality assessment was carried out by two authors (AA, MAM). The authors were not blinded to the source of the document or authorship for the purpose of data extraction. Standardized data extraction forms were used by authors to independently and blindly summarize all the data available in the RCTs meeting the inclusion criteria [Moher et al., 1999]. The data were compared and discrepancies were addressed with discussion until consensus was achieved. The analysis was prepared in accordance with the Preferred Reporting of Systematic Reviews and Meta-Analyses (PRISMA) statement [Moher et al., 2009]. Random effect model was used for analysis of all the variables.

The included RCTs must have reported on at least one clinically relevant outcome pertaining to the intraoperative and postoperative period. Only adult (>18 years) patients requiring elective surgical intervention purely for the repair of incisional hernia either by open or laparoscopic method were the target population for this meta-analysis. Exclusion criteria included studies that investigated the effect of open versus laparoscopic repair in a mixture of primary and incisional hernia repair and duplicate publications. The 12 outcome variables analyzed included (a) hernia diameter; (b) operative time; (c) length of hospital stay; (d) overall complication rate; (e) bowel complications; (f) reoperation; (g) wound infection; (h) wound hematoma or seroma; (i) time to oral intake; (j) back to work; (k) recurrence rate; and (l) post-operative neuralgia. We used the Jadad scoring system to evaluate the methodological quality of the identified RCT's [Haynes et al., 2006, Jadad et al., 1996].

Statistical Analysis and Risk of bias across Studies

Meta-analyses were performed using odds ratios (ORs) for binary outcome and standardized mean differences (SMDs) for continuous outcome measures. The slightly amended estimator of OR was used to avoid the computation of reciprocal of zeros among observed values in the calculation of the original OR [Agresti et al., 1996]. Random effects model based on the inverse variance weighted method approach was used to combine the data [Sutton et al., 2000]. Heterogeneity among studies was assessed using the Q statistic and I^2 index [Higgins et al., 2002, Hedges et al., 1985, Cochran et al., 1954, Huedo-Medina et al., 2006, Sutton et al., 2000]. If the observed value of Q was greater than the associated x^2 critical value at a given significant level, in this case 0.05, we conclude the presence of statistically significance between-studies variation. In order to pool continuous data, mean and standard deviation of each study is required. However, some of the published clinical trials did not report the mean and standard deviation, but rather reported the size of the trial, the median and range. Using these available statistics, estimates of the mean and standard deviation were obtained using formulas proposed by Hozo [Hozo et al., 2005]. Funnel plots were created in order to determine the presence of publication bias in the present meta-analysis. Both total sample size and precision (reciprocal of standard error) were plotted against the treatment effects (OR for dichotomous variables and SMD for continuous variables) [Egger et al., 1997, Tang et al., 2000, Span et al., 2006]. All estimates were obtained using a computer program written in R [R: Language and Environment for Statistical Computing [Computer Program]. All plots were obtained using the metafor-package [Viechtbauer et al., 2010]. In the case of tests of hypotheses, the paper reports p-values for different statistical tests on the study variables. In general, the effect is considered to be statistically significant if

the p-value is small. If one uses a 5% significance level then the effect is significant only if the associated p-value is ≤5%.

RESULTS

The six studies, [Asencio et al., 2009, Eker et al., 2013, Itani et al., 2010, [Navarra et al., 2007, Olmi et al., 2007, Rogmark et al., 2013] that met the inclusion criteria are detailed in Table 1, Fig 1. The pooled data for the 12 outcomes are summarized in Table 2. Statistically significant reductions in bowel complications was noted with open surgery compared to the laparoscopic repair based on five studies namely [Asencio et al., 2009, Eker et al., 2013, Itani et al., 2010, Olmi et al., 2007, Rogmark et al., 2013] (OR 2.56, 95% CI 1.15, 5.72, p=0.02) (Fig 2). Comparable effects were noted for other variables which include hernia diameter (SMD -0.27, 95% CI -0.77, 0.23, p=0.29) (Fig 3), operative time (SMD -0.08, 95% CI -4.46, 4.30, p=0.97) (Fig 4), overall complications (OR -1.07, 95% CI -0.33, 3.42, p=0.91) (Fig 5), wound infection (OR 0.49, 95% CI 0.09, 2.67, p=0.41) (Fig 6), wound hematoma or seroma (OR 1.54, 95% CI 0.58, 4.09, p=0.38) (Fig 7), reoperation rate (OR 0.32, 95% CI 0.07, 1.43, p=0.14) (Fig 8), time to oral intake (SMD -0.16, 95% CI -1.97, 2.28, p=0.89) (Fig 9), length of hospital stay (SMD -0.83, 95% CI -2.22, 0.56, p=0.24) (Fig 10), back to work (SMD -3.14, 95% CI -8.92, 2.64, p=0.29) (Fig 11), recurrence rate (OR 1.41, 95% CI 0.81, 2.46, p=0.23) (Fig 12), and postoperative neuralgia (OR 0.48, 95% CI 0.16, 1.46, p=0.20) (Fig 13). The RCTs collectively demonstrated moderate methodological quality based on Jadad score with an average score of 2.7 (out of five), with a range of 2 to 3 (Table 1).In general there was a high degree of heterogeneity detected for most of the outcomes in the included studies except for bowel complications, recurrence rate, reoperation and neuralgia (Table 2). Most of the funnel plots demonstrate asymmetry and thus suggest the presence of publication bias for a majority of outcomes (Fig 14).

Table 1
Salient Features of Various RCTs

Authors/Year	Pt	Open	Lap	Follow-up	Jadad Score		
	n	n	n	months	Randomized	Blinding	Dropouts/ Withdrawals
Olmi at al./2006	170	85	85	24	1	0	0
Navara et al./2007	24	12	12	6	2	0	0
Asencio et al./2008	84	39	45	12	2	0	1
Itani et al./2010	146	73	73	2	2	0	1
Eker et al./2013	194	100	94	35	2	0	1
Rogmark et al./2013	133	69	64	2	2	0	1

Lap= Laparoscopic, n= number, Pt= Patient

Table 2
Polled Statistics

Clinical Variable	Pt	Pooled Statistics	Overall effect Test		Test for heterogeneity		
	n	SMD or OR [CI]	Z	Pr	*Q*	Pr	I^2 [CI] in %
Hernia Diameter	751	-0.27 [-0.77; 0.23]	-1.06	0.29	56.88	<0.0001	90.64 [75.14; 98.37]
Operative Time	605	-0.08 [-4.46; 4.30]	-0.03	0.97	456.7	<0.0001	99.73 [NA; NA]
Bowel Complications	751	2.56 [1.15; 5.72]	2.30	0.02	1.38	0.93	0 [0; 42.56]
Complications	751	1.07 [0.33; 3.42]	0.11	0.91	47.22	<0.0001	90.64 [72.87; 98.53]
Wound Infection	751	0.49 [0.09; 2.67]	-0.83	0.41	21.11	<0.0001	74.07 [30.43; 94.84]
Wound Hematoma/ Seroma	751	1.54 [0.58; 4.09]	0.87	0.38	16.99	0.0045	74.03 [25.06; 96.09]
Reoperation	411	0.32 [0.07; 1.43]	-1.49	0.14	0.91	0.82	0 [0; 73.66]
Oral Intake	108	0.16 [-1.97; 2.28]	0.14	0.89	19.45	<0.0001	94.86 [NA; NA]
LOS	751	-0.83 [-2.22; 0.56]	-1.17	0.24	226.4	<0.0001	98.64 [96.45; 99.77]
Back To Work	316	-3.14 [-8.92; 2.64]	-1.06	0.29	217.1	<0.0001	99.54 [NA; NA]
Recurrence	751	1.41 [0.81; 2.46]	1.21	0.23	0.22	0.99	0 [NA; NA]
Neuralgia	303	0.48 [0.16; 1.46]	-1.28	0.20	0.01	0.94	0 [0; 84.94]

N= number, NA= Not available, OR= Odds ratio, SMD= Standardized mean difference

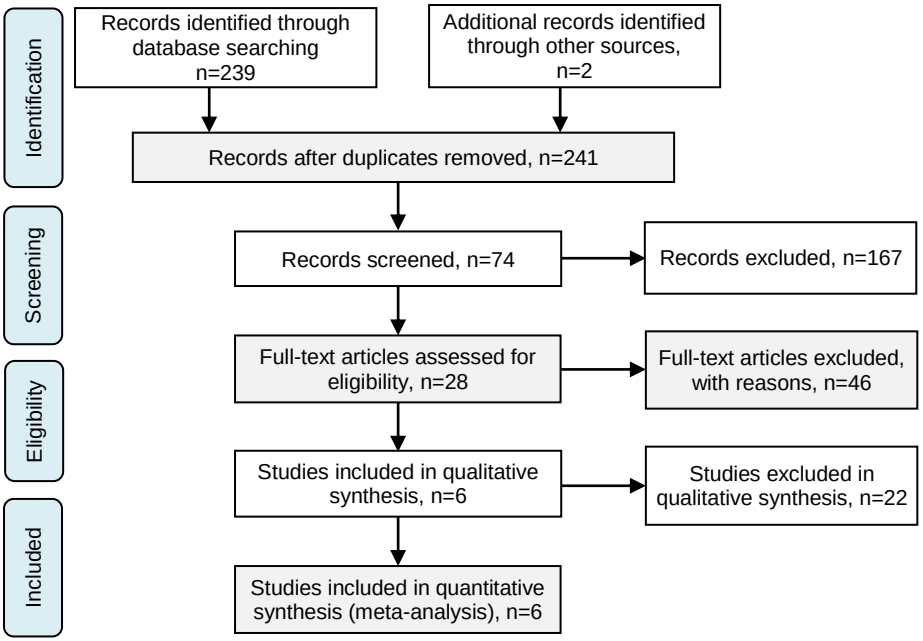


Figure 1: PRISMA Flow Diagram

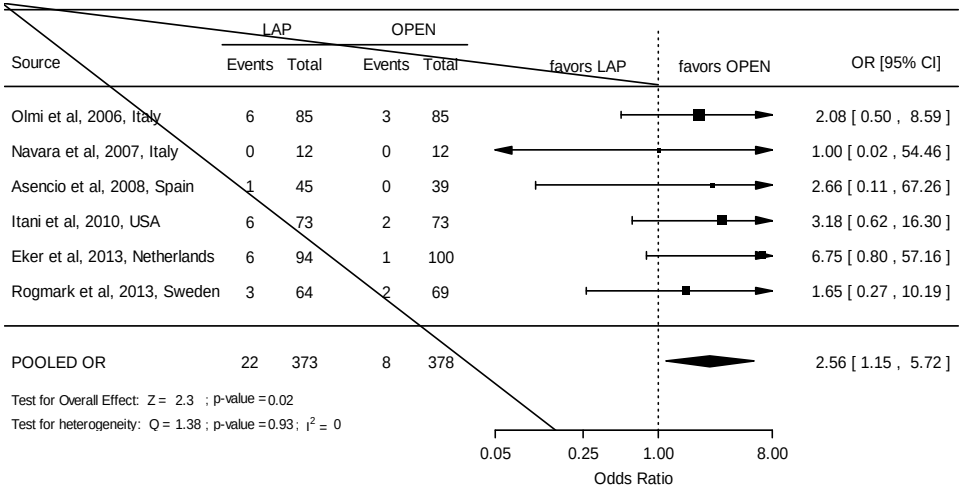


Figure 2: Forest Plot of Bowel Complications

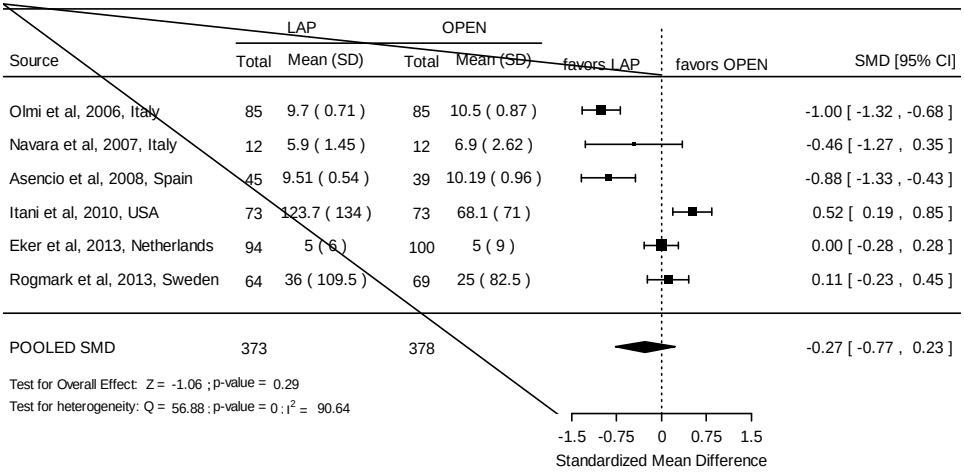


Figure 3: Forest Plot of Hernia Diameter

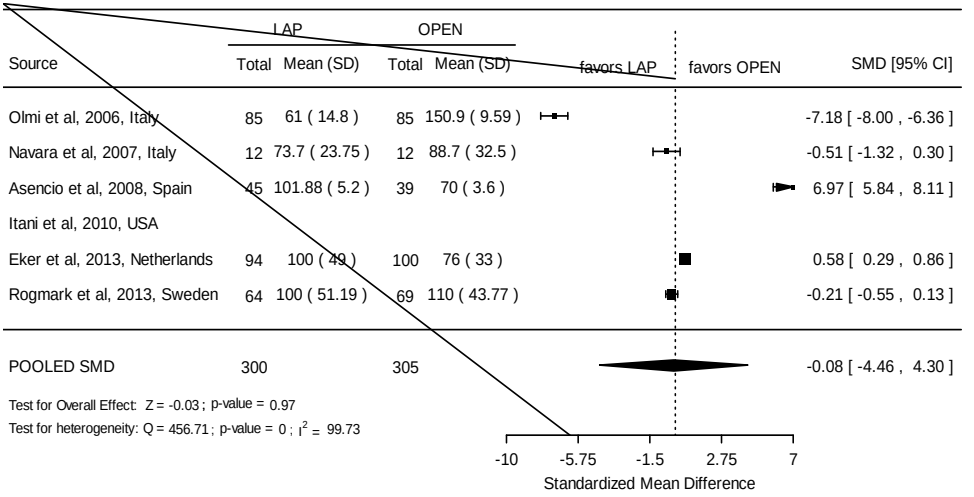


Figure 4: Forest Plot of Operative Time

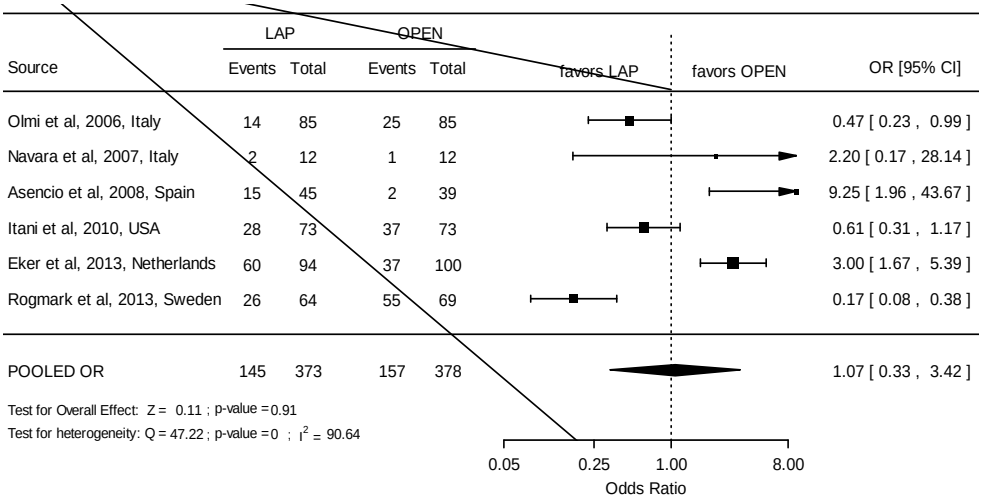


Figure 5: Forest Plot of Overall Complications

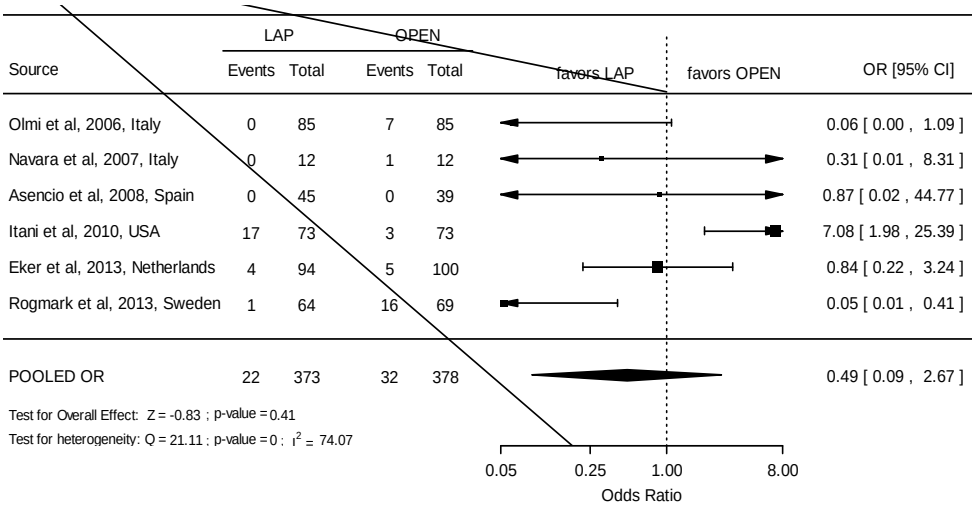


Figure 6: Forest Plot of Wound Infection

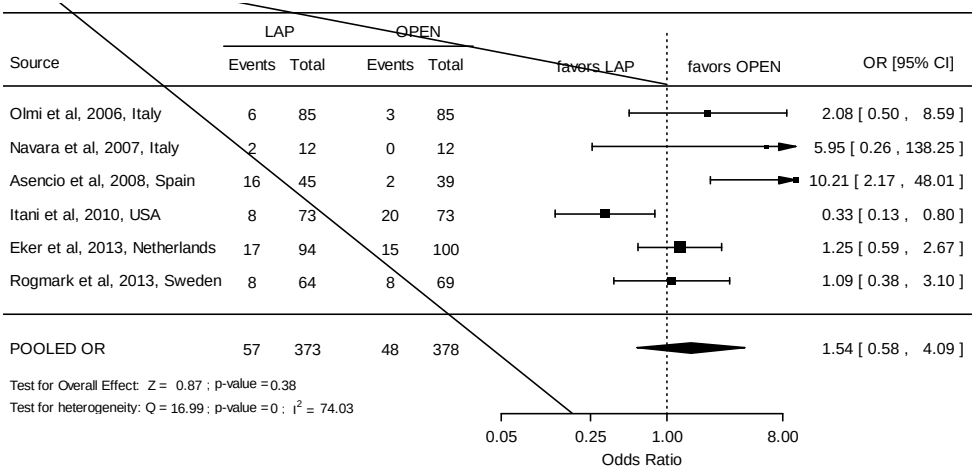


Figure 7: Forest Plot of Wound Haematoma or Seroma

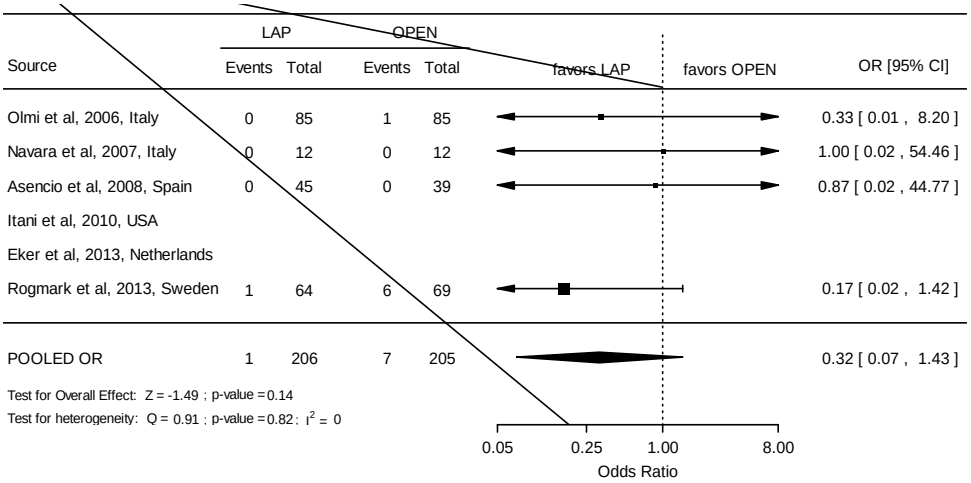


Figure 8: Forest Plot of Reoperation

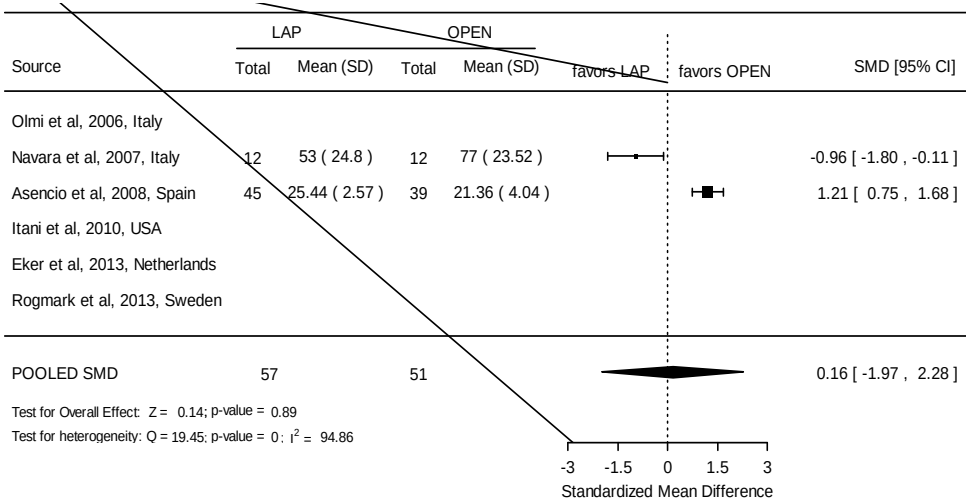


Figure 9: Forest Plot of Time to Oral Intake

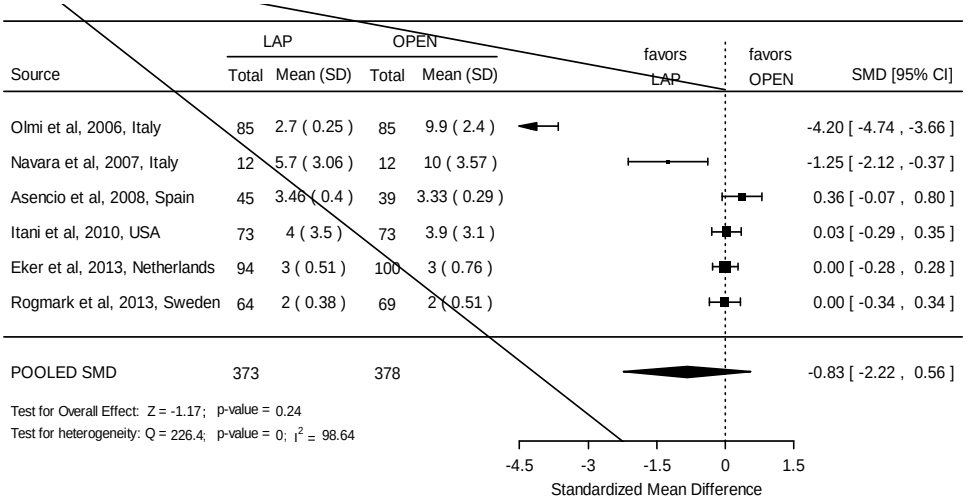


Figure 10: Forest Plot of Length of Hospital Stay

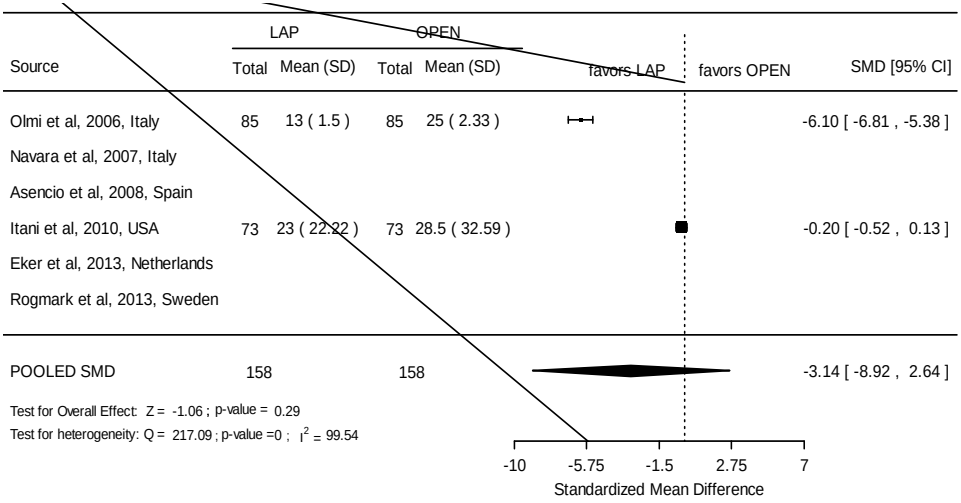


Figure 11: Forest Plot of Back to Work

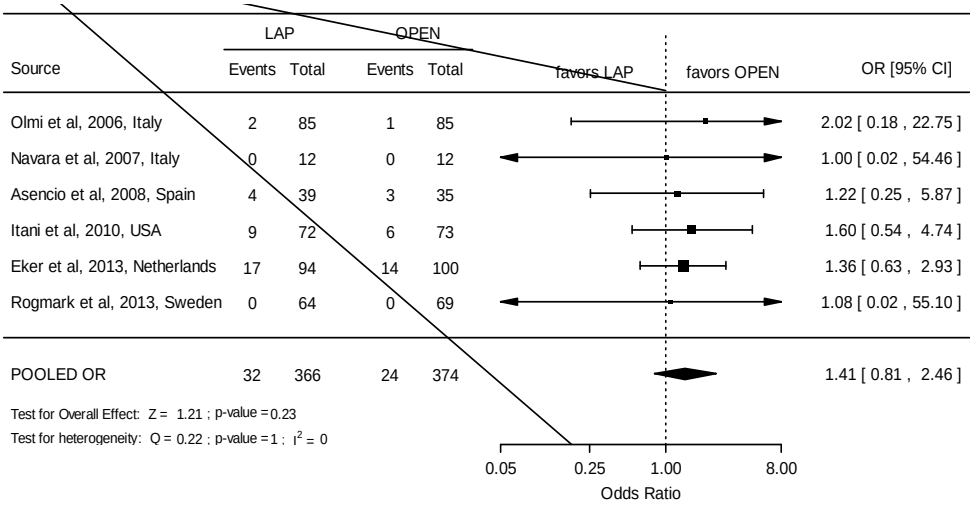


Figure 12: Forest Plot of Recurrence

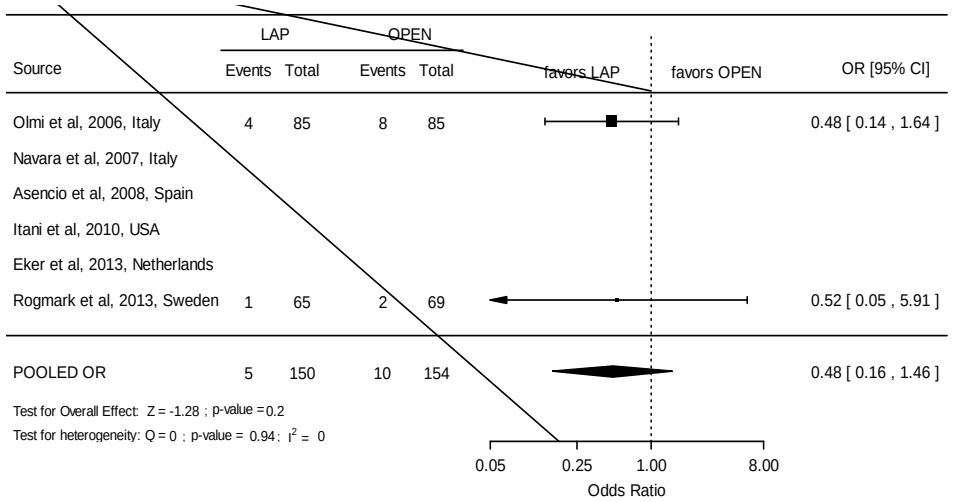
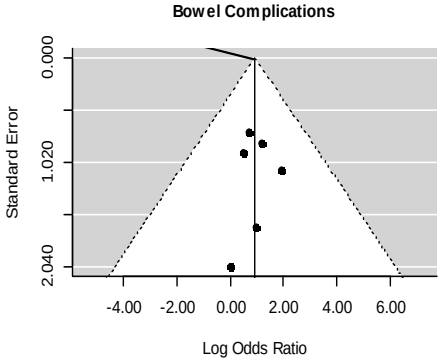
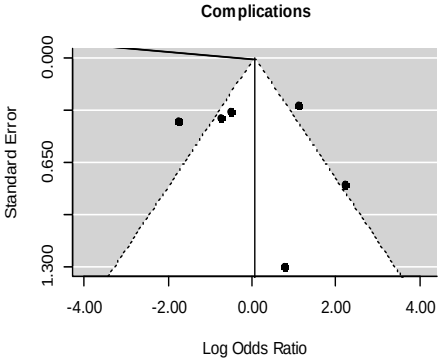
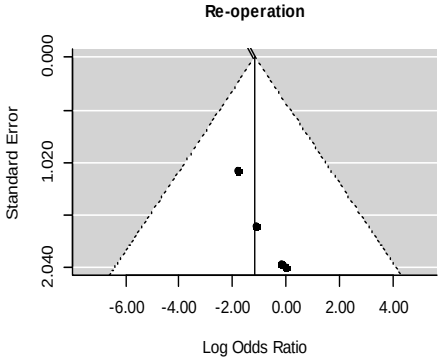
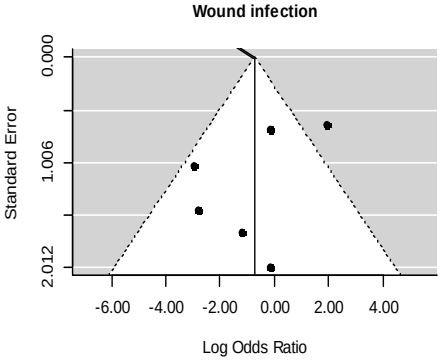


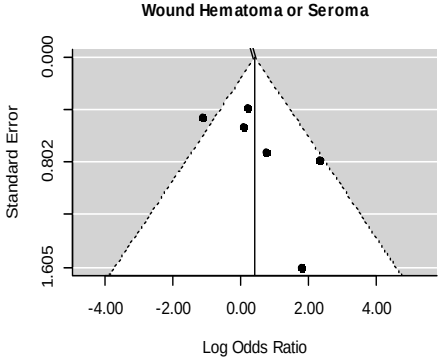
Figure 13: Forest Plot of Neuralgia

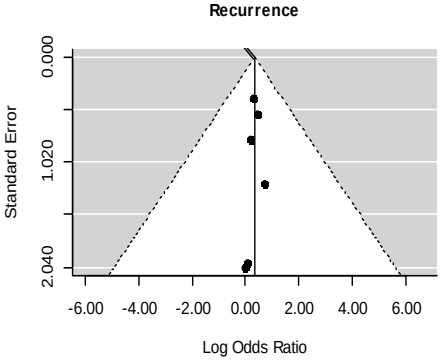

Bowel Complications
Standard Error
0.000
1.020
2.040
-4.00
-2.00
0.00
2.00
4.00
6.00
Log Odds Ratio


Complications
Standard Error
0.000
0.650
1.300
-4.00
-2.00
0.00
2.00
4.00
Log Odds Ratio


Re-operation
Standard Error
0.000
1.020
2.040
-6.00
-4.00
-2.00
0.00
2.00
4.00
Log Odds Ratio


Wound infection
Standard Error
0.000
1.006
2.012
-6.00
-4.00
-2.00
0.00
2.00
4.00
Log Odds Ratio


Wound Hematoma or Seroma
Standard Error
0.000
0.802
1.605
-4.00
-2.00
0.00
2.00
4.00
Log Odds Ratio


Recurrence
Standard Error
0.000
1.020
2.040
-6.00
-4.00
-2.00
0.00
2.00
4.00
6.00
Log Odds Ratio

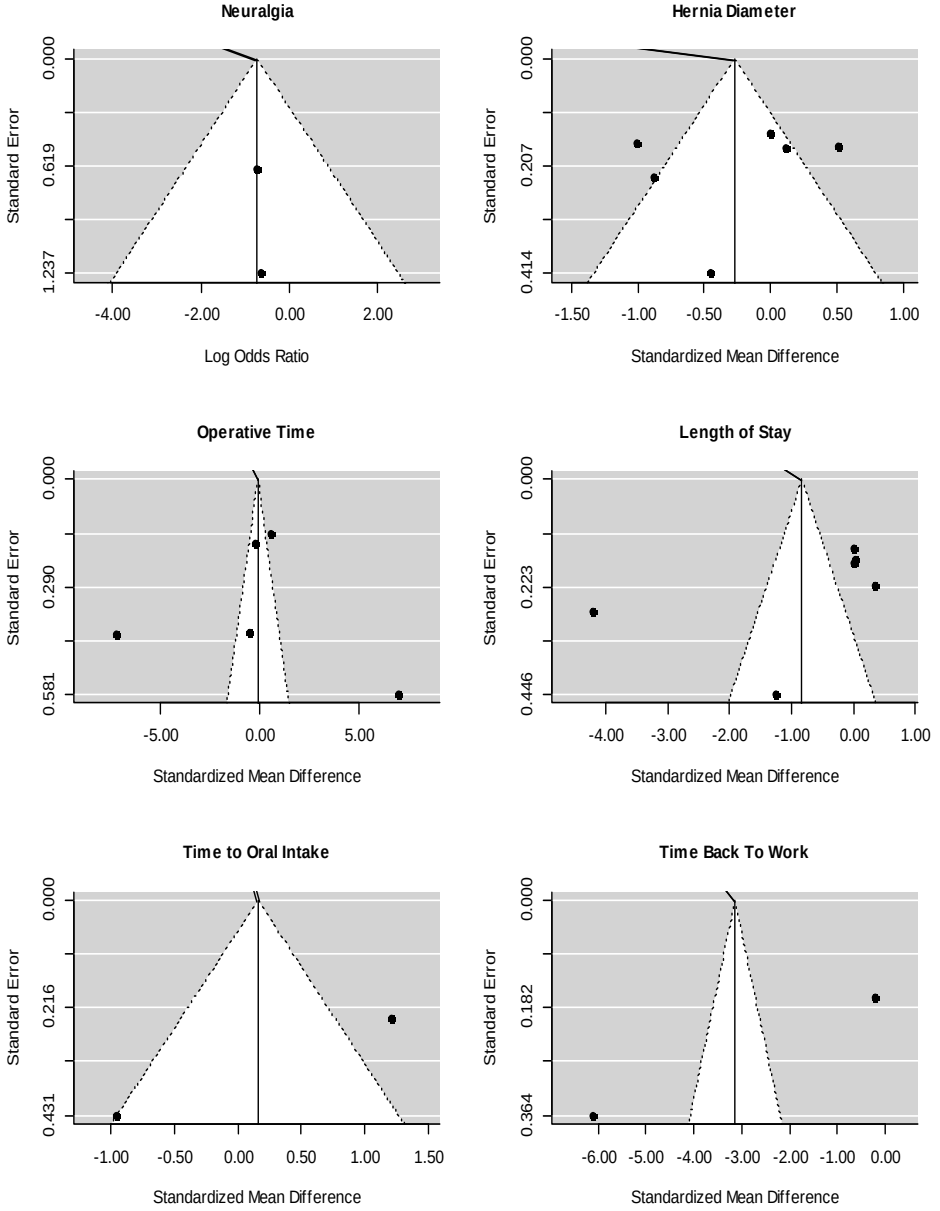


Figure 14: Funnel Plots

DISCUSSION

In the modern surgical era, laparoscopic repair has increasingly been utilized in the management of incisional hernia. First described by Le Blanc {LeBlanc et al., 1993], the technique has evolved and is now replacing open repairs where possible. Large multi-centered series [Bencini et al., 2004, Ben-Haim et al., 2002, Moreno-Egea et al., 2004, Rosen et al., 2003, Ujiki et al., 2004] have described outstanding outcomes with laparoscopic techniques citing less complications and recurrence rates of less than 10%.

We observed that laparoscopic technique was used to repair larger hernia diameters at times in our meta-analysis (Fig 3). There could be a number of explanations for this discrepancy. First of all the laparoscopic technique quite often detects more than one hernia defects whether large or small with ease. Secondly it is entirely possible that by inflating the abdomen in the laparoscopic technique, the size of these defects may become exaggerated. Therefore by measuring the size of all visible defects during laparoscopy, small or large, and documenting it as a combined defect, large diameters hernias are reported during laparoscopic repair. Whereas an open repair in a non-distended abdomen only measures the largest defect which the surgeon can feel at the time of dissecting the tissue and possibly missing the adjacent smaller defects. Itani and Rogmark's studies [Itani et al., 2010, Rogmark et al., 2013] showed markedly large hernias were repaired using laparoscopic techniques compared to their open counterpart.

The operative time taken by laparoscopic as well as the open repair was comparable in our meta-analysis based on five out of six studies [Asencio et al., 2009, Eker et al., 2013, Itani et al., 2010, Navarra et al., 2007, Rogmark et al., 2013].

Bowel complications in a variety of forms were reported by all the six RCTs [Asencio et al., 2009, Eker et al., 2013, Itani et al., 2010, Navarra et al., 2007, Olmi et al., 2007, Rogmark et al., 2013]. Pooling of this data revealed a statistically significant increase in bowel complications in the laparoscopic group. The severity of bowel injury is determined by the type of intestine injured, i.e. small or large, the time delay between the occurrence, detection and treatment, and the amount of soiling that occurs [Bishoff et al., 1999, Henniford et al., 2003]. Unrecognized enterotomies or recognized bowel injuries lead to conversion to open repair [Ascenio et al., 2009, Itani et al., 2010]. Rogmark [Rogmark et al., 2013][26] also reported bowel injuries but this did not directly lead to conversion.

The overall complication rate was comparable in the two groups based on six RCTs [Asencio et al., 2009, Eker et al., 2013, Itani et al., 2010, Navarra et al., 2007, Olmi et al., 2007, Rogmark et al., 2013] as also highlighted by other authors [Bencini et al., 2004, Ben-Haim et al., 2002]. However, surgical site infections, hematomas, seromas and superficial wound infections etc. were noted more often in the open group than the laparoscopic group. Nonetheless when all these variables (i.e. wound infection, wound hematoma and seroma) were analyzed separately and the results were once again comparable for both groups. Olmi [Olmi et al., 2007][24] reported that subcutaneous drain placement was required by 97.6% of the open group patients, as was also highlighted in all the other trials [Asencio et al., 2009, Eker et al., 2013, Itani et al., 2010, Navarra et al., 2007, Rogmark et al., 2013]. However, very few drains were used in the laparoscopic group. A number of authors [Olmi et al., 2007, Itani et al., 2010, Rogemark et al., 2013]

have shown significantly higher wound infection rates for open repairs compared to laparoscopic repairs.

Reoperation rate was reported by four studies [Asencio et al., 2009, Navarra et al., 2007, Olmi et al., 2007, Rogmark et al., 2013] out of six studies under consideration. Analysis showed comparable outcomes for both groups.

The time taken to oral intake was statistically insignificant for both groups based on only two studies [Asencio et al., 2009, Navarra et al., 2007]. As the number of patients analyzed for this variable is so small, any meaningful conclusion is not possible.

Only two authors [Navarra et al., 2007, Olmi et al., 2007] have documented shorter length of hospital stay following laparoscopic repair compared to the open group. However, four out of six RCTs [Asencio et al., 2009, Eker et al., 2013, Itani et al., 2010, Rogmark et al., 2013] found comparable length of hospital stay for both these procedures.

Two RCTs [Olmi et al., 2007, Itani et al., 2010] reported that patients in the laparoscopic group took less time to recover and went back to work quicker. Rogmark [Rogmark et al., 2013] on the other hand reported time taken to full recovery, instead of time taken to return to work. In our meta-analysis, only two RCTs [Itani et al., 2010, Olmi et al., 2007] reported back to work data which failed to show any difference between the two groups.

All six RCTs namely [Asencio et al., 2009, Eker et al., 2013, Itani et al., 2010, Navarra et al., 2007, [Olmi et al., 2007, Rogmark et al., 2013] reported the recurrence rate. Pooling of the data revealed no difference between the two groups. Still, the data available on the recurrence rate may be erroneous due to short follow-up in all of these RCTs. Furthermore as the number of patients recruited in all the RCTs are very small, the true recurrence rate may be underestimated.

Our analysis based on two studies [Olmi et al., 2007, Rogmark et al., 2013] showed no significant difference in the post-operative neuralgia between laparoscopic and open repair groups. This finding was not in line with other laparoscopic procedures like appendectomy or cholecystectomy where less pain is observed following laparoscopic techniques. Once again, a small number of patients analyzed for this variable may be responsible for obscuring the true difference between the two procedures.

CONCLUSIONS

On the basis of our meta-analysis, we conclude that laparoscopic and open repair of incisional hernia is comparable. We strongly feel that objective assessment is required to evaluate the long term effectiveness of the two procedures. Recurrence rates should be measured for a lengthier period of time (e.g. 5 and 10 years) and not just for two years. Also, larger RCTs recruiting greater numbers of patients with strict inclusion and exclusion criteria and standardized techniques are crucial for meaningful comparison, effectiveness of the procedures and accuracy of results.

ACKNOWLEDGEMENTS

We gratefully acknowledge and thank Dr Matthew John Burstow for presenting the abstract of this paper at The Royal Australasian College of Surgeons Annual Scientific Congress, **Sands Expo and Convention Centre, Marina Bay Sands, Singapore** in 2014. The citation for published abstract is "ANZ Journal of Surgery 2014; 84 (Suppl 1): 67."

AUTHORS' CONTRIBUTIONS

1. AA, BM and MAM were responsible for the concept and design of this meta-analysis. Furthermore they take full responsibility for the integrity of the work as a whole, from the inception to published article.
2. AA and MAM were responsible for the acquisition and interpretation of the data.
3. SK, FR, MBH and RMY were responsible for analyzing and interpretation of the data in depth from the statistical point of view.
4. All authors were involved in drafting the manuscript and revising it critically for important intellectual content and have given final approval of the version to be published. Furthermore all authors have participated sufficiently in the work to take public responsibility for its content.

REFERENCES

1. Agresti, A. (1996). An Introduction to Categorical Data Analysis. Wiley & Sons; New York.
2. Asencio, F., Aguiló, J., Peiró, S., Carbó, J., Ferri, R., Caro, F. and Ahmad, M. (2009). Open randomized clinical trial of laparoscopic versus open incisional hernia repair. *Surgical Endoscopy*, 23, 1441-8.
3. Bencini, L. and Sanchez, L.J. (2004). Learning curve for laparoscopic ventral hernia repair. *Am J Surg.*, 187, 378-82.
4. Ben-Haim, M., Kuriansky, J., Tal, R., Zmora, O., Mintz, Y., Rosin, D., Ayalon, A. and Shabtai, M. (2002). Pitfalls and complications with laparoscopic intraperitoneal expanded polytetrafluoroethylene patch repair of postoperative ventral hernia. *Surg Endosc.*, 16, 785-8.
5. Bishoff, J.T., Allaf, M.E., Kirkels, W., Moore, R.G., Kavoussi, L.R. and Schroder, F. (1999). Laparoscopic bowel injury: incidence and clinical presentation. *J Urol*., 161, 887-90.
6. Burger, J.W., Luijendijk, R.W., Hop, W.C., Halm, J.A., Verdaasdonk, E.G. and Jeekel, J. (2004). Long-term follow-up of a randomized controlled trial of suture versus mesh repair of incisional hernia. *Ann Surg.*, 240, 578-585.
7. Cochran, W.G. (1954). The combination of estimates from different experiments. *Biometric*, 10, 101-29.
8. Egger, M., Smith, G.D., Schneider, M. and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Br. Med. J.*, 315, 629-34.
9. Eker, H.H., Hansson, B.M., Buunen, M., Janssen, I.M., Pierik, R.E., Hop, W.C., Bonjer, H.J., Jeekel, J. and Lange, J.F. (2013). Laparoscopic vs. open incisional hernia repair: a randomized clinical trial. *JAMA Surg.*, 13, 259-63.

10. Haynes, R.B., Sackett, D.L., Guyatt, G.H. and Tugwell, P. (2006). *Clinical epidemiology: how to do clinical practice research.* 3rd edition. Philadelphia, PA: Lippincott Williams & Wilkins.
11. Hedges, L.V. and Olkin, I. (1985). *Statistical Methods for Meta Analysis*: Academic Press; Orlando, Florida.
12. Henniford, B.T., Park, A., Ramshaw, B.J. and Voeller, G. (2003) Laparoscopic repair of ventral hernias: nine years' experience with 850 consecutive cases. *Ann. Surg.*, 238, 391-400.
13. Higgins, J.P.T. and Thompson, S.G. (2002). Quantifying heterogeneity in a meta-analysis. *Stat Med.*, 21, 1539-1558.
14. Hozo, S.P., Djulbegovic, B. and Hozo, I. (2005). Estimating the mean and variance from the median, range and size of a sample. *BMC Med Res Methodol*. 5,13.
15. Huedo-Medina, T.B., Sanchez-Meca, J., Marin-Martinez, F. and Botella, J. (2006). Assessing heterogeneity in meta analysis: Q Statistic or I^2 Index? *Am Psychol Assoc.*, 11, 193-206.
16. Itani, K.M., Hur, K., Kim, L.T., Anthony, T., Berger, D.H., Reda, D. and Neumayer, L. (2010). Comparison of laparoscopic and open repair with mesh for the treatment of ventral incisional hernia: A randomized trial. *Archives of Surgery*, 145, 322-8.
17. Jadad, A.R., Moore, R.A. and Carroll, D. (1996). Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials*, 17, 1-12.
18. LeBlanc, K.A. and Booth, W.V. (1993). Laparoscopic repair of incisional abdominal hernias using expanded polytetrafluoroethylene: preliminary findings. *Laparosc Endosc.*, 3, 39-41.
19. LeBlanc, K.A., Whitaker, J.M., Bellnager, D.E. and Rhynes. V.K. (2003). Laparoscopic incisional and ventral hernioplasty: lessons learned from 200 patients. *Hernia*, 7, 118-24.
20. Moher, D., Cook, D.J. and Eastwood, S. (1999). Improving the quality of reports of meta-analyses of randomized controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. *Lancet*, 354, 1896-1900.
21. Moher, D., Liberati, A. and Tetzlaff, J. (2009). The PRISMA Group. Preferred reporting items for systematic reviews and meta-analysis: the PRISMA statement. *PLoS Med*., 6:e1000097.
22. Moreno-Egea, A., Torralba, J.A., Girela, E., Corral, M., Bento, M., Cartagena, J., Vicente, J.P., Aguayo, J.L. and Canteras, M. (2004). Immediate, early, and late morbidity with laparoscopic ventral hernia repair and tolerance to composite mesh. *Surg Laparosc Endosc Percutan Tech*., 14, 130-5.
23. Navarra, G., Musolino, C., De Marco, M.L., Bartolotta, M., Barbera, A. and Centorrino, T. (2007). Retromuscular sutured incisional hernia repair: a randomized controlled trial to compare open and laparoscopic approach. *Surg Laparosc Endosc Percutan Tech*., 17, 86-90.
24. Olmi, S., Scaini, A., Cesana, G.C., Erba, L. and Croce, E. (2007). Laparoscopic versus open incisional hernia repair: an open randomized controlled study. *Surg Endosc.*, 21, 555-9.
25. R: A Language and Environment for Statistical Computing [Computer Program]. Version 1. Vienna: R Foundation for Statistical Computing; 2008.

26. Rogmark, P., Petersson, U., Bringman, S., Eklund, A., Ezra, E., Sevonius, D., Smedberg, S., Osterberg, J. and Montgomery, A. (2013). Short-term outcomes for open and laparoscopic midline incisional hernia repair: a randomized multicenter controlled trial: the ProLOVE (prospective randomized trial on open versus laparoscopic operation of ventral eventrations) trial. *Ann Surg.*, 258, 37-45.
27. Rosen, M., Brody, F., Ponsky, J., Walsh, R.M., Rosenblatt, S., Duperier, F., Fanning, A. and Siperstein, A. (2003). Recurrence after laparoscopic ventral hernia repair. *Surg Endosc*., 17, 123-8.
28. Span, J., Carière, E., Croockewitt, S., Smits, P. (2006). Publication bias, effects on the assessment of rosiglitasone. *Br J Clin Pharmacol*., 62, 732.
29. Sutton, A.J., Abrams, K.R., Jones, D.R., Sheldon, T.A. and Song, F. (2000). *Methods for Meta-analysis in Medical Research*. London: John Wiley.
30. Tang, J.L. and Liu, J.L.Y. (2000). Misleading funnel plot detection of bias in meta-analysis. *J Clin Epidermiol*., 53, 477-484.
31. Tsereteli, Z., Pryor, B.A., Heniford, B.T., Park, A., Voeller, G. and Ramshaw, B.J. (2008). Laparoscopic ventral hernia repair (LVHR) in morbidly obese patients. *Hernia*, 12, 233-8.
32. Ujiki, M.B., Weinberger, J., Varghese, T.K., Murayama, K.M. and Joehl, R.J. (2004). One hundred consecutive laparoscopic ventral hernia repairs. *Am J Surg*., 188, 593-597.
33. Viechtbauer, W. (20010). Conducting Meta-Analyses in R with the metaphor Package, *Journal of Statistical Software*, http://www.metafor-project.org/doku.php/metafor

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 263-276

TESTING THE EQUALITY OF THE TWO INTERCEPTS FOR THE PARALLEL REGRESSION MODEL

Budi Pratikno[1] and **Shahjahan Khan**[2]

[1] Department of Mathematics and Natural Science
Jenderal Soedirman University, Purwokerto, Jawa Tengah,
Indonesia. Email: bpratikto@gmail.com

[2] School of Agricultural, Computational and Environmental Sciences,
International Centre for Applied Climate Sciences,
University of Southern Queensland, Toowoomba, Australia.
Email: khans@usq.edu.au

ABSTRACT

Testing the equality of the two intercepts of two parallel regression models is considered when the slopes are suspected to be equal. For three different scenarios on the values of the slope parameters, namely (i) unknown (unspecified), (ii) known (specified), and (iii) suspected, we derive the unrestricted (UT), restricted (RT) and pretest (PTT)tests for testing the intercept parameters. The test statistics, their sampling distributions, and power functions of the tests are obtained. Comparison of power functions and sizes of the tests are provided.

KEYWORDS AND PHRASES

Linear regression; intercept and slope parameters; pre-test test; non-sample prior information; and power function.

***2010 Mathematics Subject Classification*:** Primary 62F03 and Secondary 62J05.

1 INTRODUCTION

Two linear regression lines are parallel if the two slopes are equal. A parallelism problem can be described as a special case of two related regression lines on the same dependent and independent variables that come from two different categories of the respondents. If the independent data sets come from two random samples ($p = 2$), researchers often wish to model the regression lines for lines groups that are parallel (i.e. the slopes of the two regression lines are equal) or whether the lines have the same intercept. To test the parallelism of the two regression equations, namely $y_{1j} = \theta_1 + \beta_1 x_{1j} + e_{1j}$ and $y_{2j} = \theta_2 + \beta_2 x_{2j} + e_{2j}$, $j = 1, 2, ..., n_i$, for the two data sets: $\boldsymbol{y} = [\boldsymbol{y}_1', \boldsymbol{y}_2']'$ and $\boldsymbol{x} = [\boldsymbol{x}_1', \boldsymbol{x}_2']'$ where $\boldsymbol{y}_1 = [y_{11}, ..., y_{1n_1}]'$ $\boldsymbol{y}_2 = [y_{21}, ..., y_{2n_2}]'$, $\boldsymbol{x}_1 = [x_{11}, ..., x_{1n_1}]'$, $\boldsymbol{x}_2 = [x_{21}, ..., x_{2n_2}]'$. We use an appropriate two-sample *t* test for testing $H_0 : \beta_1 = \beta_2$ (parallelism). This *t* statistic is given as

$$t = (\tilde{\beta}_1 - \tilde{\beta}_2) / S_{(\tilde{\beta}_1 - \tilde{\beta}_1)},$$

where $\tilde{\beta}_1$ and $\tilde{\beta}_2$ are estimate of the slopes β_1 and β_2 respectively, and $S_{(\tilde{\beta}_1 - \tilde{\beta}_1)}$ is estimate of the standard error of the estimated difference between slopes (Kleinbaum, 2008, p. 223). The parallelism of the two regression equations above can be expressed as a single model of matrix form, that is,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\Phi} + \boldsymbol{e},$$

where $\boldsymbol{\Phi} = [\theta_1, \theta_2, \beta_1, \beta_2]'$, $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]'$ with $\boldsymbol{X}_1 = [1, 0, x_1, 0]'$ and $\boldsymbol{X}_2 = [0, 1, 0, x_2]'$ and $\boldsymbol{e} = [e_1, e_2]'$. The matrix form of the intercept and slope parameters can be written, respectively, as $\boldsymbol{\theta} = [\theta_1, \theta_2]'$ and $\boldsymbol{\beta} = [\beta_1, \beta_2]'$ (cf Khan, 2006). In this model, p independent bivariate samples are considered such that $y_{ij} \approx N(\theta_i + \beta_i x i_j, \sigma^2)$ for $i = 1, ..., p$ and $j = 1, ..., n_i$. The parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)'$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ are the intercept and slope vectors of the p lines. See Khan (2003, 2006, 2008) for details on parallel regression models and analyses.

To explain the importance of testing the equality of the intercepts (parallelism) when the equality of slopes is uncertain, we consider the general form of the PRM of a set of $p (p > 1)$ simple regression models as

$$\boldsymbol{Y}_i = \theta_i \mathbf{1}_{ni} + \beta_i \boldsymbol{x}_{ij} + \boldsymbol{e}_{ij}, \; i=1,2,...,\text{p, and } j = 1, 2, ..., n_i, \qquad (1.1)$$

where $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{in_i})'$ is a vector of n_i observable random variables, $\mathbf{1}_{n_i} = (1, ..., 1)$ is an n_i-tuple of $1'$s, $\boldsymbol{x}_{ij} = (x_{i1}, ..., x_{in_i})'$ is a vector of n_i independent variables, θ_i and β_i are unknown intercept and slope, respectively, and $\boldsymbol{e}_i = (e_{i1}, ..., e_{in_i})'$ is the vector of errors which are mutually independent and identically distributed as normal variable, that is, $\boldsymbol{e}_i \approx N(\mathbf{0}, \sigma^2 \boldsymbol{I}_{n_i})$ where $\boldsymbol{I}_{n_i}$ is the identity matrix of order n_i. Equation (1.1) represent p linear models with different intercept and slope parameters. If $\beta_1 = ... = \beta_p = \beta$, then there are p parallel simple linear models if θ_i's are different. Here, the parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)'$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ are the intercept and slope vectors of the p lines.

Bancroft (1944) introduced the idea of pretesting NSPI to remove uncertainty. The outcome of the pretesting on the uncertain NSPI is used in the hypothesis testing to improve the performance of the statistical test (Khan and Saleh, 2001; Saleh, 2006, p.55-58; Yunus and Khan, 2011a). The suspected value of the slopes may be (i) unknown or unspecified if NSPI is not available, (ii) known or specified if the exact value is available from NSPI, and (iii) uncertain if the suspected value is unsure. For the three different scenarios, three different of statistical tests, namely the (i) unrestricted test (UT), (ii) restricted test (RT) and (iii) pre-test test (PTT) are defined. In the area of estimation

with NSPI there has been a lot of work, notably Bancroft (1944, 1964), Hand and Bancroft (1968), and Judge and Bock (1978) introduced a preliminary test estimation of parameters to estimate the parameters of a model with uncertain prior information. Khan (2003, 2008), Khan and Saleh (1997, 2001, 2005, 2008), Khan et al.(2002), Khan and Hoque (2003), Saleh (2006) and Yunus (2010) covered various work in the area of improved estimation using NSPI, but there is a very limited number of studies on the testing of parameters in the presence of uncertain NSPI. Although Tamura (1965), Saleh and Sen (1978, 1982), Yunus and Khan (2007, 2011a, 2011b), and Yunus (2010) used the NSPIfor testing hypotheses using nonparametric methods, the problem has not been addressed in the parametric context.

The study tests the equality of the intercepts for $p \geq 2$ when the equality of slopes is suspected. We test the intercept vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)'$ when it is uncertain if the *p* slope parameters are equal (parallel). We then consider the three different scenarios of the slope parameters, and define three different tests:

for the UT, let ϕ^{UT} be the test function and T^{UT} be the test statistic for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against $H_a : \boldsymbol{\theta} > \boldsymbol{\theta}_0$ when $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ is unspecified,

for the RT, let ϕ^{RT} be the test function and T^{RT} be the test statistic for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against $H_a : \boldsymbol{\theta} > \boldsymbol{\theta}_0$ when $\boldsymbol{\beta} = \beta_0 \mathbf{1}_p$ (fixed vector),

for the PTT, let ϕ^{PTT} be the test function and T^{PTT} be the test statistic for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against $H_a : \boldsymbol{\theta} > \boldsymbol{\theta}_0$ following a pre-test (PT) on the slope parameters. For the PT, let ϕ^{PT} be the test function for testing $H_{0}^{*} : \boldsymbol{\beta} = \beta_0 \mathbf{1}_p$ (a suspected constant) against $H_{a}^{*} : \boldsymbol{\beta} > \beta_0 \mathbf{1}_p$ (to remove uncertainty). If the H_{0}^{*} is rejected in the PT, then the UT is used to test the intercept, otherwise the RT is used to test H_0. Thus, the PTT depends on the PT which is a choice between the UT and RT.

The unrestricted maximum likelihood estimator or least square estimator of intercept and slope vectors, $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)'$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$, are given as

$$\tilde{\boldsymbol{\theta}} = \overline{\boldsymbol{Y}} - \boldsymbol{T}\tilde{\boldsymbol{\beta}}^{UT} \text{ and } \tilde{\boldsymbol{\beta}} = \frac{(\boldsymbol{x}_i'\boldsymbol{y}_i) - \left(\frac{1}{n_i}\right)\left[\mathbf{1}_i'\boldsymbol{x}_i\mathbf{1}_i'\boldsymbol{y}_i\right]}{n_i Q_i}, \tag{1.2}$$

where $\theta = (\theta_1, ..., \theta_p)'$, $\tilde{\beta} = (\tilde{\beta}_1, ..., \tilde{\beta}_p)'$, $\boldsymbol{T} = \text{Diag}(\overline{x}_1, ..., \overline{x}_p)$, $n_i Q_i = \boldsymbol{x}_i'\boldsymbol{x}_i - \left(\frac{1}{n_i}\right)\left[\mathbf{1}_i'\boldsymbol{x}_i\right]$, and $\tilde{\theta}_i = \overline{Y_i} - \tilde{\beta}_i \overline{x}_i$ for $i = 1, .., p$.

Furthermore, the likelihood ratio (LR) test statistics for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against $H_a : \boldsymbol{\theta} > \boldsymbol{\theta}_0$ is given by

$$F = \frac{\tilde{\theta}' \boldsymbol{H}' \boldsymbol{D}_{22}^{-1} \boldsymbol{H} \tilde{\theta}}{(p-1)s_e^2}, \tag{1.3}$$

where $\boldsymbol{H} = \boldsymbol{I}_p - \frac{1}{nQ}\mathbf{1}_p \mathbf{1}_p'$, $\boldsymbol{D}_{22}^{-1} = \text{Diag}(n_1 Q_1, .., n_p Q_p)$, $nQ = \sum_{i=1}^{p} n_i Q_i$, $n_i Q_i = \boldsymbol{x}_i' \boldsymbol{x}_i - \frac{1}{n_i}(\mathbf{1}_i' \boldsymbol{x}_i)^2$ and $S_e^2 = (n-2p)^{-1} \sum_{i=1}^{p} (\boldsymbol{Y} - \tilde{\theta}_i \mathbf{1}_{n_i} - \tilde{\beta} \boldsymbol{x}_i)'(\boldsymbol{Y} - \tilde{\theta}_i \mathbf{1}_{n_i} - \tilde{\beta} \boldsymbol{x}_i)$ (Saleh, 2006, p. 14-15).

Under H_0, F follows a central F distribution with $(p-1, n-2p)$ degrees of freedom (d.f.), and under H_a it follows a noncentral F distribution with degrees of freedom and noncentrality parameter $\Delta^2 / 2$, where

$$\Delta^2 = \frac{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \boldsymbol{D}_{22} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)}{\sigma^2} \tag{1.4}$$

and $\boldsymbol{D}_{22} = \boldsymbol{H}' \boldsymbol{D}_{22}^{-1} \boldsymbol{H}$. When the slope $(\boldsymbol{\beta})$ is equal to $\beta_0 \mathbf{1}_p$ (specified), the restricted mle of intercept and slope vectors are given as

$$\hat{\theta}_i = \tilde{\theta}_i + \boldsymbol{T}\boldsymbol{H}\tilde{\beta}_i \text{ and } \hat{\beta}_i = \frac{\mathbf{1}_k \mathbf{1}_k' \boldsymbol{D}_{22}^{-1} \tilde{\beta}_i}{n\boldsymbol{Q}} \tag{1.5}$$

The following section provides the proposed tests. Section 3 derives the distribution of the test statistics. The power function of the tests are obtained in Section 4. An illustrative example is given in Section 5. The comparison of the power of the tests and concluding remarks are provided in Sections 6 and 7.

2. THE THREE TESTS

To test the equality of the intercepts when the equality of slopes is suspected, we consider three different scenarios of the slopes. The test statistics of the UT, RT and PTT are then defined as follows.

For $\boldsymbol{\beta}$ unspecified, the test statistic of the UT is given by

$$T^{UT} = \frac{\tilde{\theta}' \boldsymbol{H}' \boldsymbol{D}_{22}^{-1} \boldsymbol{H} \tilde{\theta}}{(p-1)s_e^2}, \tag{2.1}$$

where $s_e^2 = (n-2p)^{-1} \sum_{i=1}^{n} (\boldsymbol{Y} - \tilde{\theta}_i \mathbf{1}_{n_i} - \tilde{\beta} \boldsymbol{x}_i)'(\boldsymbol{Y} - \tilde{\theta}_i \mathbf{1}_{n_i} - \tilde{\beta} \boldsymbol{x}_i)$.

The T^{UT} follows a central F distribution with $(p-1, n-2p)$ degrees of freedom. Under H_a, it follows a noncentral F distribution with $(p-1, n-2p)$ degrees of freedom and noncentrality parameter $\Delta^2 / 2$. Under normal model we have

$$\begin{pmatrix} \tilde{\theta} - \boldsymbol{\theta} \\ \tilde{\beta} - \boldsymbol{\beta} \end{pmatrix} \approx N_{2p} \left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \sigma^2 \begin{pmatrix} \boldsymbol{D}_{11} & -\boldsymbol{T}\boldsymbol{D}_{22} \\ -\boldsymbol{T}\boldsymbol{D}_{22} & \boldsymbol{D}_{22} \end{pmatrix} \right], \tag{2.2}$$

where $\boldsymbol{D}_{11} = \boldsymbol{N}^{-1} + \boldsymbol{T}\boldsymbol{D}_{22}\boldsymbol{T}\boldsymbol{\beta}$ and $\boldsymbol{N} = \text{Diag}(n_1, ..., n_p)$.

When the slope is specified to be $\boldsymbol{\beta} = \beta_0 \mathbf{1}_p$ (fixed vector), the test statistic of the RT is given by

$$T^{RT} = \frac{(\hat{\theta}' \boldsymbol{H}' \boldsymbol{D}_{22}^{-1} \boldsymbol{H} \hat{\theta}) + (\tilde{\beta}' \boldsymbol{H}' \boldsymbol{D}_{22}^{-1} \boldsymbol{H} \tilde{\beta})}{(p-1)s_e^2}, \tag{2.3}$$

where

$$s_r^2 = (n-p)^{-1} \sum_{i=1}^{p} (\boldsymbol{Y} - \hat{\theta}_i \mathbf{1}_{n_i} - \hat{\beta} \boldsymbol{x}_i)'(\boldsymbol{Y} - \hat{\theta}_i \mathbf{1}_{n_i} - \hat{\beta} \boldsymbol{x}_i) \text{ and } \hat{\beta} = \beta_0 \mathbf{1}_p$$

The T^{RT} follows a central *F* distribution with $(p-1, n-2p)$ degrees of freedom. Under H_a, it follows a noncentral *F* distribution with $(p-1, n-2p)$ degrees of freedom and noncentrality parameter $\Delta^2 / 2$. Again, note that

$$\begin{pmatrix} \hat{\theta} - \boldsymbol{\theta} \\ \hat{\beta} - \boldsymbol{\beta} \end{pmatrix} \approx N_{2p} \left[\begin{pmatrix} \boldsymbol{TH}\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix}, \sigma^2 \begin{pmatrix} \boldsymbol{D}^*_{11} & \boldsymbol{D}^*_{12} \\ \boldsymbol{D}^*_{12} & \boldsymbol{D}_{22} \end{pmatrix} \right], \tag{2.4}$$

where $\boldsymbol{D}^*_{11} = \boldsymbol{N}^{-1} + \dfrac{\boldsymbol{T}\mathbf{1}_p \mathbf{1}_p' \boldsymbol{T}\boldsymbol{\beta}}{nQ}$ and $\boldsymbol{D}^*_{12} = \frac{1}{nQ} \mathbf{1}_p \mathbf{1}_p' \boldsymbol{T}$.

When the value of the slope is suspected to be $\boldsymbol{\beta} = \beta_0 \mathbf{1}_p$ but unsure, a pre-test on the slope is required before testing the intercept. For the preliminary test (PT) of $H^*_0 : \boldsymbol{\beta} = \beta_0 \mathbf{1}_p$ against $H^*_a : \boldsymbol{\beta} > \beta_0 \mathbf{1}_p$, the test statistic under the null hypothesis is defined as

$$T^{PT} = \frac{\tilde{\beta}' \boldsymbol{H}' \boldsymbol{D}_{22}^{-1} \boldsymbol{H} \tilde{\beta}}{(p-1)s_e^2}, \tag{2.5}$$

which follows a central *F* distribution with $(p-1, n-2p)$ degrees of freedom. Under H_a, it follows a noncentral *F* distribution with $(p-1, n-2p)$ degrees of freedom and noncentrality parameter $\Delta^2 / 2$. Again, note that

$$\begin{pmatrix} \tilde{\theta} - \beta_0 \mathbf{1}_p \\ \tilde{\beta} - \hat{\beta} \end{pmatrix} \approx N_{2p} \left[\begin{pmatrix} (\tilde{\beta}^* - \beta_0)\mathbf{1}_p \\ \boldsymbol{H}\boldsymbol{\beta} \end{pmatrix}, \sigma^2 \begin{pmatrix} \mathbf{1}_p \mathbf{1}_p' / nQ & \mathbf{0} \\ \mathbf{0} & \boldsymbol{H}\boldsymbol{D}_{22} \end{pmatrix} \right], \tag{2.6}$$

where $\tilde{\beta}^* \mathbf{1}_p = \dfrac{\mathbf{1}_p \mathbf{1}_p' \boldsymbol{D}_{22}^{-1} \boldsymbol{\beta}}{nQ}$ (Saleh, 2006, p. 273).

Let us choose a positive number $\alpha_j (0 < \alpha_j < 1, \text{ for } j = 1, 2, 3)$ and real value $F_{v1,v2,v3}$ (v_1 be numerator d.f. and v_2 be denominator d.f.) such that

$$P(T^{UT} > F_{p-1,n-2p,\alpha 1} \mid \boldsymbol{\theta} = \boldsymbol{\theta}_0) = \alpha_1, \tag{2.7}$$

$$P(T^{RT} > F_{p-1,n-2p,\alpha 2} \mid \boldsymbol{\theta} = \boldsymbol{\theta}_0) = \alpha_2, \tag{2.8}$$

$$P(T^{PT} > F_{p-1,n-2p,\alpha 3} \mid \boldsymbol{\beta} = \beta_0 \mathbf{1}_p) = \alpha_3. \tag{2.9}$$

Now the test function for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against $H_a : \boldsymbol{\theta} > \boldsymbol{\theta}_0$ is defined by

$$\Phi \begin{cases} 1, \text{ if } (T^{PT} \le F_c,\ T^{RT} > F_b) \text{ or } (T^{PT} > F_c,\ T^{UT} > F_a); \\ 0, \text{ otherwise,} \end{cases} \tag{2.10}$$

where $F_a = F_{\alpha 1, p-1, n-2p,}$ $F_b = F_{\alpha 2, p-1, n-2p}$ and $F_c = F_{\alpha 3, p-1, n-2p.}$

3. DISTRIBUTION OF TEST STATISTICS

To derive the power function of the UT, RT and PTT, the sampling distribution of the test statistics proposed in Section 2 are required. For the power function of the PTT the joint distribution of (T^{UT}, T^{PT}) and (T^{RT}, T^{PT}) is essential. Let $\{N_n\}$ be a sequence of alternative hypotheses defined as

$$N_n : (\boldsymbol{\theta} - \boldsymbol{\theta}_0, \boldsymbol{\beta} - \beta_0 \mathbf{1}_p) = \left(\frac{\boldsymbol{\lambda}_1}{\sqrt{n}}, \frac{\boldsymbol{\lambda}_2}{\sqrt{n}} \right) = \boldsymbol{\lambda}, \tag{3.1}$$

where $\boldsymbol{\lambda}$ is a vector of fixed real numbers and $\boldsymbol{\theta}$ is the true value of the intercept. Under N_n the value of $(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ is greater than zero and under H_0 the value of $(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ is equal zero.

Following Yunus and Khan (2011b) and equation (2.1), we define the test statistic of the UT when $\boldsymbol{\beta}$ is unspecified, under N_n, as

$$T_1^{UT} = T^{UT} - n \left\{ \frac{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \boldsymbol{H}' \boldsymbol{D}_{22}^{-1} \boldsymbol{H} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)}{(p-1) s_e^2} \right\}. \tag{3.2}$$

The T_1^{UT} follows a noncentral F distribution with noncentrality parameter which is a function of $(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ and $(p-1, n-2p)$ degrees of freedom, under N_n.

From equation (2.3) under N_n, $(\boldsymbol{\theta} - \boldsymbol{\theta}_0) > 0$ and $(\boldsymbol{\beta} - \beta_0 \mathbf{1}_p) > 0$, the test statistic of the RT becomes

$$T_2^{RT} = T^{RT} - n \left\{ \frac{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \boldsymbol{H}' \boldsymbol{D}_{22}^{-1} \boldsymbol{H} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) + (\boldsymbol{\beta} - \beta_0 \mathbf{1}_p)' \boldsymbol{H}' \boldsymbol{D}_{22}^{-1} \boldsymbol{H} (\boldsymbol{\beta} - \beta_0 \mathbf{1}_p)}{(p-1) s_r^2} \right\} \tag{3.3}$$

The T_2^{RT} also follows a noncentral *F* distribution with a noncentrality parameter which is a function of $(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$ and $(p-1, n-2p)$ degrees of freedom, under N_n. Similarly, from the equation (2.5) the test statistic of the PT is given by

$$T_3^{PT} = T^{PT} - n\left\{\frac{(\boldsymbol{\beta}-\beta_0\mathbf{1}_{\mathrm{p}})'\boldsymbol{H}'\boldsymbol{D}_{22}^{-1}\boldsymbol{H}(\boldsymbol{\beta}-\beta_0\mathbf{1}_{\mathrm{p}})}{(p-1)s_e^2}\right\} \tag{3.4}$$

Under H_a, the T_3^{PT} follows a noncentral *F* distribution with a noncentrality parameter which is a function of $(\boldsymbol{\beta}-\beta_0\mathbf{1}_{\mathrm{p}})$ and $(p-1, n-2p)$ degrees of freedom.

From equations (2.1), (2.3) and (2.5) the T^{UT} and T^{PT} are correlated, and the T^{RT} and T^{PT} are uncorrelated. The joint distribution of the T^{UT} and T^{PT}, that is,

$$\left(T^{UT}, T^{PT}\right)' \tag{3.5}$$

is a correlated bivariate *F* distribution with $(p-1, n-2p)$ degrees of freedom. The probability density function (pdf) and cumulative distribution function (cdf) of the correlated bivariate *F* distribution is found in Krishnaiah (1964), Amos and Bulgren (1972) and El-Bassiouny and Jones (2009). Later, Johnson et al. (1995, p. 325) described a relationship of the bivariate *F* distribution with the bivariate beta distribution. This is due to the pdf of the bivariate *F* distribution has a similar form with the pdf of *beta distribution of the second kind*.

Following El-Bassiouny and Jones (2009), the covariance and correlation between the T^{UT} and T^{PT} are then given as

$$\begin{aligned} Cov(T_1^{UT}, T_3^{PT}) &= \frac{2f_1f_2}{(f_1-2)(f_2-2)(f_2-4)} \\ &= \frac{2(n^2-4np+4p^2)}{(n-2p-2)^2(n-2p-4)}, \text{ and} \end{aligned} \tag{3.6}$$

$$\begin{aligned} \rho^2_{T_1^{UT}T_3^{PT}} &= \frac{d_1d_2(f_1-4)}{(f_1+d_1-2)(f_2+d_2-2)(f_2-4)} \\ &= \frac{(n^2-2np+p^2)(n-2p-4)}{(2n-3p-2)^2(n-2p-4)}. \end{aligned} \tag{3.7}$$

Note in the above expressions $d_1 = d_2 = p-1$ and $f_1 = f_2 = n-2p$ are the appropriate degrees of freedom for the T^{UT} and T^{PT} respectively.

4. THE POWER AND SIZE OF TESTS

The power function of the UT, RT and PTT are derived below. From equation (2.1) and (3.2), (2.3) and (3.3), and (2.5) and (3.4), the power function of the UT, RT and PTT are given, respectively, as:

The power of the UT

$$\begin{aligned}\pi^{UT}(\boldsymbol{\lambda}) &= P(T^{UT} > F_{\alpha_1,p-1,n-2p} \mid N_n)\\ &= 1-P(T_1^{UT} \le F_{\alpha_1,p-1,n-2p} - k_1\delta_1),\end{aligned} \tag{4.1}$$

where $\delta_1 = \boldsymbol{\lambda}_1' \boldsymbol{D}_{22} \boldsymbol{\lambda}_1$ and $k_1 = \frac{1}{(p-1)s_e^2}$.

The power of the RT

$$\begin{aligned}\pi^{RT}(\boldsymbol{\lambda}) &= P(T^{RT} > F_{\alpha_1,n-1,n-2p} \mid N_n)\\ &= 1-P\left(T_2^{RT} \le F_{\alpha_2,p-1,n-2p} - \frac{(\boldsymbol{\lambda}_1'\boldsymbol{H}'\boldsymbol{D}_{22}^{-1}\boldsymbol{H}\boldsymbol{\lambda}_1)+(\boldsymbol{\lambda}_2'\boldsymbol{H}'\boldsymbol{D}_{22}^{-1}\boldsymbol{H}\boldsymbol{\lambda}_2)}{(p-1)s_r^2}\right)\\ &= 1-P\left(T_1^{RT} \le F_{\alpha_1,p-1,n-2p} - k_2(\delta_1+\delta_2)\right),\end{aligned} \tag{4.2}$$

where $\delta_2 = \boldsymbol{\lambda}_2' \boldsymbol{D}_{22} \boldsymbol{\lambda}_2$ and $k_2 = \frac{1}{(p-1)s_r^2}$.

The power function of the PT is

$$\begin{aligned}\pi^{PT}(\boldsymbol{\lambda}) &= P(T^{PT} > F_{\alpha_3,p-1,n-2p} \mid K_n)\\ &= 1-P\left(T_3^{PT} \le F_{\alpha_3,p-1,n-2p} - \frac{\boldsymbol{\lambda}_2'\boldsymbol{H}'\boldsymbol{D}_{22}^{-1}\boldsymbol{H}\boldsymbol{\lambda}_2}{(p-1)s_e^2}\right)\\ &= 1-P(T_3^{PT} \le F_{\alpha_3,p-1,n-2p} - k_1\delta_2).\end{aligned} \tag{4.3}$$

The power of the PTT

$$\begin{aligned}\pi^{PTT}(\boldsymbol{\lambda}) &= P(T^{PT} < F_{\alpha_3,p-1,n-2p}, T^{RT} > F_{\alpha_2,p-1,n-2p})\\ &\quad + P\left(T^{PT} \ge F_{\alpha_3,p-1,n-2p}, T^{UT} > F_{\alpha_1,p-1,n-2p}\right)\\ &= (1-\pi^{PT})\pi^{RT} + d_{1r}(a,b),\end{aligned} \tag{4.4}$$

where $d_{1r}(a,b)$ is bivariate F probability integrals, and it is defined as

$$\begin{aligned}d_{1r}(a,b) &= \int_a^\infty \int_b^\infty f(F^{PT}, F^{UT})\,dF^{PT}\,dF^{UT}\\ &= 1-\int_0^a \int_0^b f(F^{PT}, F^{UT})\,dF^{PT}\,dF^{UT},\end{aligned} \tag{4.5}$$

in which

$$a = F_{\alpha_3, p-1, n-2p} - \frac{\boldsymbol{\lambda}_2' \boldsymbol{H}' \boldsymbol{D}_{22}^{-1} \boldsymbol{H} \boldsymbol{\lambda}_2}{(p-1)s_e^2} = F_{\alpha_3, p-1, n-2p} - k_1 \delta_2 ,$$

and $$b = F_{\alpha_1, p-1, n-2p} - \frac{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \boldsymbol{H}' \boldsymbol{D}_{22}^{-1} \boldsymbol{H} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)}{(p-1)s_e^2} = F_{\alpha_1, p-1, n-2p} - k_1 \delta_1 .$$

The $\int_0^a \int_0^b f(F^{PT}, F^{UT}) dF^{PT} dF^{UT}$ in equation (4.5) is the cdf of the correlated bivariate noncentral *F*(BNCF) distribution of the UT and PT.

From equation (4.4), it is clear that the cdf of the BNCF distribution involved in the expression of the power function of the PTT. Using equation (4.7), we use it in the calculation of the power function of the PTT. R codes are written, and the R package is used for computations of the power and size and graphical analysis.

Furthermore, the size of the UT, RT and PTT are given respectively as:

The size of the UT

$$\begin{aligned} \alpha^{UT} &= P(T^{UT} > F_{\alpha_1, p-1, n-2p} \mid H_0 : \boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &= 1 - P(T_1^{UT} \le F_{\alpha_1, p-1, n-2p}), \end{aligned} \tag{4.8}$$

The size of the RT

$$\begin{aligned} \alpha^{RT} &= P(T^{RT} > F_{\alpha_2, p-1, n-2p} \mid H_0 : \boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &= 1 - P(T_2^{RT} \le F_{\alpha_2, p-1, n-2p} - k_2 \delta_2), \end{aligned} \tag{4.9}$$

The size of the PT is given by

$$\begin{aligned} \alpha^{PT}(\boldsymbol{\lambda}) &= P(T^{PT} > F_{\alpha_3, p-1, n-2p} \mid H_0) \\ &= 1 - P(T_3^{PT} \le F_{\alpha_3, p-1, n-2p}). \end{aligned} \tag{4.10}$$

The size of the PTT

$$\begin{aligned} \alpha^{PTT} &= P(T^{PT} \le a|_{H_0}, T^{RT} > d|_{H_0}) + P(T^{PT} > a, T^{UT} > h|_{H_0}) \\ &= (1 - P(T^{PT} > F_{\alpha_3, p-1, n-2p}))P(T^{RT} > F_{\alpha_2, p-1, n-p}) + d_{1r}(a, h), \end{aligned}$$

(4.11)

where $h = F_{\alpha_1, p-1, n-2p}$.

5. POWER COMPARISON BY SIMULATION

To compare the tests graphically we conducted simulations using the R package. For $p = 3$, each of three independent variables $(x_{ij},\ i = 1, 2, 3,\ j = 1, ..., n_i)$ are generated from the uniform distribution between0 and 1. The errors $(e_i,\ i = 1, 2, 3)$ are generated from the

normal distribution with $\mu = 0$ and $\sigma^2 = 1$. In each case $n_i = n = 100$ random variates were generated. The dependent variable (y_{ij}) is determined by $y_{1j} = \theta_{01} + \beta_{11} x_{1j} + e_1$ for $\theta_{01} = 3$ and $\beta_{11} = 2$. Similarly, define $y_{2j} = \theta_{02} + \beta_{12} x_{2j} + e_2$ for $\theta_{02} = 3.6$ and $\beta_{12} = 2$; $y_{3j} = \theta_{03} + \beta_{13} x_{3j} + e_3$, for $\theta_{03} = 4$ and $\beta_{13} = 2$, respectively. For the computation of the power function of the tests (UT, RT and PTT) we set $\alpha_1 = \alpha_2 = \alpha_3 = \alpha = 0.05$. The graphs for the power functionof the three tests are produced using the formulas in equations (4.1), (4.2) and (4.4).The graphs for the size of the three tests are produced using the formulas in equations(4.8), (4.9) and (4.11). The graphs of the power and size of the tests are presented in the Figures 1 and 2.

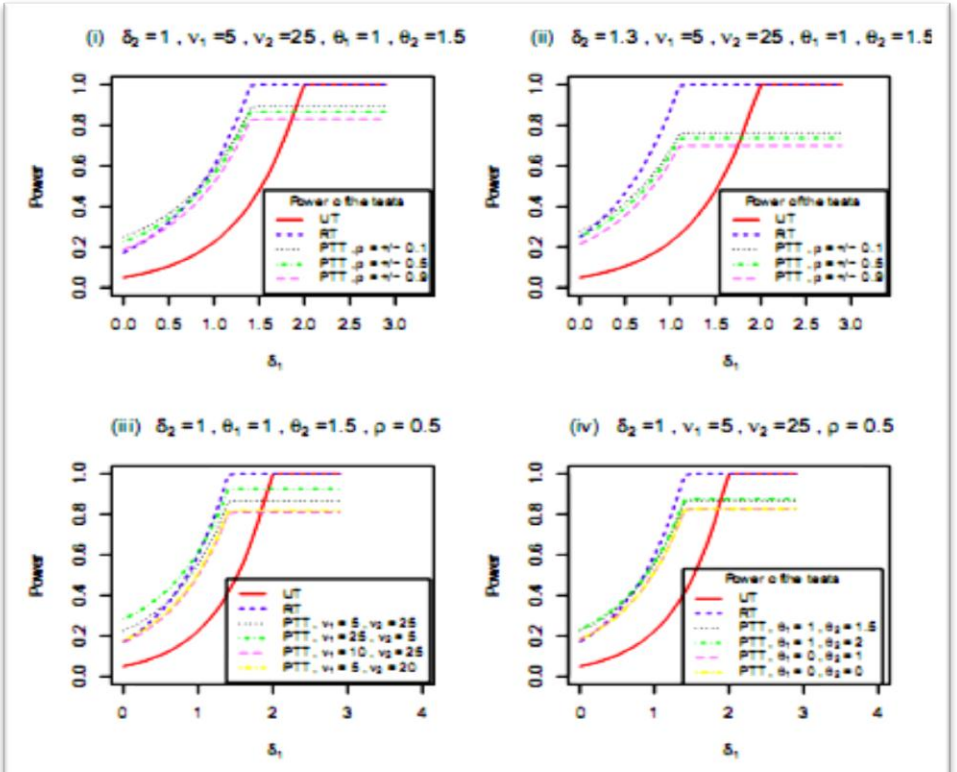


Figure 1: The power function of the UT, RT and PTT against δ_1 for some selected ρ, degrees of freedom and noncentrality parameters.

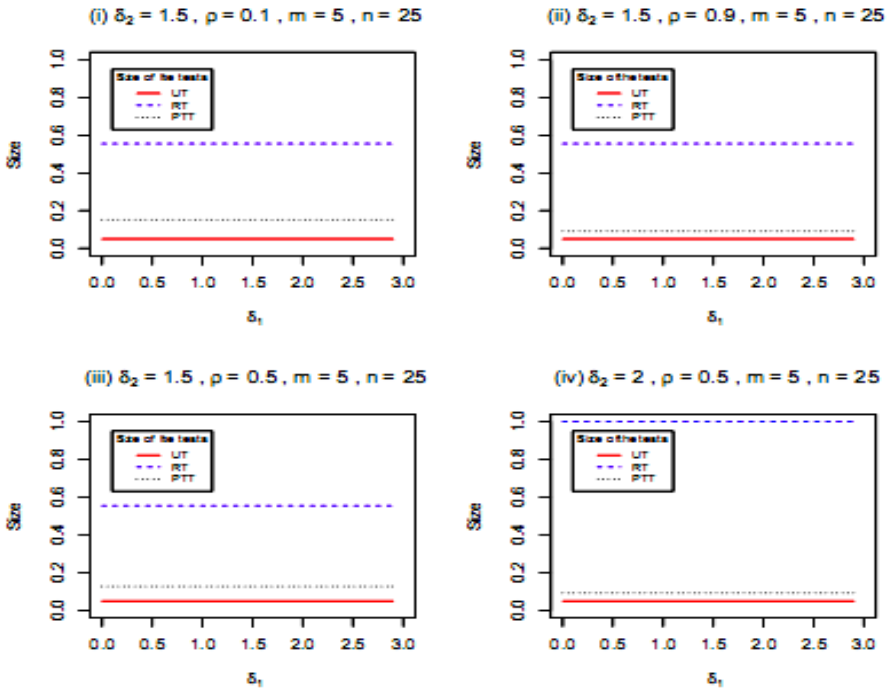


Figure 2: The size of the UT, RT and PTT against δ_1 for some selected ρ and δ_2.

6. COMPARISON AND CONCLUSION

The form of the power curve of the UT in Figure 1 is concave, starting from a very small value of near zero (when δ_1 is also near 0), it approaches 1 as δ_1 grows larger.The power of the UT increases rapidly as the value of δ_1 becomes larger. The shape of the power curve of the RT is also concave for all values of δ_1 and δ_2 . The power ofthe RT increases as the values of δ_1 and/or δ_2 increase (see Figures 1(i) and 1(ii), and equation (4.2)).

The power of the PTT (see Figure 1) increases as the values of δ_1 increase. Moreover, the power of the PTT is always larger than that of the UT and RT for the values of δ_1 around 0.7 to 1.5.

The size of the UT does not depend on δ_2 . It isa constant and remains unchanged for all values of δ_1 and δ_2 . The size of the RT increases as the value of δ_2 increases. Moreover, the size of the RT is always larger than that of the UT, but not for PTT for the smaller values of the δ_1 (not far from 0).

The size of the PTT is closer to that of the UT for larger values of $\delta_2 = 2$. The difference (or gap) between the size of the RT and PTT increases significantly as the value of δ_2 and ρ increases. The size of the UT is $\alpha^{UT} = 0.05$ for all values of δ_1 and δ_2. For all values of δ_1 and δ_2, the size of the RT is larger than that of the UT, $\alpha^{RT} > \alpha^{UT}$. For all the values of ρ, $\alpha^{PTT} \leq \alpha^{RT}$.

Based on the above analyses, the power of the RT is always higher than that of the UT for all values of δ_1 and δ_2. Also, the power of the PTT is always larger than that of the UT for all values δ_1 (see the curves for interval values of $0.7 < \delta_1 < 1.5$), δ_2 and ρ. The size of the UT is smaller than that of the RT and PTT for all δ_1. The power of the PTT is higher than that of the UT and tends to be lower than that of the RT. The size of the PTT is less than that of the RT but higher than that of the UT.

REFERENCES

1. Amos, D.E. and Bulgren, W.G. (1972). Computation of a multivariate *F* distribution. *Journal of Mathematics of Computation*, 26, 255-264.
2. Bancroft, T.A. (1944). On biases in estimation due to the use of the preliminary tests of significance. *Annals of Mathematical Statistics*, 15, 190-204.
3. Bancroft, T.A. (1964). Analysis and inference for incompletely specified models involving the use of the preliminary test(s) of significance. *Biometrics*, 20(3), 427-442.
4. El-Bassiouny, A.H. and Jones, M.C. (2009). A bivariate *F* distribution with marginals on arbitrary numerator and denominator degrees of freedom, and related bivariate beta and *t* distributions. *Statistical Methods and Applications*, 18(4), 465-481.
5. Han, C.P. and Bancroft, T.A. (1968). On pooling means when variance is unknown. *Journal of American Statistical Association*, 63, 1333-1342.
6. Johnson, N.L., Kotz, S. and Balakrishnan, N. (1995). Continuous univariate distributions, Vol. 2, 2nd Edition. John Wiley and Sons, Inc., New York.
7. Judge, G.G. and Bock, M.E. (1978). *The Statistical Implications of Pre-test and Stein-rule Estimators in Econoetrics*. North-Holland, New York.
8. Khan, S. (2003). Estimation of the Parameters of two Parallel Regression Lines under Uncertain Prior Information. *Biometrical Journal*, 44, 73-90.
9. Khan, S. (2005). Estimation of parameters of the multivariate regression model with uncertain prior information and Student-*t* errors. *Journal of Statistical Research*, 39(2), 79-94.
10. Khan, S. (2006). Shrinkage estimation of the slope parameters of two parallel regression lines under uncertain prior information. *Journal of Model Assisted and Applications*, 1, 195-207.
11. Khan, S. (2008). Shrinkage estimators of intercept parameters of two simple regression models with suspected equal slopes. *Communications in Statistics - Theory and Methods*, 37, 247-260.

12. Khan, S. and Saleh, A.K. Md. E. (1997). Shrinkage pre-test estimator of the intercept parameter for a regression model with multivariate Student-*t* errors. *Biometrical Journal*, 39, 1-17.
13. Khan, S. and Saleh, A.K. Md. E. (2001). On the comparison of the pre-test and shrinkage estimators for the univariate normal mean. *Statistical Papers*, 42(4), 451-473.
14. Khan, S., Hoque, Z. and Saleh, A.K. Md. E. (2002). Improved estimation of the slope parameter for linear regression model with normal errors and uncertain prior information. *Journal of Statistical Research*, 31(1), 51-72.
15. Khan, S. and Hoque, Z. (2003). Preliminary test estimators for the multivariate normal mean based on the modified W, LR and LM tests. *Journal of Statistical Research*, 37, 43-55.
16. Khan, S. and Saleh, A.K. Md. E. (2005). Estimation of intercept parameter for linear regression with uncertain non-sample prior information. *Statistical Papers*, 46, 379-394.
17. Khan, S. and Saleh, A.K. Md. E. (2008). Estimation of slope for linear regression model with uncertain prior information and Student-*t* error. *Communications in Statistics-Theory and Methods*, 37(16), 2564-2581.
18. Kleinbaum, D.G., Kupper, L.L., Nizam, A. and Muller, K.E. (2008). Applied regression analysis and other multivariable methods. Duxbury, USA.
19. Krishnaiah, P. R. (1964). On the simultaneous anova and manova tests. Part of Ph.D. thesis, University of Minnesota.
20. Saleh, A.K. Md. E. (2006). Theory of preliminary test and Stein-type estimation with applications. John Wiley and Sons, Inc., New Jersey.
21. Saleh, A.K. Md. E. and Sen, P.K. (1978). Nonparametric estimation of location parameter after a preliminary test on regression. *Annals of Statistics*, 6, 154-168.
22. Saleh, A.K. Md. E. and Sen, P.K. (1982). Shrinkage least squares estimation in a general multivariate linear model. *Proceedings of the Fifth Pannonian Symposium on Mathematical Statistics*, 307-325.
23. Schuurmann, F.J., Krishnaiah, P.R. and Chattopadhyay, A.K. (1975). Table for a multivariate *F* distribution. *The Indian Journal of Statistics*,37, 308-331.
24. Tamura, R. (1965). Nonparametric inferences with a preliminary test. *Bull. Math. Stat.*, 11, 38-61.
25. Yunus, R.M. (2010). Increasing power of M-test through pre-testing. Unpublished PhD Thesis, University of Southern Queensland, Australia.
26. Yunus, R.M. and Khan, S. (2007). Test for intercept after pre-testing on slope - a robust method. *In: 9th Islamic Countries Conference on Statistical Sciences (ICCS-IX)*: *Statistics in the Contemporary World - Theories, Methods and Applications*.
27. Yunus, R.M. and Khan, S. (2011a). Increasing power of the test through pre-test - a robust method. *Communications in Statistics-Theory and Methods*, 40, 581-597.
28. Yunus, R.M. and Khan, S. (2011b). M-tests for multivariate regression model. *Journal of Nonparamatric Statistics*, 23, 201-218.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 277-288

EXPLORATION INTEREST OF JAMBI COMMUNITY ABOUT BAITUL MAAL WATTAMWIL (BMT) BY USING REGRESSION LOGISTICS BINARY ANALYSIS

Titin Agustin Nengsih
Islamic Economical Departement of Syariah Faculty of IAIN STS Jambi
Email: titin_ipb@yahoo.com

ABSTRACT

BMT is expected to give real contribution in developing real economical sector, especially for industrial activity which hasn't fullfilled many requirements for gaining monetary fund from Syariah Bank. Actually, BMT has essensial part among society especially for UMK doer, since it is as a micro monetary syariah institution which is able to solve fundamental financial problem. Therefor, it needs a research in order to explore many factors which influence the public interest toward BMT.

Regression logistics method is used for establishing statistical variable which is associated within establish interest of society or not about BMT. These analysis gets that almost of whole factors influence interst of society about BMT. Those factors are type of work, level of education, level of earning, level of expenditure, and gaze about bank interest with usury. Meanwhile the ability factor to existence of BMT is the factor who not related significantly to BMT. The level of accuracy of the model is made, it is proven with clarification between the similarity of observation and prediction. Where is 73.68 % respondence interested in and 87.88% respondence isn't interested in to BMT has capabled is predicted correctly with the level of accuracy as big as 82.69 %. Therefore, overall prediction accuracy from this model in the amount of 82.69 with cutting the probability by 0,5.

The relationship between unimpeded variable and impeded variable can be seen from nagelkerke R Square value as big as 57.17%. Thus, research is only capable to explain the unimpeded variable case to impeded variable in amount of 57.17%. These cases are categorized sufficient to assosiated or related among unimpeded variable that is type of work, level of education, level of earning, level of expenditure, the knowladge about existence of BMT, and opinion about bank interest can be related by interested in BMT to interest of BMT.

KEYWORDS

BMT, Regresi Binnery Logistic

1. INTRODUCTION

Baitul Mal is a financial institution established by the Caliph (Islamic government) which is dealing with the regulation of financial activities, ranging from accepting, saving, until distributing money for public interest oriented in welfare and justice. The

concept of Baitul mal wat Tamwil (BMT) in Indonesia has been rolling over for decade, precisely in 1992. This concept is an implementation and reduplication of existed Baitul Mal in classical Islamic history, by developing innovation and creative creations toward conventional concept considering to the needs and the development of nowadays modern management of financial institutions.

The term of BMT is getting popular when in September 1994 *Dompet Du'afa* of *Republika* together with the association of Indonesian Syariah Banks held ZIS and economical Shariah management training in Bogor. Then other trainings by Dompet Du'afa held in Semarang and Yogyakarta. Then the term of BMT became more famous among society.

The presence of BMT in Indonesia is aimed to increase life's standard and welfare of people, this also has important role in improving small and medium enterprises in their working area, based on the vision of BMT that the development of economy should be built from scratch by considering business partnerships. As an economical institution based on humanity, BMT organized its active ties in accordance with the laws set by the government. The Laws are: law No. 7 of 1992 concerning in Banking, Law No. 10 of 1998 and No. 72 of 1992 concerning Bank based on the principle of sharing[1].

Economic agents in Jambi is still dominated by Micro and Small Enterprises (MSEs). MSE has contributed over 69.06% of job vacancy. However, MSE still face their classical problem, the low managerial capability in business management, low of competitive products, and law of eligibility in accessing capital from formal financial institutions such as the conventional banks.

Capital is an important issue. Considering that most of MSEs do not have access to formal financial institutions as result from the requirement of guarantee/collateral. Actually, the majority of MSEs has potential developing business, but for formal financial institutions, it's becoming risk because of high transaction costs, the availability of collateral. Finally, to solve the problem of capital, MSEs do tend to use the services of moneylenders. Moneylenders lend money to customers with some tight requirements, one of them is the high interest in short time. It doesn't give solution of the problem, but lead them into more complicated problem.

Therefore, it is necessary to have professional microfinance institution to fulfill the needs of MSEs in accessing capital for their business. It can also be the bridge between the formal financial institutions to MSEs. The role of BMT is very important for society, especially MSEs doer, because it is an Islamic microfinance institutions which is able to solve fundamental problems of capital[2].

The development of BMT in Jambi isn't so fast, from the period 1995 – 2012. There are only 14 BMT in Jambi city and until now there are only five active BMT. This indicates that BMT is not yet familiar among Jambi society. So that its role hasn't not yet optimally giving access to MSEs capital problem.

Based on the description above, it is necessary to have study to explore the development of BMT in Jambi city and how big the interest of public toward the existence of BMT as well as the factors that could affect a person's interest to participate in the BMT.

In accordance with the problem above, the authors formulate the following hypothesis:

H1: Type of Job affect a person's interest to BMT
H2: The level of education affects a person's interest to BMT
H3: The earn level affect a person's interest to BMT
H4: The consuming level affect a person's interest to BMT
H5: Knowledge of the existence of BMT affect a person's interest to BMT
H6: The assumption of usury system affect a person's interest to BMT

2. MATERIALS AND METHODS

Data Collection: questionnaires distributed among 104 respondents. 104 respondents are randomly chosen among eight sub district in Jambi city. They are sub districts of South Jambi, East Jambi, TelanaiPura, DanauTeluk, Pasar, Jelutung, Kota Baruand Pelayangan.

Research Variables

No	Variable	Code
1	Interests Associated with BMT	0 = Interested 1 = Not interested
2	Type of Job	1 = civil service jobs 2 = Private Employees 3 = Traders 4 = Labour 5 = Retired 6 = TNI/Police 0 = Self Employed
3	Level of Education	1 = SD 2 = junior high 3 = high school 0 = S1
4	Income Levels	1 = <500 2 = 501- 1.000 3 = 1.001- 1.500 4 = 1.501-2.000 5 = 2.001-2.500 0 = more than 2.501
5	The Expenditure Level	1 = 500 2 = 501- 1.000 3 = 1.001- 1,500 4 = 1.501-2.000 5 = 2.001-2.500 0 = more than 2.501
6	Knowledge of the Existence of BMT	1 = Know 0 = Do not Know
7	Opinions of Usury System	1 = Yes, the same 2 = doubtful 3 = Not at 0 = Do not know

Binary Logistic Regression:

Logistic regression modeling procedure applied for modeling the response variable (Y) which is categorized based on one or more predictor variables (X), both category and continue [3]. Unlike in the linear regression model, which can be obtained directly predicted values **y** because the shape model is **y** function of explanatory variables, in logistic regression, modeled values is the probability of occurrence of a specific category (generally probability Y=1) so that later on the obtained models is model of the relationship between p(Y=1) with some X explanatory variables. If the response variable consists of two categories, for example: Y = 1 (success) and Y = 0 (fail) then the logistic regression method which can be applied is a binary logistic regression.

If $\pi(x) = P_i$ states the probability of an individual has a value Y = 1, then the logistic regression model with k independent variables can be written as

$$\text{logit}(P_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots\ldots + \beta_k X_k$$

with

$$\text{logit}(\pi(x)) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right)$$

Odd Ratio:

If the model has a positive slope coefficient, the probability of occurrence increases linear with the increase of the value of the explanatory variables. On the contrary, If the coefficient is negative, the probability of occurrence will decrease for the higher value of the explanatory variables. One of criterion to find the relation between the value of the explanatory variable and the probability occurrence of a category on the response variable is the odds-ratio. This value will be one of output standards in the logistic regression produced many kind of software.

$$\text{odds ratio} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

The odds ratio indicates how big the probability certain categories may occur in the first than the second person. Although the precise definition is not that simple, as the discussion is the ratio of the odds of two individuals, not two individual risk ratio. Because the value of the odds ratio is obtained by dividing the two odds which is never negative, so that the odds ratio is always more than or equal to zero. The value odds ratio = 1 occurs only if both odds value are equal. Thus, if the odds ratio = 1, we say that the two groups of equal risk.

Binner logistic regression techniques are used to test the research problem, which factors determine significantly the public interest towards BMT in Jambi city. The variable in this study is the public interest towards BMT. Public interest becomes the dependent variable. While the explanatory variable are the type of job, level of education, level of income, expenditure levels, knowledge of the existence of BMT, and knowledge of the usury system.

3. COMMENTS AND CONCLUSION

Descriptive Analysis

The general overview of the respondents characteristics can be seen from the number of respondents on the terms of education, age, occupation, family income per month, and monthly family expenses.

By age, respondents can be seen in Figure 1. In Figure 1 indicates that respondents are most numerous in the age category 35 to 45 years 57 people. While respondents where at least in the category of less than 25 years are 6 people.

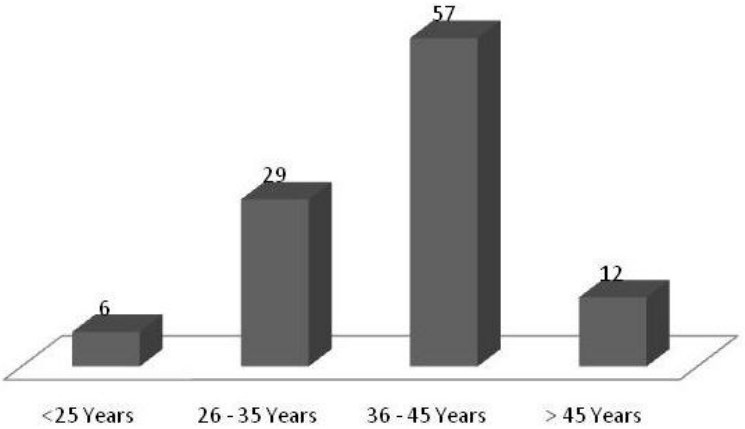


Fig. 1: Number of Respondents by Age

From the cross-tabulation Table 1, indicates the number of respondents by type of occupation and education of the respondents. Occupation is an activity of respondents which gradually done each day. From all of survey respondents (104 respondents), the occupation are varied. However, the type of job (occupation) which stands out in the study are 34 entrepreneurship, 24 traders and 16 laborer. Type of education is the latest education level of respondents. Most of the mare high school-graduated respondents (54 people), view of them are S.1 degree graduated respondents (10 people).

That is also indicates that half of high school-graduated respondents work as entrepreneur (28 people), as traders (9 people) and as laborers (8 people). While junior school-graduated respondents mostly work as a trader (11 people). Most of the elementary graduated respondents work as traders and laborers. While respondents with S.1 graduated work so varies. They are Civil Cervants, Private Employees, Retirees, Police/Army and the entrepreneur. From the table it can be seen that with low level education, the occupation they gained are more difficult as unskilled laborers. In the other hand, the higher of education level, the easier to get proper job.

Table 1
Number of Respondents by Education and Employment

Occupation	Elementary	Junior high	Senior high	S.1	Total
Civil Cervant	0	1	3	3	7
Private Employee	1	0	3	2	6
Trader	4	11	9	0	24
Laborer	4	4	8	0	16
retiree	1	0	0	1	2
Army/Police	2	7	3	3	15
Entrepreneur	2	3	28	1	34
Total	14	26	54	10	104

Respondent's income level is measured by the main and additional income for a month. Monthly income level of the respondents ranged from less than 500,000, - . there are 3 people (2.88%) and among Rp. 1.501.000, - s / d Rp. 2.000.000, - there are 31 persons (29.8%). The description of the amount of monthly income of the respondents is shown in Table 2. From the amount of income, we can conclude that most of the respondents has proper income level. It can be seen from only a small part respondents whose income less than Rp. 1.000.000 (14.4%).

Table 2
Number of Respondents by Income Level

Income Level	f	%
< Rp.500.000	3	2.88
Rp.501.000 – Rp.1.000.000	12	11.54
Rp.1.001.000 – Rp.1.500.000	23	22.12
Rp.1.501.000 – Rp.2.000.000	31	29.81
Rp.2.001.000 – Rp.2.500.000	19	18.27
> Rp.2.501.000	16	15.38
Total	104	100.00

Most of respondents monthly expenditure level is ranged Rp.1.001.000, - until Rp.1.500.000, (41 people (39.4%)). While respondents who spent more than Rp.2.500.100, only 6 people (5.7%). Table 3 shows that the most of the respondents who are the people of Jambi city has monthly expenses above Rp.1.000.000, -.

Table 3
Number of Respondents by Level of Expenditure

Level of Expenditure	f	%
< Rp.500.000	10	9.62
Rp.501.000 – Rp.1.000.000	21	20.19
Rp.1.001.000 – Rp.1.500.000	41	39.42
Rp.1.501.000 – Rp.2.000.000	17	16.35
Rp.2.001.000 – Rp.2.500.000	9	8.65
> Rp.2.501.000	6	5.77
Total	104	100.00

Factors Which Influences Public Interest toward BMT

In this study, variables thought to influence a person's interest of BMT are the type of work (X1), education level (X2), income level (X3), the level of expenditure (X4), knowledge of the existence of BMT (X5), opinions about the system of usury (X6).

Binary logistic regression is one of analysis to determine the relationship between the dependent variable (interest) with the independent variable (X). Logit model used in this study for guessing the public interest or not toward BMT. The dependent variable is not interested (y = 1) and interested (y = 0). Binary logistic regression model as follows:

$$\text{logit}(P_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_1 X_3 + \beta_2 X_4 + \beta_1 X_5 + \beta_2 X_6$$

where:

logit (Pi) = probability of the i-th individual interested in BMT
X1 = Type of Job
X2 = Level of Education
X3 = Income Levels
X4 = level of expenditure
X5 = Knowledge of the existence of BMT
X6 = Opinions on system of usury

Logistic regression method was used to determine associated or influenced variables in determining public interest they are interested or not interested in BMT. This can be determined by the Wald test which can be seen in Table 4. This analysis result is that almost all of the factors affects the public interest toward BMT. These factors are type of job, education level, income level, expenditure levels, and the opinion on the system usury.

These are each categories for each variables, the categories directly related to interest BMT are as follows:

1. Type of Job has positive affect to the interest of BMT. From the Wald can be found that type of job which affect public interest toward BMT are civil servants and army / police. This is because the Wald test p-value is smaller than the specified α of 10%. Thus, variable categories of civil servants and army / police are not interested in BMT.
2. Level of education has positive affect to interest in BMT. The level of education will significantly not interested in BMT. The overall level of education is

elementary, junior high school, high school, and S.1 significant effect on interest in BMT. From Wald test obtained from all levels of education found that the Wald test p-value is smaller than the specified α of 10%. It can be said that all levels of education have a significant effect on the interest of BMT.

3. The level of income is income Rp.501.000 - Rp.1.000.000, income Rp.1.001.000 - Rp.1.500.000, income Rp.2.001.000 - Rp.2.500.000 are positive and significant effect on the interest of BMT. This is because the Wald test p-value is smaller than the specified α of 10%. Variable income is the income level category Rp.501.000 - Rp.1.000.000, income Rp.1.001.000 - Rp.1.500.000, Rp.2.001.000 - Rp.2.500.000 are significantly not interested in BMT.
4. The level of expenditure; the expenditure Rp.501.000 - Rp.1.000.000, and expenditure more than Rp.2.501.000 significantly affect the interest of BMT. Expenditure Rp.501.000 - Rp.1.000.000 negatively affect the interests of BMT. While expenditure more than Rp.2.501.000 positively affect the interest of BMT. It can be said that expenditure Rp.501.000 - Rp.1.000.000 significantly interested in BMT. While expenditure more than Rp.2.501.000 is significantly not interested in BMT.
5. The public opinion of usury system of which the bank interest category is equal with usury, for whom do not know or assume that is not equal bank interest with usury, has negative affect on interest in BMT. This is because the Wald test p-value is smaller than the specified α of 10%. Category variables "not knowing the usury is or the assumption that the bank interest is not equal with usury can be associated that these categories are not interested in BMT.

Table 4
Wald Test Results

Variable	B	S.E.	Uji Wald	Sig.
Occupation			6.80	0.34
Occupation (1)	3.32	1.93	2.94	**0.09**
Occupation (2)	28.35	11957.79	0.00	1.00
Occupation (3)	0.42	1.18	0.13	0.72
Occupation (4)	0.70	1.18	0.35	0.55
Occupation (5)	21.90	27196.79	0.00	1.00
Occupation (6)	2.28	1.16	3.85	**0.05**
Education			11.72	**0.01**
Education (1)	6.85	2.29	8.92	**0.00**
Education (2)	7.15	2.30	9.68	**0.00**
Education (3)	5.01	2.03	6.09	**0.01**
Income Level			6.47	0.26
Income Level (1)	13.32	13692.07	0.00	1.00
Income Level (2)	7.31	2.96	6.07	**0.01**
Income Level (3)	5.29	2.45	4.68	**0.03**
Income Level (4)	4.13	2.17	3.63	**0.06**
Income Level (5)	3.41	2.14	2.54	0.11
Expenditure Level			12.88	**0.02**
Expenditure Level (1)	13.11	13692.07	0.00	1.00

Variable	B	S.E.	Uji Wald	Sig.
Expenditure Level(2)	-5.43	2.83	3.69	**0.05**
Expenditure Level(3)	-3.43	2.55	1.81	0.18
Expenditure Level(4)	0.50	2.44	0.04	0.84
Expenditure Level(5)	2.65	2.07	1.65	0.20
Knowledge Of BMT	0.34	0.87	0.15	0.70
Usury			7.25	**0.06**
Usury(1)	1.32	0.91	2.08	0.15
Usury(2)	-0.98	1.02	0.92	0.34
Usury(3)	-3.28	1.73	3.60	**0.06**
Constant	-7.63	2.92	6.80	0.01

Note: bold = significant at $\alpha = 10\%$

Variables established logistic regression model can be interpreted by using odds ratio value. Odds ratios value can be seen in Table 5. The detail results for each variables according to the odds ratio are as follow:

1. Variable types of jobs has odd ratio more than 1. Category civil servants has the odds ratio 27.54. This may imply that the civil servants are not likely interested in 27.54 or 28 times greater than the entrepreneur. While the category Military / Police have odds ratio 9.75. This may imply that the army / police are not likely interested in 9.75 or 10 times greater than the entrepreneur. This is consistent with the fact that civil servants and army / police already have a fixed salary so they'd tend to save and to make money loan from the bank rather than from BMT.
2. The overall level of education; junior high school, senior high school, and S.1 has odds ratio more than 1. Junior high school category has odds ratio of 947.69. This may imply that the junior high school is not likely interested, 947.69 or 947 times larger than Elementary school. While the senior high school category has the odds ratio of 1276.79. This may imply that the high school is not likely interested, 1276.79 or 1277 times larger than elementary school. While Category S.1 has odds ratio of 150.106. This may imply that S.1 is not likely interested 150.106 or 150 times larger than elementary school.
3. Level of income Rp.501.000 - Rp.1.000.000, income Rp.1.001.000 - Rp.1.500.000, and income Rp.2.001.000 - Rp.2.500.000 have odds ratio more than 1. The Rp.501.000 - Rp.1.000.000 ,- category has odds ratio of 1490.03. This may imply that Rp.501.000 - Rp.1.000.000 is not likely interested 1490.03 or 1490 times greater than those who have income more than Rp.2.501.000, -. While the income category Rp.1.001.000 - Rp.1.500.000 has odds ratio of 198.91. This may imply that the income Rp.1.001.000 - Rp.1.500.000 is not likely interested 198.91 atau 199 times greater than those who have income more than Rp.2.501.000, -. While the income category Rp.2.001.000 - Rp.2.500.000 has odds ratio of 62.19. This may imply that the income Rp.2.001.000,- –Rp.2.500.000,- is not likely interested 62.19 or 62 times greater than those who have income more than Rp.2.501.000, -.
4. The level of expenditure. Those who have expenditure level Rp.501.000 - Rp.1.000.000 have odds ratio close to zero. Expense categories of Rp.501.000 - Rp.1.000.000 has odds ratio of 0.004. This may imply in reverse that expense

more than Rp.2.501.000, - has the odd ratio of 1 / 0.004 = 229.11. It can be concluded that the expenditure level more than Rp.2.501.000, - is more likely interested than the expenditure level Rp.501.000 - Rp.1.000.000 by 229 times. This is in line with our daily circumstances that the more expense of society the more likely be interested in the BMT in terms of borrowing money.

5. The public opinion about the usury is equal with Bank interest, the category of "do not know" and " the opinion that usury is not equal with an " interest Bank" have odds ratio close to zero. The odds ratio is 0:04. This may imply in reverse that category "do not know" has odds ratios of 1 / 0:04 = 25. It can be concluded that the category of "not knowing" is more likely interested than "the opinion that usury is not equal with an interest" category by 25 times. People who do not know the greater will likely be interested in BMT from the same category who expressed no interest with usury. People who do not know would be easier to join BMT by giving some lecture or spiritual motivation that will explain that the interest is equal with the practice of usury. If they are given the correct understanding, they tend to choose the shari'ah economic institutions rather than conventional institutions, one of them is BMT.

Table 5
Logistic Regression Odds Ratio Value

Variable	Odd Ratio
Occupation(1)	**27.54**
Occupation(2)	2058200920033.94
Occupation(3)	1.52
Occupation(4)	2.01
Occupation(5)	3244865488.76
Occupation(6)	**9.75**
Education(1)	**947.70**
Education(2)	**1276.80**
Education(3)	**150.11**
Income Level(1)	0.00
Income Level(2)	**1490.04**
Income Level(3)	**198.92**
Income Level(4)	**62.19**
Income Level(5)	30.32
Expenditure Level(1)	492144.14
Expenditure Level(2)	**0.004**
Expenditure Level(3)	0.03
Expenditure Level(4)	1.65
Expenditure Level(5)	14.22
Knowledge of BMT	1.41
Usury(1)	3.73
Usury(2)	0.37
Usury(3)	**0.04**

Note: bold = significant at $\alpha = 10\%$

The determination of the best model from the model that has been formed by using cutting criteria with an error rate of classification errors are fairly balanced between type 1 and type II errors. Determination of probability value is selected in the value of 0.5. Precision matrix can be seen in Table 6. The accuracy of the model is proven by the clarification between the observations and the predictions which are equal, where 73.68% of respondents are interested in BMT and the 87.88% of respondents are not interested in BMT has been able to be predicted with total accuracy rate of 82.69%.

In the table below shows that of the 38 respondents interested in BMT is categorized correctly by 28 (73.68%). Meanwhile, of the 64 respondents not interested in BMT is properly categorized as 58 (87.88%). Overall accuracy of the predictions of this model is 82.69% with the cutting of the probability of 0.5.

Table 6
Matrix Model Accuracy

Actual Group	Prediction		
	Interested	Not Interested	Corect Presentage
Interested	28	10	73.68
Not Interested	8	58	87.88
Overall Percentage			82.69

From the results of binary logistic regression analysis we can conclude that the independent variable; type of job, level of education, level of income, expenditure levels, knowledge of the existence of BMT, and opinion on the interest rate can be associated with the interest in BMT significantly. The value of the association can be seen in Table 5. The association between the independent and dependent variables can be seen from the value of Nagelkerke R Square of 57.17%. Thus, this research is only able to explain the occurrence of independent variables toward the dependent variable in 57.17%. It is categorized there is an association between the independent variable is the type of job, level of education, level of income, expenditure levels, knowledge of the existence of BMT, and opinions on the flowers can relate to public interest in BMT. 42.83% of other variables that might also explain the dependent variable so that the model can be explained as a whole.

Table 7
The Strength of Association between the Dependent Variable with the Independent Variable

-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
80.2676	**0.4179**	**0.5717**

4. CONCLUSION

Logistic regression method is used to determine variables associated with or influenced in determining the interest or not interest in BMT. This analysis result is that almost all of the factors affects the public interest in BMT. The factors are type of job, education level, income level, expenditure level, and the opinion of the system of usury.

While knowledge of the existence of BMT factor is not related significantly to BMT. The accuracy of the model is proven by the clarification between the observations and the predictions which are equal, where 73.68% of respondents are interested in BMT and the 87.88% of respondents are not interested in BMT has been able to be predicted with total accuracy rate of 82.69%.

The association between the independent and dependent variables can be seen from the value of Nagelkerke R Square of 57.17%. Thus, this research is only able to explain the occurrence of independent variables toward the dependent variable in 57.17%. It is categorized there is an association between the independent variable is the type of job, level of education, level of income, expenditure level, knowledge of the existence of BMT, and opinions on the flowers can relate to public interest in BMT.

5. REFERENCES

1. Agresti, A. (2007). *An Introduction Categorical Data Analysis*, John Wiley and Sons. Inc., New York.
2. Dawam, R. and Islam, D. (1999). Transformasi Sosial Ekonomi. Yogyakarta: Pustaka Pelajar.
3. Heri Sudarsono (2003). Bank Dan Lumbago Kelantan Syria's: Deskripsidan Ilustrasi. Jogjakarta: Ekonisia.
4. Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logistic Regression*, second edition, John Wiley & Sons, USA.
5. Mannan, M.A. (1992). Ekonomi Islam: Teoridan Praktek, terj. Islamic Economis: Theory and Practice. Jakarta: PT. Intermasa.
6. Muhammad Ridwan (2004). Manajemen Baitul Mal waTamwil. Jogjakarta: UII Press.
7. Suhendi, H. (2004), BMTdan Bank Islam, Bandung.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 289-298

FORECASTING SPARE PARTS DEMAND: A CASE STUDY AT AN INDONESIAN HEAVY EQUIPMENT COMPANY

Ryan Pasca Aulia, Farit Mochamad Afendi and **Yenni Angraini**
Statistics Department, Bogor Agricultural University
Email: rpa160993@gmail.com

ABSTRACT

Forecasting spare parts demand is a common issue dealt by inventory managers at maintenance service organization. The large number of items held in stocks and the random demand occurrences make most of the difficulties. An Indonesian heavy equipment company targets to advance its forecast accuracy. Accordingly, this study has two main goals. Firstly, all Stock Keeping Units are classified based on their demand patterns, utilizing their average inter-demand interval and squared coefficient of variation of demand sizes as the classifiers. After that, four simple forecasting methods are applied to each demand class and the best forecasting method in term of its forecast errors is chosen. Evaluation of forecast accuracy is made by means of the Mean Absolute Scaled Error, MAD-to-Mean ratio, and Percentage Best. The forecasting competition results show the dominance of Syntetos-Boylan Approximation for erratic, smooth, and intermittent demand, and Simple Moving Average for lumpy demand.

KEYWORDS

Coefficient of variation, inter-demand interval, simple moving average, single exponential smoothing, Croston's method, Syntetos-Boylan Approximation, MASE, MAD-to-Mean, percentage best

1. INTRODUCTION

Spare parts in automotive industry posses a wide range of characteristics. They are highly varied in costs, service requirements, and demand patterns (Boylan and Syntetos, 2008). Many of them are slow-moving as they are only ordered in small quantities occasionally. Thousands of these items are stored in the warehouse and they are valuable investments for companies providing maintenance service.

Demands of spare parts exhibit infrequent and irregular patterns, most of which are intermittent characterized by high frequency of zero values in demand history. This condition is sometimes accompanied by large variations of demand sizes when they occur (erraticity), which creates lumpy demand. These make the forecast of spare parts demand more difficult, hence a challenging task. Even so, considerable improvements on forecast accuracy are possibly converted to reduction of inventory costs and raised customer service levels.

Forecasts of future demands are vital input to an inventory model. They determine how much stocks held and how much to order from vendors to meet customer demand.

Producing inaccurate forecasts can lead to unfulfilled demands or stock-outs. Therefore, careful managerial decision must be made in order to achieve satisfactory customer service at minimum inventory costs.

Many forecasting techniques have been proposed in literature. However, it is difficult to decide the most superior one and generalize it to a particular case. This is due to the type of data and how forecast errors are measured. Different accuracy metrics can pick different forecasting methods as the most accurate one for the same data. Moreover, optimal parameter value for a certain method may vary from one condition to another.

The heavy equipment company which provided the data exercised in this study aims to improve their forecast accuracy. The company applies both deterministic and stochastic model to generate forecasts. As for the statistical model, which becomes the focus of this study, they use Simple Moving Average (SMA) of the previous twelve monthly demands. The forecasts made by those models act as direct input to determine the maximum inventory level on a max-min system.

Trying to provide some alternatives, simple forecasting methods usually found in practice are Single Exponential Smoothing (SES), Croston's method, and Syntetos-Boylan Approximation (SBA). These methods are easy to apply in the software package used by the company and have been reported to produce reasonable results in previous studies. Their performances are going to be compared with SMA, which is treated as the benchmark method.

To examine forecast accuracy, the company utilizes the Mean Absolute Percentage Error (MAPE). Unfortunately, this measure is very problematic in situation where demands are highly intermittent as it causes degeneracy and asymmetry issues. For that reason, this study employs different accuracy measures: the Mean Absolute Scaled Error (MASE), MAD-to-Mean ratio, and Percentage Best (PBt). Each accuracy measure provides its own unique information.

A comparative study by Syntetos and Boylan in 2005 is used as the primary reference of this paper. They compared the performance of SMA, SES, Croston's method and their new estimator (SBA) on demand series from automotive industry. The results suggested the superiority of SBA. Moreover, Syntetos and Boylan made some comments about the behavior of the error measures adopted in their study.

Due to the large number of Stock Keeping Units (products kept in stock) or SKUs, and their wide range of characteristics, a classification scheme ought to be built. To serve this purpose, Syntetos *et al.* (2005) provided a demand categorization scheme to select the most appropriate forecasting procedure (Figure 1). They compared the theoretical Mean Squared Error (MSE) of SES, Croston's method, and SBA, and established regions of superior performance of each method. As the scheme borders, Syntetos *et al.* constructed the cut-off values of the average inter-demand interval (ADI) and squared coefficient of variation (CV^2) of non-zero demand to four discrete demand categories (erratic, lumpy, smooth, and intermittent).

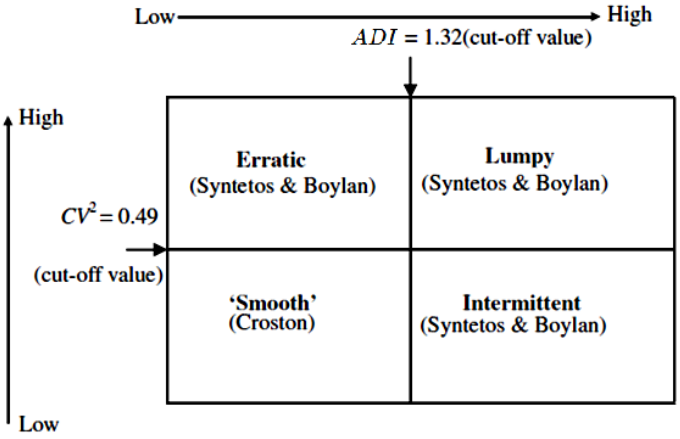


Fig. 1: Syntetos *et al.* Demand Categorization Scheme (Boylan *et al.* (2008))

Overall the objective of this study is to categorize every spare parts demand history into four demand patterns and the best forecasting method on each demand category is decided. Hopefully, this study could offer some recommendations about the most appropriate forecasting approach to those four demand categories.

2. METHODOLOGY

The data exercised in this study was queried from a branch office of a heavy equipment company in Indonesia. They were in forms of sales transaction record from January 2010 to June 2013. At first, they were pivoted for each parts number to obtain monthly order quantity (demand). All SKUs that were ordered at least once during 2012 were further considered. They consisted of bolt, cartridge, filter, piston, ring, valve, etc.

The demand series were divided into in-sample and out-of-sample set, where demands in 2013 were reserved for data validation process. To allow for fair comparisons with SMA which only made use of the latest twelve months data, the fitting process for all other methods started at January 2012. At the end of December 2012, one to six months ahead forecasts were made. Each forecasting method would generate flat forecasts for all horizons.

The procedures involved in this study are:

1. **Demand Categorization**
   Compute for all SKUs the ADI and CV^2 of the demands when they occur. Subsequently, each SKU is categorized based on the cut-off values suggested by Syntetos *et al.* (2005).

2. **Demand Forecasting**
   Apply the four forecasting methods described in Table 1 to every demand series. The smoothing constants exercised are identical to those in Syntetos and Boylan (2005). It should be noted that the same smoothing constant is applied to update

demand sizes and intervals for the last two methods. In regard to the initial value of SES, the average demand over the first 24 months is used. Similarly, the average of demand size and inter-demand interval over the first 24 months are taken to be the initial SES estimates of demand size and inter-demand interval.

Table 1
Forecasting Methods

Forecasting Methods	Smoothing Constants (α)
Simple Moving Average	12-months span
Single Exponential Smoothing	0.05, 0.10, 0.15, 0.20
Croston's Method	0.05, 0.10, 0.15, 0.20
Syntetos-Boylan Approximation	0.05, 0.10, 0.15, 0.20

3. Forecasts Evaluation

Calculate the forecast errors (MASE, MAD-to-Mean, and PBt) of all demand series and asses the performance of the four forecasting methods. The descriptive measures, MASE and MAD-to Mean, are aggregated across series and the method that generates the lowest median is considered as the best. To determine the PBt across series, the minimum MASE and MAD-to-Mean of every forecasting method on each demand series are first calculated and then ranked. When ties occur on a series, all methods with minimum MASE or MAD-to-Mean value are tallied. The method that produces the highest PBt is regarded as the best method.

Finally, the best forecasting method on each category is compared with SMA. Two-sided Wilcoxon signed rank tests are conducted to test the median of pair-wise MAD-to-Mean difference between the best forecasting method and SMA across series. The null hypothesis is the median of differences between the two corresponding methods is equal to zero.

3. RESULTS AND DISCUSSION

A total of 9,308 SKUs have been forecasted, 7,432 of them fall into the intermittent category. The other 1,320 items are categorized as lumpy demand, while the remaining 279 and 277 items fall into the erratic and smooth category. Overall, the ADI ranges from 1 to 35 months with median 8.75 months and the CV^2 ranges from 0 to 6.8 with median 0.12. The average demand per unit time ranges from 0.03 to 21,380.33 units per month.

The properties of SKUs on each demand category are shown in Table 2. SKUs on intermittent category are highly varied in inter-order interval, some are only ordered once in a year while the other are ordered more frequently. They, together with those on lumpy category, are very slow-moving and yet make the majority of items held in stocks. Meanwhile, SKUs on smooth and erratic category are ordered almost every month although they are ordered with very variable quantities.

Table 2
Properties of Each Demand Category

Demand Categories	Average Demand per Period (unit per month)		Squared Coefficient of Variation of Demand Sizes		Average Inter-demand Interval (month)	
	Median	IQR	Median	IQR	Median	IQR
Erratic	8.083	14.236	0.744	0.466	1.129	0.177
Lumpy	1.278	2.090	0.750	0.464	2.917	3.056
Smooth	11.500	26.722	0.353	0.141	1.029	0.167
Intermittent	0.139	0.333	0.080	0.210	11.667	12.500

The MASE results generally confirm those of the MAD-to-Mean. For erratic and smooth category they pick SBA (α=0.20) as the best method, while for the lumpy category SMA is considered as the best. Slight disagreement is observed on intermittent category. The MASE of SBA (α=0.05) is the lowest across series, while SES (α=0.05) results in the lowest MAD-to-Mean followed by SBA (α=0.05).

On erratic category (Figure 2), SMA is better than smoothing methods for lower smoothing constants (α=0.05, 0.10) but worse for higher ones (α=0.10, 0.15). It can be seen from Figure 2 that SBA continually performs better than the rest smoothing methods, with one exception for its MASE at α=0.05 where SES is better. On lumpy category where SMA is the best, SES constantly performs as the second best followed by SBA and Croston as shown in Figure 3.

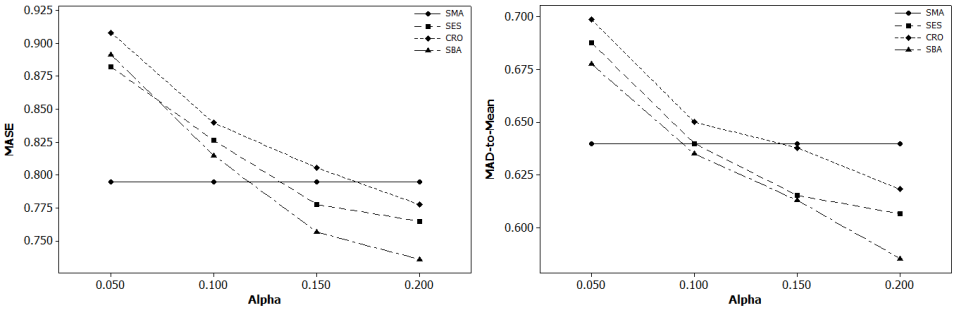


Fig. 2: The median of MASE and MAD-to-Mean Across Erratic Category Series

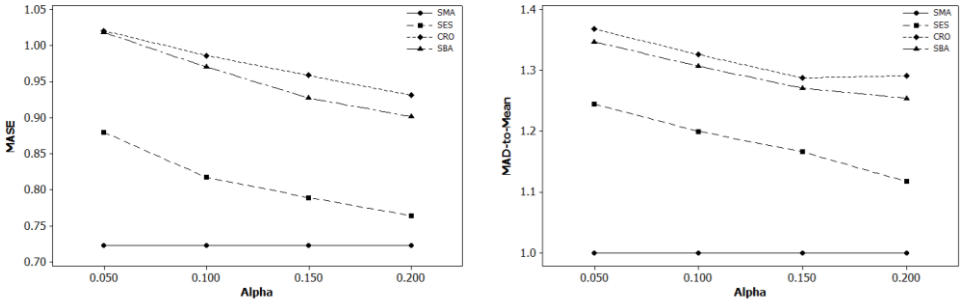


Fig. 3 The median of MASE and MAD-to-Mean Across Lumpy Category Series

SBA once again performs better than the remaining forecasting methods on smooth category apart from its MASE at α=0.05 (Figure 4). This is also the case with intermittent category excluding its MAD-to-Mean at α=0.05 (Figure 5). It should be noticed from Figure 5 below that SES performance on intermittent category deteriorates greatly for alphas higher than 0.05.

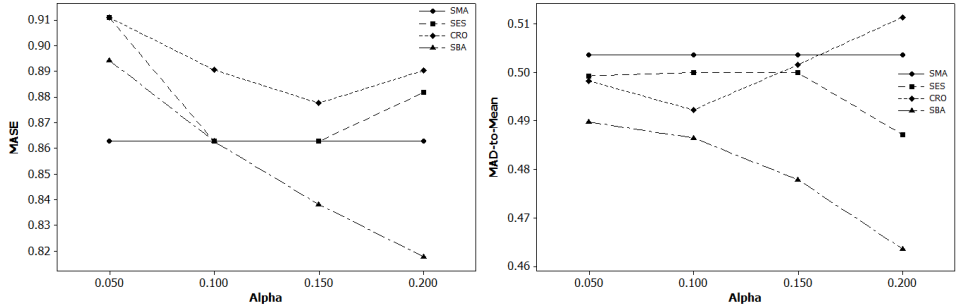


Fig. 4 The median of MASE and MAD-to-Mean Across Smooth Category Series

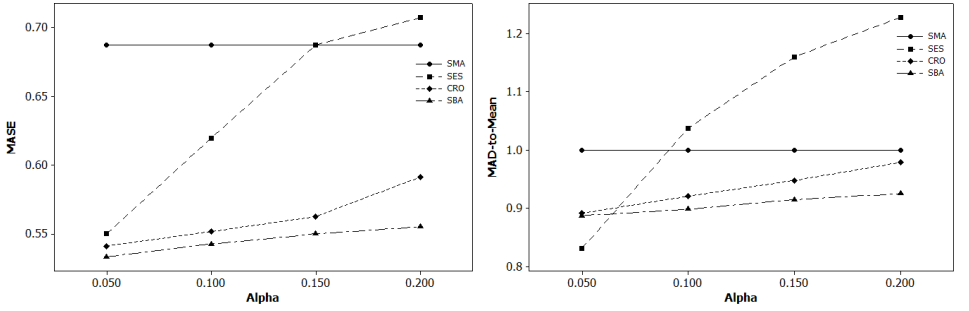


Fig. 5 The median of MASE and MAD-to-Mean Across Intermittent Category Series

Higher smoothing constant is preferred for erratic, lumpy, and smooth category, where α=0.20 generally produces the best result. In contrast, lower smoothing constant (α=0.05) is found to be optimal for intermittent category. This study also finds that, for the same alpha values, SBA performs better than Croston in all cases. These outcomes can be verified from the line plots shown above.

On the whole, the MAD-to-Mean of lumpy category is the highest among other demand categories which makes this category hardest to forecast. As expected, the MAD-to-Mean of smooth category is found to be the lowest. Regarding the MASE results, intermittent category produces MASE across series which is considerably below the other three demand categories. This means that on this category the performance of all corresponding methods is the finest when compared to the naive forecasts.

Talking about the variation of MASE and MAD-to-Mean across series, the inter-quartile range (IQR) of MASE and MAD-to-Mean on lumpy and intermittent category is much higher than those of erratic and smooth category. This can be interpreted as SKUs on these categories are forecasted with various levels of accuracy. Perhaps this is contributed by the large number of items which fall in these categories.

As mentioned earlier, the performance of the best forecasting method on each category is going to be compared with SMA which is viewed as the standard method. Since SMA is the best on lumpy category, it is going to be compared with the second best method (SES with α=0.20). Exception is made on intermittent category, where SES (α=0.20) generates the lowest MAD-to-Mean. In its place, SBA (α=0.20) is going to be matched with SMA because of its more stable forecast errors across alpha values.

The two-sided Wilcoxon signed rank test results are reported in Table 3. All methods with exemption on lumpy category show significant improvement over SMA at 1% significance level. In addition, there is not enough evidence that SMA is better than SES (α=0.20) on lumpy category. Nevertheless, SMA is the one recommended here owing to its simple calculation.

Table 3
The Wilcoxon Signed Rank Test Results of Each Demand Category

Demand Categories	Wilcoxon Signed Rank Tests	Test Results at 1% Significance Level
Erratic	SBA (α=0.20) vs. SMA	Reject the null hypothesis
Lumpy	SES (α=0.20) vs. SMA	Retain the null hypothesis
Smooth	SBA (α=0.20) vs. SMA	Reject the null hypothesis
Intermittent	SBA (α=0.05) vs. SMA	Reject the null hypothesis

The Percentage Best results based on MASE and MAD-to-Mean are alike, so they are averaged and rounded to the nearest half percentages in Table 4. This is in line with Syntetos and Boylan (2005) finding that PBt seems to be insensitive to the descriptive measure chosen. Additionally, the accuracy differences on MASE and MAD-to-Mean across series are not necessarily reflected on PBt, which only counts the number of series

for which one method performs better than all other methods based on their MASE or MAD-to-Mean values.

Table 4
The Percentage Best Results of Each Demand Category

Forecasting Methods	Demand Categories			
	Erratic	Lumpy	Smooth	Intermittent
SMA	18.0%	24.5%	16.0%	13.5%
SES	25.5%	31.0%	24.5%	34.5%
CRO	19.0%	16.5%	21.0%	8.0%
SBA	37.5%	28.0%	38.5%	44.0%

Table 4 suggests the superiority of SBA, with the exception for lumpy category where SES performance is slightly better than SBA. This indicates that SBA is suitable for majority of the SKUs though different smoothing constant should be implemented depending on their demand characteristics.

The forecast errors of all methods are still considerably high, which is probably attributable to the long forecast horizon. The benefits obtained from the above-discussed methods seem lost when forecasting is done for more than one period ahead. This is related to the ability of those methods which can only produce flat forecasts no matter how long the forecast horizon is. The best forecasting method for each demand category based on three error measures is recapped in Table 5.

Table 5
The Best Forecasting Method of Each Demand Category

Demand Categories	**Forecast Accuracy Measures**		
	MASE	**MAD-to-Mean**	**PBt**
Erratic	SBA (α=0.20)	SBA (α=0.20)	SBA
Lumpy	SMA	SMA	SES
Smooth	SBA (α=0.20)	SBA (α=0.20)	SBA
Intermittent	SBA (α=0.05)	SES (α=0.05)	SBA

Based on this table, SBA (α=0.20) is recommended for both erratic and smooth category. Meanwhile, SBA (α=0.05) is advised for intermittent category considering SES deteriorating performance on higher alpha values. As for lumpy category, the hardest category to forecast, SMA is preferred than SES (α=0.20) on account of its simplicity. This is of course a rather surprising finding, which calls for more investigation. It should be pointed out that these outcomes may only be applicable to the current data set.

6. COMMENTS AND CONCLUSION

This study attempts to provide alternative forecasting strategy for an Indonesian heavy equipment company. To help achieving this goal, SKUs are classified into four demand categories, making use of their average inter-demand interval and squared coefficient of variation of demand sizes as the classification parameters with the cut-off values suggested from literature. The forecasting results imply that Syntetos-Boylan Approximation is the most appropriate forecasting method for erratic, smooth, and intermittent demand. On the other hand, Simple Moving Average should be maintained as the standard forecasting method for items with lumpy demand. Nonetheless, the search for a more powerful forecasting method which is straightforward to apply on the company software package is not over yet.

Several extensions can be made for future studies. The optimization of the smoothing constants used to update forecasts has not been taken into consideration. Also, the effect of forecast lead time on forecast accuracy necessitates further assessment. More importantly, the forecasting implication of employing the recommended method on the company inventory system needs to be evaluated. To practitioners, what matters the most are stock control performance metrics such as inventory turnover and customer service level.

REFERENCES

1. Boylan, J.E. and Syntetos, A.A. (2008). Forecasting for Inventory Management of Service Parts. In: Kobbacy KAH, Murthy DNP, editors. *Complex System Maintenance Handbook*. London: Springer-Verlag, 479-508.
2. Boylan, J.E., Syntetos, A.A. and Karakostas, G.C. (2008). Classification for Forecasting and Stock Control: A Case Study. *J. Opl. Res. Soc.*, 59, 473-481.
3. Syntetos, A.A. and Boylan, J.E. (2005). The Accuracy of Intermittent Demand Estimates. *Int. J. Forecast*, 21, 303-314.
4. Syntetos, A.A., Boylan, J.E. and Croston, J.D. (2005). On the Categorization of Demand Patterns. *J. Opl. Res. Soc.*, 56, 495-503.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 299-310

ESTIMATION OF DEMOGRAPHIC STATISTIC OF PEST *APHIS GLYCINES* BY LESLIE MATRIX AND LOTKA-EULER EQUATION BASED ON *JACKKNIFE* RESAMPLING

Leni Marlena[1], Budi Susetyo[1] and **Hermanu Triwidodo[2]**
[1] Department of Statistics, Bogor Agricultural University,
Bogor, Indonesia. Email: stat.leni@yahoo.co
buset008@yahoo.com
[2] Department of Plant Protection, Bogor Agricultural University,
Bogor, Indonesia. Email: petanimerdeka@gmail.com

ABSTRACT

Estimation of demographic statistic is needed to estimate the growth of a population. The population based on the growth of the female in reproducing the off springs. Gross reproductive rate (*GRR*), net reproductive rate (R_0), intrinsic rate of increase (*r*), and generetion time (*T*) are the demographic statistics which are important in determining the level of pest population density. The cohort data from biological research of *A. glycines* of soybean plant control and soybean plant with PGPR was given. Based on the existing cohort data, a life table was constructed, following the estimation of *GRR*, R_0, *r*, and *T*. Based on this biological observation, only one value of those statistic was obtained. It means that the data does not have variability, whereas variability is needed on parameter estimation. In statistics, there are known resampling methods such as *jackknife.* Using those method, new set cohort data can be generated from the original data set so that some demographic statistic could be obtained. There are two approximations to estimate *r*, which are by Leslie matrix and Lotka-Euler equation. Based on the computation techniques and computation time proved that the estimation of *r* by Leslie Matrix is much easier and faster than by Lotka-Euler Equation. The t-test at $\alpha = 5\%$ shows that PGPR can dencrease the intrinsic rate of increase.

KEYWORDS

Demographic Statistic, *Jackknife*, Leslie Matrix, Lotka-Euler Equation, Pest.

1. INTRODUCTION

Estimation of demographic statistic is needed to estimate the growth of a population. Insect demographic statistic is a quantitative analysis of insect populations which related to survival, fecundity, and population growth patterns (Andrewartha and Birch 1982). Gross reproductive rate (*GRR*), net reproductive rate (R_0), intrinsic rate of increase (*r*), and generation time (*T*) are the demographic statistics which are important in determining the level of pest population density. Gross reproduction rate (*GRR*) is the total number of offspring produced by all female during a single generation. Net reproduction rate (R_0) the total number of offspring produced per female during the interval of time (Poole 1974). Intrinsic rate of increase (*r*) is of values as a means of describing the growth

potensial of a population under given climate ad food condition (Messenger 1964; Watson 1964 in Poole 1974). Then generation time (*T*) is the mean time from birth of parents to birth of offspring (Price 1997 and Begon *et al.* 2008).

The method to collect the data of which was used to obtain the statistic, was done by preserving some pests to then each pest was observed on daily basis and any growth to be noted. The observation result is then to be called as cohort data. Based on the existing cohort data, a life table was constructed, following the estimation of *GRR*, R_0, *r*, and *T*. An example of the life table can be seen at Table 1. Based on this biological observation, only one value of those statistic was obtained. It means that the data does not have variability, whereas variability is needed on parameter estimation. If a researcher would like to obtain the variability of those demographic statistic, so the biological observation ought to be repeated many times. However, to conduct such method, much cost and time will be needed.

In statistics, there are known resampling methods such as *bootstrap* and *jackknife*. But in this paper, just focused on *jackknife* resampling method. Using those method, many new set cohort data can be generated from the original data set so that some demographic statistic could be obtained. In this research, t-test was applied to determine whether or not the PGPR can be used to decrease the intrinsic rate of increase of *A. glycines.* Intrinsic rate of increase (*r*) commonly is estimated using Lotka-Euler equation optimization. Optimizing Lotka-Euler equation is done iteratively so it needs pretty high computation technic and needs much computation time. Estimation *r* using Leslie matrix will be easier and faster computationally. In this paper will be proven that estimating *r* using Lotka-Euler Equation gives same result with using Leslie matrix.

2. LITERATURE REVIEW

Population Growth Model

As time passes the population density increases at a faster and faster rate until (preserving the myth of the unlimited environtment) with time the number of individuals approaches infinity (Poole 1974). The equation of population growth is

$$N_t = N_0 e^{rt} \quad (1)$$

where N_t is the number of individuals in the population at some time *t*, and N_0 is the population density at some arbitrarily set time t=0. The letter *t* represent units of time, and *r* is a constant of proportionality.

The rate at which *N*, the number of individuals, changes as time changes can be represented as *dN/dt*,

$$\frac{dN}{dt} = rN$$

at time *t* the rate of increase in density is postulated to be equal to the constant *r* times the number of individuals present. The constant *r* is a measures of the rate of multiplication of the populationat during an interval of time equal to 1. The value of *r* is a constant for a constant set of environmental conditions. If these conditions change, usually so will *r* (Poole 1974).

The instrinsic rate of increase is defined only for a population with stable age distribution. Eq. (1) is valid only of there is a stable age distribution and only if r do not change. Rearranging Eq. (1) slightly considering the change in number of individuals for only one time interval t=1 gives

$$\frac{N_{t+1}}{N_t} = e^r = \lambda$$

the finite rate of increase. While r is the rate of increase of a population with stable age distribution, λ is the multiplication of the population in one interval of time.

The Life Table

The basic information needed to study density changes and rates of increase or decrease is contained in a life table. A life table contains such vital statistics as the probability of a certain age dying or, conversely, the average number of offspring produced by a female of given age (Poole 1974). Life table analysis is the most reliable method to account for survival and reproduction of a population (Price *et al*. 2011). Females are the focus of life table budgets because of their reproductive potential. A life table be constructed with following columns: (1) x is the pivotal age in units of time (days, week, etc); (2) l_x is the probability of females surviving at age x; (3) m_x is the number of offspring produced by each female surviving at age x; then (4) Columns l_x and m_x are then multiplied together to give the total number of female offspring (female eggs laid) in age interval (the pivotal age being x); this is $l_x m_x$ column (Southwood 1978; Begon *et al*. 2008). The example of life table can be seen in Table 1.

The total number of female offspring produced by all female during a single generation (*GRR*) is equal to sum of m_x, or

$$GRR = \sum m_x.$$

The most important summary term that can be extracted from a life table and fecundity schedule is the basic reproductive rate, denoted by R_0 (Begon *et al.* 2008). This is the mean number of offspring produced per female by the end of the cohort. The more usual way of calculating R_0, however, is from the formula:

$$R_0 = \sum l_x m_x . \qquad (2)$$

Table 1
Example: *A. glycines* life table

x (pivot age)	l_x	m_x	$l_x m_x$
0.5	0.9512	0	0
1.5	0.9512	0	0
2.5	0.9268	0.2105	0.1951
⋮	⋮	⋮	⋮
d^* – 0.5	0	0	0
	$\sum$	*GRR*	R_0

*length days of observation

Estimation of The Intrinsic Rate of Increase (*r*)

The calculation of the parameter *r*, is based only on the females of population, and its assumed that there are enough males to ground. The net reproduction rate, R_0 is analogous to λ, the finite rate of increase, except that λ is defined for an interval of time equal to 1, and R_0 for a lenght of time equal to the mean lenght of generation, *T*. If the time intervals *t* equals to *T*, R_0 equals to λ, and in general

$$R_0 = e^{rT}.$$

The mean generation time can be roughly estimated by dividing the log to the base *e* of R_0 by the intrinsic rate of increase *r*.

$$T = \frac{\ln R_0}{r}. \quad (3)$$

If λ is close to 1.0, as simple approximation to Eq. (3) (Dublin dan Lotka 1925) is

$$T \approx \frac{\sum x l_x m_x}{\sum l_x m_x}.$$

If *T* is given, the rough estimate of *r* is

$$r = \frac{\ln R_0}{T}. \quad (4)$$

This is only a rough, crude approximation but may be fairly accurate if λ is close to 1.0.

If mortality and fertility remain the same and the population has a stable age distribution, the value of *r* may be estimated from the equation (Poole 1974)

$$\sum_{x=0}^{n} e^{-rx} l_x m_x = 1. \quad (5)$$

The possible value of *r* are subsituted into Eq. (5), possibly beginning with the rough estimate calculated from Eq. (4), until the left side of the equation equals 1. This method of computation is called iteration or an iterative process.

Leslie Matrix

Matrix population models are as a result of studies by Bernadelli (1941), P.H Leslie (1945, 1948) and Lewis (1942). In his 1945 paper, Leslie expressed the basic age-specific projection equations in matricial form and applied matrix analysis to determine the stable age distribution. Leslie matrix model, named after P.H Leslie due to his discovery,generally makes use of age specific rates of fertility and mortality of a population. In its simplest form, Leslie matrix model only considers female population. This assumption is from the fact that there will always be enough males to fertilize the females.

Three basic statistics necessary to construct Leslie matrix is (Poole 1974):

1. $n_{x,t}$: number of females alive in the group age *x* to *x* + 1 at time-*t*;
2. p_x : probability of a female surviving into the next age class;
3. F_x : average number of female offspring produced at interval time *t* to *t*+1 at age group *x* to *x*+1.

If k is maximum reproduction age of an organism at cohort data, based on the definition before, so:

$$n_{0,t+1} = F_0 n_{0,t} + F_1 n_{1,t} + \cdots + F_k n_{k,t} \quad (6)$$

and

$$\begin{aligned} n_{1,t+1} &= p_0 n_{0,t} \\ n_{2,t+1} &= p_1 n_{1,t} \\ &\vdots \\ n_{k,t+1} &= p_{k-1} n_{k-1,t} \end{aligned}$$

the equation above can be written in the following matrix:

$$\mathbf{n}_{t+1} = \begin{bmatrix} n_{0,t+1} \\ n_{1,t+1} \\ n_{2,t+1} \\ \vdots \\ n_{k,t+1} \end{bmatrix} = \begin{bmatrix} F_0 & F_1 & \cdots & F_{k-1} & F_k \\ p_0 & 0 & \cdots & 0 & 0 \\ 0 & p_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & p_{k-1} & 0 \end{bmatrix} \begin{bmatrix} n_{0,t} \\ n_{1,t} \\ n_{2,t} \\ \vdots \\ n_{k,t} \end{bmatrix} = \mathbf{M}\mathbf{n}_t \quad (7)$$

or can be writtern as:

$$\mathbf{n}_{t+1} = \mathbf{M}\mathbf{n}_t,$$

Vector $\mathbf{n}_t$ indicates a population size by age category were counted at time t, and $\mathbf{M}$ is the Leslie matrix.

Leslie matrix and life table is a demographic model commonly used to evaluate the viability of a population. Leslie matrix and life tables can be used to estimate the intrinsic rate of increase. Leslie matrix has at most one positive eigen value (the other is negative numbers or complex), say λ_1. For the ecologist, λ_1 called the dominant eigen value.

The dominant eigen value is also called the finite rate of increase, and is related to the intrinsic rate of increase, r, by

$$\ln \lambda_1 = r.$$

Jackknife

The jackknife or “leave one out" procedure is a cross-validation technique, first developed by Quenouille to estimate the bias of an estimator. John Tukey then expanded the use of the jackknife to include variance estimation and tailored the name of jackknife because like a jackknife – a pocket knife akin to a Swiss army knife and typically used by boy scouts – this technique can be used as a ”quick and dirty" replacement tool for a lot of more sophisticated and specific tools.

Jackknife method is based on sequentially deleting points x_i and recomputing $\hat{\theta}$. Let $\hat{\theta}_n$ be an estimator of θ based on n i.i.d random vectors X_1, X_2, ..., X_n, i.e., $\hat{\theta}_n = f_n(X_1, X_2, ..., X_n)$, for some function f_n. Let

$$\hat{\theta}_{n,-1} = f_{n-1}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$$

be the corresponding recomputed statistic based on all without the i-th observation. Let (Efron 1982)

$$\hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{n,-i}.$$

The jackknife estimate of bias is

$$\widehat{Bias}_{jack} = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}_n) \quad (8)$$

leading to the bias – corrected "jackknife estimate" θ_{jack} of θ,

$$\theta_{jack} = \hat{\theta}_n - \widehat{Bias} = n\,\hat{\theta}_n - (n-1)\hat{\theta}_{(\cdot)}. \quad (9)$$

Tukey's formula for estimating standard error of $\hat{\theta}_n$,

$$\widehat{se}_{jack} = \left[\frac{n-1}{n}\sum_{i=1}^{n}\{\hat{\theta}_{n,-i} - \hat{\theta}_{(\cdot)}\}^2\right]^{1/2}. \quad (10)$$

3. DATA AND METHODOLOGY

Data that will be used in this study is cohort data of *A. glycines* soybean plant control and soybean plant with PGPR (*Plant Growth Promoting Rhizobacteria*), result study of the Undergraduate Student of Plant Protection Departement, Bogor Agricultural University. The study was conducted from January to April 2013. Observations were made every day on each of *A. glycines* were still alive or dead, change skin, and the number of nymphs born. Cohort data of *A. glycines* soybean plant control consist of 46 individuals with length days of observation were 37 days. Cohort data of *A. glycines* soybean plant with PGPR consist of 41 individuals with length days of observation were 26 days.

The following is the steps to calculate demographic statistic of *A. glycines* using *jackknife*:

1. Remove the *i*-th row of original cohort data to form a new cohort data with $k = b - 1$, and $i = 1, 2, ..., b$;
2. a. Calculate m_x, l_x, and $l_x m_x$;
   b. Calculate $F_x = m_x$ and p_x;
3. a. Construct a life table based on the result value from step 2a;
   b. Construct a Leslie matrix based on the result value from step 2b;
4. Calculate *GRR*, R_0, *r*, and *T*. The demographic statistic, *r* and *T* is calculated twice. The first is calculated based on a life table (3a) and the second is calculated based on Leslie matrix (3b);
5. Repeat step 1, 2, 3, and 4 until the *n*-th row of original cohort data is deleted;
6. Calculate the estimation of *GRR*, R_0, *r*, and *T* based on *jackknife* resampling;
7. Calculate the standard error for those each demographic statistic;
8. Construct the 95% confidence interval of *GRR*, R_0, *r*, and *T*.

The algorithm flowchart of intrinsic rate of increase can be seen on Figure 2. If θ_{jack} is the estimator of θ based on *jackknife* resampling, so the 95% of confidence interval will be calculated by this formula:

$$\theta_{jack} \pm z_{\alpha/2} se_{jack}..$$

The t-test is applied to determine whether or not the PGPR can be used to decrease the intrinsic rate of increase of *A. glycines.*

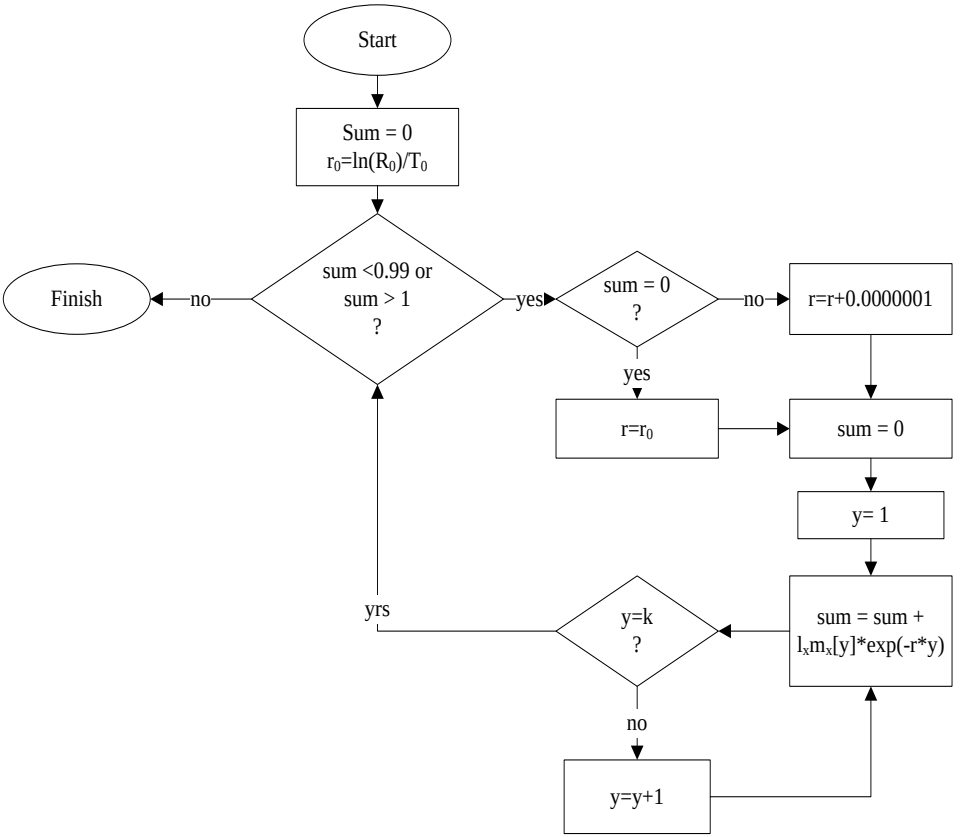


Figure 2: Algorithm of r Estimation using Lotka-Euler Equation

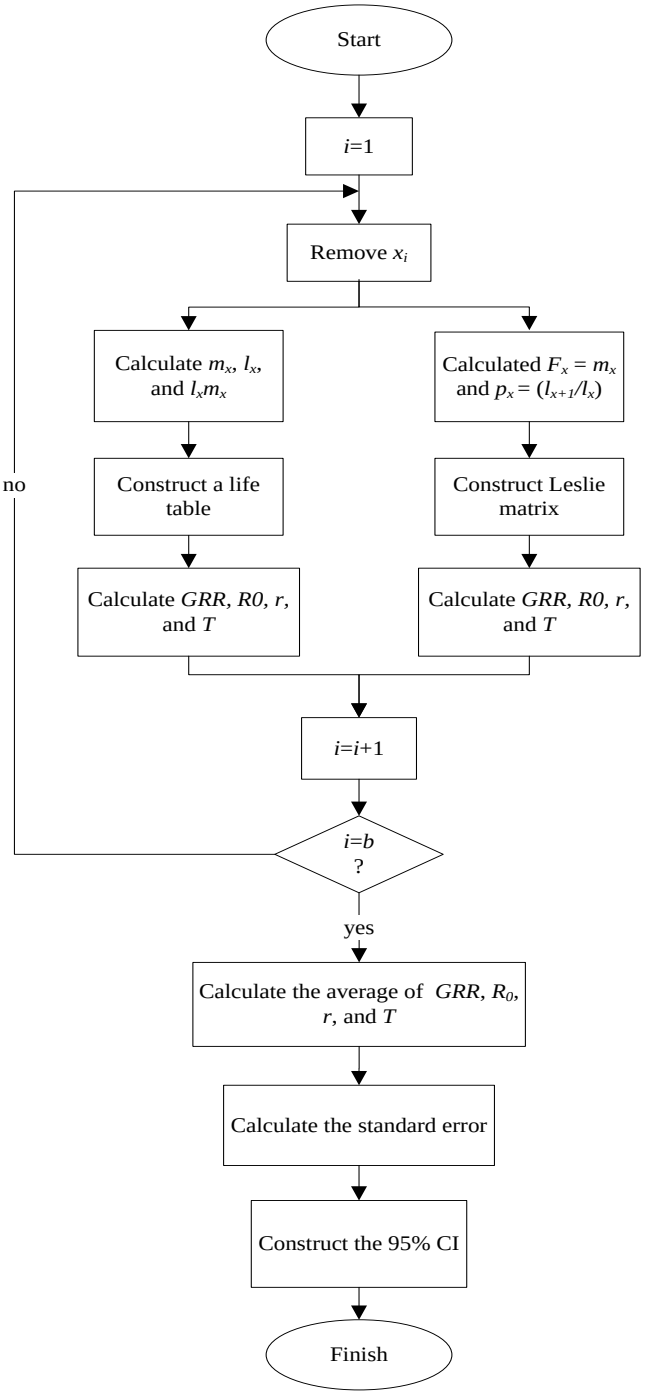


Figure 1: Algorithm of Jackknife Resampling on A. Glycines Cohort Data

4. RESULT AND DISCUSSION

The demographic statistic of pest based on original cohort data of *A. glycines* that was calculated by life table and Leslie matrix can be seen on Table 2. The value of gross reproduction rate (*GRR*) of *A. glycines* soybean plant control is 104.86. It means number offspring produced by all *A. glycines* female during the cohort generation is about 105 offspring. Then the net reproduction rate (R_0) is 63.326, R_0, can be interpreted that the population of *A. glycines*has capacity to multiply about 63 times in each generation under the given set of environmental condition and in an unlimited environment (Poole 1974). The intrinsic rate of increase, *r*, is 0.537, it means the number of individuals added to populations *A. glycines* are 0.537 pest per day (Andrewartha dan Birch 1982). Then the generation time (*T*) is 8.958 days.

Table 2
Demographic Statistic of Original Cohort Data

Cohort	GRR	R_0	r	T
A. glycines soybean plant control	104.861	63.326	0.537	7.724
A. glycines soybean plant with PGPR	71.834	57.780	0.513	7.911

For soybean plant with PGPR, the *GRR* is 71.834. It means that the number offspring produced by all *A. glycines* female during the cohort generation is 72 offspring. Then the population of *A. glycines* of soybean plant has capacity to multiply 57.781 times in each generation under the given set of environmental condition and in an unlimited environment. The intrinsic rate of increase estimation, *r*, is 0.513, it means the number of individuals added to populations *A. glycines*are 0.513 pest per day. The cohort generation time (*T*) estimation of *A. glycines* is 7.911 days.

Table 3
The 95% Confidence Interval of Each Demographic Statistic

Demographic statistic	*A. glycines* soybean plant	
	Control	PGPR
GRR	112.687 ± 50.490	71.834 ± 6.551
R_0	63.326 ± 10.834	57.780 ± 8.998
r	0.537 ± 0.016	0.513 ± 0.018
T	7.720 ± 0.209	7.904 ± 0.248

Table 3 shows 95% confidence interval of each demographic statistic. In general, the cohort data of *A. glycines* soybean plant control has greater standard error than the cohort data of *A. glycines* soybean plant with PGPR. So the confidence interval for estimation of demographic statistic of *A. glycines* soybean plant control wider than the cohort data of *A. glycines* soybean plant with PGPR.

Table 4
The t-test Effectiveness of PGPR to Intrinsic Rate of Increase (*r*) of *A. glycines*

Method	H_0	H_1	Nilai-p
Jackknife	$r_A \leq r_B$	$r_A > r_B$	0.000^*

A: *A. glycines* soybean plant control, B: *A. glycines* soybean plant with PGPR; *significant at α=5%.

Decreasing value of both intrinsic rate of increase of *A. glycines* soybean plant control and PGPR shows that PGPR application is effective to obstruct the *A. glycines* rate of increase. T-test is needed to show it. Based on resampling *jackknife,* we have 46 estimator for each demographic statistic of *A. glycines* soybean plant control and 41 estimator for demographic statistic of *A. glycines* soybean plant with PGPR. Using 5% level of significance, the t-test shows that PGPR can be used to decrease the intrinsic rate of increase of *A. glycines.*

Table 5 shows that computation time to estimate the demographic statistics of *A. glycines* soybean plant with PGPR is faster than *A. glycines* soybean plant control. Its because the sample size of *A. glycines* soybean plant with PGPR is less than the sample size of *A. glycines* soybean plant with PGPR. On Table 5 also shows that to estimate the demographic statistics of *A. glycines* by Lotka-Euler equation needs too much time. It can caused by iterative process on this method. On Leslie matrix there is no iterative process, so the computational time is very fast.

Table 5
Computational time to estimate demographic statistic of A. glycines

Data	Method	Time (second)
Soybean plant control	Lotka-Euler	1639.580
	Leslie matrix	0.280
Soybean plant with PGPR	Lotka-Euler	801.630
	Leslie matrix	0.180

6. CONCLUSION

Estimation of statistical demographic of *A. glycines* either using Leslie Matrix or Lotka-Euler equation produces exactly the same value. Estimating demographic statistic of *A. glycines* by Leslie Matrix is much easier and faster than by Lotka-Euler equation based on the computational time dan techniques. The t-test at α = 5% shows that PGPR can dencrease the intrinsic rate of increase.

REFERENCES

1. Agustini A. (2013). Pengaruh Plant Growth Promoting Rhizobacteria Terhadap Biologi dan Statistik Demografi *Aphis glycines* Matsumura (Hemiptera: Aphididae) pada Tanaman Kedelai [Undergraduate Thesis]. Bogor (ID): Institut Pertanian Bogor.
2. Andrewartha HG and Birch LC. (1982). *Selections from The Distributions and Abundance of Animals*. Chicago (USA): The University of Chicago.
3. Begon M, Townsend CR, and Herper JL. (2008). *Ecology: From Individuals to Ecosystems*. 4th edition. Oxford (UK): Blackwell Puslishing.
4. Chernick MR. (2008). *Bootstrap Methods: A Guide for Practitioners and Researchers.* Second Edition. New York (USA): John Wiley & Sons.
5. Efron B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Philadelphia (PA
6. Poole RW. (1974). *An Introduction to Quantitative Ecology*. USA: McGraw-Hill.
7. Price PW. (1997). *Insect Ecology*. 3th ed. New York (USA): John Wiley & Sons.
8. Price PW, Denno RF, Eubank MD, Finke DI, Kaplan I. (2011). *Insect Ecology Behavior, Population and Communities*. Cambridge (UK): Cambridge University Press.
9. Southwood TRE. (2000). *Ecological Methods*. 3th ed. New York (USA): Wiley-Blackwell.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 311-318

SIMULTANEOUS ANALYSIS OF THE LECTURERS POSITIONING AND STUDENTS SEGMENTATION IN THE SELECTION OF THESIS SUPERVISOR

Yusma Yanti[1], Bagus Sartono[2] and **Farit M Afendi[2]**
[1] Graduate School, Bogor Agricultural University
Bogor, Indonesia. Email: yusmayanti.fn@gmail.com
[2] Department Statistics, Bogor Agricultural University
Bogor, Indonesia.

ABSTRACT

Segmentation is a strategy to understand the structure of the market, so people clustered based on certain characteristics. While positioning is the way an individual point of view of an object. The classical approach regarding segmentation and positioning is to them separately. DeSarbo (2008) propose Clusterwise Bilinear Spatial Multidimensional Scaling Model (CBSMSM) to simultaneously analysis to segmentation and positioning. The CBSMSM parameters are estimated by using alternating least squares (ALS) algorithm. This analysis is apply to thesis supervisor preference undergraduate Statistics IPB. Results are expected in this model will simultaneously estimating the number of segments, segment-forming attributes, a objects space and a preference vector into segments for each segment. One of the results is that the students are divided into four segments and each segment has different characteristics. The characteristics of each segment can be described by its constituent attributes, and is also known lecturer position in each segment.

KEYWORD

Alternating Least Square, Thesis Supervisor, Clusterwise Analysis.

1. INTRODUCTION

Segmentation is basically a strategy to understand the market structure. Statistical methods that are often used include: regression based segmentation, latent class analysis and the analysis of the Q factor (Kasali, 1998). Regression based segmentation states that sometimes clusters formed is not related at all to the behavior of the respondent. So could every cluster has two or three groups with different average. Latent class analysis is developed from the analysis of the factors that has two functions, namely to identify the theoretical constructs of the variables and to reduce the number of variables. This method helps to estimate the latent parameters that provide information such as the number of segments and opportunities for a product enter one segment. While the Q factor analysis to identify groups of respondents (segment).

Apart from segmentation, anoke important thing to be known in the marketing process is positioning. Positioning is not the things we do on the product, but something that will be done to the brain of potential customers. In this process the respondent is

expected to position the product and attributes that respondents had specific assessment and identify themselves with the product. There are several techniques that can be used for positioning analysis. First, the perception map technique, which is often used is the analysis of discriminant analysis and multidimensional scaling (MDS). Second, preference mapping techniques, this technique using factor analysis, discriminant analysis and MDS. Third, the laddering technique is a technique that identifies the attributes that make up the preferences.

During this time, the two processes do not overlap. In practice people do only perform either segmentation analysis, or positioning analysis. Empirical modeling approach that has been done on marketing segmentation and positioning are using MDS and cluster analysis. The analysis of segmentation and positioning are performed relationship of positioning the product with the attributes. While cluster analysis for classifier market segments based on existing attributes. But there are some problems in this method. First, each type of analysis typically optimize different loss function, different aspect of the data are often ignored by the disjointed application of such sequential procedures. Second, there are many types of procedures, as documented in the vast psychometric and classification literature, and each procedure can render different results (DeSarbo 1994).

To solve this problem, we could apply an approach of MDS and cluster sequential analysis using parametric finite mixture or better known as latent class MDS. In this model the vector or ideal point segment each respondent not come from individuals but by groups of respondents. DeSarbo (2008) stated drawbacks latent class MDS. First, the latent class MDS is a parametric model that requires a certain distribution assumption. Often, continuous support distributions are used in the finite mixture and are applied to discrete response scale, it is unlikely that the assumptions used. Second, if the data distribution deviates from the exponential family, when multivariate normal distributions are employed, problems will arise in the estimation of separate full covariance matrices by derived segment. Third, latent class MDS are highly nonlinear, so it takes a long computation time when estimation procedure. Fourth, there is a local optimum solution requires the analysis is repeated for each value dimension and groups. Fifth. The solution is obtained by heuristic means that the solution is not the most optimal, but closer to the optimum solution. These properties may give different results for each information. Finally, latent class MDS can cause fuzzy opportunities that are difficult to interpret because it is a solution that consists of multiple partitions.

The solution to solving the above problem is with a method of data reduction and classification simultaneously which can be applied to the segmentation and positioning. The procedure used also does not require the assumption of distribution as well as the existing MDS. In addition, it is also expected to estimate the global optimum parameters in the faster estimation stage for each iteration and more time efficient. This analysis is known as the Clusterwise Bilinear Spatial Multidimensional Scaling Model. Application of this analysis will be done on the field of education, the process apply to thesis supervisor preference undergraduate statistics IPB. This data is based on several lecturer selection criteria used by the student in undergraduate statistics IPB, the students can propose a lecturer to be their thesis supervisor. But in reality, there are many students who apply for piling on just a few lecturers. So that, the departments must divided number of students guided by each lecturer. The purpose of this study was to obtain

simultaneously between segmentation and positioning students in the selection of adviser scientific work, where the segment representation by vectors, the product by the coordinates.

2. CLUSTERWISE BILINEAR SPATIAL MULTIDIMENSIONAL SCALING

DeSarbo (2008) illustrated the proposed Clusterwise Bilinear Spatial Multidimensional Scalling Model (CBSMSM) in the context of simultaneous positioning and segment estimation. Suppose that there are J product and each product is evaluated by N respondent using K attributes. The evaluation scores then are collected into a matrix of $\boldsymbol{Z_{jk}}$. Further, each respondent also stated the preference score for each product and represented by $\Delta_{\boldsymbol{ij}}$ matrix**.** Then, the model written by:

$$\Delta_{\boldsymbol{ij}} = \sum_{s=1}^{S} \boldsymbol{P_{is}} \sum_{r=1}^{R} \boldsymbol{X_{jr}}\, \boldsymbol{Y_{sr}} + \boldsymbol{b} + \boldsymbol{\varepsilon_{ij}} \qquad \textbf{(1)}$$

This equation can be used if K < J. Y_{sr} describe coordinates for the segment s. P_{is} segment membership is a binary indicator variable, provided that:

$$P_{is} = \begin{bmatrix} 0, & \text{if respondent is not classifield in segment s,} \\ 1, & \text{if otherwise} \end{bmatrix}$$

$P_{is} \in \{0,1\}, \sum_{s=1}^{S} P_{is} = 1$, if a respondent for partitions

$0 < \sum_{s=1}^{S} P_{is} \leq S$, for overlapping segment

Visually, the vector MDS is shown on the data structure (space products and segments) and simultaneously group the respondents into segments, allowing to do the partition or segment members overlapping. The projection of a vector product segment to indicate the magnitude of the respondent wishes to choose a product on the segment. The addition of the normal vector typical attributes and $\boldsymbol{\alpha_{kr}}$ plot or suitable methods such as regression of each attribute in two dimension is used in marketing application because the direction of the vector indicates the movement of a product improved predictions. Some attribute vectors are plotted without normalization, so that its length gives an indication of how the relationship between the attributes of the dimension. Or it can provide a correlation table between product attributes and product coordinates on the dimensions. So as to provide a brief summary of the data structure associated with segmentation-targeting-positioning.

The Estimation Procedure

Given Δ, and values of S and R, our goal is to estimate $\boldsymbol{P} = (P_{is})$, $\boldsymbol{X} = \left(X_{jr}\right)$ a or $\boldsymbol{\alpha} = (\alpha_{kr})$, b, and $\boldsymbol{Y} = (Y_{sr})$ to minimize the following sum of square error (SSE)

Equation (1) has been obtained: $\Delta_{ij} = \sum_{s=1}^{S} P_{is} \sum_{r=1}^{R} X_{jr}\, Y_{sr} + b + \varepsilon_{ij}$. So it can obtained: $\varepsilon_{ij} = \Delta_{ij} - \sum_{s=1}^{S} P_{is} \sum_{r=1}^{R} X_{jr}\, Y_{sr} - b$. To estimate $\boldsymbol{P, X, Y,}$ and b then we minimize Φ, where:

$$\begin{aligned} \Phi &= \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\Delta_{ij} - \sum_{s=1}^{S} P_{is} \sum_{r=1}^{R} X_{jr} Y_{sr} - b\right)^2 \\ &= \sum_{i=1}^{I} \sum_{j=1}^{J} \varepsilon_{ij}^2 \end{aligned} \qquad (2)$$

If $\Delta_{IJ} = (\Delta_{ij})$, $\Delta_{ij} = \sum_{s=1}^{S} P_{is} \sum_{r=1}^{R} X_{jr} Y_{sr} + b + \varepsilon_{ij}$, if we let $\Delta = \Delta^* + b$, then $\Delta^* = \boldsymbol{PYX}' + \boldsymbol{\varepsilon}$, so that:

$$\begin{aligned} \text{Min } \Phi &= Min\ tr(\boldsymbol{\varepsilon}'\,\boldsymbol{\varepsilon}) = tr\,[(\Delta^* - \boldsymbol{PYX}')'(\Delta^* - \boldsymbol{PYX}')] \\ &= tr\,[\Delta^*\Delta - \Delta^{*\prime}\,\boldsymbol{PYX}' - \boldsymbol{XYP}'\Delta^* + \boldsymbol{XY}'\boldsymbol{P}'\boldsymbol{PYX}] \\ &= tr(\Delta^*\Delta) - 2\,tr\,(\Delta^{*\prime}\,\boldsymbol{PYX}') + tr(\boldsymbol{XY}'\boldsymbol{P}'\boldsymbol{PYX}) \end{aligned}$$

Alternating Least Squares (ALS) algorithm to estimate the model complete clusterwise estimation is done by five cycle estimation steps:

1. Estimate X
   Beginning with the first order conditions, we calculate the partial derivatives of SSE expression (equation 2) with respect to $\boldsymbol{X}$. Solving for $\boldsymbol{X}$,

   $$\widehat{\boldsymbol{X}} = \Delta^{*\prime}\,\boldsymbol{PY}\,(\boldsymbol{Y}'\boldsymbol{P}'\boldsymbol{PYX})^{-1} \tag{3}$$

   which is estimable only for $R \leq S$ (one of the identification restrictions for overlapping segment) with equation (3), then $\boldsymbol{X} = \boldsymbol{Z}\,\boldsymbol{\alpha}$, when we use the chain rule for derivatives equation reduce to the following:

   $$2\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\alpha}(\mathbf{Y}'\boldsymbol{P}'\boldsymbol{PY}) - 2\boldsymbol{Z}'(\Delta^{*\prime}\boldsymbol{PY}) = 0 \tag{4}$$

   $$\widehat{\boldsymbol{\alpha}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\,(\Delta^{*\prime}\boldsymbol{PY})\,(\mathbf{Y}'\boldsymbol{P}'\boldsymbol{PY})', \tag{5}$$

   This value exists if the terms of R and S in the same order condition, full rank matrix $\boldsymbol{Z}$ and $K < J$ met.

2. Estimate Y
   Beginning with the first order conditions, we calculate the partial derivatives of SSE expression (equation 4) with respect to $\boldsymbol{Y}$. Which follows from the properties of the trace operator and its derivatives, thus:

   $$\widehat{\boldsymbol{Y}}' = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\Delta^{*\prime}\boldsymbol{P}(\boldsymbol{P}'\boldsymbol{P})^{-1} \tag{6}$$

3. Estimate P
   Note that $P_{is} \in \{0,1\}$ this represents the segment membership indicator binary variables such that: $\sum_s P_{is} \geq 1, \forall i$ and $\sum_s P_{is} > 1, \forall S$, Thus: $\Phi = tr(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}) = tr(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon})$ and $\Phi = \sum_{i=1}^{I} H_{ii}$ with $\boldsymbol{H} = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ thus:

   $$\Phi = \sum_{i=1}^{I}(\Delta_i^* - P_i\boldsymbol{YX}')'(\Delta_i^* - P_i\boldsymbol{YX}') \tag{7}$$

   Because Δ_i^* and P_i only affect Φ in the ith observation, the optimization here is separable over i. That is, $\sum_s P_{is} > 1, \forall S$ can be conditionally minimized by observation to obtain a conditionally global optimal $\boldsymbol{P}$,given $\boldsymbol{X}$, $\boldsymbol{Y}$, Δ^*. For each i, we minimize $\Phi_i = \varepsilon_i\,\varepsilon_i'$ with respect to P_i. here, we enumerate over all solution options (S options for the case of partitions: $2^S - 1$ solutions options for overlapping segments) for each P_i (ignoring the $\mathbf{0}$ solution of no membership in any derived segment) to minimize $\Phi_i\ \forall_i$.

4. Estimate b

We first define $\hat{\Delta}_{ij} = \sum_{s=1}^{S} \hat{P}_{is} \sum_{r=1}^{R} \hat{X}_{jr} \hat{Y}_{sr}$, with: $\boldsymbol{L} = (\Delta_{ij})$, $\boldsymbol{K} = (\mathbf{1}, \boldsymbol{M})$, $\mathbf{1}' = (1, 1, \cdots, 1)$, $\boldsymbol{M} = (\hat{\Delta}_{ij})$. Then we can formulate this estimation problem as s simple least square one and calculate the following:

$$\begin{pmatrix} \hat{b} \\ \hat{a} \end{pmatrix} = (\boldsymbol{K}'\boldsymbol{K})^{-1}\boldsymbol{K}'\boldsymbol{L} \tag{8}$$

Constant (a) is not identifiable, because it can be directly embedded into $\boldsymbol{X}$ or $\boldsymbol{Y}$ and then set equal to 1.

5. Test for Convergence
VAF (*Variance Accounted For*) can calculate as well as the value of R^2 by using the formula:

$$VAF = \frac{\sum_{i=1}^{N}\sum_{j=1}^{J}(\Delta_{ij} - \hat{\Delta}_{ij})^2}{\sum_{i=1}^{N}\sum_{j=1}^{J}(\Delta_{ij} - \overline{\Delta}_{..})^2}, \text{ dimana } \overline{\Delta}_{..} = \frac{1}{NJ}\sum_{i=1}^{N}\sum_{j=1}^{J}\Delta_{ij} \tag{9}$$

IF $\boldsymbol{VAF^{(IT)} - VAF^{(IT-1)} \leq}$ **error** output all parameters estimated and stop otherwise, increase TI = IT + 1 and return to step 1.

Parameter estimation process in ALS algorithm has the properties of interdependence. Parameters that would supposedly require other variables in the estimation process, while the variable used is also influenced by the parameters to be suspected. So at the beginning of the initialization process of the analysis performed and the results of the analysis will be affected by the initialization. This leads to results that are not always the same when the computing process. To obtain the optimum solution analyzed repeatedly. Each analysis will certainly have different results because ALS is very dependent on the initial initials are used. The analysis is said to be consistent if the difference between the value of SSE at iteration always close together and even always the same. To obtain a conviction results, so the results are expressed consistent conducted replications of 50 to 60 times.

But in the process, although it has been obtained the best value, not necessarily the result of this analysis is quite good. Should be checked first crucial point values for X, Y and α. There is a possibility the entire coordinate Y is concentrated in one group, as well as for X and α. The size of the segment will also be an obstacle, sometimes one segment has very little number of respondents, while the other segments in the accumulation of the respondents. In case of these results, it must be done again until consistent repetition. The best results of the analysis if the value for the Y matrix dispersed, while X and α groupings always stable. For a one-time computational process requires 1-3 minutes, then for each combination of segments and dimensions require a computing time of about 3 hours. Other obstacles also occur during the process of computing a program that is used takes up to 20 minutes or more for a single repetition. And usually results underestimate or overestimate. The results of this analysis certainly cannot be used because it does not fit with the purpose of analysis.

3. RESULT AND DISCUSSION

The CBSMSM was the implemented to the data of a survey on the preference of students in selecting supervisor for their thesis. There are 23 supervisor whom was scored by 65 students. Each student evaluated each of supervisor according to as may as 10 attributes. The list of the attributes as follows:

1. Easy to meet.
2. Has a good reputation on the timeline.
3. Has a good reputation an academic record.
4. Has a good reputation an non-academic record.
5. Familiar to students.
6. Has a good relationship with student.
7. Has an excellent teaching style.
8. Keen and easy going.
9. Competent on topic of the research.
10. Charismatic.

The test step of the analysis in to determination the number of segments and the number of dimentions. İt was determined using trial-and-error approach by trying all combination of both parameters and correcting the SSE (Sum Square Error). The one with 4 segments and 2 dimentions resulted the lowest SSE and considered as the final configuration. It can illustrated by Figure 1.

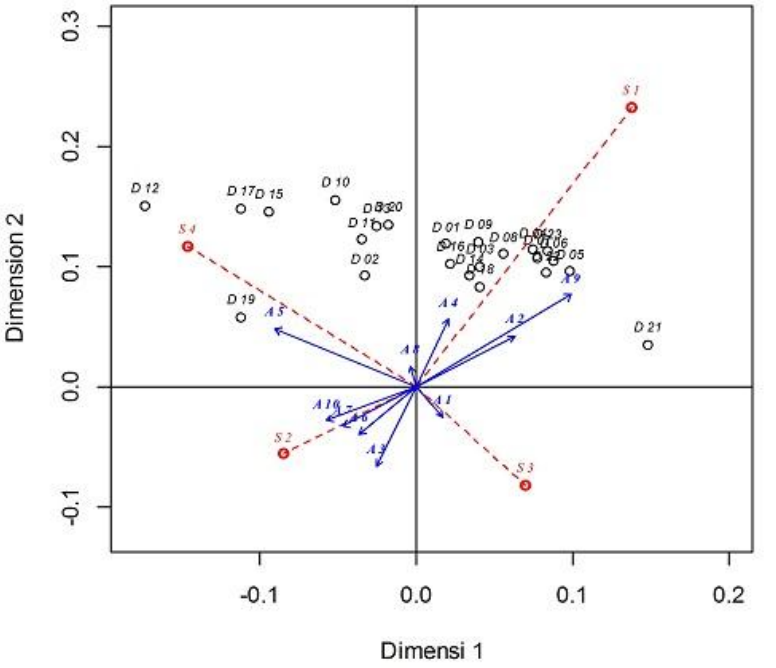


Figure 1: Clusterwise Solution: Dimension 1 versus Dimension 2

Figure 1 illustrates that there is a significant grouping lecturer characteristics. however, there are some lecturer who were driven from large groupings, such as lecturer 12, lecturer 21 and lecturer 19 who looks straight out of different positions of the other lecturers that accumulate in some groups. However, this does not preclude lecturer were able to join in a segment with the other lecturers. Some lecturers do not be a shaper of the four segments, such as lecturer 01, lecturer 02, lecturer 03, lecturer 14 and lecturers 16. This is because, the projected value of the lecturer is to segment smaller than the average value is added to the standard half deviation of the entire lecturer for each segment. Segments s1 formed by the attribute "has a good reputation on the timeline, has a good reputation an non-academic record, keen and easy going, and competent on topic of the research". Lecturers who became formers s1 is lecturer 04, lecturer 05, lecturer 06, lecturer 07, lecturer 08, lecturer 09, lecturer 22 and lecturer 23, the number of respondents consisted of 6 male and 6 respondents female respondents with a median cumulative grade point 3.08. Broadly speaking, s1 formed by the attributes of a general nature. s1 segment is the segment with the lowest median value ipk respondents. the respondents in this segment will indirectly seek information in advance of each lecturer. this segment could also be said that the segment formed by male respondents as the number of male respondents in this segment is greater than the average in general.

Segment s2 with lecturer 12 and lecturer 17 lecturer 19. this segment respondents consists of 8 male respondents and 13 female respondents with a median cumulative grade point 3:47. the characteristics of this segment is formed because the respondents considered the lecturer has a more intellectual level. it is formed by attributes "has a good reputation an academic record, has a good relationship with student, has an excellent teaching style, and charismatic".

S3 segment there is only one lecturer, this is lecturer 21 attribute shaper, which simplifies the lecturer easy to meet is forming this segment. Respondents with a median IPK 3.58 consisting of 14 respondents male and 5 female respondents is forming S3. Similarly, S3, S4 segment is also formed by only one attribute, namely the lecturer feels more familiar to students. Lecturers who in this segment consists of lecturer 10, lecturer 11, lecturer 12, lecturer 13, lecturer 15, lecturer 17, lecturer 19 and lecturers 20. This segment is a segment that is dominated by women respondents consisting of 2 male respondents and 10 female respondents.

6. CONCLUSION

Based on the analysis of data simultaneously between positioning lecturer in the election commission head supervisor of scientific papers by 65 respondents to the 23 lecturer in the Department of Statistics IPB based on 10 attributes, it can be concluded that there are four-segments. Each lecturer has the possibility to become a supporter of several segments. While the attribute will only be one of the supporters of each segment. Attributes of conformity with the interest of the respondent lecturer expertise is an attribute to the distribution of the smallest value. Overall lecturer accumulate in a group, but there are some professors who have quite different values such as lecturer 12, lecturer 19 and lecturer 21, so it looks like an outlier to the other.

REFERENCES

1. DeSarbo, Wayne S. Radjeep Grewal and Crystal J. Scott. (2008). A Clusterwise Bilinear Multidimentional Scaling Methodology for Simultaneous Segmentation and Positioning Analysis. *Journal of Marketing Research*, XLV(June), 280-292.
2. Johnson, A Richard and Wichern W. Dean. (1998). *Applied Multivariate Statistical Analysis 2nd edition*. Prentice Hall International: New Jersey
3. Kasali, Rhenald. (1998). *Membidik Pasar Indonesia Segmentasi Targeting Positioning*: Jakarta.
4. Kotler, Philip (2008). *Manajemen Pemasaran Jilid 1*. Erlangga: Jakarta.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 319-326

CLASSIFICATION OF DROPOUT STUDENT IN SULAWESI WITH BAGGING CART METHODS

Dina Srikandi and **Erfiani**
Departemen Statistika, FMIPA, Institut Pertanian Bogor, Indonesia
Email: dina.srikandi@yahoo.com

ABSTRACT

In 2012, the dropout rate for children aged 7-17 years throughout the province on the island of Sulawesi is still higher than the national average rate. The aim of this research to trace the swampy distribution and characteristic of dropout student require to clasify the children age 7-17 years old according to its characteristics. Classification and regression trees (CART) use in this research with aims to produce several group of dropout student which relative more detailed according to their characteristics. To improve the stability and accuracy of predictive CART then applied the technique Bootstrap aggregating (Bagging) on methods of CART. To measure the accuracy of the classification of the two algorithms are used misclassification rate. The result of this research indicate that the continuity education of the child influenced by the child's age, mother's education level, the child's sex, household size, employment status of household head, and economic status. Application of bagging techniques produce higher classification accuracy than the CART algorithm.

KEYWORD

Accuracy of Classification Rate, Bootstrap aggregating (Bagging), Classification and Regression Trees (CART), Classification of dropout student.

INTRODUCTION

Education has a strategic role in national development to become a developed nation, independent and civilized. Therefore, the increase community access to quality education services is one of the important agenda in national development as contained in the National Medium Term Development Plan 2010-2014 as well as a top priority in the work plan of the Government. However, in reality, not all children have the opportunity to obtain a proper education and the widest possible to cause them to drop out of school. Low levels of education will encourage the emergence of a variety of social problems that increasingly the more troubling. One the factors that can be measured by the low level of education is high dropout rates. If in an area having a high school dropout rate, it can be said the region has a low level of education.

According to Statistics Indonesia (BPS), in 2012 the percentage of population aged 7-17 years who never went to school with school dropout status in Indonesia by 2.72 percent, meaning that out of every 1000 inhabitants aged 7-17 years there were 27 dropouts. Based on the distribution of school dropout rate in Indonesia, there are some

areas that the school dropout rate above the national average. One interesting phenomenon is the whole province on the island of Sulawesi has a dropout rate above the national average, i.e. South Sulawesi (3.84 percent), North Sulawesi (4.52 percent), Southeast Sulawesi (4.54 percent), Central Sulawesi (4.71 percent), west Sulawesi (7.01 percent) and Gorontalo (7.09 percent).

Attempt to resolve the issue one is to identify students drop out of school and out of school by tracing the distribution and characteristics of school dropouts through grouping/classification of school children aged 7-17 years according to its characteristics. This condition is a form of classification of data with many variables that scale variable mixture of both nominal, ordinal, interval and ratio. The classification is usually difficult to meet the assumption of normality and homogeneous variance and more precisely performed with nonparametric approach.

CART approach for classifying statistical data has been widely used in various fields. The purpose of CART is to classify an observation or an observation group into a sub-group of the known classes. Compared with the classical clustering methods, CART has some advantages such as the results easier to interpret, more accurate and faster computation. This method is a method that can be applied to data sets that have a large number of variables that are very much and with variable scale binary mixture through a sorting procedure. According to John and Webb (1999), the level of trust that can be used to classify new data on the accuracy of CART is generated by pure classification tree formed from the data that have the same condition (learning data). The resulting classification tree CART unstable, due to small changes in the learning data will affect the results of prediction accuracy. To overcome these problems, Breiman (1996) introduced the technique of bagging (bootstrap aggregating). Bagging is a technique that can be used with various methods of classification and regression methods to improve the stability and predictive power of CART. For it, in this study will be carried out classification characteristics of school children in addition to the methods of CART also the bagging method CART [1].

CART (CLASSIFICATION AND REGRESSION TREES)

Classification and regression tree (CART) analysis recursively partitions observations in a matched data set, consisting of a categorical (for classification trees) or continuous (for regression trees) dependent (response) variable and one or more independent (explanatory) variables, into progressively smaller groups (De'ath and Fabricius 2000, Prasad et al. 2006).

Main Steps for making a decision tree using CART Algorithm (Soni Sneha. 2010):

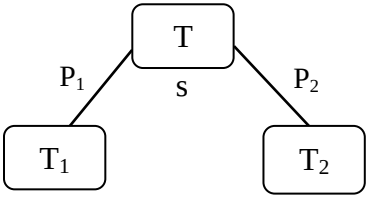
1. The first is how the splitting attribute is selected.
2. The second is deciding upon what stopping rules need to be in place.
3. The last is how nodes are assigned to classes.

Step 1: Splitting a Node: The goal of splitting up a sample is to get sub-samples that are more pure than the original sample. If there are M attributes, there will be a total of M splits to consider. For numerical attributes the splits are binary in nature and the test is of the form {is X_m <= c ?}. Commonly used technique is to choose a split that will create

the largest and purest child nodes by only looking at the instances in that node. This technique is referred to the 'greedy' or the 'local optimization' approach.

In Greedy Approach following steps are used (i) Search each attribute to find the best split for it (ii) Each of these splits is a candidate split, and there will be M candidate splits for the M attributes being considered (iii) Compare the M splits and pick the best split (iv) Some algorithms keep the second and third best splits as surrogate splits in reserve. These splits are ranked in order in how close they resemble the behavior of the primary splitting rule. These surrogate splits are used when predicting new instances that have missing values for the splitting attribute. Splitting attributes are chosen based on a goodness of split. If we define an impurity function l(t) where t is any given node, then the goodness of split is defined to be the decrease in impurity resulting from the split.

The next picture shows a candidate split s that will create child node T_1 and T_2 from T. The goodness of split will be the difference between the impurity of node T and the sum of the impurities for the child nodes of T (in this case T_1 and T_2). The goal is to find the split with the greatest reduction in impurity.



In the split shown above, the goodness of split for split s defined as:

$$\Delta I(S,T) = I(t) - P_1 I(t_1) - P_2 I(t_2) \quad (1)$$

P_1 and P_2 = Proportions of the instances of t that go into t_1 and t_2 respectively

$I(t)$ = Impurity Function

Impurity function can be defined as by using the concept of conditional probability p(j|t). If there are j classes in all, the conditional probability p(j|t) is the probability of having a class j in node t. An estimate of this conditional probability is N_j/N. Where N is the total number of instances in the node and N_j is the number of class j instances in the node. The impurity of a node is a function of this conditional probability. CART uses Gini Index for defining the impurity function, its formula is:

$$l(t) = \sum_{i \neq j} P(i|t)P(j|t) \quad (2)$$

Like the entropy function, this measure will reach a maximum when all classes are evenly distributed in the node and it will be at a minimum if all instances in the node belong to one class. There are no issues with the bias discussed previously as the CART algorithm that uses the Gini index is a binary split algorithm and does not have to deal with highly branching splits.

Step 2: Stopping Rules and Building the Final Model CART uses backward pruning algorithms. This means that they will grow a tree until it is not possible to grow it any further and thus the only stopping rule is when there are only 2 instances left in a node.

When all nodes are like this, the tree growing process will end. Pruning will be necessary to build smaller tree models that perform better on new data and not just on the training data. The idea is to remove leafs that have a high error rate. CART uses Pruning in which each node in the tree model has a certain number of instances that are misclassified, say E out of a total of N instances in the node. The training error rate (we will call f) for each node is then simply E/N.

Step 3: Assigning Classes to Tree Nodes: Everynode in a tree carries with it a particular classification. This classification is usually determined by a simple majority. In a given node, the class attached to it will be the class that is most well represented by the instances in the node. Leaf nodes are different in that their classification is final, and it is from them that a model's predictive performance is determined. Each node will have an error rate, say e, which is the proportion of misclassified instances in it. The probability that a particular classification will be correct is then simply 1-e. The probability of a correct prediction from the model is then the weighted average of these probabilities from each leaf. These estimates can be based on training data or on a separate and independent test data used to validate the model.

BOOTSTRAP AGGREGATING

Bagging (Breiman, 1996), a name derived from "bootstrap aggregation", was the first effective method of ensemble learning and is one of the simplest methods of arching[1]. The meta-algorithm, which is a special case of the model averaging, was originally designed for classification and is usually applied to decision tree models, but it can be used with any type of model for classification or regression. The method uses multiple versions of a training set by using the bootstrap, i.e. sampling with replacement. Each of these data sets is used to train a different model. The outputs of the models are combined by averaging (in case of regression) or voting (in case of classification) to create a single output. Sutton (2004) recommends to replicate as much as 25 or 50 times.

VARIABLES USED IN THE RESEARCH

Response variable used in this study is the status of children drop out of school at the age of 7-17 years. The predictor variables were used as follows: the child's sex (X1), the child's age (X2), household size (X3), mother's education level (X4), employment status of household head (X5), economic status (X6), and classification of residence (X7)

EXPERIMENTAL RESULTS

In the classification of dropout student with CART method used Gini as the goodness of split. In the process of the maximum pruning to obtain the optimal tree is used method of test sample estimate. Termination criteria selected tree formation by the number of observations in each child node of at least 5.

[1] Arching (adaptive reweighting and combining) is a generic term that refers to reusing or selecting data in order to improve classification.

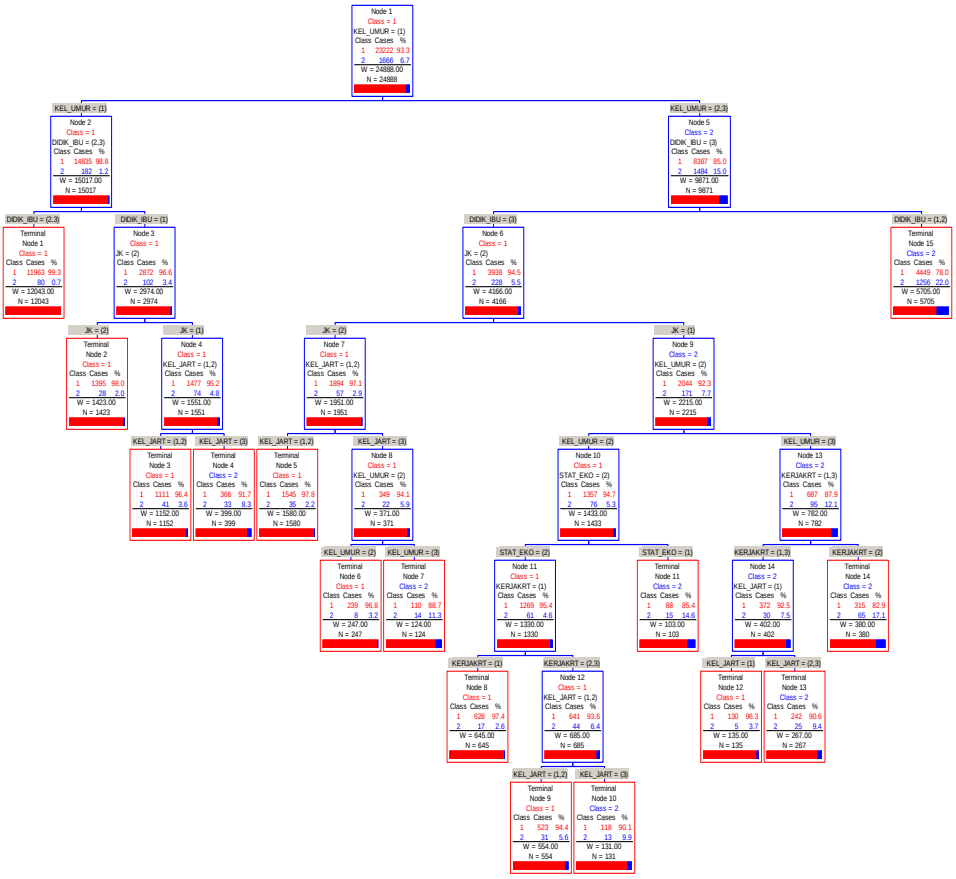


Fig. 1: Tree Details

Optimal classification tree generated from pruning process is not built by all predictor variables, only 6 variables are entered into the classifier in building an optimal tree models, i.e. the child's age, mother's education level, the child's sex, household size, employment status of household head, and economic status. Based on the optimal tree obtained that 7 groups predicted as a group of children in school and 8 groups were predicted as a group of children out of school.

Table 1
Variable Importance

Variable	Score
Child's age	100.0000
Mom's educ	63.1515
Child's sex	3.8711
Household size	1.8571
Employment Stat	1.4851
Economic stat	1.3961
classif	0.3915

Predictor variables into the main splitting the optimal tree is the mother's education level variable, it looks at the role of these variables are ranked first important variables that form the optimal classification tree. It also means that the mother's education level variable is the most dominant variable in the formation of classification models. More detailed information about the nodes of the optimal classification tree can be seen in Figure 1.

To see the accuracy of CART method, used new data from outside the data forming the model by 10 percent of data for testing. The results of testing the data yield an accuracy value of 76.59 percent.

Table 2
Prediction Success – Test

Actual Class	Total Class	Percent Correct	1 N =4514	2 N = 1735
1	5,817	76.22%	4,434	1,383
2	432	81.48%	80	352
Total:	6,249			
Average:		78.85%		
Overall % Correct:		76.59%		

APPLICATION OF BAGGING CART

Effect of bagging on CART can be seen by comparing the results of the classification accuracy without bagging and after bagging technique is applied. The results of the comparison can be seen in the following table:

Table 3
Comparison of the Classification Accuracy

Techniques	% Correct
First Tree Only (CART)	77.519%
Bagging CART with Replication 50 times	78.262%

Based on the output above shows that the application of bagging technique on CART improve the classification accuracy from 77.519 percent to 78.262 percent on bagging CART. In other words, the application of bagging CART can improve the classification accuracy of 0.743 percent.

CONCLUSION

Classification of Dropout Student in Sulawesi with CART Methods indicate that the continuity education of the child influenced by the child's age, mother's education level, the child's sex, household size, employment status of household head, and economic status.

Application of Bagging techniques produce higher classification accuracy than the CART algorithm.

REFERENCES

1. Agresti, A. (2002). *Categorical Data Analysis*. John Wiley and Sons, Canada.
2. Brieman, L. (1996). Bagging predictors. *Machine Learning.* [Cited by 25] (2.44/year).
3. Budiani S. (2014). *Determinants of Children's Dropout school (Analyzed from Susenas Data 2012)*. Thesis. Program Of Magister University of Indonesia. Depok.
4. Chen, R. and DesJardins, S. (2007). *Exploring the Effects of Financial Aid on the Gap in Student Dropout Risks by Income Level*.
5. De'ath and Fabricius (2000). Prasad et al. (2006), *CART Analysis: Details*
6. http://wekadocs.com/node/2
7. http://www.epa.gov/caddis/da_basic_4.details.html
8. Lewis dan Roger, J. (2000). *An Introduction to Classification and Regression Trees (*CART*) Analysis*, presented at the 2000 Annual meeting of society for Academic Emergency medicine of Sanfransisco, California
9. Oey-Gardiner, M. (1991). Gender Differences in Schooling in Indonesia. *Bulletin of Indonesian Economic Studies* 27(1), 57-79.
10. Soni Sneha (2010). Implementation of Multivariate Data Set By CART Algorithm *International Journal of Information Technology and Knowledge Management*, 2(2), 455-459.
11. Rogers, C.C. (2005). Rural Children at A Glance. *Economic Information Bulletin Number 1*. Washington, DC: U.S. Department of Agriculture Economic Research Service. 23 April 2013. http://www.ers.usda.gov/media/525929/eib1_2_.pdf
12. Statistics Indonesia (2013). *Profile of Indonesian Child 2013*. Jakarta.
13. Sutton, C.D. (2005). Classification and regression trees, Bagging, and Boosting, *Handbook of Statistics*, 24, 303-329.

14. Sutton, C.D. (2012). *Human Development Index 2011*. BPS: Jakarta.
15. Sumarmi (2009). *Bagging CART Approach to Clasify Characteristic of Drop Out Student in Jambi*. Thesis. Program of Magister Institut of Technology Sepuluh Nopember Surabaya.
16. Top 10 Algorithm in Data Mining According to the Survey Paper of Xindong Wu et al know (inf syst (2008) @ springer Verlog London Limited 2007, 14:1-37, 4th Dec'2007.
17. Yohannes, Y. and Webb, P. (1999). *Classification and Regression Trees, A User Manual for Identifying Indicators of Vulnerability to Famine and Chronic Food Insecurity*, Microcomputers in Policy Research, International Food Policy Research Institute, Washington, D,C, USA.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 327-336

OPTIMIZING CLASSIFICATION URBAN / RURAL AREAS IN INDONESIA WITH BAGGING METHODS IN BINARY LOGISTIC REGRESSION

Shafa Rosea Surbakti[1], Erfiani[2] and **Bagus Sartono[3]**
Statistics Department, FMIPA, Bogor Agricultural University, Indonesia
Email: [1]rosea.shafa@gmail.com
[2]erfiani_ipb@yahoo.com
[3]bagusco@gmail.com

ABSTRACT

Classification of "kelurahan" and rural area into urban/rural status basically meant to form a layer (stratum) were used in the survey sampling techniques. With the status of urban and rural areas, the sample can represent the entire population correctly. Logistic regression is one method of non-parametric regression where the response variable is categorical data. Binary logistic regression was used when the response variable consists of two categories. This method can also be used for data classification. The accuracy of the classification of binary logistic regression can be improved by using the bootstrap aggregating (bagging). Considered bagging method can improve the predictive power of some estimators or specific algorithms such as regression or classification trees. The purpose of this study was to see how much improvement the accuracy of the classification of urban / rural using Binary Logistic Regression with or without bagging process. The data used in this case is data Potensi Desa (PODES) 2011. Previous studies showed that the variable density of population, the number of farm households, and the presence of tertiary facilities provide real influence as a distinctive urban/rural classifier. Based on the same research in different cases logistic regression classification accuracy is 82.52%. Bagging the logistic regression method is expected to increase the accuracy of classification at 2:57% compared to the classical binary logistic regression.

1. INTRODUCTION

The Development of Indonesia is essentially aims to improve people's lives, both material and spiritual enhancement. To create targeted development required careful planning. One aspect of support in development planning is the availability of detailed data on the level of the smallest region. Information to the smallest area can be used as a guide in making policy more conical.

Administrative territorial division according to the Ministry of Interior of the Republic of Indonesia consists of the Neighborhood (RT), Rukun Warga (RW), Village / Village, District, City / County, Province until the government at the national level. Zoning is intended for the management of local government within the boundaries of each region according to the principle of autonomy, the deconcentration, decentralization, and assistance. Central Board of Statistics (Badan Pusat Statistik) seeks to provide information to the level of the smallest region in this case at the level of the urban / rural.

Due to the use of the term "kelurahan"/rural more accurately geared to the interests of the administration, the Central Board of Statistics perform classification "kelurahan"/rural into urban or rural status. Classification of area "kelurahan"/ rural into urban/rural status basically meant to form a layer (stratum) which were used in the survey sampling techniques. With the status of urban and rural areas, the expected sample can be drawn to represent the entire population well. In the analysis, classification of urban / rural areas will give better results depict the actual situation compared with the classification of the urban/ rural (Wynandin, 1986).

Due to the differences in each country to determine the distinguishing characteristics of urban and rural, the variables distinguishing between urban and rural areas cannot be summarized into a single definition for all countries (United Nations, 2014). In the classification of urban/rural, regional characteristics are often used as the main reference. Striking difference from the urban / rural population can be seen from the density, the fulfillment of the local economy, the existence of facilities and so forth. Results of development helped change the criteria that was used in the classification of urban / rural, so it needs to be a review of what criteria can now be used as a differentiator between urban and rural areas. Science which continues to grow also enrich the urban area classification method / rural. The method is considered more appropriate one in classifying urban / rural than other methods.

Many previous studies have done the classification of the area. In 2005 Siti Wahyuningrum been using MARS approach to classification of rural / poor village in East Kalimantan. Furthermore Miftahudin Arif (2008) have used Neural Network in determining the rating analysis villages. In 2008, Masykuri grouping Regency / City by Numbers HDI by using a weighted metric scaling multiple dimensions. Developments in the science of data mining also provide a substantial contribution in the data classification method. Considered bagging method can improve the predictive power of some estimators or specific algorithms such as regression or classification trees. Galih Widhi (2013) has conducted research using logistic regression bagging to increase classification accuracy on logistic regression for Classification Level Household Welfare Farmers chili. Based on these studies, the development of a binary logistic regression bagging method would be applied to the classification of urban / rural and accuracy of classification can be compared to the classical logistic regression to determine the most optimal method.

The purpose of this study is:

1. To evaluate the variables used in the real distinction between urban and rural areas.
2. To apply bagging on binary logistic regression method and see how much improvement the accuracy of the classification of urban / rural.
3. To compare the optimization of the classification of urban / rural areas in Indonesia with binary logistic regression bagging method with classical Binary logistic regression.

The benefits of this research is enrich the knowledge of statistics with the adoption of bagging on binary logistic regression to optimize the classification of urban / rural areas in Indonesia. In the future, this research is expected to be used as a pilot to develop this method on other issues.

2. LITERATURE REVIEW

The simplest difference between urban and rural areas in a country is usually based on the assumption that urban society has a way of life and usually have a standard of living that is different than the rural population. In countries that are developing industry sector, these differences become less visible again and appear more striking differences such as the population density in the region. Some countries felt the need to add additional criteria which are believed to be able to distinguish urban and rural areas such as the percentage of the population employed in agriculture, or the availability of electricity and running water availability, and ease of access to health facilities, schools, and recreation areas. Even in industrialized countries also added distinguishing criteria such as agricultural areas, trade centers, industrial centers, community centers and other services are considered able to distinguish between urban and rural areas (UN, 2008).

United Nations also stated that no recommendation that can be used to explain the meaning of the urban or rural areas clearly. This is because each country has a different view on the urban / rural. So that each country must make their own definitions according to the needs of the country and due to this reason also each state should decide which region in the category of urban and rural categorized (UN, 2014).

Urban Area

United Kingdom Government classify an area as urban if the population who live in the area proficiency level has more than 10,000 people. Census of India (2011) defines an urban area if the area has a city government and meet the requirements, among others, has a resident population living in the region of at least 5,000 people, at least 75% of male labor force working in non-agricultural sectors, and has a density population of at least 400 people per km2.

In Indonesia, the urban definition according to Undang-undang No. 22/1999 on Regional Autonomy, urban areas are areas with major non-farm activities with the composition as a function of the area of government services, social services and economic activity. Urban is the status of a village-level administrative area / villages that meet the classification criteria of urban areas.

Rural Area

According to World Bank (2008) rural areas can be defined by the number of settlements, population density, distance to the metropolitan area, administrative segregation and the role of the agricultural sector. The Organization for Economic Co-operation and Development uses a population density of 150 people per km^2 for defining rural areas.

According to Paul H. Landis (1948) is a village of less than 2,500 inhabitants. With characteristic feature as follows:

a) Have known each other socially familiar life among thousands of lives.
b) There is a linkage similar feelings of affection towards habits
c) How to trying (economic) is the most common agrarian highly influenced by nature such as: climate, natural conditions, natural resources, while non-agricultural employment is to be odd.

Rural areas, according to Undang-undang No. 26 of 2007 on the National Spatial Plan is a region that has a major agricultural activity, including natural resource management with the composition as a function of the area of rural settlements, government services, social services and economic activity. In Kamus Besar Bahasa Indonesia (2005), the village is a unit area inhabited by a panel of families who have a system of self-government (headed by a village head) or a group of houses in the villages outside the city which is unity.

From the overall definition can be concluded that the two main criteria commonly used to distinguish urban and rural, among others: population density per km^2 and livelihoods of the majority of the population in agriculture or non-agriculture sectors.

Binary Logistic Regression

Logistic regression is one method of non-parametric regression where the response variable is categorical data. Binary logistic regression was used if the response variable consists of two categories (Agresti, 2002). In Binary Logistic Regression method can be used for data classification. Response variable Y = 1 states that the incidence of "successful" (in the category), while for Y = 0 indicates that the incidence of "fail" (not in the category). Variables Y will follow Binomial distribution.

The general form of logistic regression models odds with k independent variables is formulated as follows:

$$\pi(x_i) = \frac{\exp(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)}{1+\exp(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)} \quad (1)$$

If the models in the above equation is transformed by using the logit transformation of π(x), then the logistic model can be written as equation (Azen, 2011):

$$g(x) = ln\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \quad (2)$$

Parameter estimation in logistic regression using Maximum Likelihood Estimation. These methods estimate the parameter β by maximizing the likelihood function (Hosmer and Lemeshow, 2000). Predicted values of parameters are then tested to determine which independent variables that significantly affect the model. Testing the independent variables can be performed simultaneously or partially.

The hypothesis in partial parameter testing is:

$H_0 : \beta_j = 0$
$H_1 : \beta_j \neq 0, j = 1, 2, \ldots, k$

Partial parameter testing using Wald's Test is (Azen, 2011):

$$X^2 = \left(\frac{\widehat{\beta-\beta}}{S_{\widehat{\beta}}}\right)^2 \sim \chi^2 \qquad \text{dengan df } = 1 \quad (3)$$

Parameter testing can be carried out simultaneously with the likelihood ratio test statistic (Likelihood Ratio Test) G by the following equation (Azen, 2011):

$$G^2 = -2\, ln\left(\frac{L_0}{L_k}\right) \sim \chi^2_{(k)} \quad (4)$$

The easiest way to measure the goodness of fit of the model is to use the test of Hosmer and Lemeshow fitness model (Azen, 2011):

$$G_{HL}^2 = \sum_{j=1}^{k} \frac{(Observed_i - Expected_i)^2}{Expected_i(1-{Expected_i}/{n_j})'} \quad (5)$$

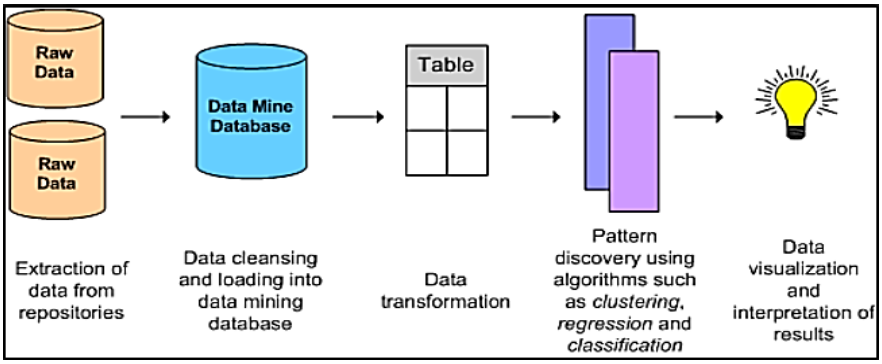
One method to measure the ability of a method of classification in predicting new data group used the opportunity of correct classification called Correct Classification Rate.

$$CCR = \frac{n_{11}+n_{22}}{n_{11}+n_{12}+n_{21}+n_{22}}$$

where $n_{11}\ dan\ n_{22}$ number of case classified correctly.

Data Mining and Ensemble Methods

Data mining is a process of computing by analyzing the pattern (pattern recognition) on a large data base which aims to solve the problem. One main goal of data mining is to build a good model of a set of data (Zhou, 2012). Models are useful form of data mining for classification and prediction. To use the method of model building ensemble (ensemble methods), several algorithms (learner) are tested to resolve the problem / the same case. Ensemble method is used to improve the accuracy in classification. Some commonly used methods of ensemble is boosting and bagging. While algorithms often used for modeling, among others, are classification trees, neural networks, naïve Bayes and etc.



Picture 1. Data Processing Scheme in Data Mining

Bagging (Bootstrap Aggregating)

Bagging (Bootstrap aggregating) is a bootstrap sampling technique to generate the data set used for the application of different methods of analysis. Suppose a data set consisting of *n* data. Examples of the measurement data is to be raised invitation *n* sampling with replacement. Some of the original value of the data *n* will appear repeatedly, or even do not appear at all. Suppose that the process is done as much as *B* times, it will get a sample size *B* of the data size *n*. Bagging two main stages in which the bootstrap which is a sampling of the data sample owned (resampling) and aggregating that incorporates many of the estimated value into a value allegation by voting.

Bagging Regresi Logistik Biner

Bagging logistic regression is a combination of data mining techniques with binary logistic regression statistical methods as a learner. The algorithm of bagging a binary logistic regression is as follows:

1. In the training data, taken bootstrap samples of n observations from the data set by means of recovery n times. Further replication bootstrap samples as B times.
2. Make a binary logistic regression model of each bootstrap sample sets, so that will form B binary logistic regression model.
3. In the testing of data, each replication calculated response opportunities or π (x) of each observation and counting accuracy of classification and a classification error or misclassification.
4. Do the voting classification produced in the first replication to replication to B.
5. Compare the results with the state that it actually voting. If Bagging proved better, it will produce high Correct Classification Rate.

Research bagging logistic regression in the case of the classification of household welfare in the city of Malang with replication bagging as many as 50 to 80 times that produce a higher level of accuracy than the classical binary logistic regression during replication was 60 (Ningrum, 2012). While research bagging logistic regression in the case of the classification accuracy rate of household welfare chili farmers to replicate as much as between 80 hingga 140 bagging produce a higher level of accuracy than the classical binary logistic regression during replication 140 (Kurniawan, 2013).

3. METHODS

The data used in this study is the result of data collection POTENSI DESA (PODES) in 2011. The response variables in this study is the urban classification (1) and rural (0). Independent variables or predictor variables in this study are:

Tabel 1
Peubah Penyusun Model

Peubah	Nama Peubah	Tipe
(1)	(2)	(3)
X1	Kepadatan Penduduk	Numerik
X2	Jumlah Rumah Tangga Pertanian	Numerik
X3	Keberadaan Taman Kanak-kanak	Kategorik
X4	Skor Keberadaan SMP	Kategorik
X5	Skor Keberadaan SMU	Kategorik
X6	Skor Keberadaan Pasar	Kategorik
X7	Skor Keberadaan Bioskop	Kategorik
X8	Skor Keberadaan Pertokoan	Kategorik
X9	Skor Keberadaan Rumah Sakit	Kategorik
X10	Skor Keberadaan Hotel	Kategorik
X11	Skor Keberadaan Telepon	Kategorik
X12	Skor Keberadaan Listrik	Kategorik

3. RESULT

Classic Binary Logistic Regression

From the partial parameters test results obtained that all predictor variables significant in distinguishing urban and rural areas. Parameter values and Wald's test statistic is shown in the following table:

Table 2
Parameter value of β, Wald's Test dan P-values Testing Parameter Partial

Variable	β	Wald's Test	p-value
(1)	(2)	(3)	(4)
X1	0.00329	2310.76957	0.00000
X2	-0.20483	3013.11032	0.00000
X3	1.95465	145.14995	0.00000
X4	2.41371	488.90575	0.00000
X5	2.95768	1199.36309	0.00000
X6	2.91611	993.24938	0.00000
X7	3.62472	114.80499	0.00000
X8	3.69364	1389.93460	0.00000
X9	3.49603	1349.45859	0.00000
X10	3.04029	161.56252	0.00000
X11	3.19981	367.49532	0.00000
X12	3.24312	980.45309	0.00000
Konstanta	-6.49225	897.23002	0.00000

Because the p-values of all independent variables significant at α = 0.05 then the whole variables included in the model. From table 2 forming a binary logistic regression model as follows:

$$\pi(x) = \frac{\exp(-6{,}492+0{,}003x_1-0{,}205x_2+1{,}955x_3(0)+2{,}414x_4(0)+\cdots+3{,}243x_{12}(0))}{1+\exp(-6{,}492+0{,}003x_1-0{,}205x_2+1{,}955x_3(0)+2{,}414x_4(0)+\cdots+3{,}243x_{12}(0))} \quad (6)$$

The accuracy of the classification of the classic binary linear regression model is equal to 97.95% with a classification error of 2.05%.

Bagging Bagging Regresi Logistik Biner

Bagging Binary Logistic Regression method considered can be used to increase or improve classification accuracy produced by classical binary logistic regression. With bootstrap replication as much as 10, 20, 30, 40, and 50, the accuracy of classification and the resulting classification error is as follows:

Table 3
Bagging Logistic Regression Classification Accuracy Comparison with the Logistic Regression Classic

Number of Bootstrap Replication	Classification Accuracy (CCR)(%)	Misclassify (ECR)(%)	Classification Accuracy (CCR) Logistic Regression Classic (%)
(1)	(2)	(3)	(4)
10	97,961	2,039	97,95
20	97,903	2,097	97,95
30	97,929	2,071	97,95
40	97,935	2,065	97,95
50	97,941	2,059	97,95

From Table 3 shows that the results of voting on 10 replication have classification accuracy (CCR), which is higher than the classic binary logistic regression. However, the results of voting 20 first replication decreased classification accuracy than classical binary logistic regression. At 30 to 50 times the value of the bootstrap replication accuracy of classification slowly began to rise, but still below the classical binary logistic regression. So it can be concluded that the classification accuracy on a small bootstrap replication quite unstable.

4. KESIMPULAN

The conclusion can be obtained from this study are:

1. All variables are used as a real differentiator between urban and rural areas significantly affect the model. This means that all the variables capable of distinguishing urban and rural areas well.
2. Application of binary logistic regression bagging with replication between 10 to 50 CCR value unstable. So from this study and from previous studies it can be concluded that the binary logistic regression bagging will produce a stable CCR values and can be a better classification methods if replication is performed over 100 times.
3. Optimizing the classification of urban / rural areas in Indonesia with binary logistic regression bagging method can only be carried out and produce better results than the classic binary logistic regression in replication in the top 100.

REFERENCES

1. Agresti, A. (2002). *Categorical Data Analysis*. John Wiley & Sons, Inc, New York.
2. Azen, R. and Walker, C.M. (2011). *Categorical Data Analysis for Behavioral and Social Science*. Routledge, New York.
3. Badan Pusat Statistik (2010). Peraturan Kepala Badan Pusat Statistik No. 37 tahun 2010 Tentang Klasifikasi Perkotaan dan Perdesaan di Indonesia. BPS, Jakarta.
4. Breiman, L. (1994). *Bagging Predictor*. *Technical Report*. Department of Statistics University of California, California.
5. Kurniawan, Galih Widhi, dkk. (2013). Metode Bagging Regresi Logistik Untuk Peningkatan Ketepatan Klasifikasi Pada Regresi Logistik (Studi Kasus: Klasifikasi

Tingkat Kesejahteraan Rumah Tangga Petani Cabai Kabupaten Blitar Dan Kabupaten Kediri. *Jurnal Mahasiswa Statistik* 1(2). Universitas Brawijaya, Malang.
6. Landis, Paul H. (1948). *Pengantar Sosiologi Pedesaan dan Pertanian*. PT. Gramedia Pustaka Utama, Jakarta.
7. Lapoliwa, Hans, dkk. (2005). Kamus Besar Bahasa Indonesia. Balai Pustaka, Jakarta.
8. Miftahudin, Arif. (2008). Analisis Rating Menggunakan Metode Klasik dan Jaringan Syaraf Tiruan Studi Kasus Klasifikasi Desa/Kelurahan di Kabupaten Enrekang. Thesis. Institut Teknologi Sepuluh Nopember, Surabaya.
9. Ningrum, E.S. (2012). Klasifikasi Kesejahteraan Rumah Tangga Di Kota Malang dengan Pendekatan Bagging Regresi Logistik. Skripsi. Institut Teknologi Sepuluh November, Surabaya.
10. Department of Economic and Social Affairs Statistics Division. (2008). *Principles and Recommendations for Population and Housing Censuses*. United Nation, New York.
11. Department of Economic and Social Affairs Statistics Division (2014). *Principles and Recommendations for a Vital Statistics System*. United Nation, New York.
12. World Bank (2008). *World Development Report 2008*; Agriculture for Development. World Bank, Washington DC.
13. Zhou, Zhi-Hua. (2012). Ensemble Methods Foundation and Algorithms. CRC Press, Florida.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 337-342

THE USE OF LAGRANGE MULTIPLIER TO ENSEMBLE TWO RETURN VALUES OF GENERALIZED PARETO AND MODIFIED CHAMPERNOWNE DISTRIBUTIONS

Aji Hamim Wigena[1], Cici Suhaeni[2] and Deby Vertisa[3]
Department of Statistics, Bogor Agricultural University, Jalan Meranti,
Kampus IPB Darmaga, Bogor 16680, Indonesia
Email: [1] ajiwigena@ymail.com
[2] chi2shaeny@yahoo.com
[3] deby.vertisa@gmail.com

ABSTRACT

Return values, considered as extreme values, can be estimated based on generalized pareto and modified champernowne distributions. The first distribution tends to give an over estimate while the second gives an under estimate. It is necessary to ensemble both estimates using optimum weights to have more accurate return value. These optimum weights are determined using lagrange multiplier. The results are compared to those determined using trial-and-error method and simple linear regression method. The three methods are applied to the daily rainfall data from January 1985 to March 2009 in Darmaga Bogor and give the same optimum weights. However, the trial-and-error method needs an iterative computation process while the other methods (simple linear regression and lagrange multiplier) are more exact and do not need any iterative process.

KEYWORDS

Ensemble, extreme, lagrange multiplier, modified Champernowne, Generalized Pareto.

1. INTRODUCTION

Extreme value estimation can be similar to the estimation of return values based on extreme value distribution with heavy tail characteristic. Two distributions which have that kind of characteristic and can be used to estimate the return values are generalized Pareto distribution (GPD) and modified Champernowne distribution (MCD). The first distribution is one of two distributions in extreme value theory [2]. The second distribution was introduced by [1]. Both distributions need a threshold in order to estimate the extreme values.

Both distributions show the behavior of the right tail of distributions [2] and [3]. MCD is a heavy tailed distribution that converges to GPD. The extreme value estimation with GPD results in an over estimate while MCD results in an under estimate. These estimates are different in terms of positive and negative deviation from the true value. In order to get more accurate estimate closed to the true value, both estimates are combined using an optimum weighted ensemble method.

The trial-and-error method has been used to determine the optimum weights to ensemble the extreme estimates based on MCD and GPD [8]. This method uses an iterative computation process based on RMSE (Root Mean Square Error). The optimum weights are determined iteratively until the minimum value of RMSE is reached. The optimum weights can also be determined using simple linear regression without intercept [9]. This method is also based on RMSE but without iterative computation process. The estimation uses least square method. One of the weights is the slope of the regression model.

Langrange multiplier can be an alternative method to find the optimum weights [7]. The RMSE can be the objective function and the values of weights which are greater than or equal zero and less than or equal one are the constraints. This paper discusses the use of lagrange multiplier to determine optimum weights to ensemble the return values based on MCD and GPD. The weights are compared to that using trial-and-error and linear regression methods.

2. DATA AND METHODS

2.1. Data

This study used the same data as in [8] and [9], which are daily rainfall data from 1 January 1985 to 31 March 2009 in Darmaga Bogor Station from the Meteorological, Climatological and Geophysical Agency of Indonesia (BMKG). Data are grouped based on different period of time for modeling and validation. Data for modeling are from 1 January 1985 to 31 December 2008 and data for validation are from 1 January 2009 to 31 March 2009.

2.2. Methods

The first step in the study is to determine a threshold (u). Chavez-Demoulin [4] suggested the value of 90th quantile or 10% of the highest value as the threshold and data more than the threshold are categorized as extreme values. This threshold was used to estimate the parameters of GPD and MCD. GPD with two parameters σ and ξ is defined as:

$$H(y) = \begin{cases} 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-\frac{1}{\xi}}, \xi \neq 0 \\ 1 - \exp\left(-\frac{y}{\sigma}\right), \xi = 0 \end{cases} \quad (1)$$

where $y = x - u, x > u$, $\sigma > 0$ and $0 < \xi < 0$. The scale (σ) describes the variance and the shape (ξ) describes the behavior of right tail.

MCD with three parameters (α, M, and c) is defined as:

$$T(x) = \frac{(x+c)^{\alpha} - c^{\alpha}}{(x+c)^{\alpha} + (M+c)^{\alpha} - 2c^{\alpha}} \quad (2)$$

where $x > u, \alpha > 0, M > 0$ and $c \geq 0$. The location (M) describes the center of data, the scale (α) describes the variance and the shape (c) describes the behavior of right tail.

The extreme values are estimated based on GPD $(\hat{x}_{GPD})$ and MCD $(\hat{x}_{MCD})$. The estimate of extreme value of GPD is the following return level:

$$\hat{x}_{GPD} = u + \frac{\sigma}{\xi}((m \cdot \delta_u)^{\xi} - 1) \quad (3)$$

and the estimate of extreme value of MCD is:

$$\hat{x}_{MCD} = [m \cdot \delta_u((M + c)^{\alpha} - c^{\alpha}) - (M + c)^{\alpha} + 2c^{\alpha}]^{\frac{1}{\alpha}} - c \quad (4)$$

where m = the number of days ahead, $\delta_u \widehat{=} k/N$, k = the number of extreme data above the threshold, and N= the number of all data [2].

Both estimates are ensembled using optimum weight (w_{opt}). The ensembled estimate is $\hat{x}_{Ens}$ as in Eq. (5).

$$\hat{x}_{Ens} = w_{opt} \cdot \hat{x}_{MCD} + (1 - w_{opt}) \cdot \hat{x}_{GPD} \quad (5)$$

Finding the optimum weight iss based on RMSE between the observation data, x_i, and the ensembled estimate, $\hat{x}_{Ens}$. The weight is optimum when RMSE is minimum.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{t}(x_i - \hat{x}_i)^2}{t}} \quad (6)$$

The three extreme value estimates, $\hat{x}_{GPD}$, $\hat{x}_{MCD}$, and $\hat{x}_{Ens}$, were compared based on RMSE (Eq. 6) between the observation data, x_i, and those estimates.

One method to find the optimum weights is based on trial-and-error using an iterative procedure [7, 8]. The procedure is the following:

1. Determine a weight (w_{opt}) with constraints that $0 \leq w_{opt} \leq 1$.
2. For each weight, compute $\hat{\boldsymbol{y}}_{ens}$ and RMSE.
3. Continue step 1 and 2 until the optimum weight when RMSE is minimum.

The value of w_{opt} is started from 0 to 1. An increment can be any number between 0 and 1. If the increment of w_{opt} is 0.001 and accordingly the value of $(1 - w_{opt})$ will be from 1 to 0 with increment -0.001. So, the pairs of weights, $(w_{opt}; 1 - w_{opt})$ are (0;1), (0.001;0.999), … , (1;0). For each pair, RMSE is computed. The pair of weights is optimum if RMSE is minimum.

The other method is the use of simple linear regression without intercept [9]. The regression coefficient, slope, can be used to estimate one (w_{opt}) of the two weights. Based on Eq.(5) we can derive to the following:

$$(\hat{x}_{Ens} - \hat{x}_{GPD}) = w_{opt}(\hat{x}_{MCD} - \hat{x}_{GPD}) \quad (7)$$

The Eq.(7) is the form of simple linear regression without intercept, where the weight, w_{opt}, is the estimate of regression coefficient, slope. So, the optimum weight, w_{opt}, is found.

Lagrange multiplier method minimizes the objective function with some constraints [6]. The objective function in this study is [7]:

$$\sum_{i=1}^{t}\left(x_i - (w_{opt} \cdot \hat{x}_{MCD} + (1 - w_{opt}) \cdot \hat{x}_{GPD})\right)^2 \quad (8)$$

and the constraint is:

$$0 \leq w_{opt} \leq 1 \tag{9}$$

The optimum weight is:

$$w_{opt} = \frac{\left[-\sum_{i=1}^{t}(\hat{x}_{MCD_i}\hat{x}_{GPD_i})+\sum_{i=1}^{t}\hat{x}_{GPD_i}^2\right]}{\left[\sum_{i=1}^{t}(\hat{x}_{MCD_i}-\hat{x}_{GPD_i})^2\right]}. \tag{10}$$

3. RESULTS

3.1. Thresholds

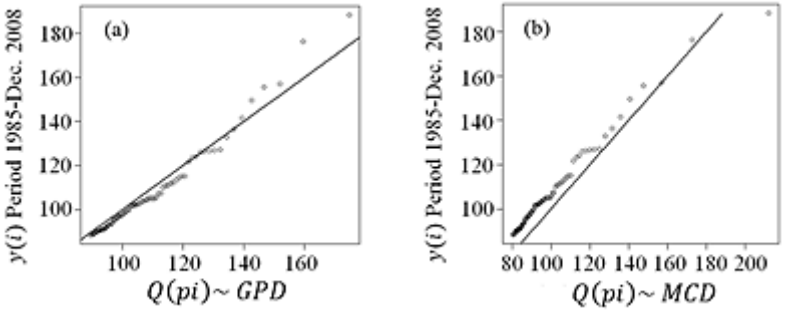
The threshold was determined based on the 90th quantile. According to data for modeling from 1 Jan. 1985 to 31 Dec. 2008, the threshold was 36.0 mm. The same threshold was used to estimate the parameters of GPD and MCD.

3.2. Parameter Estimates of GPD and MCD

The parameter estimates of GPD and MCD for data from 1 January 1985 to 31 December 2008, the estimate of scale (σ) and shape parameters (ξ) of GPD, are 25.22 and -0.0862 respectively. The estimate of ξ is a negative value which indicates the density function has a light tail so the high extreme value is most likely not going to happen [2]. The location parameter estimates of MCD is 53.00. The scale parameter estimates of MCD is 5.51 and the shape parameter estimates of MCD is 0.0000284.

3.3. Ensemble of Extreme Estimates

Based on the QQ plot in Figure 1 shows that GPD and MCD were fit to the data and able to estimate the extreme values well. However, these are different in whether the estimation is over or under estimate. The QQ plot in Figure 1 (a) shows that most of the estimates are below the line which are more than the theoretical values or over estimate. A few estimates are above the line which indicate these estimates are less than the theoretical values or under estimate. Therefore, the extreme rainfall estimates based on GPD tends to be over estimate, while those based on MCD tends to be under estimate shown in Figure 1 (b). Both estimates were ensembled to find the new estimate.



(a) (b)

Figure 1: QQ-plot of estimates based on GPD (a) and MCD (b) on 90th quantile for period from 1 January 1985 to 31 December 2008

The two extreme estimates, based on GPD and MCD, were combined to give an estimate ($\hat{x}_{Ens}$). The experiment using trial-and-error method was implemented to find

the optimum weight according to the smallest RMSE in order to get better estimate. This process gives the optimum weight $w_{opt} = 0.243$ at the minimum RMSE = 2.638. Figure 2 shows that most of the estimates are in the line which are very closed to the theoritical values even though a few estimates are above the line which are under estimate.

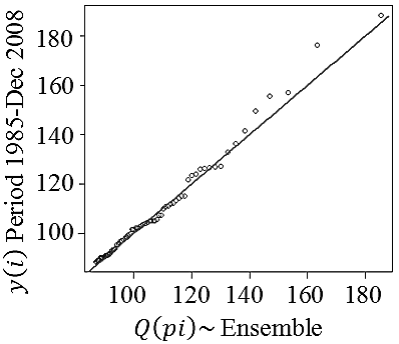


Figure 2: QQ-plot of ensembled estimate with w_{opt} =0.243 on 90th quantile for period from 1 January 1985 to 31 December 2008

Compared to the extreme estimates of GPD and MCD in Figure 1, the estimate of the weighted ensemble is more accurate. Therefore, the case of over estimate in GPD and under estimate in MCD can be solved by weighted ensemble method. The optimum weight obtained from period of data can be used to forecast extreme value based on the ensemble values [5].

The optimum weight estimated using simple linear regression without intercept in Eq.(7) and using lagrange multiplier in Eq.(10) are 0.243. This weight is the same as the weight resulted from the trial-and-error method. The three methods can be used to ensemble the return values estimated based on GPD and MCD.

4. COMMENTS AND CONCLUSION

The trial-and-error method needs iteratively computation process because in this method there is no closed form of the optimum weight. Different from the trial-and-error method, in the other two methods, simple linear regression and lagrange multiplier, do not need any iterative process. The closed forms of optimum weight based on both methods can be derived.

The extreme rainfall was predicted well based on MCD that converges to GPD, even though the estimate using MCD tends to be under estimate while using GPD tends to be over estimate. The ensemble of these two estimates needs an optimum weight. The optimum weight can be estimated using lagrange multiplier method as well as the trial-and-error and simple linear regression methods. The three methods result in the same optimum weight.

REFERENCES

1. Buch-Larsen, T., Nielsen, J.P., Guillen, M. and Bolance, C. (2005). Kernel density estimation for heavy-tailed distribution using the Champernowne transformation. *Statistics*, 39, 503-518.
2. Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*, London: Springer.
3. Gilli, M. and Kellezi, E. (2003). An Application of Extreme Value Theory for Measuring Financial Risk. *Computational Economics*, 27, 1-23.
4. Chavez, D.V. and Sardy, S. (2006). A bayesian non parametric peaks over threshold method to estimate risk measures of nonstationary financial time series. [*online* journal]. http://www.math.ethz.ch/riskometer/ npot.pdf. [29 April 2011].
5. Hafid, M. (2013) Pendugaan Nilai Ekstrim Menggunakan Sebaran Champernowne Termodifikasi, Sebaran Pareto Terampat, dan Nilai Gabungan (Studi Kasus Curah Hujan Harian Darmaga Bogor). [skripsi]. Institut Pertanian Bogor.
6. Hillier, F.S. and Lieberman, G.J. (2000). *Introduction to Operations Research.* New York (US): McGraw-Hill.
7. Vertisa, D. (2013). Penentuan Bobot Optimum dengan Pengganda Lagrange untuk Penggabungan Nilai Dugaan Ekstrim Curah Hujan, [skripsi]. Institut Pertanian Bogor.
8. Wigena, A.H., Djuraidah, A. and Hafid, M. (2013). Ensemble of Extreme Estimates Based on Modified Champernowne and Generalized Pareto Distributions. *Proceeding International Seminar on Sciences 2013*, Bogor Agricultural University, 15-17th 2013.
9. Wigena, A.H., Djuraidah, A. and Mangku, I.W. (2014). Ensemble Two Return Levels of Generalized Pareto and Modified Champernowne Distributions using Linear Regression. *Advances and Applications in Statistics*, 40(2), 157-167.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 343-352

DEMOGRAPHIC TRANSITION IN BANGLADESH: EVIDENCE FROM POPULATION AND HOUSING CENSUS 1981-2011

Md. Mashud Alam, Md. Shamsul Alam and **Amjad Hossain**
Population and Housing Census-2011, Bangladesh Bureau of Statistics
Dhaka, Bangladesh.

ABSTRACT

Bangladesh is now in demographic transition. The population growth is reducing while the working age population is increasing. This is a great widow of opportunity for the economic development of the country.

According to the Population Census 1981, the population in the working age 15-59 was 41.6 million which increased to 52.6 million in 1991, 68.1 million in 2001 and 83.4 million in 2011. Thus we see that the volume of working age population is increasing tremendously. If these potential labour force can be used in gainful economic activities then the volume of Gross Domestic Product (GDP) will increase at a very higher rate which will in turn help in overall economic development of the country. The main input for the growing labour force should be education and skill. If the growing labour force, particularly the youths, are provided with proper education and skill then they can contribute towards a tangible output for the country. It may be mentioned that many countries of the world particularly China and India have attained economic growth through this demographic bonuses and Bangladesh is now at the door step of the demographic dividend.

The present paper will highlight the growth of population & labour force in the country during the last decade and shed light on the widow of opportunity for Bangladesh.

1. INTRODUCTION

Demographic transition refers to transition from high birth and death rates as the country develops from a pre-industrial economic system to an industrial economic system. Bangladesh economy is now shifting from agriculture to no-agricultural direction. The contribution of agriculture in our national GDP is diminishing while the contribution of industry and service are increasing. The contribution of agriculture in our GDP was more than 20.0% in the beginning of the current millennium which reduced to 16.0% in 2013-14. On contrary, the contribution of industry sector has increased to around 30.0% from 20.0% in the same period. This economic transition resulted in the decline in fertility and mortality situation of the country. As a result major changes has been occurred in the in the age structure of population. This change has created scope for more working age population which is a great window of opportunity for economic development of the country.

2. CHANGE IN POPULATION SIZE AND GROWTH

The change in population size and growth since independence has been presented in Table-1. It is observed that in 1974 the total population of the country was 76.4 million which increased to 149.8 million in 2011. Thus the population increased almost double in the last 37 year. But in the recent year due to many positive steps by the government the growth of population decreased. On the other hand, the death rate also decreased as many deadly diseases have been controlled by global initiatives. The annual population growth of the country reduced almost 50% during the period.

Table 1
Population Size and Growth 1974-2011

Census ear	Total Population (Million)	Growth Rate
1974	76.4	2.48
1981	89.9	2.35
1991	111.5	2.17
2001	130.5	1.59
2011	149.8	1.37

Source: Population Census Report in different year

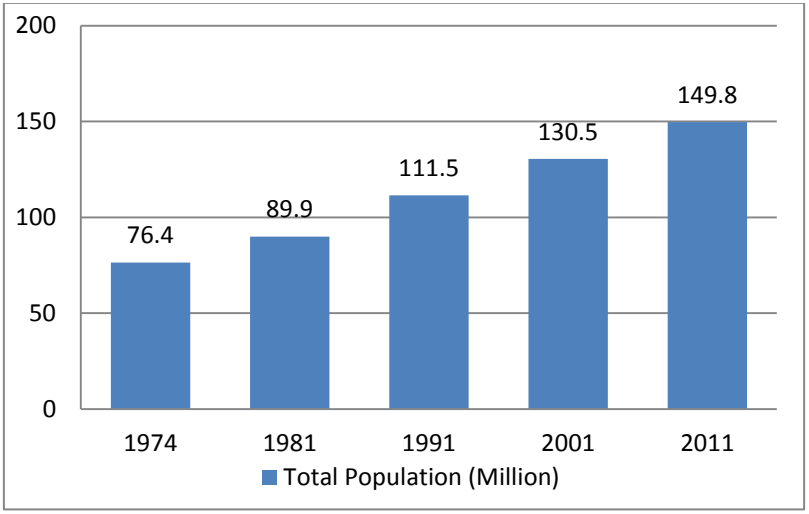


Figure 1: Population Size 1974-2011

3. FERTILITY DECLINE

Bangladesh has experienced a high growth of population in early eighties when the growth of population was 2.48% which reduced to 1.59% in 2001 and further reduced to 1.37% in 2011. Due to the success of family planning the level of total fertility has declined rapidly. The declining trend in forced in fertility can be seen from the table-2.

Table 2
Decline in total Fertility Rate in the last three Decades, 181 to 2011

Year	National	Rural	Urban
1981	5.0	5.3	3.2
1985	4.8	5.1	3.1
1991	4.2	4.5	2.9
2001	2.6	2.8	1.7
2005	2.5	2.7	1.9
2011	2.1	2.2	1.7

Source: SVRS, BBS

From the table we see that TFR declined 5.0 per women in 1981 to 2.1 per women in 2011. The decline in fertility has positive impact in the demographic transition which resulted the growth of youth and working age population.

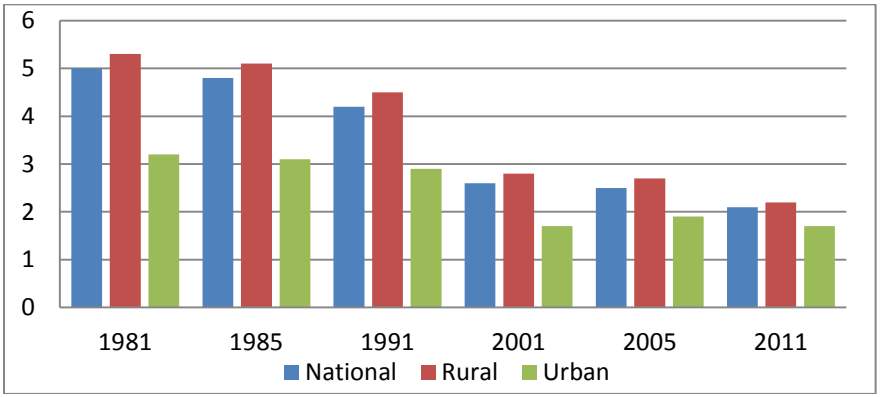


Figure 2: Fertility Rate in the last There Decades, 1981 to 2011

4. DECLINE IN MORTALITY

Following the decline in fertility, the mortality also declined due to control of communicable diseases and improvement in the water and sanitation system. As a result the life expectancy of the people also increased. The trends in mortality over the last three decades 1981 to 2011 and the expectation of life have been presented in Table-3.

Table 3
Crude Death Rate by Residence and the Life Expectancy at Birth 1981-2011

Year	Crude Death Rate			Life Expectancy at Birth		
	National	Urban	Rural	National	Urban	Rural
1981	11.5	1.2	7.2	54.8	54.3	60.3
1985	12.0	1.9	8.3	55.1	54.7	60.1
1991	11.2	11.5	7.8	56.1	55.8	60.2
2001	4.8	5.2	4.3	64.2	63.2	66.4
2005	5.8	6.1	4.9	65.2	64.5	67.9
2011	5.5	5.8	4.8	69.0	68.6	69.9

Source SVRS, BBS

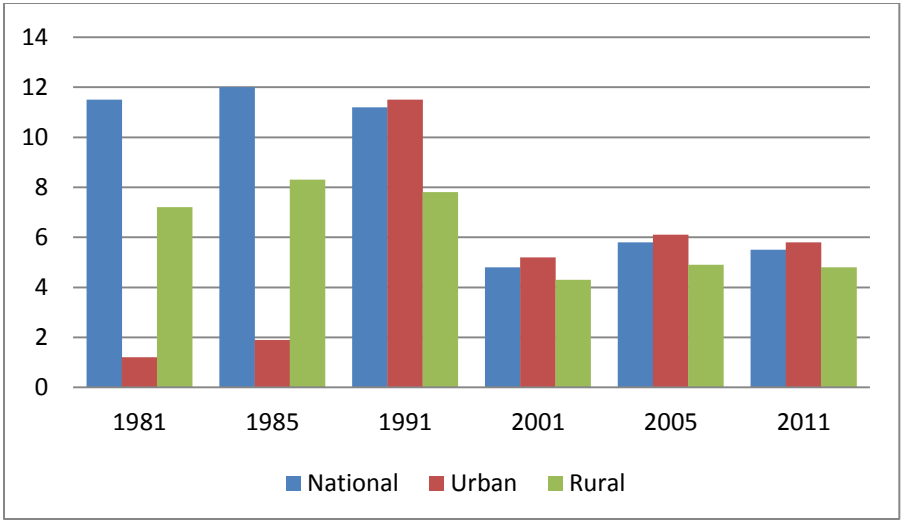


Figure 3: Crude Death Rate by Residence in 1981-2011

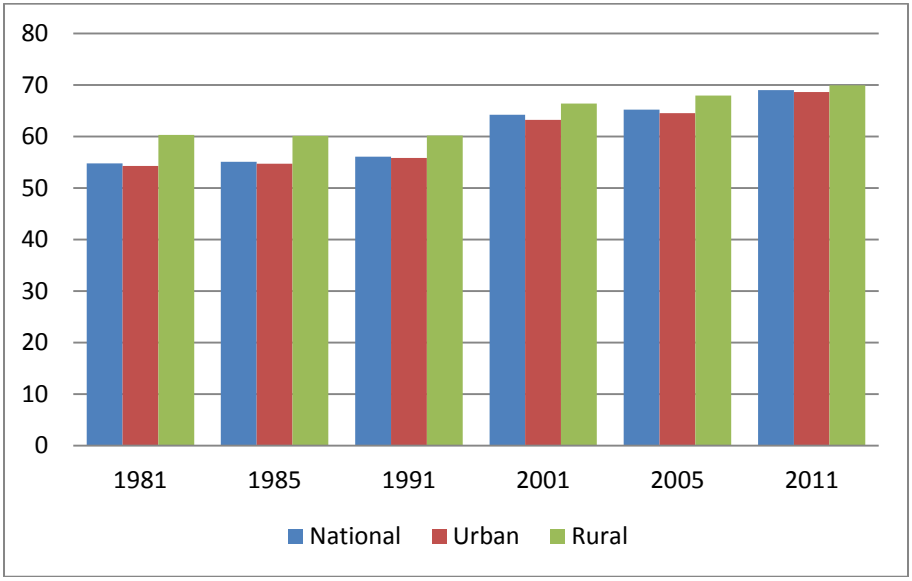


Figure 4: Life Expectancy at Birth 1981-2011

It is observed that crude death rate reduced from 11.5 per thousand in 1981 to 5.5 per thousand in 2011 (Table-3). Consequently, the life expectancy at birth increased substantially during the period. The life expectancy at birth was 54.3 years in 1981 which increased to 69.0 years in 2011 which is commendable.

5. POPULATION DISTRIBUTION BY BROAD AGE GROUP

Due to reduction in fertility and mortality substantial changes occurred in the population distribution by broad age group (Table-4). The broad age distribution by sex shows that in 1981 the male population in the age group 0-14 was 46.5% which reduced to 45.3% in 1991, 38.6% in 2001 and 35.5% in 2011. Similarly, for the female population it was 46.8% in the age group 0-14 in 1981 which reduced to 44.9% in 1991, 38.4% in 2001 and 33.8% in 2011.

On the other hand, working age population (15-59) increased in the subsequent years compared to 1981. For the male population it was 47.4% in 1981 which increased to 48.8% in 1991, 55.7% in 2001 and 56.6% in 2011. In case of female population, the population in the working age population was 48.1% in 1981 50.3% in 1991, 55.2% in 2001 & 59.2% in 2011.

Table 4
Population Distribution in the Broad Age Group and Dependency Ratio by Sex

Sex	Age group	Population Proportion in Census years			
		1981	1991	2001	2011
Male	0-14	46.5	45.3	38.6	35.5
	15-59	47.4	48.8	55.7	56.6
	60+	6.1	5.9	5.7	7.9
Female	0-14	46.8	44.9	38.4	33.8
	15-59	48.1	50.3	55.2	59.2
	60+	5.1	4.8	6.4	7.0

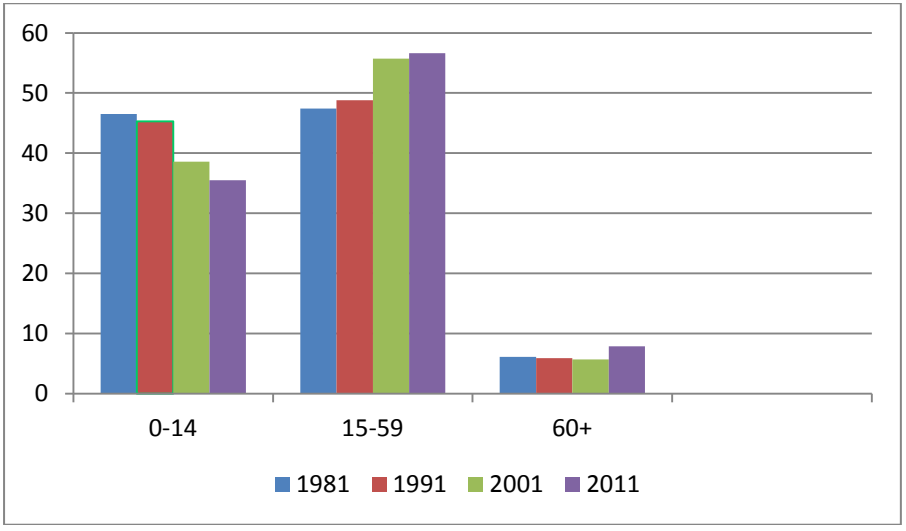


Figure 5: Population distribution in the Broad Age group

6. DEMOGRAPHIC DEPENDENCY RATIO BY SEX:

Demographic dependency ratio by sex has been presented in Table-4. Demographic dependency ratio is defined by the ratio of population 0-14 and 60 years and above to the population 15-59 expressed in percentage.

The demographic dependency ratio male population was 111.0 in 1981 which reduced to 105.0 in 1991, 79.5 in 2001 and 76.7 in 2011.

Similarly, the demographic dependency ratio for female population was 107.9 in 1981, 98.8 in 1991, 79.5 in 2001 & 68.9 in 2011.

The substantial reduction of the demographic dependency ratio indicate that the burden of dependency to the working age population is reducing over the years which lead to improvement in the economic condition of the population

Table 5
Demographic Dependency Ratio by Sex in Different Census Years

Sex	Demographic Dependency Ratio in Different Census Years			
	1981	1991	2001	2011
Male	111.0	105.0	79.5	76.7
Female	107.9	98.8	79.5	68.9

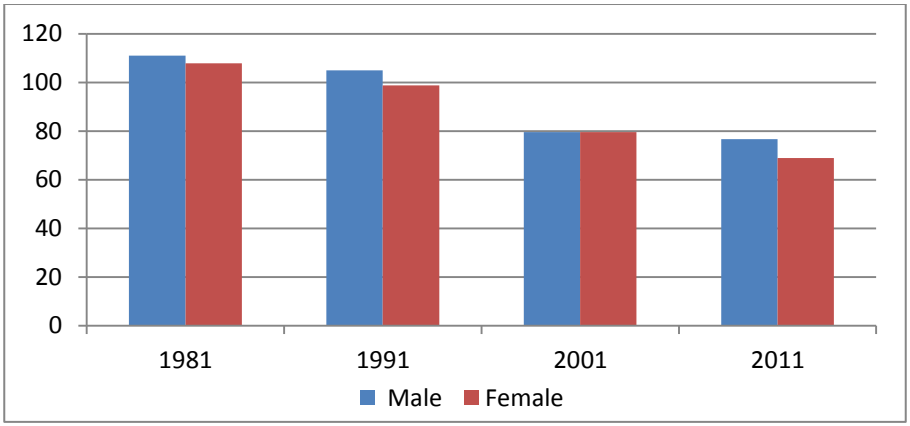


Figure 6: Demographic Dependency Ratio by Sex in Different Census Years

7. GROWTH OF YOUTH POPULATION

Youth are the most potential segment of population. Their contribution in the economy can play positive role in the growth of GDP. It may be noted that due demographic transition, the volume as well as proportion of youth population are increasing in Bangladesh over the last three decades. The volume of youth labour force and their proportion in the total population of the country has been presented in table-6.

Table 6
Youth Population of Age 15-29 in Different Census Years (1981-2011)

Year	Youth Population (Million)			Percent of Total Population		
	Both Sex	Male	Female	Both Sex	Male	Female
1981	21.3	10.6	10.7	24.5	23.6	25.4
1991	26.8	13.0	13.8	25.2	23.7	26.8
2001	34.2	16.2	18.0	27.4	25.3	29.9
2011	39.6	18.5	21.1	27.5	25.7	29.4

Source: Population Census, 1981, 1991. 2001 & 2011

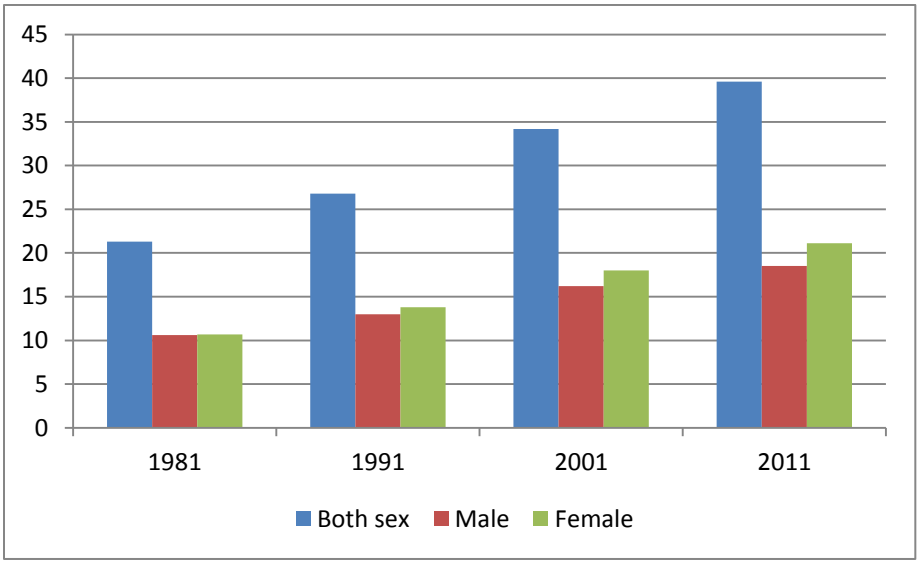


Figure 7: Youth Population of Age 15-29 in Different Census Years (1981-2011)

It is notable that youth population increased from 21.3 million to 39.6 million in the last three decades. The increase is 86.4% in the period and 2.9% annually which is much higher than the total population growth. The proportion of youth population to total population also increased during the period, it was 24.5% in 1981 and increased to 27.5% in 2011. Interestingly, the growth of female youths are comparatively higher than their male counterpart. This may be explained by the outmigration of males for overseas employment in the last three decades. The female youths increased from 10.7 million in 1981 to 21.1 million in 2011. The increase is 97% in the three decades with an annual increase of 3.2% which is higher than the male population.

8. GROWTH OF LABOUR FORCE

The growth of civilian labour force is expected due to demographic transition which is termed as demographic dividend or demographic bonus. The growth of labour force in Bangladesh obtained from the labour force survey has been presented in table-7.

Table 7
Civilian Labour Force in Different Years and Growth Rate

Year	Total Civilian Labour Force (Million)			Growth Rate		
	Both Sex	Male	Female	Both Sex	Male	Female
2002-2003	46.3	36.0	10.3	4.4	3.8	6.5
2005-2006	49.5	37.3	12.1	2.2	1.2	5.5
2010	56.7	39.5	17.2	3.4	1.4	8.7

Source: LFS, 2002-03, 2005-06 and 2010

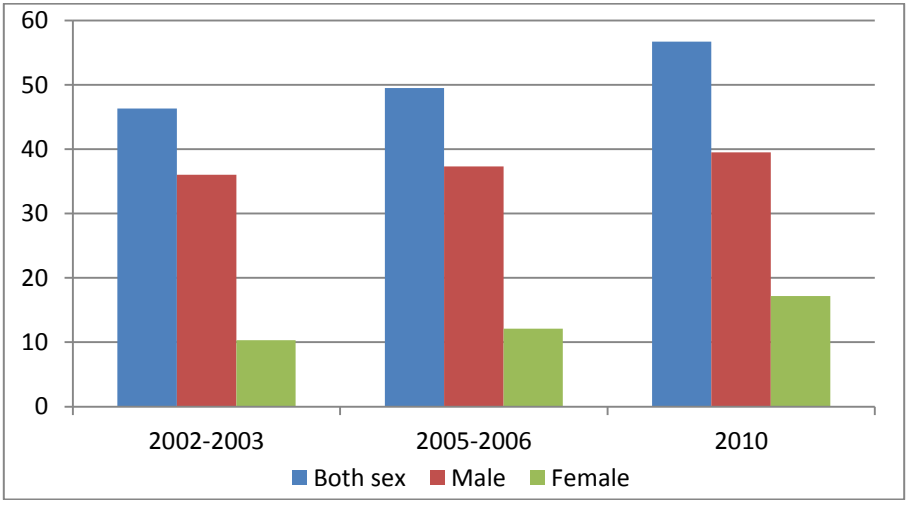


Figure 8: Civilian Labour Force in Different Years

It is observed from the table that the laour force is increasing over the year. In 2002-03, the size of labour force was 46.3 million which increased to 56.7 million in 2010. Therefore, an additional 10.4 million labour force added between the period 2002-03 to 2010 which is commendable. Interestingly, the growth of female labour force is much higher than its male counterpart. The size of female labour force was 10.3 million in 2002-03 which increased to 17.2 million in 2010. The increase is 6.9 million during the period. The increasing number of female labour force may be explained by the participation of large number of female workers in the manufacturing sector, particularly in the Readymade Garments (RMG) industries. The low increase of male laobur force may be explained by the migration of potential labour force to other countries of the world for employment.

9. GROWTH OF YOUTH LABOUR FORCE:

Youth labour force are the driving force for the economic development of the country. It is praise worthy that due to demographic transition youth labour force of the country are increasing. It is observed that (table-8) youth labour force of the country increased from 17.8% million in 2005-06 to 20.9 million in 2010, that is 3.1 million youths added to the labour force in 5 years' time with an annual increase of 0.62 million. It may be

noted that during this period the female youths increased tremendously with an increase of 3.2 million during the period and male youth decreased by 0.1 million which may be explained by the migration of male youths abroad for employment. The data obtained from Bureau of Manpower Employment and Training (BMET) shows that about 0.3 million people went abroad in 2010 most of these out migrants are male which resulted the downward trend of male labour force.

Table 8
Youth Labour Force 2005-06 to 2010

Year	Youth Labour Force (15-29) Million		
	Both sex	Male	Female
2005-06	17.8	13.2	4.6
2010	20.9	13.1	7.8

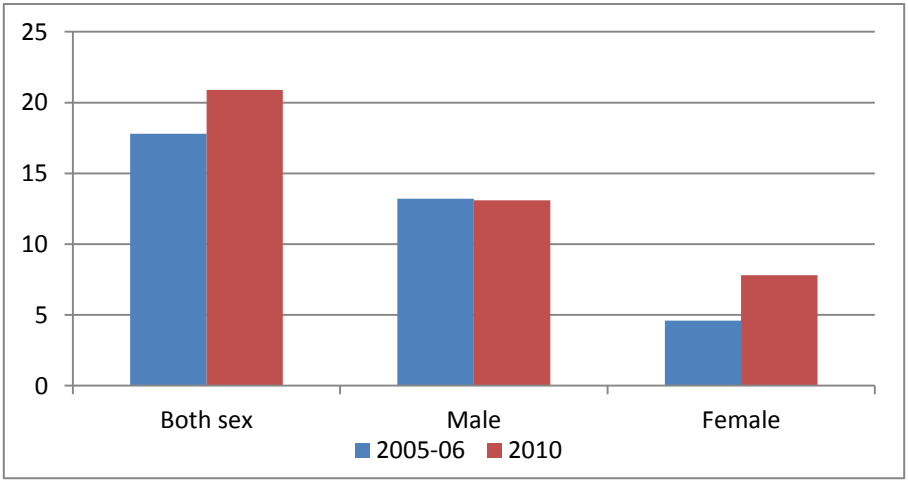


Figure 9: Youth Labour Force 2005-06 to 2010

10. CONCLUDING REMARKS

The paper highlighted the demographic transition in Bangladesh in a nutshell. The demographic transition can be utilized in a very useful manner in order to accelerate the growth. It may be noted that higher proportion of working age population leads to relatively higher per capita income, increased saving, increased investment, higher growth and higher employment. It is praise worthy that Bangladesh is maintaining a GDP growth rate of over 6.0 percent in the last few years. The benefit of demographic transition can be better harnessed by using the higher number of working age population in productive employment. The country's policy makers and planners should think how the increased labour force can be deployed in gainful economic activities. More attention should be given for the development of their skill so they become asset for the country. Adequate productive jobs-domestic as well as foreign will need to be created to benefit from the demographic dividend.

REFERENCES

1. Bangladesh Bureau of Statistics (BBS) (2014). Population and Housing Census 2011, National Volume-I Analytical Report (Draft).
2. Bangladesh Bureau of Statistics (BBS) (2007). Population and Housing Census 2001, National Volume-I Analytical Report
3. Bangladesh Bureau of Statistics (BBS) (1994). Population and Housing Census 1991, National Volume-I Analytical Report
4. Bangladesh Bureau of Statistics (BBS) (1984). Population and Housing Census 1981, National Volume-I Analytical Findings and National Tables.
5. Bangladesh Bureau of Statistics (BBS) (2011). Report of Labour the Force Survey 2010.
6. General Economic Division, Planning Commission (2013). National Sustainable Development Strategy 2010-2021.
7. Bangladesh Bureau of Statistics (BBS) (2013). Report of the Sample Vital Registration System, 2011 Centre for Policy Dialogue, Demographic Transitional Policy & Policy Changes.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 353-362

CLUSTERWISE LINEAR REGRESSION BY LEAST SQUARE CLUSTERING (LS-C) METHOD

Megawati Suharsono Putri[1], Bagus Sartono[2] and **Budi Susetyo[2]**
[1] Graduate School, Bogor Agricultural University, Indonesia. Email: megawatisuharsonoputri@gmail.com
[2] Department Statistics, Bogor Agricultural University, Indonesia.

ABSTRACT

Linear regression analysis has problems if the data distribution is nonlinear or heterogeneous. A sample sometimes is taken from population that has an unknown subpopulation. It is possible to make data distribution becomes nonlinear and heterogeneous, so that more than one of regression model needed to solve these problems. If the standard linear regression is used to estimate the data that has an unknown subpopulation, it will cause an error prediction model. Therefore, clustering is needed to estimate the unknown subpopulation and model regression for each subpopulation. Clusterwise linear regression is a clustering technique based on parameters regression characteristic, to find and reconstruct the hidden structure of sample that taken from the population that has a subpopulation of the unknown, randomly.

KEYWORDS

Regression Analysis, Least Square Estimation, Exchange Algorithm, Robust Regression, Clusterwise.

1. INTRODUCTION

Regression analysis is a statistical technique that used to examine the functional relationship from one or more independent variables to one dependent variable, and especially to find the relationship of the model that unknown yet perfectly (Aunuddin 1989). Set of points that can be connected by a line or a particular curve is called the regression line. Set of points sometimes there is more than one, so that when the sets of points are formed into a regression line, wrong estimation will be occur. Sets of points that make up more than one regression line thought to be caused by the presence of a unknown subpopulation. DeSarbo and Cron (1988) states that if the standard linear regression is used to estimate the data which has unknown subpopulation, it will cause a wrong estimation models that have a small coefficient of determination.

As an illustration, the marketing department want to see the relationship among the price and the purchase of an item. The scatter plot among the price and the purchase of items making two sets points or clusters. The first cluster has large negative coefficient regression (slope), while the second one has small negative coefficient regression. The

intercept of first cluster also greater than the second cluster. So, it can be concluded that the first cluster is weak economic cluster and the second cluster is strong economic cluster. Information obtained when using standard linear regression analysis is limited to the regression coefficient (slope) becomes negative and inaccurate estimation.

Estimation of one set of regression coefficients in a population which consist of unknown subpopulations will be a problem and potentially to be misleading. Clustering based on the characteristics of the regression parameters is needed so can estimate the unknown subpopulation yet (DeSarbo et al. 1989). According DeSarbo and Cron (1988), clusterwise linear regression is one of the important regression for model estimation for the data which have unknown subpopulation. Clusterwise linear regression is a clustering technique based on parameters regression characteristic, to find and reconstruct the hidden structure of sample that taken from the population that has a subpopulation of the unknown, randomly. (Qian and Wu 2010).

Estimation method for clusterwise linear regression analysis that used in this research is least square clustering (LS-C). LS-C method use exchange algorithm to obtain the optimum clusters by the ordinary least squares method (OLS) for the estimation of the regression parameters. Optimum criteria that used in LS-C is the minimum of sum of the sum of square error (SSSE). The initialize in this method using one of robust regression, that is least median of squares.

2. CLUSTERWISE LINEAR REGRESSION

Clusterwise linear regression is a clustering technique based on parameters regression characteristic, to find and reconstruct the hidden structure of sample that taken from the population that has a subpopulation of the unknown, randomly. Clusterwise linear regression was first introduced by Spath in 1979 using exchange algorithm. The initial step in most clustering techniques including clusterwise linear regression is determine the number of clusters in data. The precise number of cluster can optimize the observations that fit into the clusters appropriately so will be minimizing the error.

Spath (1979) determines the exact number of clusters by using over-fitting from the smallest cluster until the cluster that has the minimum error which is the decrease of error is significantly among the estimate number of clusters. The initial that used by Spath (1979) to determine the initial observations that fit into certain clusters is by $p_j = 1 + \mathrm{mod}(j-1, k)$ and $j_0 = n$. Precision of clustering is by the exchange algorithm after the initialization.

General clusterwise linear regression model is:

$$y_j = \sum_{i=1}^{k} \sum_{l=1}^{p} a_{ji} x_{jl} b_l^i + e_j \qquad (1)$$

where $j = 1, 2, \ldots, n$; $l = 1, 2, \ldots, p$; $i = 1, 2, \ldots, k$; y_j = the value of the dependent variable for subject/ovservation j; x_{jl} = the value of the *l-th* independent variable for

subject/observation j; b_l = the l-th ordinary least square (OLS) regression coefficient ; e_j = error for subject/observation j, $e_j \overset{iid}{\sim} N\left(0,\sigma_i^2\right)$ for all $j \in C_i$, $i = 1,...,k$ and

$$a_{ji} = \begin{cases} 1, & \text{if observation } j \text{ is assigned to cluster } i \\ 0, & \text{else} \end{cases}$$

The propose of clusterwise linear regression is to estimate a_{ik} dan b_j^k to minimizes:

$$\Phi = \sum_{j=1}^{n}\sum_{l=1}^{p}\left[y_j - \sum_{i=1}^{k}\sum_{l=1}^{p} a_{ji}x_{jl}b_l^i \right]^2 \quad (2)$$

In this research, LS-C method will be compared with Spath modification method. The initialization method of Spath modification method is by using the random initialization. At the beginning of the random initialization, the number of clusters is determined in stages i.e. calculate SSSE from the smallest of the cluster which is one cluster until all the possible number of clusters are formed. Random initialization will produce a lot of different possibilities for initialization of each possibility cluster, so the estimate of regression parameters can be different though not significant. Therefore, estimation of parameters in each cluster possibility would be as much as 100 times the repetition. Criteria for selection of the best regression model in each cluster possibility is the most minimum SSSE for 100 times repetitions. Determination of the number of clusters is determined if the selected SSSE or the most minimum SSSE decreased significantly in certain clusters and clusters tend to be constant in the next.

Determination of the number of clusters are incrementally by trying possibilities clusters makes a long random initialization process. Determination of random inizialiation was also made into a long process of computing. Wu and Qian (2011) proposed the initialization by using robust regression which is least median of squares (LMS). Robust regression has the advantages of less sensitive to outliers, so the regression line for initialization has a small error. It makes the iteration process in exchange algorithm is fewer so that can reduce computing time. Wu and Qian (2011) also stated that robust regression procedure has a high consistency to determine the observations in clusters so can accelerate the process of iteration.

LS-C method estimated a number of clusters and regression parameters in each cluster simultaneously. At initialization by using robust regression, the optimal number clustering (k) is only done in a single process that is at the initial stage so that can reduce the computation time. Initialize by using robust regression has the disadvantage that it can not determine the number of clusters.

An Algorithm of Least Square Clustering (LS-C)

An algorithm of a clusterwise linear regression method in by the least squares clustering is:

1. Data Exploration
   Make a scatter plot of each independent variable on the dependent variable. If the scatter plot indicates the presence of clusters, it can be used clusterwise linear regression analysis.

2. Initialization

In the method of least squares penggerombolan (P-KT), the optimal number of penggerombolan (*k*) is only done in a single process that is at the stage of initialization. The criteria of LS-C is used as a rule the selection of the best number of clusters (Wu and Qian 2011). All observations are given initial partition by $\varphi = \{1, 2, ..., n\}$. A cluster is denoted by *C* and the complement of clusters denoted by C^c. Initialization steps on LS-C as follows:

Step 1. Consider the linear model

$$\boldsymbol{y_i} = \boldsymbol{x_i'\beta} + \boldsymbol{e} \tag{3}$$

Based on the whole dataset, one estimates $\boldsymbol{\beta}$ by a robust method, i.e. least median of squares method.

Step 2. Put all data points, whose distances to the regression hyper plane estimated in Step 1 less than a predetermined number, say δ, into a set C1. If the $|C_1|$ and $|C_1^c|$ are both larger than a predetermined integer, say p, set ℓ=1 and go to the next step, otherwise, set ℓ = 0 and go to Step 5.

Step 3. Based on the dataset $\bigcap_{i=1}^{\ell} C_i^c$, one estimates $\boldsymbol{\beta}$ by the same robust method used in Step 1.

Step 4. Put all data points in $\bigcap_{i=1}^{\ell} C_i^c$, whose distances to the regression hyper plane estimated in Step 3 is less than δ, into a set $C_{\ell+1}$. If $|C_{\ell+1}|$ and $\left|\bigcap_{i=1}^{\ell+1} C_i^c\right|$ are both larger than p, set $\ell = \ell + 1$ and repeat Step 3; otherwise, go to Step 5.

Step 5. The initial partition is $\left\{C_1, ..., C_\ell, \bigcap_{i=1}^{\ell} C_i^c\right\}$ if ℓ >1 or just the whole dataset itself if ℓ = 0.

3. EXCHANGE ALGORITHM

An algorithm to obtain the optimum clusters based on smallest SSSE criteria as follows:

Step 1. Label all the data points in the sample as 1 to n. Given an initial partition $\prod_k = \{C_1, ..., C_k\}$ of $\varphi = \{1, 2, ..., n\}$, fit regression models for each of the *k* clusters by ordinary least square (OLS) and calculate the sum of the sum of squares error $SSSE_0$ for this partition. Initialize *i* = 0.

Step 2. Set $i = i + 1$ and *i* = 1 if *i* > *n*. Suppose $i \in C_j$, Then move *i* into C_h, $h = 1, ..., k$ and $h \neq j$ respectively. For each of these *k – 1* relocations, re-fit the regression models by OLS for the changed clusters and calculate the overall sum of the squared

residuals accordingly. Denote the smallest one by $SSSE_h$. If $SSSE_h < SSSE_0$, redefine $C_j = C_j - \{i\}$, $C_h = C_h + \{i\}$, and set $SSSE_h = SSSE_0$. Otherwise keep i in C_j.

Step 3. Repeat Step 2 until the objective function (*SSSE*) could not be reduced any further, which means no observation relocation is necessary and the optimal clustering is achieved for this *k*.

3. RESULT AND DISCUSSION

Human development index (HDI) is a measure of the achievement of human development based on a number of basic components of quality of life. HDI calculation consists of three index e.g. health index such as life expectancy (LE) in units of years, education index such as the percentage of the population and the average number of school (ANSY) in years, and buying power index in the form of spending per capita in the currency.

One of the main basis for the capital of the region is to improve the quality of human development which the indicator is the HDI, is a development fund set out in the regional government budget (RGB). Therefore, in this research will use the data HDI of district/city in East Java province in 2013 as the dependent variable (Y). Based on the calculation of HDI index, the independent variables used is the percentage of the economy (X_1), health (X_2) and education (X_3) on RGB districts / cities in East Java province in 2013. The sample size of the district / city in East Java is 38. According to Mankiw (2003), measuring instruments of economic growth a major area is the gross regional domestic product (GRDP). By using the data GRDP, it can be seen the distribution of development of a region. Therefore, GRDP data is used to compare the percentage of equity in each business field in each district/city GRDP data East Java province.

3.1 Data Exploration

Data description among each independent variable on the dependent variable needs to be done as early information to determine indications of clusters. Scatter plot can help determine the presence of clustering (DeSarbo et al. 1989).

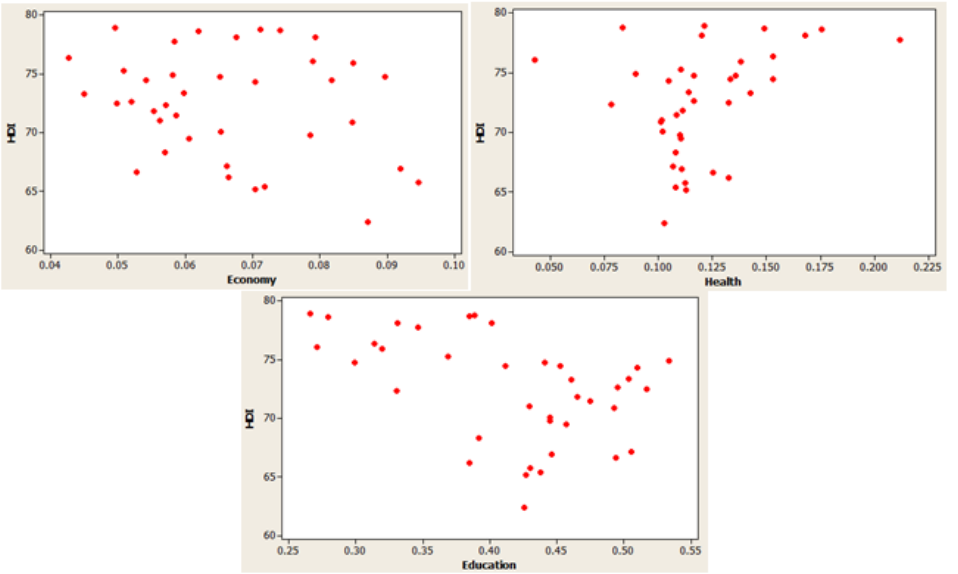


Figure 1: Scatter plot among Economy (X_1), Health (X_2) and Education (X_3) to IPM (Y)

In Figure 1, the scatter plot among the economy with the HDI does not form a linear line and indicate clusters. In the scatter diagram among the HDI health also does not form a linear line and visible indications of the clusters which on the top tend to have a positive slope, while the bottom is tend to have a negative slope. In the scatter diagram among education and HDI is looked data does not form a linear pattern and tend to converge at some point so indicate clusters. If using the standard linear regression, the R-Square 31.5%. That is very small.

3.2 Spath Modification Method

Spath (1979) determines the exact number of clusters by using over-fitting from the smallest cluster until the cluster that has the minimum error which is the decrease of error is significantly among the estimate number of clusters. The initialization of the spath modification method is determines the exact number of clusters by using over-fitting from the smallest cluster until the cluster that has the minimum error which is the decrease of error is significantly among the estimate number of clusters.. Initial random initialization will make a lot of different possibilities for each initialization of cluster. Therefore, estimation of parameters in each cluster would be as much as 100 times the repetition. Then find the minimum of SSSE at every possible clusters as shown in Table 1

Table 1
Minimum SSSE at Every Possible Clusters

Cluster	SSSE
1	519. 265
2	114.5403
3	18.15009
4	11.04469
5	6.841521
6	4.19561
7	0.2875541

Based on Table 1, that shows a decrease significantly SSSE in the second and third cluster. SSSE does not decrease significantly in the fourth clusters and so on. This is evident in Figure 2.

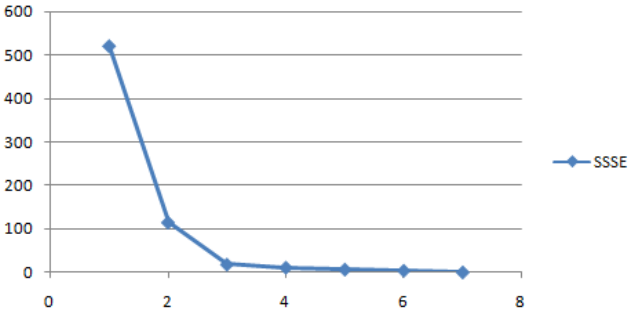


Fig. 2: Scree Plot between each Cluster Possibility with a Minimum SSSE

In Figure 2, looked the SSSE of second and third clusters decrease significantly and after three clusters tend to decrease constantly or insignificantly. Therefore, set the number of clusters in 3 clusters. Then the linear regression model in each clump that is:

$$\begin{aligned} y_1 &= 78.47 + 78.32x_1 - 5.28x_2 - 30.77x_3 + e \\ y_2 &= -21.25 - 42.22x_1 + 279.29x_2 + 137.23x_3 + e \\ y_3 &= 80.097 + 111.489x_1 - 2.824x_2 - 25.764x_3 + e \end{aligned} \tag{3}$$

The size sample of each clusters are n_1= 16, n_2=10 and n_3=12.

3.3 Least Squares Clustering Method (LS-C)

LS-C method estimates a number of clusters and regression parameters in each cluster simultaneously so that the determination of the number of clusters is not gradual. The initialization is using $\delta = 1.645\sigma$ for separating and forming the cluster. By using the LS-C, the clusters obtained by two clusters. Linear regression analysis model in each cluster is:

$$\begin{aligned} y_1 &= 86.44 - 104.95x_1 - 160.87x_2 + 15.02x_3 \\ y_2 &= 82.931 + 18.624x_1 - 1.967x_2 - 20.269x_3 \end{aligned} \tag{4}$$

The size sample of each clusters are n_1= 19 and n_2=19.

3.4 Comparison between Spath Modification Method and LS-C Method

Spath modification method gives three clusters with SSSE is 18.15 while the LS-C method produces two clusters with SSSE 114.5403. SSSE will be smaller if the clusters is increase, so the adjusted R-square will also increase. At 2 clusters, SSSE and the adjusted R-Square in spath modification and LS-C methods has the equal value. However, the iteration in Spath modification method more than the LS-C method. This is due to the random initialization in Spath modification method. In the method of the LS-C, the initialization is by using robust regression so that the regression line formed fairly consistent so that the point move in exchange algorithm is less the in Spath modification method. It is listed in Table 2.

Table 2
Iterations, SSE and Adjusted R-Square in Spath Modification and LS-C Method

	2 Clusters			3 Clusters		
	Iteration	**SSE**	**Adj R-Square**	**Iteration**	**SSE**	**Adj R-Square**
	245	73.048	60.82%	235	11.107	84.77%
Spath		41.492	46.16%		5.145	98.52%
Modification					1.897	92.76%
		114.54			18.1501	
PKT	189	73.048	60.82%			
		41.492	46.16%			
		114.54				

3.5 Cluster Description

The first cluster is a city cluster because 90% of the city in East Java Province is in the first cluster, while the second one has only one city. That is Batu.

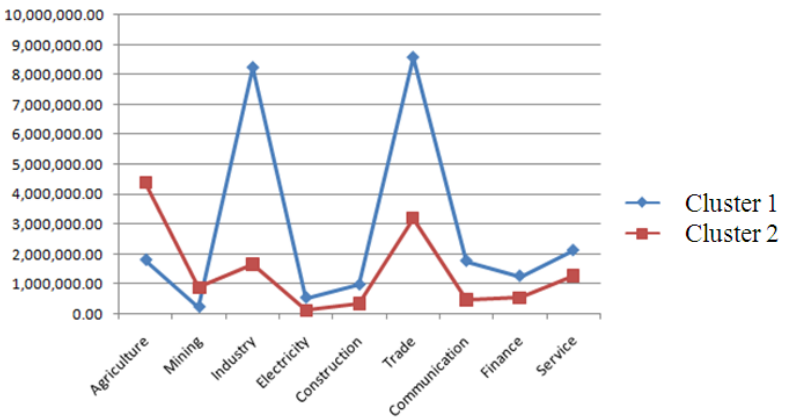


Figure 3: The Average of GRDP of each Cluster in 2 Cluster

Figure 3 is a graph of the average GDRP of each cluster in each business field. Figure is using the LS-C method. In Figure 3 the first cluster is dominated by the city is significantly superior in industry, electricity, construction, trade, communications, finance and services. But the second cluster is dominated by the district is superior in agriculture and mining.

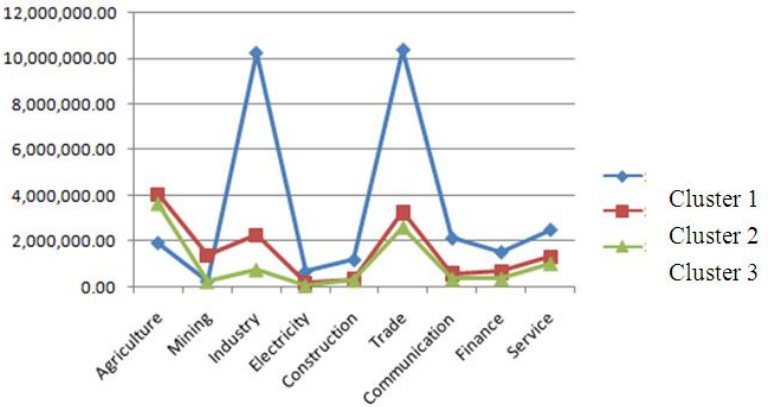


Figure 4: The Average of GRDP of each Cluster in 3 Cluster

Figure 4 using a spath modification method. A number of clusters are formed on the Spath modification method is 3 clusters. In the Spath modification method, the first clusters look more superior than the second and third clusters in almost the entire field of business, except in the field of agriculture and mining. The pattern of the second and third clusters also have similar and tend to almost near. Therefore, will be investigated using two clusters.

In first cluster by two clusters, the adjusted R-square is 46.16%. 80% observation in first cluster by LS-C is the first cluster overall observations by Spath modification. 20% other observation are Trenggalek district, Pasuruan and Mojokerto. If using the linear regression analysis on the first cluster by LS-C, Trenggalek district, Pasuruan and Mojokerto is an unusual observation. That is because the level of the economy of the observation are much higher than other observation in the cluster. Therefore, the economic variables Trenggalek, Pasuruan and Mojokerto will be replaced with the average complement. The adjusted R Square increase significantly from 46.16% to 73.6% if the data area of economic variables Trenggalek, Pasuruan and Mojokerto replaced with the average complement. Therefore, the cause of the formed of three clusters only because of the proportion of economic allocation in the RGB is higher than observation complement of its cluster. Thus, it can be concluded that the two clusters is the right number of clusters.

6. CONCLUSION

The initialization method of Spath modification method is by using the random initialization. At the beginning of the random initialization, the number of clusters is determined in stages i.e. calculate SSSE from the smallest of the cluster which is one cluster until all the possible number of clusters are formed. Random initialization will

produce a lot of different possibilities for initialization of each possibility cluster, so the estimate of regression parameters can be different though not significant. Therefore, estimation of parameters in each cluster possibility would be as much as 100 times the repetition. Criteria for selection of the best regression model in each cluster possibility is the most minimum SSSE for 100 times repetitions. Determination of the number of clusters is determined if the selected SSSE or the most minimum SSSE decreased significantly in certain clusters and clusters tend to be constant in the next. Gradual process and random initialization make long process of computing.

In the method of the LS-C, the initialization by using robust regression which is least median of squares (LMS). LS-C method estimated a number of clusters and regression parameters in each cluster simultaneously, so the iteration process is faster. The consistency to determine the observations in clusters can reduce computing time. Determination of the number of clusters based robust regression has the disadvantage which is can not determine number of clusters from beginning. If the number of clusters has been known in theory, the LS-C method can not be used.

REFERENCES

1. Aunuddin. (1989). Analisis Data. Bogor: Depdikbud Dirjen Pendidikan Tinggi Pusat Antar Universitas Ilmu Hayat IPB.
2. DeSarbo, W.S. and Cron, W.L. (1988). A maximum likelihood methodology for clusterwise linear regression, *Journal of Classification.* 5, 249-282.
3. DeSarbo, W.S., Oliver, R.L. and Rangaswamy, A. (1989). A simulated annealing methodology for clusterwise linear regression. *Phychometrika.* 54(4), 707-736
4. Qian, G., and Wu, Y. (2011). Estimation and selection in regression clustering. *European JPAM*. 4(4), 455-466.
5. Spath, H. (1979). Algorithm 39 clusterwise linear regression. *Computing*. 22(4), 367-373.
6. Spath, H. (1982). A fast algorithm for clusterwise linear regression. *Computing*. 29(2), 175-181.
7. Mankiw, N.G. (2003). Teori Makro Ekonomi, 5th Edition. Erlangga, Jakarta.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 363-370

FORECASTING OF PADDY PRODUCTION IN INDRAMAYU REGENCY WITH TRANSFER FUNCTION MODEL

Rena Foris Windari[1] and **Erfiani**[2]
Departement of Statistics, Bogor Agricultural University, Indonesia
Email: renaforis89@gmail.com

ABSTRACT

Paddy is one of crucial commodities in Indonesia because it is the main food for Indonesian peoples. The percentage of them who are consume rice is about 97 %. It means that only 3% of households who not consume rice. Indramayu Regency as granary hasimportant role as a supplier of rice at the surrounding areas. Therefore, paddy production in that area has been enhanced to achive self-sufficiency in rice. paddy production is influenced by the amount of farmers, fertilizer, pesticides, irrigation systems and areage. The influence of acreage for paddy production can be explained by the transfer function model. The transfer function is a model that combines the characteristics of regression model and time series model.

The result of the transfer function model indicate that paddy production is influenced by paddy production previous month, acreage three previous month. MAPE value for forcasting 12 month is 12.5%. Forcasting value of transfer function model is closer to actual value for all month except on Juni 2012, September 2012 and November 2012.

Furthermore, the aplication of transfer function model for forcasting paddy production is better that ARIMA model.

KEYWORD

Paddy production, transfer function, ARIMA.

1. INTRODUCTION

Paddy is one of crucial commodities in Indonesia because it is the main food for Indonesian peoples. The percentage of them who are consume rice is about 97 %. It means that only 3% of households who not consume rice.These conditions indicate that Indonesian peoples are still depend on rice. Indramayu Regency as granaryhas important role as supplier of rice for the surrounding areas. Therefore, the enhansment of paddy production in that region has beencultivated in order to achieve self-sufficiency in rice.According to the research results of Triyono (2010), paddy production is influenced by the amount of farmers, fertilizer, pesticides, irrigation systems and land area.

One of the necessary information which is related with achievement self-sufficiency in rice is the forecasting of paddy production. Paddy production is one of variables that can be observed as time series data. Forecasting of time series datawhich is based on behavior of the past data can be performed byAutoregresive Integrated Moving Average

(ARIMA) model. However, if we want to see the effect of the explanatory variables, so we can be used transfer function model. Transfer function model is a multivariate time series model which is combines characteristics of the ARIMA model and regression model (Wei 1990). Regression is use data which is independent of each observations, while transfer function use data which is not independent of each period. Liu and Hanssens (1982) explains that the transfer function model is usually applied in engineering, economics and management. The aim of this research is to implement transfer function model in agricultural. Transfer function model is used to forcast paddy production which is influenced by acreage in Indramayu Regency.

2. METHODOLOGY

Transfer function model is an improvement of the ARIMA model and commonly called the multivariate ARIMA. If the time series Y_t is correlated with one or more time series X_t, so it can be made a time series model to estimate the value of Y_t based on information from X_t. This model is called the transfer function model. Y_t is called output series and Xt is input series (Makridakis et al. 1983). The general form of transfer function model as follow(Wei 1990):

$$Y_t = \delta_r^{-1}(B)\omega_s(B)X_{t-b} + a_t$$

The error component (a_t) may be modeled by the ARIMA (p, d, q) (P, D, Q) ^{s} so that the combination of transfer function model with error is:

$$Y_t = \delta_r^{-1}(B)\omega_s(B)X_{t-b} + \frac{\theta_q(B)\Theta_Q(B^S)(B^S)a_t}{\phi_p(B)\Phi_Q(B^S)(1-B)^d(1-B)^D}$$

with:

$$\omega_s(B) = \omega_0 - \omega_1 B - \omega_2 B^2 - \ldots - \omega_s B^s$$
$$\delta_r(B) = 1 - \delta_1 B - \delta_2 B^2 - \ldots - \delta_r B^r$$
$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$$
$$\Phi_P(B) = 1 - \Phi_1 B - \Phi_2 B^2 - \ldots - \Phi_P B^P$$
$$\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q$$
$$\Theta_Q(B) = 1 - \Theta_1 B - \Theta_2 B^2 - \ldots - \Theta_Q B^Q$$

b,s,r,p,q,P,Q is constans

This study use secondary data from the Directorate General of Food Crops, Ministry of Agricultural and the BPS-RI.Paddy production (tons) in Indramayu is used as the response variable (Y_t) andacreage (hectare) is used a the predictor variables (X_t). Data on January 2003 to December 2011 is used to modeling, while the data in 2012 is used as a validation of the model.

Procedure analysis steps which were conducted in this study is as follows:

1. Data exploration.
2. Identification of transfer function model.

   Step 1 : Model identification
   a. Preparedinput seriesand output series.
   b. Pre-whiteningof input series.
   c. Pre-whitening ofoutput series.
   d. Calculated cross correlation betweeninput series and output series that had been pre-whitened.
   e. Determined theorde of transfer function tentative model (b, s, r) which is connect input series and output series.
   f. DeterminedARIMA modelforerror component (a_t)

   Step 2 : Parameters estimation of transfer function model
   a. Estimated of initial parameters.
   b. Estimated of final parameters.

   Step 3 : Diagnostics checking of transfer function model
   a. Calculated autocorelation of residual of the model.
   b. Calculated cross correlation betweeninput series andresidual.
3. Forcasting of paddy production for 12 period.
4. Calculatedof *Mean Percentage Error (MAPE)*
5. Compared the result of forecasting with ARIMA model and transfer function model.

3. RESULT AND DISCUSSION

Data Exploration

Figure 1 show that the paddyproduction data was fluctuate that tend to increase every years. Highest paddy production values is 282237.53 tons which was occurred on April 2007while the lowest production 13524.04 tons which isoccurred in December 2006.If it was observed every year, the pattern of paddy production data was increase continuously from january until April, then fell back to December. Thus, the highest paddy production was occurred on April, while the lowest production occurred in December. Such fluctuations tend to be similar every year. That was indicated that the paddy production contains seasonal patterns with the orde of seasonal is 12.Seasonal patterrn not only shown by paddy production data but also shown by acreage data. Figure 2 indicate that the fluctuation of acreage datahad same pattern which was repeated over 12 months.Futhermore, correlation value which is resulted by pearson correlation is 0.26. That is indicate that the acreage has little direct influence for paddy production.

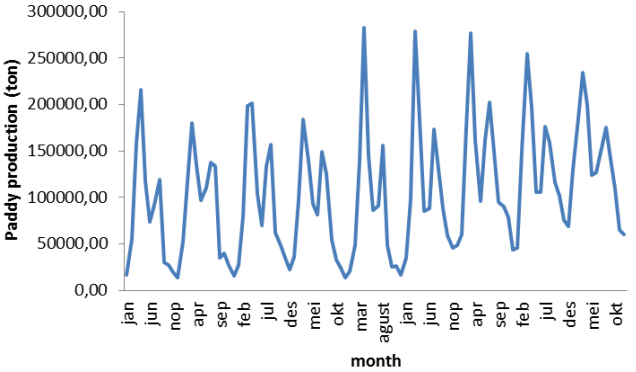


Fig. 1: Paddy production January 2002 - Desember 2011

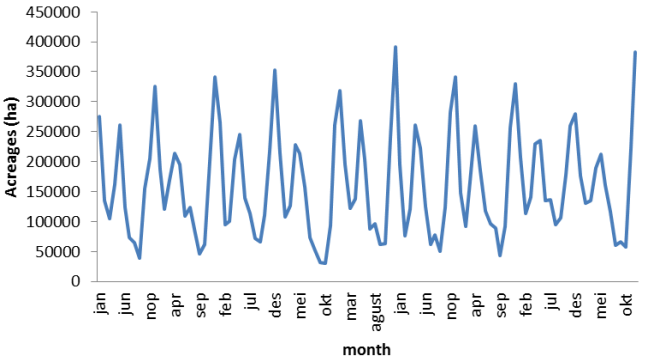


Fig. 2: Acreage Januari 2003 - Desember 2011

Transfer Function Model Identification

The first step of the transfer function analysis is prepared the input and output series. The series have to be stasionary in variety and means.Figure 1 and Figure 2 shows that paddy production and acreage are not stationary. Thats is evident from the fluctuation of the data which is not constant. Time series data is require differencing of regular or seasonal pattern to achieve stationary condition. In this case, differencing of 12 periods was appliedto paddy production (output series) and acreage (input series). Table 1 provides information about the result of Augmanted Dicky-Fuller test. Paddy production and acreage are not stationary because its has a p-value greater than 0.05. Then, differencing orde 12 had made both series were in stationary condition (p-value < 0.05).

Table 1
The result of Augmented Dickey-Fuller test

Deret	P-value	
	d=0	d=12
Produksi Padi	0.356	<0.000
Luas Tanam	0.083	< 0.000

The next step is calculate autocorrelation (ACF) and partial autocorrelation (PACF) of the stasionary series. The pattern had been formed by the ACF and PACF plots isused toidentification of ARIMA models. ARIMA model for output series and imput series is ARIMA(2,0,0)(0,1,1)^{12}.

After the ARIMA model for the input series had been obtained, the next step is pre-whitening of input and output series. Pre-whitening input seriesis transformation processes from the series which is correlated to be white noise which is uncorrelated. Then, pre-whitening of output series is obtained by performing the same transformation with the input series.

The series which is resulted by pre-whtening proses are used to determine the orde of transfer function tentative model. orde b and s is obtained based on the pattern of cross corelation (CCR) plot between theinput series and output series which had been pre-whitened, while r is obtained base on ACF plot of stasionary output series . b is determined based on the first lag that significant in CCR plot. The value of b indicate the starting point of input series isinfluence the output series. Orde s is determine base on the next significantlag which is show a particular pattern. It is represent the lenght of$\mathbf{X_t}$is influence$\mathbf{Y_t}$after b**.**Order is base on the next significant lag after the firts lagin ACF plot of stasionary output series (Y_t). Based on figure 4, orde of the initial transfer model are b = 3, s = 0 and r = 0. Then, that model is used to estimate the initial parameters of the model transfer function.

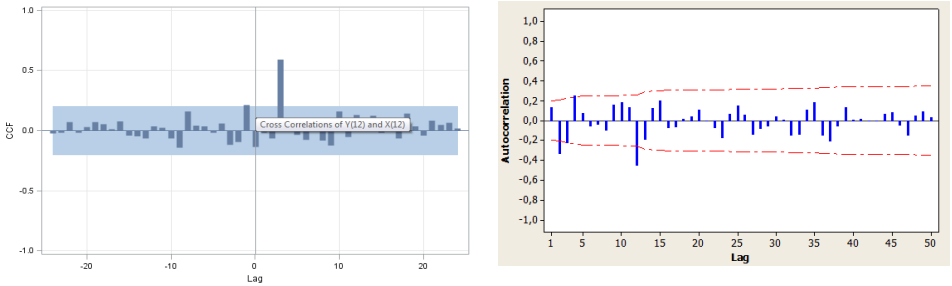


Fig. 4; CCR Plot of Input and Output Series and ACF Plot of Stasionary Output Series

Identification of the final parameters of transfer function model is combining the initial model with the error component model. Based on the significant parameters and uncorelatted error, the orde of final transfer function model is b = 3, s = 0, r = 0 with error component model is ARIMA (1,0,0)(0,0,1)^{12}. Then, the results of diagnostics check indicate that the model has residual which is uncorelated each other and independency between the input series and the residual. This condisionis shown by table 2 which is show hat p-value of Box-pierce test is greater than 0.05 and table 3 which is show that p-value of cross-correlation betweeen error component and the input series is not significant.The final model can be written as follow:

$$Y_t = \omega_0 X_{t-b} + \frac{(1-\Theta_1 B^{12})a_t}{(1-\varphi_1 B)}$$

$$Y_t = 0.567X_{t-3} + \frac{(1-0.614B^{12})a_t}{(1-0.233B)}$$

The result of the transfer function model indicate that paddy production is influence by paddy production previous month, acreage three previous month and interaction of the random effect of paddy production and acreage twelve previous month.The effect of acreage on paddy production is delayed for 3 months because its require around 2-3 months from planting to harvest time.

Table 2
The Result of Box-Pierce Tests of Error Component

Lag	P-value
5	0.37
11	0.10
17	0.07
23	0.12

Table 3
Cross-Correlation between Error and Input Series

Lag	P-value
6	0.23
12	0.14
18	0.17
24	0.26

Forcasting of Paddy Production

Table 4 show the result of forcasting of paddy production for 12 period with tranfer function model.

Table 4
Forecasting of paddy production

Month	Prediction
Januari 2012	54047.5
Februari 2012	155167.3
Maret 2012	295094.8
April 2012	229692.9
Mei 2012	115027.4
Juni 2012	121180.5
Juli 2012	180897.2
Agustus 2012	182794
September 2012	127640.4
Oktober 2012	101382.6
November 2012	77597.9
Desember 2012	68669

Comparison of Transfer Function Models and ARIMA

The accuracy of the forcasting value is determined by model validation.The concept of model validation is comparing of forcasting value and actual value. One of criterion which is used to model validate is MAPE.The best model is the model with smallest MAPE value.Comparison of MAPE value of the transfer function and ARIMA model are presented by Table 5. Table 5inform that MAPE value of transfer function model is 12.53% which is more smaller than MAPE value from ARIMA model (25.75%). Futhermore, figure 5 show that prediction value of paddy production with transfer function model are more closer to actual value than prediction value of ARIMA. Therefore, the transfer function model is better than ARIMA model when used to forcast paddy production in Indramayu Regency.

Table 4
Comparison of Transfer Function Models and ARIMA

Month	Actual	Transfer function	ARIMA
Januari 2012	45579.7631	54047.5	93442.83
Februari 2012	134535.007	155167.3	164888.1
Maret 2012	296613.605	295094.8	256240.3
April 2012	229239.616	229692.9	219731.5
Mei 2012	106254.905	115027.4	136484.4
Juni 2012	94709.6708	121180.5	137880.9
Juli 2012	194287.42	180897.2	182495.1
Agustus 2012	176694.487	182794	179554.8
September 2012	107174.393	127640.4	123137.1
Oktober 2012	90882.3933	101382.6	104244.8
November 2012	68808.091	77597.9	82320.33
Desember 2012	54624.159	68669	72214.27
MAPE		12.53%	25.70%

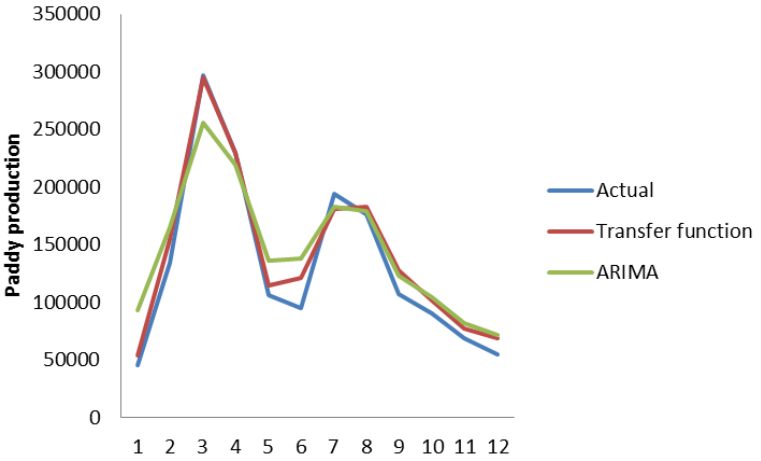


Fig. 5: Comparison Graph of Actual Value and Prediction Value of Paddy Production

4 CONCLUSION

The research was conclued that transfer function model can explain the relationship between paddy production and acreage three previous months. Forcasting value is closer to actual value for all month except on Juni, September, and November.

REFERENCES

1. Liu, L-M. and Hanssens, D.M. (1982). *Identification of Multiple-Input Transfer Function Models.* Graduate School of Management University of California Los Angeles. California.
2. Makridakis, S., Wheelwright, S.C. and Megee, V.E. (1983). *Forecasting Method and Applications*. 2nd edition. New York: John Wilcy and Sons.
3. Triyanto (2010). *Analisis Produksi Padi di Jawa Tengah*. Semarang: Universitas Diponegoro.
4. Wei, W.W.S. (1990). *Time Series Analysis Univariate and Multivariate Methods*. Canada: Addison-Wesley.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 371-374

E-M ALGORITHMS METHOD FOR ESTIMATING MISSING DATA IN REGRESSION ANALYSIS

Septian Rahardiantoro[1] and **Bagus Sartono**[2]
Department of Statistics, Bogor Agricultural University, Bogor, Indonesia
Email: [1] rahardiantoro.stk@gmail.com
[2] bagusco@gmail.com

ABSTRACT

In this paper would be explained the problem of missing data,and the developing of E-M (Expectation-Maximization) algorithms for handling the missing data. The specified case is in regression analysis with missing data in the independent variables, or in dependent variables. The proposed method is based on imputation the missing value using random number (expectation step), then it would be evaluated by maximum likelihood estimator (maximization step) for estimating the parameter model which use to estimate the missing value from the model. This proccess would be repeated iterativelyuntil the number of estimated missing value convergent in a value.Review the methods used and the illustrated case using simulation in two general frames, missing data in the independent variables, and missing data in the dependent variables. From the simulation result, E-M algorithms could be used for estimating missing value when the data in regression analysis case contained missing value.

KEYWORDS

E-M algorithm, missing value, missing value imputation, regression.

1. INTRODUCTION

In regression analysis, the problem of non-response to one or more questions in the specified studies may be very troublesome. In example, nonresponse in the survey is a distraction to our main goal of studying trends in attitudes andeconomic conditions, and we would like to simply clean the dataset so it could beanalyzed as if there were no missingness.There are two general methods for dealing this problem. One is to discard all incomplete observations and apply the OLS procedure only to the complete observations [1]. The alternative is using imputation methods to fill the missing observations [2].

This paper would be discussed the second alternative method, missing-data imputation, in the applicationof E-M algorithm. The principle of E-M algorithm is an iterative procedure that finds the MLE of the parameter vector by repeating the expectations (E) step and the maximization (M) step. The two steps are iterated until the iterations converge. [3]

Thegeneral procedures in this paper contains two general frames. First frame is designed in condition with missing value in independent variable. In this case, there are

many missing value for one independent variable. Simulation designed by simple linear regression withmissing value in independent variable. Second frame is designed in condition with missing value in dependent variableusing multiple regression. Evaluation from the estimation of missing data using correlation from data complete and data estimated. Simulation using Rsoftware.

2. E-M ALGORITHMS IN REGRESSION ANALYSIS

The E-M algorithm defined by starting with cases that have strong restrictions on the complete-data specification$\boldsymbol{f}(\boldsymbol{x}|\boldsymbol{\Phi})$**,** then presenting more general definitions applicable when these restrictions are partially removed in two stages. Suppose first that $\boldsymbol{f}(\boldsymbol{x}|\boldsymbol{\Phi})$hasthe regular exponential-family form

$$\boldsymbol{f}(\boldsymbol{x}|\boldsymbol{\Phi}) = \frac{b(\boldsymbol{x})e^{\boldsymbol{\Phi}\boldsymbol{t}(\boldsymbol{x})'}}{a(\boldsymbol{\Phi})}$$

where $\boldsymbol{\Phi}$denotes a 1 $\times r$ vector parameter, $\mathbf{t}(\mathbf{x})$ denotes a 1$\times$ r vector of *complete-data* sufficient statistics. Then suppose that $\boldsymbol{\Phi}^{(p)}$denotes the current value of$\boldsymbol{\Phi}$ after p cycles of the algorithm. The next cycle can be described in two steps, as follows:

E-*step* : Estimate the complete data sufficient statistics $\boldsymbol{t}(\boldsymbol{x})$ by finding

$$\boldsymbol{t}^{(p)} = E\left(\boldsymbol{t}(\boldsymbol{x})|y, \boldsymbol{\Phi}^{(p)}\right)$$

M-*step*: Determine $\boldsymbol{\Phi}^{(p+1)}$ as solution of the equations

$$E(\boldsymbol{t}(\boldsymbol{x})|\boldsymbol{\Phi}) = \boldsymbol{t}^{(p)}$$

Equations in M-*step* are the familiar of the likelihood equations for maximum-likelihood estimation given data from a regular exponential family [3].

So, from the algorithm above, the procedural steps in application of E-M algorithms when dealed with multiple regression, follows the steps below:

Assume the regression model behave with the errors in standard normal distribution with model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y}$denotes a $n \times 1$ vector dependent variable, $\boldsymbol{X}$ denotes a $n \times k$matrix of *complete-data*, $\boldsymbol{\beta}$ denotes a $k \times 1$ vector parameters, and also $\boldsymbol{\varepsilon}$denotes a $n \times 1$ vector of errors.

E-*step*: in this step contains the estimation of missing data, include in dependent or independent variable, with criteria $t(x) \in$ X, for independent variable, and $t(x) \in$ Y, for dependent variable (X defined as a set of independent variable values, and also Y defined as a set of dependent variable values).

M-*step*: from the assumption above, the maximum likelihood estimator for $\boldsymbol{\beta}$is equal to find the $\widehat{\boldsymbol{\beta}}$ for maximize the equations below

$$L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}ln\sigma^2 - \frac{1}{2\sigma^2}[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})]$$

So, the result is

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \quad [4]$$

Finally, it could be resumed that the main algorithm of E-M algorithm for estimating missing value in the case regression analysis, as follows

1. Find the initial value to fill missing value $\boldsymbol{t}^p(\boldsymbol{x})$
2. Use the initial value for estimating the parameter Φ^p with maximum likelihood estimator function.
3. Count the estimating of $t(x)$ based on the parameter Φ^p and the model in use.
4. Repeat the step 2 and 3 until there are sequens of $\boldsymbol{t}^0(\boldsymbol{x}), \boldsymbol{t}^1(\boldsymbol{x}), \boldsymbol{t}^2(\boldsymbol{x}), \dots, \boldsymbol{t}^l(\boldsymbol{x})$ that fulfill $\left|\boldsymbol{t}^i(\boldsymbol{x}) - \boldsymbol{t}^j(\boldsymbol{x})\right| \leq \alpha;\ i > j;\ \alpha \approx 0$.

3. SIMULATION CRITERIA

The simulation criteria in this paper would be contained in two general frames with 5 replications in each frame. First frame is designed in condition with missing value in independent variable. Simulation designed with 1 independent variable that contains missing value randomly in specify observations.Second frame is designed in condition with missing value in dependent variable. In this case, using model with 3 independent variables with complete data.

Both, in two frames specified, also would be explained the effect of the number of observation and the proportion of missing value in the data. In this paper, the number of observations for this study are $\boldsymbol{n_1 = 10}$, $\boldsymbol{n_2 = 50}$, and $\boldsymbol{n_3 = 100}$, also the proportion of missing value from 10 until 40 pencent.Evaluation from the estimation of missing data using correlation from data complete and data estimated. Simulation design using R software.

4. RESULT AND DISCUSSION

First session will discuss about E-M algorithm for estimating missing value in independent variable using simple linear regression with model

$$\boldsymbol{y = 10 + 9x + e}$$

Result of this simulation presented in table 1. From this table, it is shown that E-M algorithm for estimating the missing value in independent variable is very accurate, the correlation more than 0.99. It also shown that the increasing of missing value percentages effect the decreasing of accuracy from estimated value of E-M algorithm. And there is no effect from the number of observations in the estimating value using E-M algorithm.

Table 1
Correlation Averages for First Frame (Missing Data in Independent Variable)

Number of Observations (n)	Proportion of missing value			
	10%	20%	30%	40%
10	0.999720	0.998447	0.997631	0.996591
50	0.999474	0.998774	0.998093	0.997679
100	0.999413	0.998886	0.997981	0.997528

In the second frame designed with model of multiple regression below

$$\boldsymbol{y = 3 + 6x_1 + 9x_2 + 5x_3 + e}$$

Table 2 presented the result of second frame. From this table, it is also shown that E-M algorithm for estimating the missing value in dependent variable is very accurate, the correlation also more than 0.99. It also shown that the increasing of missing value percentages effect the decreasing of accuracy from estimated value of E-M algorithm.

Table 2
Correlation Averages for Second Frame (Missing Data in Dependent Variable)

Number of Observations (n)	Proportion of missing value			
	10%	20%	30%	40%
10	0.999648	0.999580	0.998422	0.995381
50	0.999608	0.998973	0.998959	0.998309
100	0.999767	0.999189	0.998985	0.998807

So, from the both results above, E-M algorithm is so good for estimating missing value in the data regression analysis, not only in dependent variable, but also in independent variable. But, it also has constrain that missing value in independent variable could be estimated when there is a independent variable in the model.

5. CONCLUSIONS

From the simulation designed in this paper, it could be concluded that although in independent or dependent variables, E-M algorithm seems very powerful for estimating the missing value until 40% the missingness. It could be find from the estimated values are very close to the real value, that the correlation more than 0.99.

Increasing the percentage of missing value in the data, decreasing the accuracy of estimated value using E-M algorithm. Then, it seems no effect for the number of observation to the accuracy of estimation using E-M algorithm.

REFERENCES

1. Haitovsky, Y. (1968). Missing Data in Regression Analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(1), 67-82.
2. Hippel, P.T.V. (2007). Regression with Missing Y's: An Improved Strategy for Analyzing Multiply Imputed Data. *Sociological Methodology*, 37, 1-54.
3. Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1-38.
4. Myers, R.H. and Milton, J.S. (1991). *A First Course in the Theory of Linear Statistical Models*. Boston: PWS-KENT Publising Company.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 375-384

THE CLASSIFICATION OF DISTRICT OR CITY OF THE POVERTY DATA BASED ON INDICATOR OF THE COMMUNITY WELFARE IN SUMATERA WITH THE VERTEX DISCRIMINANT ANALYSIS AND FISHER DISCRIMINANT ANALYSIS

Nurmaleni [1], I Made Sumertajaya[2] and **Bagus Sartono[2]**
[1] Graduate School, Bogor Agricultural University
Bogor, Indonesia. Email: nurmaleni@ymail.com
[2] Department Statistics, Bogor Agricultural University
Bogor, Indonesia.

ABSTRACT

Poverty is a more serious problem for in the world. Classifying the people of poor with appropriate based on the factor causative was a very important issue to support the poverty reduction strategy. Discriminant analysis is one of the methods that used for object classification. One of the latest is vertex discriminant analysis (VDA) which has advantages over Fisher discriminant analysis. VDA is to classify object with the number of variables *p* is larger than the number of observations *n*. Discriminant function is formed similar to regression analysis which is involving $\in$-insentive euclidean distance and a quadratic penalty on the coefficient of the linear predictors. In this study, the performance of the method would be examined to the number of observations *n* which is larger than the number of variables *p*. The performance of Fisher discriminant analysis and Multicategory Vertex Discriminant Analysis were compared by value of APER. The results of simulation showed that Fisher discriminant analysis was better than VDA when the assumption of normality and homogeneity of variance-covariance matrix were satisfied. VDA was to solve multicategory classification problems on the poverty data based on community welfare indicators in Sumatera. Application VDA to the poverty data showed that the analysis was better than Fisher discriminant analysis. The information of this study could contribute to the government in minimizing the problem of poverty in Sumatera.

KEYWORDS

Poverty, Fisher discriminant analysis, multicategory vertex discriminant analiysis.

1. INTRODUCTION

Poverty is a more serious problem in the world. Every country of the world have a poverty reduction agenda, Indonesia is one of them. Indonesia has a long history about poverty. Since the Asian financial crisis in 1997, the number of poor has been a sharp increase[8]. On September 2013, BPS (Badan Pusat Statistik) shows that the number of poor in Indonesia reached 28.55 million or 11,47% [1]. The number of poor had been sharp increase which was compared per Mart 2013. This showed that the number of poor increased by 480,000 for 6 months.

The Indonesia Government had taken various programs to reduce the number of poor, but the problem of poverty in Indonesia could not solve. Errors in classifying the people of poor could lead to poverty alleviation programs was not on target. Classifying the people of poor with appropriate based on the factor causative was a very important issue to support the poverty reduction strategy. One of the method of statistical analysis for classifying object is a discriminant analysis.

Discriminant analysis is a statistical technique for classifying object and allocating a new object into a class. Discriminant analysis generates differentiation of function for separating group. Discriminant analysis is one of statistical technique which is used on response variables as qualitative data and explanatory variables as quantitative data[9]. Discriminant function is linear combination of original variables that will generate the best way in group separation[10]. This function gives value as close as possible in same groups and as far as possible for objects within a group. Discriminant function was first introduced by Ronald A. Fisher in 1936[5]. Fisher discriminant analysis (FDA) can be classified of two categories or more based on some assumptions. FDA cannot classify object with the number of variables *p* that larger than the number of observations *n.* In this case, FDA gives singularity in the variance-covariance matrix that affecting the existence of inverse.

Lange and Wu (2008) introduced a new supervised learning method for multicategory classification. It is called the *vertex discriminant analysis* (VDA). VDA can classify object for the number of variables *p* which is larger than the number of observations *n.* Each *vertex* at VDA represents different categories in each group. VDA classification is performed by minimizing the objective of functions involving ϵ*-insensitive loss* and quadratic penalty. Addition of quadratic penalty has the purpose minimize the estimated value to the zero [3]. The objective function of VDA can be minimized by algorithm Majorize-Minimize [4,6].

In this study, the performance of VDA would be examined with the number of observations n which is larger than the number of variables p. The classification error (APER) of VDA would be compared with APER value of FDA. In addition, the VDA and FDA would be used in resolving multicategory classification problems on the poverty rate of district or city based on indicator of the community welfare in Sumatera.

2. METHODOLOGY

2.1 Data

The performance of Fisher discriminant analysis and vertex discriminant analysis were applied in the poverty data. VDA is ideally suited to handle multicategories and classify object with the number of variables *p* which is larger than the number of observations *n*. FDA gives singularity in the variance-covariance matrix that affecting the existence of inverse. This study, the performance of VDA and FDA compared for the number of observations *n* is larger than the number of variables *p*. The best method has value of APER smallest.

2.2 Applied Data

The poverty data (SUSENAS 2010) was used to test the performance of two classifying methods (Fisher discriminant analysis and vertex discriminant analysis). Predictor variable (X) as percentage of the data population for each district or city of the poverty data based

on indicator of the community welfare in Sumatera. Theory or previous research underlying the poverty level classes based on district or city had not found yet. Therefore, this study would classify the poverty data into three classes based on empiric distribution. District or city which had poverty rates below 7.5% was group 1, districts or cities with poverty rates between 7.5% to 22.5% was group 2 and districts or cities with poverty rates above 22.5% was group 3. Predictor variables in this research were: [2]

X1 : % open unemployment rate
X2 : % labor force participation rates
X3 : % workers who work for less than 14 hours a week
X4 : % workers who work for less than 35 hours a week
X5 : % informal sector employment
X6 : % underweight children
X7 : the infant mortality rate (infant death per 1,000 live births)
X8 : % Births Attended by Skilled Health Personnel
X9 : % Residents with Health Complaints
X10 : % morbidity
X11 : % average length of stay in hospital
X12 : % Residents take good care of themselves
X13 : % population 7-15 years who are not schooling
X14 : primary school enrolment rates (APM)
X15 : secondary education enrolment rate(APM)
X16 : High school enrolment rate (APM)
X17 : % population without access to safe water
X18 : % population without access to basic sanitation
X19 : % life expectancy at birth per years

2.3 Method of Analysis

Steps of data analysis:

1. Procedure for VDA function [7];
   a. Set the iteration counter m=0 and initialize A(0)=**0** and b(0)=**0**
   b. Define $\mathbf{y}_i = \boldsymbol{\upsilon}_j$ if the *i* th subject belong to category *j*, where

$$\boldsymbol{\upsilon}_j = \begin{cases} (k)^{-\frac{1}{2}}\mathbf{1} \text{ jika } j=1 \\ c\mathbf{1}+d\mathbf{e}_{j-1} \text{ jika } 2 \le j \le k+1 \end{cases},$$

$$c = -\frac{1+\sqrt{k+1}}{(k)^{\frac{3}{2}}} \text{ and } d = \sqrt{\frac{k+1}{k}}$$

   c. Majorize the regularized loss function as indicated in

$$R(A,b) \le \frac{1}{n}\sum_{i=1}^{n} w_i \left\|\mathbf{r}_i - \mathbf{s}_i\right\|^2 + \lambda \sum_{j=1}^{k} \left\|\mathbf{a}_j\right\|^2 + d,$$

$$= \sum_{j=1}^{k}\left[\frac{1}{n}\sum_{i=1}^{n} w_i \left(r_{ij} - s_{ij}\right)^2 + \lambda \sum_{j=1}^{k} \left\|\mathbf{a}_j\right\|^2\right] + d$$

   with ith current residual $\mathbf{r}_i^{(m)} = \mathbf{y}_i - \mathbf{A}^{(m)}x_i - \mathbf{b}^{(m)}$,

$$w_i = \begin{cases} \dfrac{1}{2\|\mathbf{r}_i^m\|}, & \text{untuk } \|\mathbf{r}_i^m\| \geq 2\epsilon \\ \dfrac{1}{4(\epsilon - \|\mathbf{r}_i^m\|)}, & \text{untuk } \|\mathbf{r}_i^m\| \leq \epsilon \\ \dfrac{1}{4(\|\mathbf{r}_i^m\| - \epsilon)}, & \text{untuk } \epsilon \leq \|\mathbf{r}_i^m\| \geq 2\epsilon \end{cases}$$

$$\mathbf{s}_i = \begin{cases} \mathbf{0} & \text{untuk } \|\mathbf{r}_i^m\| \geq 2\epsilon \\ \mathbf{r}_i^m & \text{untuk } \|\mathbf{r}_i^m\| \leq \epsilon \\ \left(\dfrac{2\epsilon}{\|\mathbf{r}_i^m\|} - 1\right)\mathbf{r}_i^m & \text{untuk } \epsilon \leq \|\mathbf{r}_i^m\| \geq 2\epsilon \end{cases}$$

d. Minimize the surrogate function and determine A(m+1) dan b(m+1) by solving k sets of linear equestions

e. If $\left\|\mathbf{A}^{(m+1)} - \mathbf{A}^{(m)}\right\| < \gamma$ and $\left|R(\mathbf{A}^{(m+1)}, \mathbf{b}^{(m+1)}) - R(\mathbf{A}^{(m)}, \mathbf{b}^{(m)})\right| < \gamma$ both hold for $\gamma = 10^{-4}$, then stop . otherwise repeat step 3 through 5.

f. After getting the discriminant function to distinguish each group, the classification is done by the formula

$$\hat{y} = \text{argmin}_{j=1,..,k+1} \left\|\mathbf{v}_j - \hat{A}x_i - \hat{\mathbf{b}}\right\|$$

g. Calculate the value of APER

2. Discriminant function with the formation stages [5]
   a. Check assumptions Fisher diskriminan analysis
      1. Check the assumption of normal multivariate distribution using the double-quantile plots chi squared
      2. The variance-covariance matrices of variables are assumed to be homogeneous across groups $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \cdots = \mathbf{\Sigma}_j$ with Box's M
      3. Multicollinearity test
      4. Absence of test outliers
   b. Form Fisher Discriminant function $y = \mathbf{a}'\mathbf{x}$, weighting coefficient vector a discriminant function $\mathbf{a}_{(px1)}$ is feature vector of $\mathbf{W^{-1}B}$, $\mathbf{x}$ is a vector of independent variables identified in the discriminant function.

$$\mathbf{W} = \sum_{j=1}^{k+1}(n_j - 1)\mathbf{S}_j,\ \mathbf{B} = \sum_{j=1}^{k+1}(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})(\bar{\mathbf{x}}_j - \bar{\mathbf{x}}),$$
$$\mathbf{S}_j = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(\mathbf{x}_i - \bar{\mathbf{x}}_j)(\mathbf{x}_i - \bar{\mathbf{x}}_j)'$$

with $j = 1,2,\ldots,(k+1); n = n_1 + n_2 + \cdots + n_j,$
$i = 1,\ldots,n_1, n_{1+1},\ldots,n_2, n_{2+1},\ldots,n_j, n_{j+1},\ldots,n$

   c. Classify new objects based on Fisher linear discriminant equation. Allocate x to group j if

$$\sum_{m=1}^{r}(\hat{\mathbf{y}}_m - \bar{\mathbf{y}}_{jm})^2 = \sum_{m=1}^{r}\left[\widehat{\mathbf{a}_m'}(\mathbf{x} - \bar{\mathbf{x}}_j)\right]^2 \leq \sum_{m=1}^{r}\left[\widehat{\mathbf{a}_m'}(\mathbf{x} - \bar{\mathbf{x}}_h)\right]^2, h \neq k,$$
$$r \leq s, s = \min(k, p)$$

d. Calculate the APER values

e. Compare the performance of two classification methods (VDA and FDA) was presented in figure 1

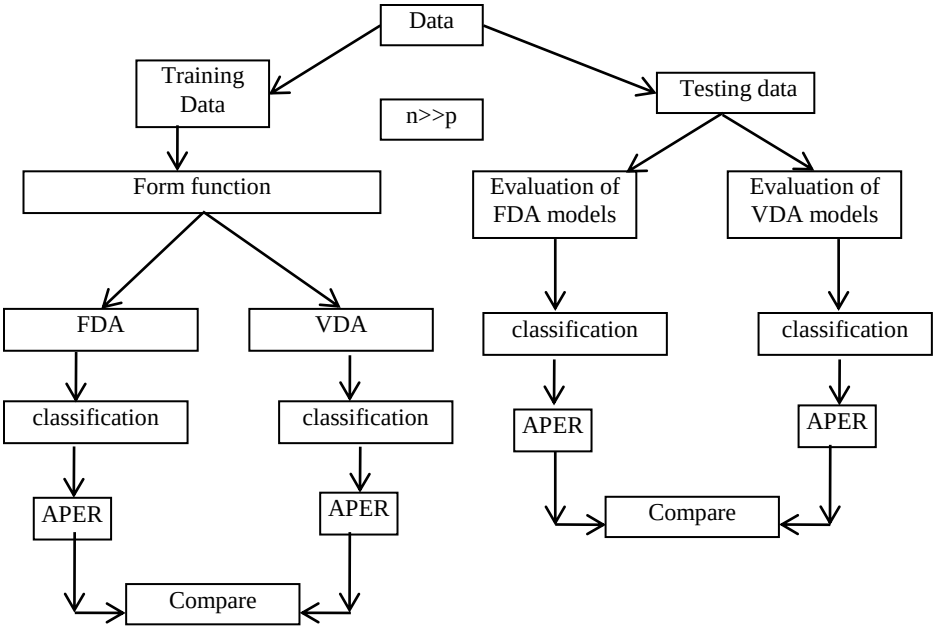


Figure 1: Comparing the performance of VDA and FDA

Steps in VDA for data are:

1. Divide poverty data into 2, training data (70%) and testing data (30%). Training data formed the discriminant function and another evaluated classification error.
2. Form VDA function
3. Classify objects based on discriminant function
4. Calculate the classification error rate based on the APER

3. RESULT AND DISCUSSION

This chapter would apply the data. Lang and Wu (2008) explained that the VDA could classify object with the number of variables *p* that larger than the number of observations *n*. In that condition, FDA showed singularity of variance-covariance matrix. Thus the possibility classification was VDA. The classification error (APER) from both would be compared. The best method had the smallest classification error.

3.1 Data Description

This section, VDA and FDA applied to percentage of the data population for each district or city of the poverty data based on indicator of the community welfare in Sumatera. Initial class of the poverty data was divided into three classes based on empiric distribution which was presented in figure 3. The figure presented normal curve over

histogram of empirical data distribution for the poverty rate of district or city in Sumatera.

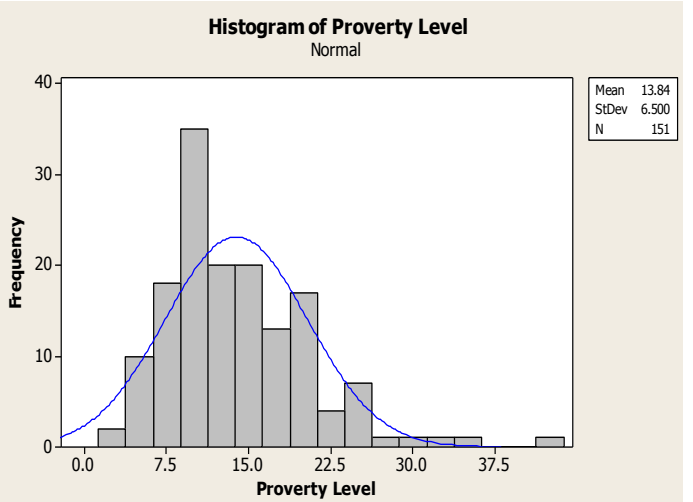


Figure 2: Normal Curve Over Histogram of Empirical Data Distribution for the Poverty Rate of District or City in Sumatera

Based on the figure 2, the data would be divided into 3 classes. Class 1 was 24 districts or cities with poverty rates below 7.5%, class 2 was 112 districts or cities with poverty rates between 7.5% to 22.5% and class 3 was 12 districts or cities with poverty rates above 22.5%. Furthermore, Figure 3 also showed average poverty rate of district or city 13.85, standard deviation 6.5, the number of observation 151 and 19 predictor variables.

Data description for each variable per class of the poverty data based on indicator of the community welfare in Sumatera was presented in Table 1. Table 1 showed that variable X_1 in class 1 was on average 7.86 with a standard deviation of 3.76. The last variable (X_{19}) in class 3 was on average 68.27 with a standard deviation of 1.90. Description of the poverty data based on indicator of the community welfare in Sumatera for each class was presented in Table 2.

Table 1
Data Description for each Variable Per Class

Variable	Class 1		Class 2		Class 3	
	Average	Standard deviation	Average	Standard deviation	Average	Standard deviation
X_1	7.86	3.76	6.42	3.23	5.50	3.56
X_2	64.80	5.09	68.73	7.57	67.54	7.54
X_3	3.79	2.24	4.76	2.60	5.71	2.46
X_4	28.21	12.72	37.32	12.13	44.88	9.29
X_5	50.79	14.37	65.03	15.89	69.39	11.69
X_6	17.48	4.98	20.51	6.80	25.55	7.24
X_7	29.27	5.37	34.07	7.18	35.78	7.59
X_8	83.98	15.46	81.16	14.08	80.36	12.83
X_9	33.61	7.62	32.90	7.71	32.53	5.23
X_{10}	18.63	4.73	18.92	5.31	20.41	3.75
X_{11}	5.37	0.69	5.56	0.86	5.27	0.55
X_{12}	66.29	8.62	72.70	8.91	74.57	8.14
X_{13}	2.15	1.64	2.33	1.43	1.55	1.19
X_{14}	92.34	4.77	94.91	2.82	95.86	2.82
X_{15}	65.45	8.36	70.38	8.41	75.13	5.24
X_{16}	54.82	14.14	50.54	12.42	60.13	8.23
X_{17}	56.62	17.60	46.59	19.73	40.15	10.12
X_{18}	16.80	12.14	28.10	18.99	33.76	16.64
X_{19}	69.56	2.60	68.78	1.84	68.27	1.90

Table 2
Description Data for Each Class

Class	N	Average	Standard Deviation	Max	Min
1	24	5.897	1.249	7.33	2.47
2	112	13.742	3.849	21.68	7.60
3	15	27.324	5.497	42.46	22.62
Total	**151**				

Table 2 showed that the poverty data average was 5.89 for class 1, the poverty in class 2 was on average 13.74, and the average was 27.32 for class 3.

3.2 FDA and VDA Modeling

The first step, the poverty rate of district or city based on indicator of the community welfare in Sumatera is divided into 2 data, training data and testing data. Training data formed the discriminant function and testing data evaluated models.

VDA and FDA formed two discriminant function for distinguishing 3 classes with the number of objects 106. Goodness of fit of FDA and VDA functions were shown by classification accuracy of each class. Those were presented in Table 3.

Table 3
Goodness of Fit of FDA and VDA Functions

Methods	Goodness of fit					
VDA			Predicted membership			
		Groups	1	2	3	Total
	Actual membership	1	11	7	0	18
		2	2	78	0	80
		3	0	4	4	8
		Total	13	89	4	106
FDA			Predicted Membership			
		Groups	1	2	3	Total
	Actual membership	1	10	8	0	18
		2	3	75	2	80
		3	0	4	4	8
		Total	13	87	6	106

Table 3 showed that 88% object was correctly classified by VDA and the apparent error rate (APER) of VDA was 12%. FDA was correctly classified as much as 84%, and the APER was 16%. Concordant and discordant values were used to see the relationship between the predicted and actual membership. Table 3 showed that concordant pairs was formed of 98% (VDA) and 96% (FDA). The relationship between the predicted and actual membership was very strong (perfect correlation). The predicted membership was very influenced by the actual membership.

3.3 Evaluation of the VDA model

VDA model was evaluated the 45 districts or city (testing data). The first object in the testing data (Mentawai) with the initial classification was class 2. The predictor variables were standardized, then they were substituted into two discriminant functions. The value of Y_1 was 0.065385 and Y_2 was -0.13688. The distance between object and vertex were presented in Table 4.

Table 4
The Distance Between Object and Vertex

Class	$\left\|\boldsymbol{v}_j\text{-}\hat{\mathbf{A}}\mathbf{x}_i\text{-}\hat{\mathbf{b}}\right\|$
1	1.050287
2	0.850226
3	1.103459

Based on the minimum distance between the object and three vertex, Mentawai was classified into class 2. Table 5 showed the VDA classification for all testing data.

Table 5
Summary of the VDA Classification for all Testing Data

	Predicted Membership				
	Groups	**1**	**2**	**3**	**Total**
Actual membership	1	2	4	0	6
	2	4	27	1	32
	3	0	5	2	7
	Total	6	36	3	45

Table 5 showed that 31 objects classified correctly for testing data with the APER value 31.1%. Table 5 showed that concordant pairs was formed of 77%. The relationship between the predicted and actual membership had correlation. The predicted membership was very influenced by the actual membership.

3.4 Evaluation of the FDA model

Similar to VDA, FDA models were evaluated the 45 districts or city (testing data). Table 6 showed the FDA classification for all testing data.

Table 6
Summary of the FDA Classification for All Testing Data

	Predicted Membership				
	Group	**1**	**2**	**3**	**Total**
Actual membership	1	4	9	1	14
	2	1	12	0	13
	3	1	11	6	18
	Total	6	32	7	45

Table 6 showed 22 objects that correctly classified for testing data with the APER value 51.1%. Concordant pairs was formed of 65%. The relationship between the predicted and actual membership had correlation. The predicted membership was very influenced by the actual membership.

4. CONCLUSION

The assumption of normal multivariate distribution on the poverty data based on indicator of the community welfare in Sumatera was not satisfied. The previously analysis was concluded that VDA was better than the FDA in classification. In the training data, the APER value of VDA (12%) had smaller than the FDA (16%). Whereas in the testing data, the result gave that VDA was better than the FDA in classification. It was showed that the APER value of VDA was 31.1% and 51.1% for FDA.

In the training data, base on empirical data distribution for the poverty rate of district or city in Sumatera, 24 districts or city had poverty rates below 7.5%, 112 districts or city had poverty rates between 7.5% to 22.5% and 12 districts or city had poverty rates above 22.5%. However, VDA produced 13 districts or city had poverty rates below 7.5%, 91 districts or city had poverty rates between 7.5% to 22.5% and 3 districts or city had poverty rates above 22.5%.

REFERENCES

1. Berita Resmi Statistika (2014). Profil Kemiskinan Di Indonesi September 2013. Www.Bps.Go.Id/Brs_File/Kemiskinan-02 January 2014.
2. BPS (2010). Survey Social Ekonomi Nasional Tahun 2010. Jakarta: BPS.
3. Hastie T, Tibshirani R, and Friedman J. (2008). The Elements of Statistical Learning. *Data Mining, Inference, and Prediction*. Ed. Ke-2. Springer. http://www.stanford.edu/~ hastie/pub.htm.
4. Hunter, D.R. and Lange (2004). Tutorial on MM Algorithms. *The American Statistician*. 58, 30-37.
5. Johnson RA and Wichern DW (2007). *Applied Multivariate Statistical Analysis*. New Jersey: Pearson Prentice Hall. Ed ke-6.
6. Lange K. (2004). *Optimization*. New York: Springer-Verlag.
7. Lange K and Wu TT (2008). An MM Algoritm For Multicategory Vertex Discriminant Analysis. *J Comput Graph Stat*., 17, 527-544.
8. Mardianto, S. (2014). Kemiskinan Di Indonesia. Universitas Syiah Kuala.
9. Mattjik AA and Sumertajaya IM. (2011). Sidik Peubah Ganda. Bogor (ID): IPB Press.
10. Rencher AC (2002). *Methods of Multivariate Analysis*. 2nd Edition. New York: Wiley.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 385-392

INSTITUTIONAL FRAMEWORK OF POVERTY REDUCTION IN PAKISTAN

Nadeem Iqbal
Faculty of Management Sciences Ghazi University
DG Khan, Pakistan

and

Rashda Qazi
Faculty of Social Sciences Ghazi University
DG Khan, Pakistan

ABSTRACT

This paper focused on institutional framework of poverty reduction programme in Pakistan. Poverty in multidimensional problem and it is also addressed by multidimensional strategies to overcome the poverty. Different strategies are adopted by Government of Pakistan to facilitate the deprived and poor people of the country. In direct strategies, Government provide money assistance and finance for human capital development to enhance the income through income generating activities. In this study primary data is collected from the beneficiaries of poverty reduction programme and non-beneficiaries. It is concluded from the data that human capital development is basic tool to pull out the people from the poverty trap. Zakat institution contributed in poverty reduction by providing the finance for human capital development to organized institution effectively. Other programme including micro-financing are not contributing effectively in poverty reduction.

INTRODUCTION

Pakistan is developing country and facing higher poverty ratio in the country. People face lack of food, shelter, education facility and other basic need. Literacy rate is low, unemployment ratio is high and economic condition is not stable in the country. Institutional framework is very weak to overcome the poverty. A study reported that poverty is originated due to institutional structure of society and Pakistan's state (Hussain 2008). That's why institutional change is necessary (change in both polity and economy) to overcome the poverty. In this paper the only focus is made on the institution as a base of sustainable development in Pakistan. By stable democratic

order, Pakistan has ability to attain the institutional condition for maintained and fair growth like equal opportunities in both polity and economy for all poor individuals, hiring and selection on the base of merit, healthy competition, effectiveness and efficiency, and right of innovative creator. With low productivity farmers are facing low productivity resources because of lower income and poverty is going to be ever increasing with the increase in numbers of children. Such circumstances entail that poor person is confronting low-productivity-income poverty trap. This trap can be burst by maximizing the production of poor person and by providing worth full resources. These requirements cannot be attained without sufficient funds which is the main problem for the poor. These resources might be used to start a new business to increase the income. These funds are also utilised for human capital development to enhance the income generating activities by providing professional training and skill to deprived and poor.

In Pakistan people have tendency to do something creative because creativity can become a cause of success of a nation. But due to lack of necessities of life like food, health, education, earning opportunities people are not able to utilize their abilities in a productive way. Hurdles in the way of human development shows poverty comes when a poor man in a split community is interlocked in to a power of groups. A poor individuals face institutions, markets and power holder groups which deprive them to access the different resources (A. Hussain, 2000). Different institutions are working for this purpose. Such institutions take some actions for rehabilitation of poor individual. To some extent these institutions are successful in achieving their goals but till now poverty prevail in Pakistan.

These institutions make huge expenditures for the health of the poor by giving donations to hospitals. Such hospitals provide free medical facilities to those who cannot afford expensive treatment and also provide free medicines. Different institutions adopt different ways to remove poverty from society like donation, zakat, hiba, education, training, interest free loans to start a business, professional trainings to get earning, medical facilities, interest free house financing to habitats the poor individual. In D.G.Khan like other cities of Pakistan NGO started a scheme to provide professional skill to poor ladies for starting a small scale business, after this training NGO provide interest free loan to start that business. This scheme worked successfully in Pakistan.

The purpose of this paper is to investigate the effectiveness of the poverty reduction institution in rural Pakistan. Although different institutions adopt different strategies for poverty reduction but the outcome of these strategies are uncertain and vague. This paper fulfils the gap by providing empirical evidence for measurement of effectiveness of these institutional frameworks for poverty reduction.

LITERATURE

Literature on poverty is very enriched as numerous studies are conducted on poverty. Most studies calculated the poverty ratio in the country and poverty line as well. As poverty is multidimensional having different facet so different causes of poverty are concluded and different strategies are suggested. In 2001 a survey was conducted by NHDR (National Human Development Report) which shows that 51% of tenants are engaged in getting debt from landlord and instead of this 57% are working on the farm of landlord without wage, only 14% labours are working for wage at a very low rate as compare to the rate of market (Hussain et al., 2003). This dominancy and power structure of state become the cause of deformation in labour and capital market which results in deprivation of poor's actual income. This unequal distribution of human and capital resources create hurdles in the growth of agriculture sector (Hussain et al., 2003). According to one study by Harbison (1964), there are several determinations in which human capital is merged with maintainable competitive edge and high performance. Investment expenditure by different institutions like NGO's and VTI on education and professional trainings can enhance the capabilities of human for earnings and also establish human capital stock (Harbison (1964),. This experience is also learnt in East Asia Centre and the secrete behind their success is human capital stock. Higher income and productivity can be achieved by capable human capital stock (Mincer 1997). For this purpose need of financing is fulfilled by zakat and micro financing institutions.

Zaman, Hassan (2000) find out the relationship between micro credit and the reduction of poverty. Meyers (2002) analyse a case study and indicate the flexibility of the microfinance product. Jansson, Torr *et al* (2003) determine indicants mentioning microfinance performance. Conroy (2003) indicate the threats faced by MFIs in South East Asia. Nghiem, H.S and J.Laurenuson (2004) find out the working capability of MFIs in Veitnam evidences from NGOs schemes. Qayyum and Ahmad (2004) find out the working capability of MFIs by using four years data and results indicate that MFIs play an important role in poverty alleviation. Stephens *et al* (Dec 2005) analyse the performance and transparency of MFIs in South Asia. Rehman W (2007) depicted roadblocks in the performance of the MFIs outreach of women in Pakistan. Haq et al (2008) corresponds the MFIs' regulatory structure. This research finds out the policy and supervision of Micro finance. Iqbal (2014) concluded Islamic financing as a basic tool for poverty reduction.

DATA SOURCE AND DATA PERIOD

The data is collected from rural households including 500 beneficiaries of institutions and 500 non-beneficiaries through questionnaire in DG Khan and Rajanpur districts.

MODELLING

The study focuses on determining the impact of institutions in managing poverty reduction programmes and to measure the determinants of rural poverty in Pakistan. The logit model has been used for the analysis of the data. In logit or binary logistics models binominal distribution is used for handling the errors associated with regression models for binary/dichotomous responses (i.e. 1 for poor and 0 for non-poor). The logistic model (logit model) is widely used and has many desirable properties (McCullagh and Nelder 1989).

Hence the Binary logistic model becomes a convenient choice when response variable can take only one of two values either 1 or 0 with formulation

$$P(y=1)= F (\beta X)$$

where X is a matrix of explanatory variables, β is vector of parameter and F(.) can be regarded as the c.d.f of a standard logistic distribution.

The goal of logistic regression is to find the best fitting model describes the relationship between dichotomous characteristics of interest (dependent variable 1 for poor and 0 for non-poor) and a set of independent variables. Logistic regression generates the coefficient to predict a logit transformation of the probability of a presence of a characteristic of interest (i.e. being poor).

$$\text{Logit } (P) = \text{Ln } (P/1-P)$$

where P is the probability of presence of characteristics of interest. The logit transformation is defined as the logged odds

$$\text{Odds}= P/1-P \text{ and}$$

$$\text{Logit } (P) = \beta_0+\beta_1X_1\beta_2X_2+\beta_3X_3\text{------ } \beta_kX_k$$

Rather than choosing parameters that minimize the sum of sequares of errors (OLS), estimation in logistic regression chooses parameters that maximize the

likelihood of observing sample values. The study uses,

P = Poverty Status (probability of being non-poor)
X_1 = Districts
X_2 = Beneficiaries
X_3 = Institutions
X_4 = Training
X_5 = Gender
X_6 = Employment Status
X_7 = Education
X_8 = Marital Status
X_9 = Age
X_{10} = Total Household Member
X_{11} = Child dependency
X_{12} = Old dependency
X_{13} = Working member Female
X_{14} = Working member Male
X_{15} = Value of animal
X_{16} = Own Land
X_{17} = Cultivated Land
X_{18} = Business assets
X_{19} = Water
X_{20} = Change in income
X_{21} = Source of Change
X_{22} = Amount Change
X_{23} = Saving
X_{24} = Loan

DATA ANALYSIS

Binary logit is used to draw the result from the raw data. Financial poverty line is calculated to determine the poverty status of household. Poor is coded 0 and non-poor is coded 1 to summarise the result in the binary logit model.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
	District	-.229	.286	.639	1	.424	.795
	Beneficiaries	4.635	1.402	10.926	1	.001	103.056
	Institution	-.769	.312	6.100	1	.014	.463
	Training	-1.553	1.160	1.792	1	.181	.212
	Gender	-1.515	.575	6.951	1	.008	.220
	EmploymentStatus	-1.571	.332	22.370	1	.000	.208
	Education	-.630	.114	30.323	1	.000	.533
	Maritalstatus	.956	.492	3.781	1	.052	2.601
	Age	-.509	.208	5.965	1	.015	.601
	TotalHHMembers	1.334	.181	54.377	1	.000	3.795
	ChildDep	.005	.211	.000	1	.982	1.005
	OldDep	.477	.280	2.909	1	.088	1.611
Step 1[a]	WorkingMembersFemales	-1.709	.525	10.587	1	.001	.181
	WorkingMembersMales	-1.343	.221	36.932	1	.000	.261
	ValueOfAnimals	.041	.064	.409	1	.522	1.042
	OwnLand	-.470	.437	1.158	1	.282	.625
	CultivatedLand1	.125	.449	.077	1	.781	1.133
	BusinessAssets	.184	.098	3.536	1	.060	1.202
	Water	.235	.157	2.229	1	.135	1.265
	ChangeInIncome	-.598	.207	8.316	1	.004	.550
	SourceOfChange	-.243	.128	3.573	1	.059	.785
	changeAmount	.075	.071	1.095	1	.295	1.078
	SavingForEmergency	-.039	.836	.002	1	.963	.962
	Loan	-.075	.237	.099	1	.753	.928
	Constant	1.815	.941	3.720	1	.054	6.142

a. Variable(s) entered on step 1: District, Beneficiaries, Institution, Training, Gender, EmploymentStatus, Education, Maritalstatus, Age, TotalHHMembers, ChildDep, OldDep, WorkingMembersFemales, WorkingMembersMales, ValueOfAnimals, OwnLand, CultivatedLand1, BusinessAssets, Water, ChangeInIncome, SourceOfChange, changeAmount, SavingForEmergency, Loan.

RESULT AND DISCUSSION

The empirically result of logit model based on cluster sample identified that beneficiaries, marital status, total member of household, old dependency, business assets, are the determinants those are significantly and positively correlated with probability of being poor in the sample area. It is also empirically evaluated from sample data that institution, gender, employment status, education, age, working male member, working female member, change and source of income are statically significant and negatively correlated with probability of being poor as shown in table. While variables including district, training, child dependency, livestock, own land,

cultivated land, water facility, change amount, saving and loan have the correct sign but are statically insignificant which may be observed from table .

It is very clear from the empirical result that human capital is basic instrument to reduce the poverty level in rural Pakistan. Education is basic factor for human capital development which is significant and negatively correlated with probability of being poor as shown in table. Thus, education is very important and vital factor for managing the poverty reduction in rural Pakistan. In addition, education also provide more chances for regular employment to male and female which also increase the income level of household and reduces the poverty level as well. It is empirically proved that education; employment, working female member and working male member of household are statically significant and negatively correlated with probability of being poor as Table revealed. Education of male and female provides them more opportunity to be employed and to earn higher income and to reduce the poverty level. Training has also negative sign with poverty. Trained person has more chance to be employed and productive and to earn higher income. So education enhances the human capital which leads to have higher chance of employment which negatively correlated with probability of being poor. Due to importance of human capital, many institutions are working for free education and training to fight against poverty in Pakistan.

It is very interesting that micro financing is not statically significant in this model as shown in table. It may be that the usage of that loan may be unproductive so monitoring and feedback is essential to achieve the desired result of micro financing. As the loaned funds could be used for consumption purpose under the pre-text of micro financing. So there is need of further research to know the causes about micro financing why is not contributing in poverty reduction in study area. This result is alarming because most of studies concluded that micro financing is basic element to get the poor out of the poverty trap as mentioned and discussed in earlier chapter. Many institutions under government arrangement and NGOs are working for managing the poverty reduction through micro financing. Thus, it is very important to research and carefully evaluate the reason about the failure to contribute in poverty reduction.

REFERENCES

1. Conroy, J.D. (2003). *Challenges of Microfinancing in South East Asia*. Financing South East Asia's Economic Development, Singapore.
2. Haq, M., Hoque, M. and Pathan, S. (2008). Regulation of Microfinance Institutions in Asia: A Comparative Analysis. *International Review of Business Research Papers* 4(4), 421-450.
3. Hussain, A. (1998). *Punjab Rural Support Programme* (PRSP), The First Four Months, Report to the Board of Directors of PRSP.

4. Hussain, A. (2003). A Policy for Pro Poor Growth, paper in Towards Pro Poor Growth Policies in Pakistan, *Proceedings of the Pro-Poor Growth Policies Symposium*, 17th March 2003, UNDP-PIDE, Islamabad.
5. Hussain, A. (2008). *Power Dynamics, Institutional Instability and Economic Growth: The Case of Pakistan, Drivers of Change Study 2007-08*, The Asia Foundation, Islamabad (Mimeo).
6. Von Stauffenberg, D., Jansson, T., Kenyon, N. and Barluenga-Badiola, M-C. (2003). *Performance Indicators for Microfinance Institutions*, Technical Guide. Washington D.C: Inter American Development Bank (IDB), Micro, Small and Medium Enterprise Division.
7. Meyer, R.L. (2002). The Demand for Flexible Microfinance Products: Lessons from Bangladesh. *Journal of International Development*, 14, 351-368.
8. Nghiem, H.S. (2004). *Efficiency and Effectiveness of Microfinance in Vietnam: Evidence from NGO Schemes in north and Central Regions*. CEPA, School of Economics, UQ, Australia.
9. Nghiem, H.S. and Laurenceson, J. (2004). *The Nature of NGO microfinance in Vietnam and Stakeholders Perception of effectiveness*. JEL-O12, O16, O17, P34, R29, UQ, Australia.
10. North, Douglass C. (2005). *Understanding the Process of Economic Change*, Princeton University Press, Princeton, N.J.
11. North, Douglass C., Wallis, John Joseph and Weingast, Barry R. (2006). *A Conceptual Framework for Interpreting Recorded Human History*, National Bureau of Economic Research, Working Paper Series, Cambridge (Mimeo).
12. *Pakistan Microfinance Network* retrieved from www.microfinanceconnect.info
13. Qayyum, A. and Ahmad, M. (2004). *Efficiency and Sustainability of Microfinance Institutions in South Asia*, PIDE, Islamabad.
14. Rehman, W. (2007). *Barriers to Microfinance Outreach for Women in Pakistan.* SLU, Institution for Ekonomi, Uppsala.
15. Stephens, B., Tazi, H., Ahmed, S., Mali, P., Wijesiriwardana, I., Athapattu, A. and Sa-Dhan (2005). *Performance and Transparency: A Survey of Microfinance in South Asia*. Washington, D.C, Microfinance Information Exchange, Inc. (MIX).
16. Zaman, H. (2000). *Assessing the poverty and Vulnerability Impact of Microcredit in Bangladesh: A Case Study of BRAC*. Washington, D.C.: World Bank.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 393-402

APPLYING SEM TO ANALYZE THE RELATIONSHIP BETWEEN LOYALTY, TRUST, SATISFACTION, AND QUALITY OF SERVICE FOR STUDENTS IN MATHEMATICS STUDY PROGRAM, STATE UNIVERSITY OF MAKASSAR

Sukarna, Aswi and **Sahlan Sidjara**
Faculty of Mathematics and Science, State University of Makasssar, Indonesia
Email: karne74@gmail.com; aswikarne@yahoo.com

ABSTRACT

Generally, students of Mathematics Study Program at State University of Makassar (UNM) still have not received optimal services, such as, the services for Practical Work (PKL=Praktek Kerja Lapangan). However, the study program has been providing them with specialized services for PKL. Variables that included in this study were service quality, trust, satisfaction, and loyalty. Various empirical studies show that service quality has a positive influence on satisfaction, and the satisfaction has a potential in building loyalty. This study aimed at testing (confirmatory) the model of the relationship between service quality, trust, satisfaction, and loyalty experienced by the students of Mathematics Study Program at UNM by employing Structural Equation Modeling (SEM).

KEYWORDS

SEM, Service Quality, and Loyalty.

1. INTRODUCTION

The development of non educational study programs at the State University of Makassar as a former Teachers' Tra ining College still has not shown significant progress. For example, until now, there is no institution at the university specificly facilitates Practical Work (PKL=Praktek Kerja Lapangan).It is different from the educational study programs for which the university has established the Unit for Field Experience Implementation (UPPL = Unit Pelaksana Pengalaman Lapangan) to organise Field Teaching Practice. For non educational study programs, PKL is managed individually by the study program, respectively. Accordingly, Mathematics Study Program has made its own initiative in managing PKL. It has been providing students with services on PKL debriefing and location setting. With special provisions, they do not need to bother finding the location. It is one of the services provided by Mathematics Study Programfor the students.

As a service organization, the main mission of a higher education institution is providing an excellent service. Starting from the admission, orientation, and ongoing academic activities in the university until graduation, students exprience dynamic nature of service where there is no standardized services resulting in some discrepancy in the service quality for the students. Satisfactory service will positively influence behavioral consequences (Indahwati, 2008).Various empirical studies show that service quality has positive influence on satisfaction, and satisfaction has the potential to build loyalty.

Therefore, the quality of service and satisfaction were investigated in terms oftheircontribution to the building loyalty to the institution(Indahwati, 2008). Students' trust to higher education institutions also affect their loyalty. Once they trust the institution, students will depend themselves on the institution and be have strong commitment to the built relationship. Commitment will grow an intention to maintain the relationship, which is represented through the loyalty to the institution.Chaudhuri and Holbrook find that trust is the antecedent of loyalty (Indahwati, 2008). Therefore, the contribution of trust also be investigated.

In analyzing the effect of service quality on loyalty of students, Structural Equation Modeling (SEM) was utilized. SEM is a combination of factor analysis, path analysis, and regression analysis (Bollen, 1989; Santoso 2007; Tiro, 2010). SEM is not only used to measure the relationship between several independent variablesand the dependent variable, but it is also used to create a graphical modeling to allow users to read the output analysis and to estimate the SEM models.This studyconducted in Mathematics Study Program aimed to test and analyze (1) the effect of service quality on satisfaction; (2) the effect of service quality on trust; (3) the effect of satisfaction on trust; (4) the effect of satisfaction on loyalty; and (5) the effect of trust on loyalty.

2. LITERATURE REVIEW

2.1. Service Quality

One indicator of professional university management is the ability of an institution to provide quality public services. Parasuraman (Manurung, 2007) argues that the service is satisfyingif the the quality of received service equals or exceeds that of the expected service. This means that there are two main factors affecting the quality of service, that is, expected service and perceived service. So the customer assessment of the quality of service depends on the capability of service providers.

According to Parasuraman (Indahwati, 2008), the quality of service has five dimensions:

a. Reliability. It is the ability of companies to provide services as promised fast, precisely, accurately, and reliably. Performance shall be in accordance with the expectations of customers.
b. Responsiveness. It can be described as a willingness to help and provide appropriate services for consumers.
c. Assurance. It is a dimension of quality of service that focuses on the ability to gain trust and belief of customers.
d. Empathy. It is the dimension of service quality that emphasizes on treating consumers with personal manner, including the ease in establishing relationships, good communication, personal attention and understanding of the needs of the individuals.
e. Physical evidence. It is the dimension of quality services that represents the physical facilities related to the ability of a company to demonstrate its existence to external parties.

Main benefits of using these five dimensions were proved empirically in a variety of research settings. However, the instrument requires adaptation, according to the context of the services under investigation (Bloemer, Ruyter & Wetzels, 1998).

2.2. Satisfaction

In general, customer satisfaction is determined by whether or not the customer's expectations are met. Customer satisfaction, according to Tjiptono (Manurung, 2007: 33), is an emotional response to the evaluation of the observation of a product or service consumption. Meanwhile, customer satisfaction as defined by Day (Manurung, 2007: 33) is the customer response to the evaluation of the discrepancy or disconfirmation perceived between prior expectations and actual performance of a product after being consumed.

This study used a conceptualization of satisfaction expressed by Oliver (1999: 34): "satisfaction is described as the fulfillment of pleasure which is the desire of consumers to meet the needs, desires, or as the fulfillment of pleasure."

2.3. Trust

Worchel (in Indahwati, 2008: 11) defines trust as an individual's willingness to rely on another party with certain risks. Trust is also defined as an individual's willingness to rely on another parties involved in the exchange because people have faith in others. These ideas emphasize the element of willingness and confidence in the trust. Morgan and Hunt (Indahwati, 2008: 12) argues that when one party has confidence that the other party involved can be trusted and has ability, it can be claimed that the trust exists.

2.4. Loyalty

Assael (1998: 130) defines loyalty as a good attitude towards a brand that generates loyalty from time to time. Meanwhile, Mowen and Minor (Indahwati 2008: 5) proposes a definition of loyalty as a condition in which consumers have a positive attitude towards a brand, a commitment to the brand, and an intention to continue purchasing in the future. Meanwhile, Boulding (Indahwati 2008: 6) suggests that the consumer brand loyalty is caused by the influence of satisfaction/dissatisfaction with the brand accumulated continuously in addition to their perception of the quality of the product. According to Zeithaml (Japarianto, 2007: 3), the ultimate objective of an organization is to establish a strong loyalty characterized by: saying positive things, recommending to friends, and continuing purchasing.

2.5. Conceptual Framework

In this research, service quality is the first construct. This construct is measured by using the dimensions developed by Parasuraman covering reliability, responsiveness, assurance, empathy, and physical evidence. The service quality influences satisfaction. If service quality received exceeds the expectation of the students, then the service quality will be perceived as an ideal quality service which eventually results in the students' satisfaction. However, if the quality of the received service is lower than the expected, then the service quality will be perceived as bad which lead to dissatisfaction.

The service quality affects trust. With good service quality, then the trust will develop to an institution. When a party has faith that other parties involved could be trusted and have capability, then the trust exists. Thus, the service quality directly influences the

satisfaction and trust, but does not directly influence the loyalty. With good service quality, the satisfaction or trust will grow and further will influence the loyalty.

The second construct is the satisfaction. It is defined as an emotional response to the evaluation of the observation of a product or service consumption. The satisfaction is measured according to two indicators developed by Oliver (1997: 34), namely, expectation and perception. Both constucts relate to the students' expectation to fulfil their needs and desires. Someone's satisfaction will affect their loyalty and improve their trust to an institution. It means that if someone feels the satisfaction to an institution, then they will be loyal to bound in honour to the institution. The satisfaction will be likely to improve the trust so that they will have good loyalty.

The third construct in this study is the trust. The level of the trust indicates the willingness of individuals to depend themselves on other parties involved in the instituion as they have faith to the parties. If faith exists, then the readiness will also exist. Conversly, if faith does not exist, then the readiness cannot be expected.

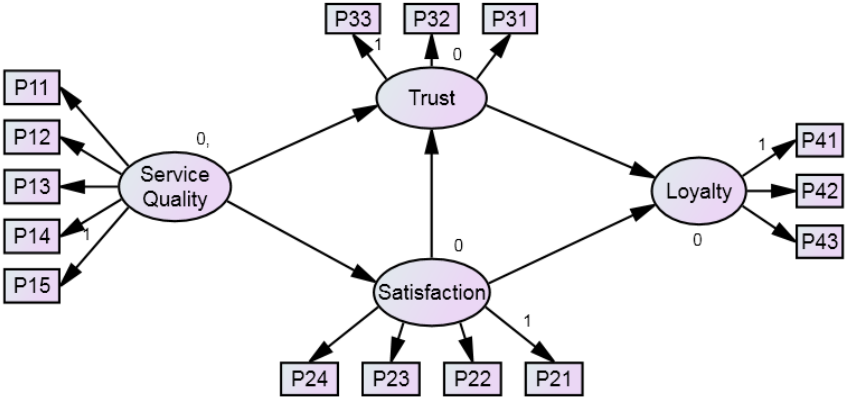
The fourth construct is the loyalty showing the final results of the achievement level attained by a hgiher education institution becuase of the service quality, satisfaction, and the students' trust. Loyalty is measured by using three indicators, namely, saying positive things, recommending to friends, and continuing purchasing.

2.6. Hypotheses

Several hypotheses are: (1) the quality of service significantly influences the students' satisfaction; (2) the quality of service has a significant effect on the students 'trust; (3) the students' satisfaction significantly affects the trust of the students; (4) the students' satisfaction has a significant effect on the loyalty of the students; and (5) the students' trust has a significant effect on the loyalty of the students.

3. METHODOLOGY

This research was conducted at Mathematics Study Program, Faculty of Mathematics and Natural Sciences, State university of Makassar. This study is a confirmatory study that explains the relevance of some of the variables that have been defined. The shape of the path tested is as follows:



The data in this study are primary data from respondents gathered by distributing questionnaires and trough Focus Group Discussion (FGD) for four variables involved, namely, the quality of service, trust, satisfaction, and loyalty. The population is all students of Mathematics Study Program. By using random cluster sample, students of Cohort 2012 and 2013 were selected.

3.1. Operational Definition of Variables

1. Quality of Service (X1) is the ability of Mathematics Study Program to provide quality service able to satisfy the students. Indicators related to the quality of service are as follows:
   a. Reliability (P11) is Mathematics Study Program's ability to provide services as promised fast, precisely, accurately, and reliably.
   b. Responsiveness (P12) is Mathematics Study Program's willingness to help and provide appropriate services for students.
   c. Assurance (P13) is Mathematics Study Program's ability to gain the trust and belief of the students.
   d. Empathy (P14) is Mathematics Study Program's ability to provide quality service emphasizing on the treating students with personal manner.
   e. Physical evidence (P15) is Mathematics Study Program's ability to demonstrate its existence to external parties.
2. Satisfaction (X2) is the result of student assessment to the overall services provided by the study program. Indicators related to the satisfaction are as follows:
   a. Satisfaction with lecturers (P21) is the feeling of being satisfied with the ability of lecturers to master of the lesson material.
   b. Satisfaction with assessment (P22) is the feeling of being satisfied with the scoring done objectively.
   c. Satisfaction with the quality of service (P23) is the feeling of being satisfied with the services provided by the study program.
   d. Satisfaction with the facilities available (P24) is the feeling of being satisfied with appropriate facilities in the study program.
3. Trust (X3) is the belief that Mathematics Study Program is reliable and has high integrity. Indicators related to the trust are as follows:
   a. Belief in intellectual abilities (P31) is the students' belief that they will be able to compete with students from other universities.
   b. Belief in the opportunity of being employed (P32) is the belief of the students that they will have a job after graduation.
   c. The belief in the reliability of Mathematics Study Program (P33) is the belief of the students that Mathematics Study Program is reliable in providing good services.
4. Loyalty (Y) is the trust of the students to continue their studies in Mathematics Study Program. Indicators related to the trust are as follows:
   a. *Saying positive things* (P41) is saying something positive about Mathematics Study Program and always think positively.

b. *Recommending to friends* (P42) is the attitude of the students recommending Mathematics Study Program to others and always inviting other students to comply with all regulations in the program.

Continuing purchasing (P43) is the behavior of students satisfied with Mathematics Study Program, which is manifested in a positive behavior.

4. RESULTS

Based on data, service quality perceived by the students of Mathematics Study Program is shown in Table 1.

Table 1
Descriptive Statistics of Service Quality

Source	P11	P12	P13	P14	P15
Mean	14.42	15.78	8.02	14.48	14.02
Median	15.00	16.00	8.00	15.00	14.00
Mode	16	17	8	16	14
Std. Deviation	2.610	2.043	1.271	2.346	2.796
Variance	6.812	4.173	1.616	5.505	7.818
Skewness	-.354	-.341	-.399	-.420	.239
Kurtosis	-.613	-.186	-.438	-.070	-.196

Based on Table 1, in general, X1 (quality of service) has a negative skewness, which means that students' perception of service quality tends to be better. Although, the value of kurtosis is negative but still close to zero. It means that the distribution is approaching the normal curve. Therefore, the quality of services provided by Mathematics Study Program to students is classified as positive category (excellent).

Table 2
Descriptive Statistics of Satisfaction

Source	P21	P22	P23	P24
Mean	15.37	13.72	10.98	14.84
Median	15.50	14.00	11.00	15.00
Mode	16	15	10	13[a]
Std. Deviation	1.916	2.789	1.880	2.260
Variance	3.670	7.779	3.535	5.105
Skewness	-.131	-.192	-.129	.031
Kurtosis	.373	-.723	-.150	-.762

a. Multiple modes exist. The smallest value is shown

In general, X2 (satisfaction) has a negative skewness, which means that students are likely to feel satisfied by the services provided by Mathematics Study Program. Although, the value of kurtosis is negative but still close to zero. It means that the distribution is approaching the normal curve. Therefore, the satisfaction of students receiving services from Mathematics Study Program is in positive category (excellent). Descriptive results of trust can be seen in Table 3.

Table 3
Descriptive Statistics of Trust

Source	P31	P32	P33
Mean	11.47	8.04	12.58
Median	12.00	8.00	12.50
Mode	12[a]	9	12
Std. Deviation	2.032	1.230	1.707
Variance	4.130	1.514	2.913
Skewness	-.295	-.410	-.127
Kurtosis	-.497	-.655	-.964

a. Multiple modes exist. The smallest value is shown

In general, X3 (trust) has negative skewness, which means that the student believes that the service will shape them into better quality. Although, kurtosis value is negative but still close to zero. It means that the distribution is approaching the normal curve. Therefore, the services provided by Mathematics Study Program can make students believe that they will succeed in their study.

Table 4
Descriptive Statistics of Loyalty

Source	P41	P42	P43
Mean	16.06	14.72	17.18
Median	16.00	15.00	17.50
Mode	16	15[a]	18
Std. Deviation	1.885	2.861	1.806
Variance	3.552	8.183	3.260
Skewness	.106	.336	-.558
Kurtosis	.136	1.498	-.402
Std. Error of Kurtosis	.478	.478	.478
Range	8	17	7

In general, Y (loyalty) has a positive and negative skewness with values close to zero (still resemble the normal curve), meaning that the students still tend to have less loyalty to the mathematics study program. Moreover, the kurtosis value is positive but it is still close to zero. It means that the distribution is approaching the normal curve. Therefore, students still do not show positive or negative loyalty to mathematics study program. So, it can be concluded that the loyalty of students still do not meet the service expectation of the organization.

5. DISCUSSION

The model tested in this study is the relationship between the four variables, namely, service quality, satisfaction, trust, and loyalty. These four variables are translated into exogenous variables (service quality), intervening variables (satisfaction and trust), and endogenous (loyalty). The shape of the model and the results of the analysis of the model are given in Figure 1.

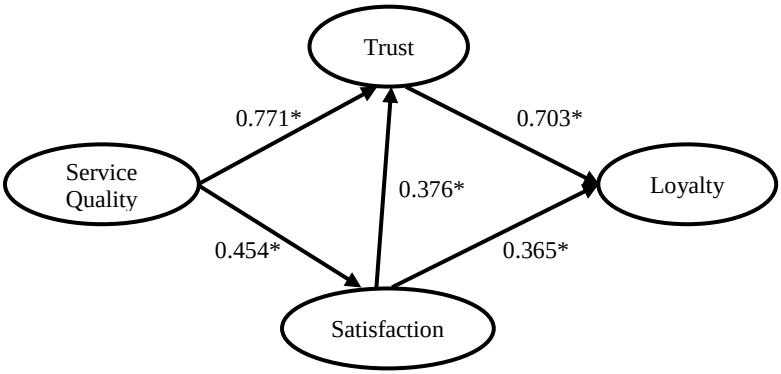


Figure 1: Relationships Simultaneously All Variables (p-value = * < 0.05)

Figure 1 describes the results of standardized effect of X1 (Service Quality) to X2 (satisfaction), that is, 0.454 and X1 to X3 (Trust), that is, 0.771. It means that, the effect of X1 to X3 is greater than the effect of the X1 to X2. Therefore, service quality has more contribution in growing the students' confidence to be successful than that of the satisfaction in receiving services. Similarly, the effect of standardized for Y (loyalty) by X2 is 0.365 and by X3 is 0.703, which is greater than the influence of X2 to loyalty.

Table 5
Regression Weights: (Group Number 1 - Default Model)

	Standardized Estimate	Unstandardized Estimate	S.E.	C.R.	P
X2 <--- X1	.454	.493	.137	3.607	***
X3 <--- X1	.771	.425	.092	4.591	***
X3 <--- X2	.376	.191	.069	2.754	.006
Y <--- X3	.703	.776	.222	3.494	***
Y <--- X2	.365	.204	.103	1.980	.048

This table shows the significance of the model in which all correlations are significant. Therefore, it can be said that the service quality has an influence on determining the students' loyalty. The service quality will directly affect the trust and Satisfaction, although it does not directly influence the loyalty.

Table 6
Squared Multiple Correlations: (Group number 1 - Default model)

	Estimate
X2	.206
X3	.999
Y	.999

Contingency coefficient 20.6% occurs in the model of ZX2 = 0.454ZX1, while contingency coefficient 99.9% occurs in two models as follows:

ZX3 = 0.771ZX1 + 0.376ZX2
ZY = 0.365ZX2 + 0.703ZX3

It is concluded that the loyalty is built by two direct influences, namely, the trust and the satisfaction. However, both these variables are variables affected directly by the service quality. The service quality is an exogenous variable, so it can be considered as the beginning of organization services. Without the service quality, the trust and the satisfaction will not exist. Consequently, there is no loyalty.

Table 7
Total Effects (Group Number 1 - Default Model)

	X1	X2	X3	Y
X2	.493	.000	.000	.000
X3	.519	.191	.000	.000
Y	.503	.352	.776	.000

Table 8
Standardized Total Effects (Group Number 1 - Default Model)

	X1	X2	X3	Y
X2	.454	.000	.000	.000
X3	.942	.376	.000	.000
Y	.827	.629	.703	.000

This table shows that the total effect of the service quality on the loyalty is 0.827, the satisfaction on the loyalty 0.703; and the trust on the loyalty 0.629, although all are greater than 50%.

6. CONCLUSION

The results of this study show that (1) the quality of service has a significant positive effect on the students' satisfaction; (2) the quality of service has a significant positive effect on the trust of students; (3) the students' satisfaction has a significant positive effect on the trust; (4) the satisfaction has a significant positive effect on the loyalty; and (5) the trust has a significant positive effect on the loyalty of the students.

The results of standardized effect of X1 (service quality) on X2 (satisfaction) is 0.454 and X1 on X3 (Trust) is 0.771. It means that the effect of X1 on X3 is greater than that of X1 on X2. Therefore, the service quality has more contribution in growing the students' confidence to be successful than that of the satisfaction in receiving services. Similarly, the effect of standardized on the loyalty by the satisfaction is 0.365 and by the trust is 0.703, which is greater than the influence of the satisfaction on the loyalty.

REFERENCES

1. Assael, H. (1998). *Consumer Behavior and Marketing Action*. 6th ed. Cincinnati, OH: South-Western College Publishing. http://www.emeraldinsight.com/ Insight/viewContentItem.do?contentType=Article&hdAction=lnkpdf&contentId=851 602. Accessed on October 18, 2008.
2. Bloemer, J., K.D. Ruyter, and M. Wetzels. (1998). Linking Perceived Service Quality and Service Loyalty: A Multi-dimensional Perspective. *European Journal of Marketing*, 33(11), 1082-1106.
3. Bollen, K.A. (1989). *Structural Equation with Latent Variables*. Wiley. New York.
4. Dirjen Dikti, Departemen Pendidikan Nasional. 2008. http://www.dikti.go.id. Accessed on 18 October 2008.
5. Indahwati, L. (2008). Hubungan Perceived Service Quality dan Loyalitas: Peran Trust dan Satisfaction sebagai Mediator. *The 2nd National Conference UKWMS*.
6. Japarianto, E. (2007). Analisis Kualitas Layanan Sebagai Pengukur Loyalitas. Universitas Kristen Petra. Surabaya.
7. Kompas Edisi (2007). http://www.kompas.com/. Accessed on October 18, 2008.
8. Manurung, M. (2007). *Pengaruh Kinerja Pelayanan Terhadap Kepuasan Nasabah Pada PT. Bank Jatim Cabang Malang*. UBM. Malang.
9. Oliver, R.L. (1999). Whence Consumer Loyalty. *Journal of Marketing*, 63, 33-44.
10. Santoso, Singgih. (2007). *Structural Equation Modeling Konsep dan Aplikasi dengan AMOS*. PT. Elex Media Komputindo. Jakarta.
11. Tiro, M.A. and Sukarna (2008). *Meluruskan Konsep Pengembangan Instrumen Tipe Skala Likert*. Makassar.
12. Tiro, M.A., Sukarna, and Aswi. (2010). *Analisis Jalur*. Makassar: Andir Publisher, Makassar.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 403-416

EMPIRICAL BAYES METHOD TO ESTIMATE POVERTY MEASURES IN A SMALL AREA

Dian Handayani[1,2], Anang Kurnia[3], Asep Saefuddin[3] and **Henk Folmer[2,4]**
[1] Department of Mathematics, State University of Jakarta, Indonesia
Email: dian99163@yahoo.com
[2] Faculty of Spatial Sciences, University of Groningen, The Netherlands
[3] Department of Statistics, Bogor Agricultural University, Indonesia
Email: anangk@ipb.ac.id, asaefuddin@gmail.com
[4] College of Economics and Management, Northwest Agriculture and Forestry University Yangling, China. Email: h.folmer@rug.nl

ABSTRACT

This research study Empirical Bayes method for estimating parameter in a small area. A small area is defined as a sub population which the sample selected from it is not large enough to yield direct estimates with specified accuracy. The standard (classic) method to estimate small area parameter assumes normality for the response variable. Furthermore, the parameter (quantity of interest) is a linear function of response variable. Nevertheless, in practice, the quantity of interest is not a linear function of response variable. In this research, we investigate Empirical Bayes (EB) to estimate small area parameter when the parameter is not a linear function of response variable. We apply the method to estimate poverty measures which is a nonlinear function of welfare variable, such as income or expenditure. To evaluate the EB method, we use bootstrap parametric. The EB method is applied to estimate poverty measures in Bogor Indonesia. Our results show that the MSE of EB is smaller than MSE of direct estimates for poverty measures in Bogor.

KEYWORDS

Small area estimation, empirical bayes, poverty measures, poverty incidence, poverty gap, poverty severity

1. INTRODUCTION

Sample surveys are usually carried out to estimate parameter for whole (larger) population. Actually, the sample survey is not only utilized to provide some estimates for the whole population but also for sub population. As a result, the sample size from some sub population is often too small or it also might be zero. In this case, direct estimation which is only based on the sample from a certain sub population is no longer valid. The high variability will be happened if it is based on very small sample size. The subpopulation which the sample selected from it is not large enough to yield direct estimates with adequate precision is called small area. The statistics method for estimating parameter in a small area is known as Small Area Estimation (SAE).

To improve the effectiveness of small sample size, one could use some auxiliary information which can be obtained from many resources, such as census or administrative data. Besides using auxiliary information, the estimation for small area parameter could employ statistical model. The SAE method is often based on mixed model which include fixed effect auxiliary information and random effect small area. SAE standard method estimates small area parameter based on linear mixed model which assumes normality on response variable and independence among small areas. The estimates of small area parameter based on a linear mixed model is called Empirical Best Linear Unbiased Predictor (EBLUP). Fay Herriot (1979) applied EBLUP to estimate PCI in some states in the US. He utilized data of Census of Population and Housing 1970. The central government in the US used Fay Herriot's results to allocate some funding to regional government.

In practice, it is often met there are some violations on EBLUP's assumption. Molina and Rao (2010) improve the EBLUP which relaxes on the assumption for response variable. In their study, the response variable is not normal distributed. It has positively skewed distribution. In standard SAE method, the parameter (the quantity of interest) that will be estimated is also often a linear function of response variable, such as a population mean. Molina and Rao (2010) proposed the EB method to estimate nonlinear parameter, such as poverty measures. They applied the EB method to estimate poverty incidence and poverty gap by sex for some provinces in Spain. In their study, poverty measures which are a function of welfare variables such as annually net income for each individual could be transformed so that the transformed variable follow normal distribution.

Poverty incidence, poverty gap and poverty severity are poverty measures which are proposed by Foster, Greer & Thorbecke (1984). Poverty incidence measures the quantity of people under poverty. Poverty gap measures degree of poverty of the people that are under poverty. Poverty severity point out to areas with severe level of poverty. The threshold poverty line is usually used to determine a household is under poverty or not. In Indonesia, the poverty line for each province is not the same. In addition, within in the same province, the poverty line for districts (or sub districts) which its category as urban is different from rural category.

To evaluate performance of EB method, Molina and Rao (2010) proposed bootstrap parametric method. In their research, the MSE for EB method is smaller than direct estimates. Following Molina and Rao (2010), we apply the EB method to estimate poverty measures for each sub district in Bogor, Indonesia. To evaluate the accuracy of estimates, we also use bootstrap parametric as Molina and Rao (2010) has done.

The paper is organized as follows. Section 2 describes direct estimation. Section 3 introduces EB predictor and bootstrap method to approximate the MSE of EB predictor. In Section 4, the EB method is applied to the estimation FGT poverty measures in Bogor. A concluding remark is reported in Section 5.

2. DIRECT ESTIMATION

Direct estimation is a method parameter estimation which is only based on sample from an interest domain. Direct estimation will yield large variance if it is based on very small ample size. Unfortunately, it is often met that sample size from interest domain is

very small particularly if the domain is categorized as unplanned domain in a certain survey. The survey is usually conducted to collect some information in larger population. However, in practice, the data survey is utilized for some planned or unplanned domains. Direct estimates for FGT poverty measures is given by:

$$\hat{P}_{\alpha i}^{Dir} = \frac{1}{n_i}\sum_{i=1}^{n_i}\left(\frac{z-R_{ij}}{z}\right)^{\alpha} I\left(R_{ij}<z\right);\ i=1,2...m, j=1,2...n_i\ ;\ \alpha=0,1,2 \qquad (1)$$

with n_i is sample size from i^{th} area , m is the number of small area, $I\left(R_{ij}<z\right)=1$ if $R_{ij}<z$ (i.e. a person under poverty) and $I\left(R_{ij}<z\right)=0$ if $R_{ij}>z$ (a person not under poverty), $\alpha=0$ for poverty incidence, $\alpha=1$ for the poverty gap, and $\alpha=2$ for poverty severity.

If weighted sampling is applied, then the direct estimator of $P_{\alpha i}$ is:

$$\hat{P}_{\alpha i}^{w} = \left(\sum_{j=1}^{n_i} w_{ij}\right)^{-1}\sum_{j=1}^{n_i} w_{ij}\left(\frac{z-R_{ij}}{z}\right)^{\alpha} I\left(R_{ij}<z\right);\ \alpha=0,1,2\ ;\ i=1,2...m \qquad (2)$$

where w_{ij} is the sampling weight of the j^{th} unit in the i^{th} area. (Note that the sampling weight w_{ij} is the inverse of the inclusion probability of the j^{th} unit from the i^{th} area.). If the sampling weights w_{ij} do not depend on the unit j, for example $w_{ij}=n_i/N_i$ under simple random sampling within areas, then (2) reduces to the un-weighted mean (1).

To improve the effectiveness of small sample size from the unplanned domain, one could use indirect estimation by "borrowing strength" from many resources, such as using census or administrative data. In other words, in this case, one should use SAE methods.

3. EMPIRICAL BEST PREDICTION

Molina and Rao (2010) propose EB method to estimate poverty measures in a small area. The estimates are produced by minimizing the Mean Squared Error (MSE). The EB method assumes that the welfare variable -which usually follows a skewed distribution- can be transformed such that the transformed variable follows a normal distribution.

Let $Y_{ij}=g\left(R_{ij}\right)$ be random variable that is a transformation of the welfare variables R_{ij}. Furthermore, let the vector $\boldsymbol{y}$ contain observations of Y_{ij} and let $\boldsymbol{y}$ be normally distributed: $\boldsymbol{y}\sim N\left(\boldsymbol{\mu},\boldsymbol{V}\right)$. The population elements or units can be classified as sampled and non-sampled units. This implies that the FGT poverty measures for the i^{th} small area can be decomposed as:

$$P_{\alpha i} = \frac{1}{N_i}\left\{ \sum_{j \in s_i} P_{\alpha ij} + \sum_{j \in r_i} P_{\alpha ij} \right\}; P_{\alpha ij} = \left(\frac{z - g^{-1}\left(Y_{ij}\right)}{z} \right)^{\alpha} I\left(g^{-1}\left(Y_{ij}\right) < z\right) = h_{\alpha}\left(Y_{ij}\right) ; j = 1,2...N_i \tag{3}$$

where s_i are the sampled and r_i the non-sampled units. $\sum_{j \in r_i} P_{\alpha ij}$ is estimated by the best predictor $\sum_{j \in r_i} \hat{P}_{\alpha ij}^{B}$ with $\hat{P}_{\alpha ij}^{B} = E_{\boldsymbol{y}_r}\left[P_{\alpha ij} \mid \boldsymbol{y}_s \right]$ where $\boldsymbol{y}_s$ is vector of values of the response variable that have been sampled while $\boldsymbol{y}_r$ is vector of values of the response variable that have not been sampled. Then the best predictor of the FGT poverty measures for the i^{th} small area is given by (Molina and Rao, 2010) : $\hat{P}_{\alpha i}^{B} = \frac{1}{N_i}\left\{ \sum_{j \in s_i} P_{\alpha ij} + \sum_{j \in r_i} \hat{P}_{\alpha ij}^{B} \right\}$, where $\hat{P}_{\alpha ij}^{EBP}$ is obtained by Monte Carlo simulation.

The variables Y_{ij} is linked to the vector of auxiliary variables $\boldsymbol{x}_{ij}$ according to the following model (Molina and Rao, 2010):

$$Y_{ij} = \boldsymbol{x}_{ij}^{T}\beta + u_i + e_{ij} \; ; j = 1,2...N_i \; ; \; i = 1,2...M \tag{4}$$

where the random effect u_i and sampling error e_{ij} are independent, $u_i \sim iid \; N\left(0, \sigma_u^2\right)$; $e_{ij} \sim iid \; N\left(0, \sigma_e^2\right)$.

Let $\boldsymbol{y}_i \sim N\left(\boldsymbol{\mu}_i, \boldsymbol{V}_i\right)$; $\boldsymbol{\mu}_i = X_i\beta$ and $\boldsymbol{V}_i = \sigma_u^2 \mathbf{1}_{N_i}\mathbf{1}_{N_i}^{T} + \sigma_e^2 \boldsymbol{I}_{N_i}$ with $\mathbf{1}_{N_i}$ a N_i unit column vector and $\boldsymbol{I}_{N_i}$ the N_i x N_i identity matrix. As above, let the vector $\boldsymbol{y}_i$ be made up of sampled and non-sampled units: $\boldsymbol{y}_i = \left(\boldsymbol{y}_{is}^{T}, \boldsymbol{y}_{ir}^{T}\right)^{T}$. Similar assumptions apply to X_i, μ_i and V_i. The distribution of $\boldsymbol{y}_{ir}$ given the sample data $\boldsymbol{y}_{is}$ is

$$\boldsymbol{y}_{ir} \left| \boldsymbol{y}_{is} \right. \sim N\left(\boldsymbol{\mu}_{ir|s}, \boldsymbol{V}_{ir|s}\right) \tag{5}$$

where

$$\boldsymbol{\mu}_{ir|s} = X_{ir}\beta + \sigma_u^2 \mathbf{1}_{N_i - n_i}\mathbf{1}_{n_i}^{T} V_{is}^{-1}\left(y_{is} - X_{is}\beta\right) \tag{6}$$

$$\boldsymbol{V}_{ir|s} = \sigma_u^2\left(1 - \gamma_i\right)\mathbf{1}_{N_i - n_i}\mathbf{1}_{N_i - n_i}^{T} + \sigma_e^2 \mathbf{I}_{N_i - n_i} \tag{7}$$

With $\boldsymbol{V}_{is} = \sigma_u^2 \mathbf{1}_{n_i}\mathbf{1}_{n_i}^{T} + \sigma_e^2 \mathbf{I}_{n_i}$ and $\gamma_i = \sigma_u^2\left(\sigma_u^2 + \sigma_e^2 / n_i\right)^{-1}$. Note that $\boldsymbol{y}_{ir} \left| \boldsymbol{y}_{is} \right.$ and $\boldsymbol{y}_r \left| \boldsymbol{y}_s \right.$ have the same distribution because of the independence of $\boldsymbol{y}_i$; $i = 1,2...M$. Molina and

Rao (2010) assume that the partition of the elements in the i^{th} population into s_i and r_i is known and that the auxiliary variables x_{ij} associated with $j \in r_i$ are known.

The steps of Monte Carlo simulation to obtain EB predictor :

- Generate L non-sampled $y_{ir}^{(l)}; i = 1, 2...L$ from the estimated conditional distribution of $\boldsymbol{y}_{ir} \,|\, \boldsymbol{y}_{is}$
- Attach the sampled elements to form simulated census vectors, $y_i^{(l)} = \left(y_{is}, y_{ir}^{(l)} \right); i = 1, 2...L$
- Calculate the desired poverty measure with each population vector : $P_{\alpha i}^{(l)} = h_\alpha \left(y_i^{(l)} \right) \; ; \, l = 1, 2...L$
- Take the average over the L simulated censuses as an approximation to the EB estimator
- Under the nested error model on y_{ij} and normality, we can generate values from the estimated predictive distribution using only univariate normal distribution (Molina and Rao, 2010) :

$$P_{\alpha i}^{EB} = E_{y_{ir}} \left[P_{\alpha i} \mid y_{is} \right] \approx \frac{1}{L} \sum_{1=1}^{L} P_{\alpha i}^{(l)}$$

The accuracy of $\hat{P}_{\alpha ij}^{EBP}$ can be estimated by the MSE as follows:

$$MSE\left(\hat{P}_{\alpha i}^{EBP} \right) = E_{\boldsymbol{y}} \left(\hat{P}_{\alpha i}^{EBP} - P_{\alpha i}^{EBP} \right)^2 \qquad (8)$$

Molina and Rao (2010) point out that analytical approximation to the MSE is complicated. They propose a parametric bootstrap to obtain the estimator of the MSE. The procedure parametric bootstrap MSE to evaluate the accuracy of $\hat{P}_{\alpha ij}^{EBP}$ as follows :

- Construct bootstrap population $y_{ij}^{*(b)}, b = 1, 2...B$ from:

  $y_{ij}^{*} = x_{ij}^{'} \hat{\beta} + u_i^{*} + e_{ij}^{*} \; ; j = 1, 2...N_i \; , \; i = 1, 2...M$, where

  $u_i^{*} \sim iid \; N\left(0, \hat{\sigma}_u^2 \right); \; e_{ij}^{*} \sim iid \; N\left(0, \hat{\sigma}_e^2 \right).$
- Calculate bootstrap population parameter $P_{\alpha i}^{*(b)}$
- From each bootstrap population, take the sample with the same indexes S as in the initial sample and calculate EBs $P_{\alpha i}^{*(b)}$ using bootstrap sample data y_s^{*} and known x_{ij} :

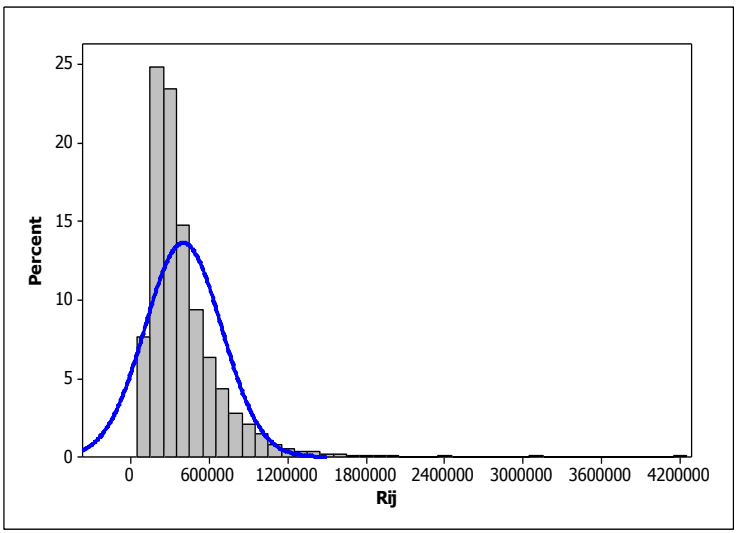
$$mse\left(\hat{P}_{\alpha i}^{EB} \right) = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{P}_{\alpha i}^{EB*(b)} - P_{\alpha i}^{*(b)} \right)^2$$

where $\hat{P}_{\alpha i}^{EB*(b)}$ is the EBP of $P_{\alpha i}^{EB}$ calculated at the b^{th} simulation, and $P_{\alpha i}^{*(b)}$ is the bootstrap population parameter of $P_{\alpha i}$ at the b^{th} simulation and '*' denotes the bootstrap outcome.

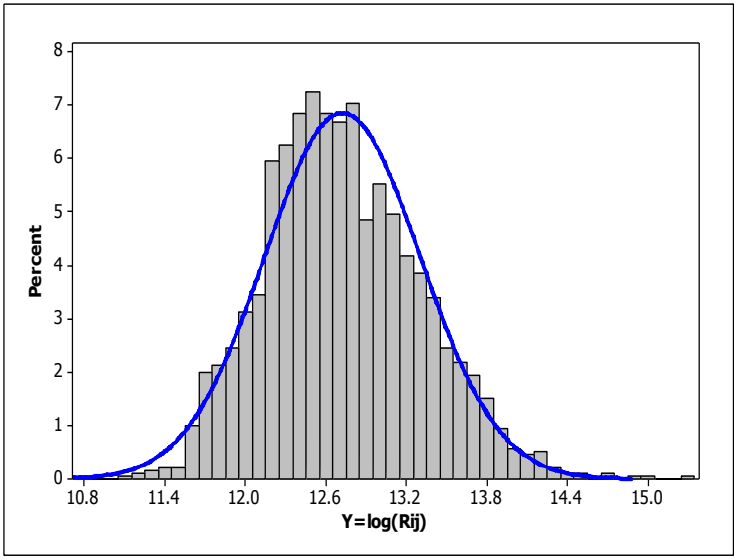
4. APPLICATION EB METHOD TO REAL DATA

In this section we apply the EB method outlined above to estimate the poverty measures at sub-district (kecamatan) level in Bogor, West Java, Indonesia. Particularly, we apply the EB method and compare it to the direct estimates. Bogor is one of the regions in the Province of West Java, Indonesia. It consists of two districts :'Kabupaten Bogor' (District I) and 'Kota Bogor' (District II). Each of the two districts consists of several 'kecamatan' (sub-districts). In the data that we analyze, there are 40 sub-districts in District I and five in District II. Each sub-district consists of several villages. The status of a village is either 'pedesaan' (rural) or 'perkotaan' (urban). Information on the status of a village is available from the 2008 Podes data set. In addition to Podes data, we also use data from the 2007 Socioeconomic Survey (also known as Susenas). The 2007 Susenas data has been designed to allow accurate estimation of poverty measures at districts level but not at the lower sub-districts level.

The variable of interest is Y_{ij} , i.e. log expenditure on basic needs (food and non-food) per capita per month R_{ij} for household- j in sub-districts- i . Hence, $Y_{ij} = \log\left(R_{ij}\right)$. Data on basic needs is available from the 2007 Socioeconomic Survey. Figure 1 (i) shows that expenditure on basic needs per capita R_{ij} does not follow a normal distribution. However, $Y_{ij} = \log\left(R_{ij}\right)$ approximately does, as shown in Figure 1 (ii).



(i)



(ii)

Figure 1. Histogram of R_{ij} (i) and of $Y_{ij} = \log\left(R_{ij}\right)$ (ii)

According to Avenzora and Karyono (2008), households in urban villages tend to have higher expenditures on basic needs than households in rural villages. Hence, the status of village where the household is located can be used as auxiliary information to estimate the poverty measures. Table 1 presents summary statistics on the number of households for the 45 sub-districts. The table shows that the median of population size is 24592 and for the sample approximately 32 which is only 0.13 % of the population size. The sample size ranges from 16 to 144.

Table 1
Summary Statistics on the Number of Households for the 45 Sub-Districts in Bogor

Variable	Mean	Min	Median	Max
Population Size (N)	28480	12047	24592	64980
Sample Size (n)	39.82	16	32	144

Sample Size per sub-district	Count	Percentage
16	15	33.33
32	16	35.56
48	7	15.56
64	2	4.44
80	1	2.22
112	2	4.44
128	1	2.22
144	1	2.22

According to Statistics Indonesia (2008), the poverty line for the Province of West Java for the period January-December 2007 is IDR 180,821 for households in urban areas and IDR 144,204 for households in rural areas.

To obtain the EBP of the FGT poverty measures, we generated 1000 Monte Carlo data sets of $(N_i - n_i)$ non-sampled Y_{ij} values. On the basis of the generated Y_{ij} values and the observed (sampled) Y_{ij}, we calculated $R_{ij} = \exp(Y_{ij})$. Then, the FGT poverty measures P0, P1 and P2 were calculated. Summary statistics for direct and EB estimates of P0, P1 and P2 for each sub district are presented in Table 2, Table 3 and Table 4 respectively.

Table 2
Summary Statistics for Direct and EB of P0 Estimates for Each Sub District

District	Direct				EB			
	Mean	Median	Min	Max	Mean	Median	Min	Max
District I	12.66	6.25	0	53.13	11.09	4.68	0.03	43.81
District II	3.20	4.17	0	4.69	0.73	0.68	0.23	1.45

Table 3
Summary Statistics for Direct and EB of P1 Estimates

District	Direct				EB			
	Mean	Median	Min	Max	Mean	Median	Min	Max
District I	2.20	0.95	0	10.33	2.17	0.92	0.0043	8.50
District II	0.36	0.45	0	0.62	0.08	0.09	0.0257	0.15

Table 4
Summary Statistics for Direct and EB of P2 Estimates

District	Direct				EB			
	Mean	Median	Min	Max	Mean	Median	Min	Max
District I	0.61	0.16	0	3.95	0.66	0.25	0.00164	3.043
District II	0.07	0.07	0	0.16	0.01	0.02	0.00471	0.023

It can be seen that the mean and the median EB estimates for P0, P1 and P2 at District II is smaller than District I. The minimum values for direct estimates of P0, P1 and P2 are zero because there is no household in the sample which is under poverty at some sub districts. On the other hand, the EB method could produce the poverty measures which are not zero although the sample size from the interest sub population is zero. This could be happened because the EB method could borrow strength from neighbor areas or other resources.

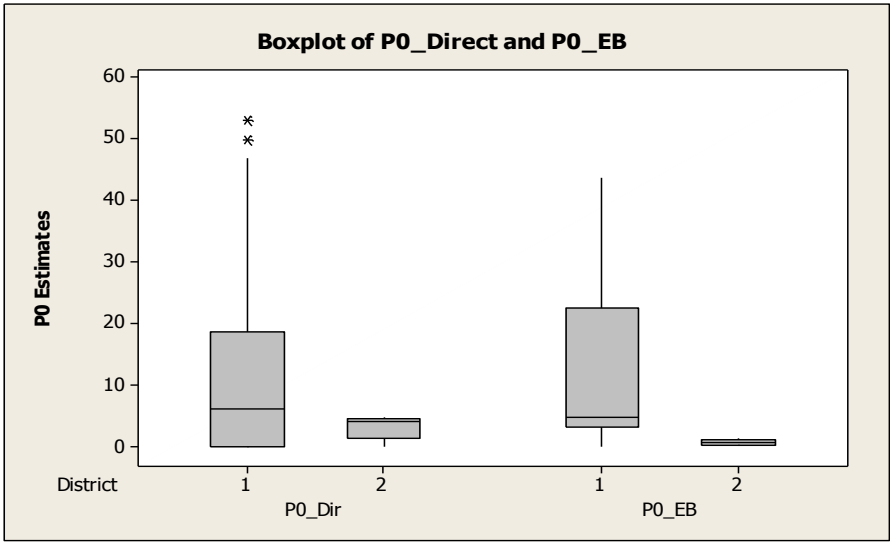


Figure 2. Boxplot of P0 Direct and EB Estimates for each District

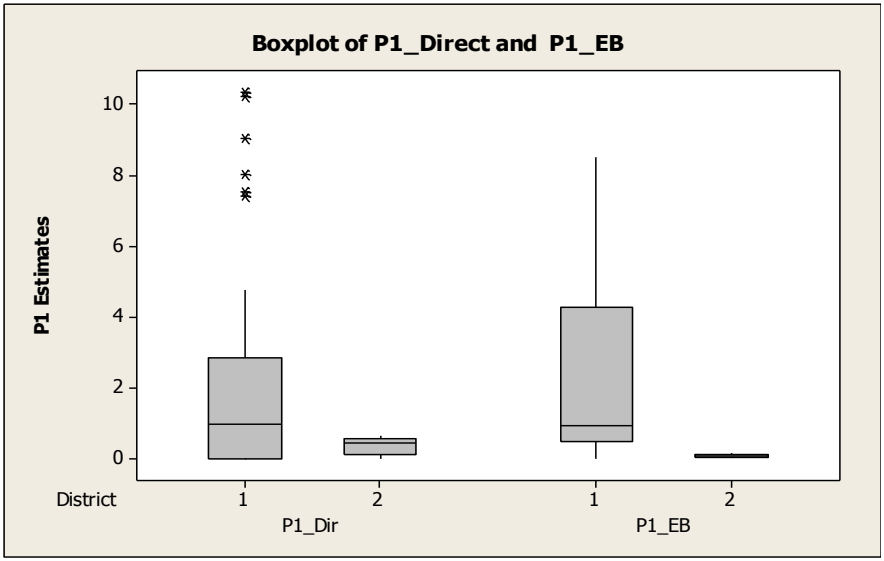


Figure 3: Boxplot of P1 Direct and EB Estimates for each District

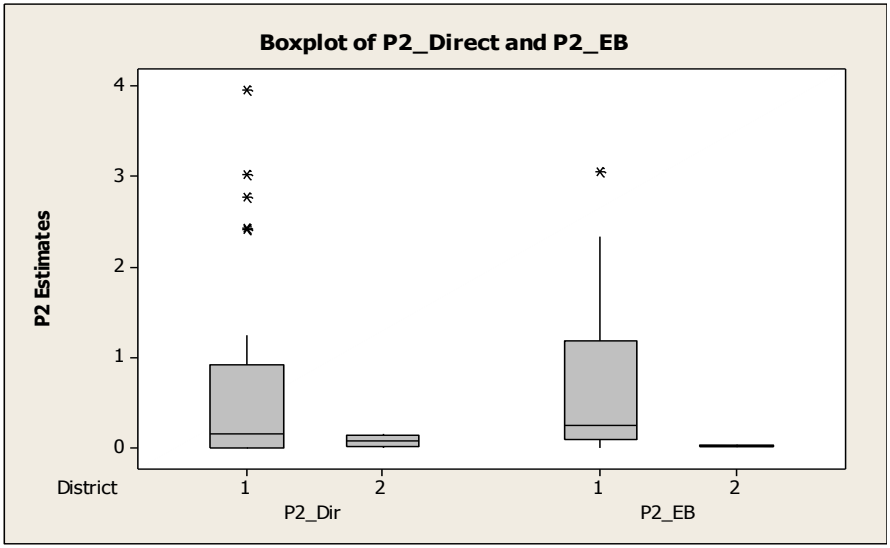


Figure 4: Boxplot of P0 Direct and EB Estimates for each District

Figure 2, 3 and 4 presents the boxplots of Direct and EB estimates of P0, P1 and P2 based on classification District I and II. Based on these figures, it can be seen that the variability of the poverty measures in District I is larger than District II. On the other hand, direct estimates and EB estimates of P0, P1 and P2 for overall sub districts in Kabupaten and Kota Bogor are shown in Figure 5,6 and 7.

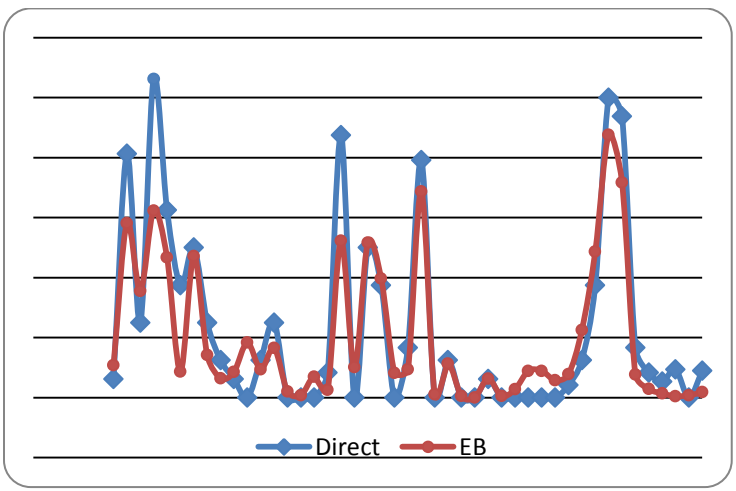


Figure 5: P0 Estimates by Direct Estimation vs EB

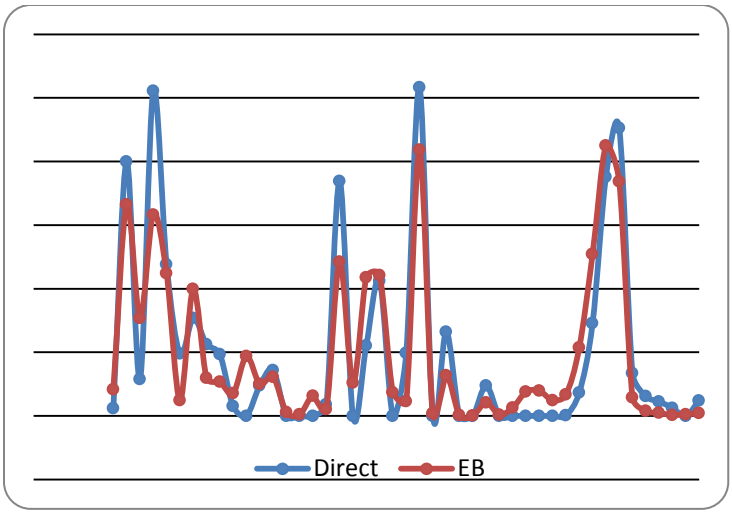


Figure 6: P1 Estimates by Direct Estimation vs EB

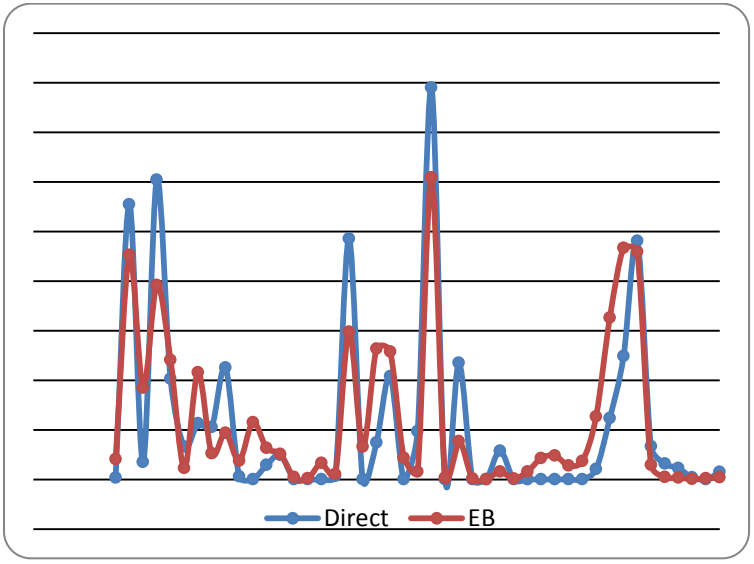


Figure 7: P2 estimates by Direct Estimation vs EB

Based on parametric bootstrap approximation, the EB method is better than direct estimates for poverty measures P0, P1 and P2. These are supported by Figure 8, 9 and 10. In these figures, it can be seen that the estimates of MSE for the EB of poverty estimates are lower than direct estimates. The MSE of EB estimates is also more stable relatively than direct estimates.

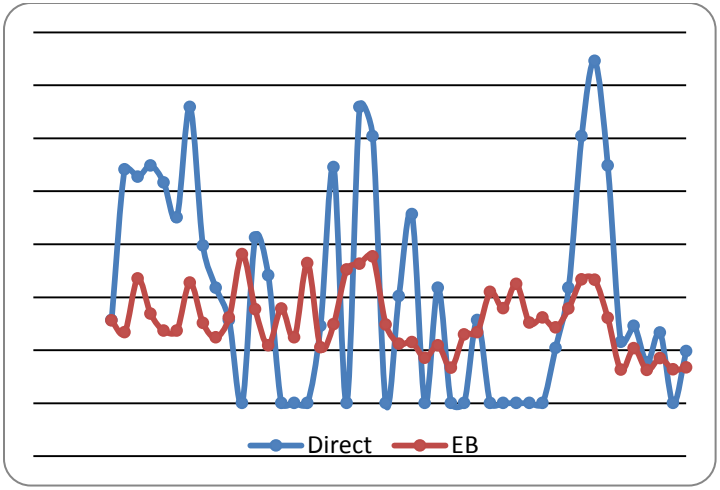


Figure 8: MSE Bootstrap Estimates for P0 Direct & P0 EB

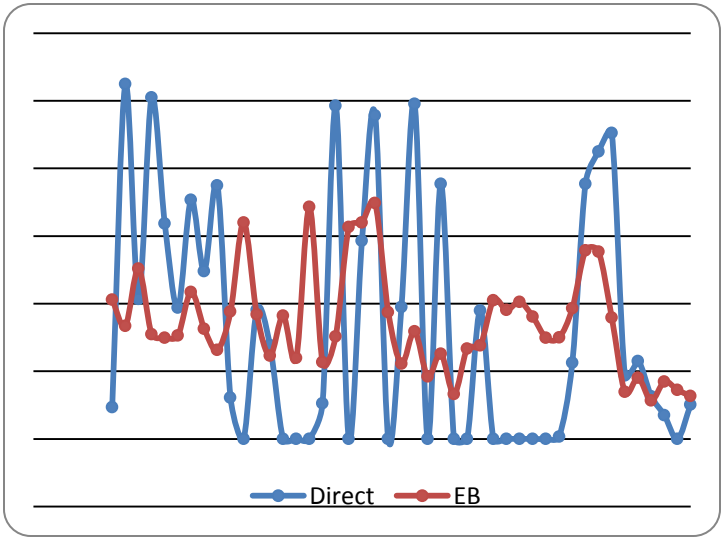


Figure 9: MSE Bootstrap Estimates for P1 Direct & P0 EB

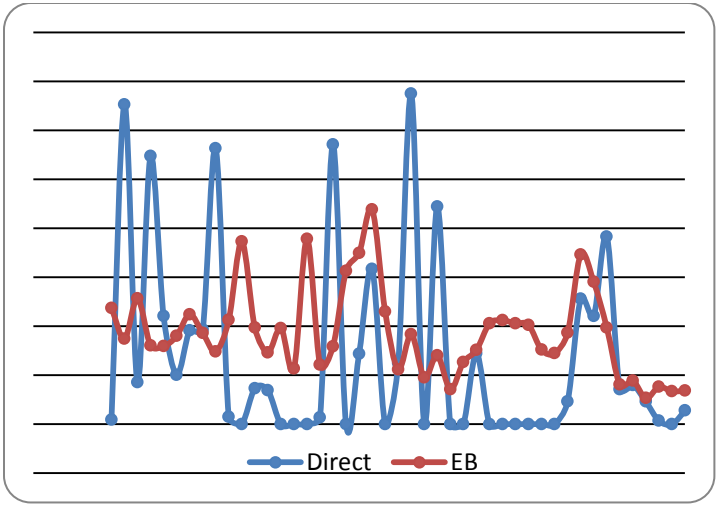


Figure 10: MSE Bootstrap Estimates for P1 Direct & P0 EB

5. CONCLUDING REMARKS

Direct estimates of poverty measures for some sub districts are zero because there is no household which is under poverty in the sample that selected from these sub districts. However, the EB estimates for these sub districts are not zero.

The estimates of poverty measures in District 2 (Kota Bogor) is smaller than District I (Kabupaten Bogor). The variability of poverty measures in Kota Bogor is also smaller than poverty measures in Kabupaten Bogor (District I).

The proposed EB method could yield poverty measures estimates which are better than the direct estimates in terms of MSE.

REFERENCES

1. Avenzora, A. and Karyono, Y. (2008). Analisisdan Penghitungan Tingkat Kemiskinan Tahun 2008. Jakarta: Badan Pusat Statistik.
2. Elbers, C., Lanjouw, J.O and Lanjouw, P. (2003). Micro-Level Estimation of Poverty and Inequality. *Econometrica*, 71(1), 355-364.
3. Fay, R.E and Herriot, R.A. (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, 74, 269-277.
4. Foster, J., Greer, J and Thorbecke, E. (1984). A Class of Decomposable Poverty Measures. *Econometrica*, 52(3), 761-766.
5. Haslett, S., Isidro, M.C. and Jones, G. (2010). Comparison of Survey Regression Techniques in the Context of Small Area Estimation of Poverty. *Survey Methodology*, 36(2), 157-170.
6. Isidro, M.C. (2010). *Intercensal Updating of Small Area Estimates*. Thesis. Massey University, Palmerstone North, New Zealand.

7. Jiang, J., Nguyen, T. and Rao, J.S. (2011). Best Predictive Small Area Estimation. *Journal of the American Statistical Association*, 106(494), 732-745.
8. Molina, I.and Rao, J.NK. (2010). Small Area Estimation of Poverty Indicators. *The Canadian Journal of Statistics*, 38(3), 369-385.
9. Neri, L., Ballini, F. and Betti, G. (2005). Poverty and Inequality Mapping in Transition Countries. *Statistics in Transition*, 7(1), 135-157.
10. Srivastava, A.K. (2009). Some Aspects of Estimating Poverty at Small Area Level. *J. Ind. Soc. Agril. Statist.*, 63(1), 1-23.
11. Tarozzi, A. and Deaton, A. (2009). Using Census and Survey Data to Estimate Poverty and Inequality for Small Areas. *The Review of Economics and Statistics*. 91(4), 773-792.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 417-430

ANALYSIS OF APPENDECTOMY IN BELGIUM USING DISEASE MAPPING TECHNIQUES

Mieke Nurmalasari[1] and **Setia Pramana**[2]
[1] Sekolah Tinggi Ilmu Ekonomi Indonesia, Jl Kayu Jati, 11A, Jakarta, Indonesia. Email: mieke_hafidz@yahoo.co.id
[2] Sekolah Tinggi Ilmu Statistik, Jl Otista 64C, Jakarta, Indonesia Email: setia.pramana@stis.ac.id

ABSTRACT

An appendectomy is the surgical removal of the vermiform appendix normally performed as an emergency, when the patient is suffering from acute appendicitis. The analysis of appendectomy related to the geographical distribution of disease hospital-admissions in Belgium was still under studied. This study is aimed to identify geographical differences in medical practices, and to investigate spatial and temporal distribution on appendectomy and incidental appendectomy incidence rate in Belgium for period 2001 to 2006.

Two different methods were applied to identify possible high incidence regions, used maps of the non-smoothed SIRs and Bayesian methods to smooth SIRs. Using Bayesian method, the best model (based on the smallest value of DIC) for appendectomy and incidental appendectomy cases are Poisson-Lognormal and Poisson-Gamma, respectively. The range of mean relative risks for appendectomy cases was smoothed between 0.81 and 1.32, and the range of mean relative risks for incidental appendectomy was smoothed between 0.32 and 2.67. Based on these models, we can conclude that the model of smoothed SIRs (mean relative risks) of appendectomy and incidental appendectomy cases among districts in Belgium for 2001-2006 periods are not related with the environment.

Based on smoothed SIR, three districts (Diskmuide, Tielt and Dinant) have higher relative risk than other districts in appendectomy cases. For incidental appendectomy, the highest relative risks are district Oudenaarde, Sint-Niklaas, Hoeiand and Bastenaken. The significant increased or decreased incidence of appendectomy and incidental appendectomy cases in districts are need to be investigated further. It might also be the expression of differences in medical practice.

Considering analysis of spatial temporal using Bernardinelli model, the risk of appendectomy cases significantly decreased in time where the risk was multiplied by approximately 0.9888 every year. In the other hand, the incidental appendectomy cases increased from 2001-2006 and the increase was significant for incidental appendectomy cases over time. The risk was multiplied by approximately 1.0392 every year.

KEYWORDS

Spatial Analysis, Appendectomy, disease mapping, Bernardinelli model.

1. INTRODUCTION

Since October 1990, all Belgian hospitals are subjected to compulsory registration with the health authorities of each admission through a standard form containing a define set of clinical data including ICD-coded diagnoses and procedures. These discharge abstracts are termed Minimal Clinical Data (MCD) and contain patient data (among which year of birth, gender, residence, and anonymous hospital and patient identifiers, stay data among which year and month of admission and discharge, length of stay, transfer to another hospital with specification of the type of hospital) and an unlimited number of diagnoses and procedures. This information (about 2,000,000 hospital-admissions per year) is transmitted to the authorities for compilation and processing. From MCD we can obtain the diagnose of two selected pathologies, appendectomy and incidental appendectomy. Nevertheless, the analysis of appendectomy related to the geographical distribution of disease hospital-admissions in Belgium was still under studied.

Both appendectomy and incidental appendectomy are a surgical removal of the vermiform appendix that differs on the purpose. An appendectomy is normally performed as an emergency procedure, when the patient is suffering from acute inflammation of the appendix known as appendicitis. In contrast, incidental appendectomy is performed incidental to other abdominal surgery, such as urological, gynecological, or gastrointestinal surgeries, intended to eliminate the risk of future appendicitis and to simply any future differential diagnoses of abdominal pain. For these cases, classical epidemiological measures such as Standardized Incidence Ratio (SIR) still can be calculated because the appendix of the patients only can be removed once.

The objectives of this study are (1) to carry out an analysis of the geographical distribution of disease hospital-admissions regarding appendectomy and incidental appendectomy, (2) to identify geographical differences in medical practices and (3) to investigate spatial and temporal distribution on appendectomy and incidental appendectomy incidence rate in Belgium for period 2001 to 2006.

2. MATERIALS AND METHODS

Data

The data used is the number of appendectomy and incidental appendectomy cases in Belgium during 2001 to 2006, and number of Belgium population from 2001 to 2007 (calculate the middle year). The population for appendectomy is all Belgian population, whereas for incidental appendectomy the population is patients older than 65 years old. District will be the unit of analysis related to appendectomy and incidental appendectomy cases. The data set comes from 43 districts. In this report, only some variables from Minimal Clinical Data (MCD) are used i.e., sex, age group, residence, patient identifiers and year.

Standardized Incidence Ratio (SIR)

Incidence rate, a number of new cases per population at risk in a given time period, is a measure of the frequency with which a disease occurs in a population over a period of time. However, in many case a region or a district with higher population may have more

cases than the one with less population. In this case the Standardized Incidence Ratio (SIR) is used instead of incidence rate.

Standardized Incidence Ratio (SIR) for each district can be defined as:

$$SIR = \frac{y_i}{E_i} = \frac{\sum_g y_{gi}}{\sum_g E_{gi}} = \frac{\sum_g y_{gi}}{\sum_g \frac{y_g^S}{n_g^S} n_{gi}} \quad (1)$$

i = 1, 2, …, m where ygi is the number of cases in age group g for study population i, yi is the total number of cases observed on the study population, Egi expected number of cases in age group g for study population i, Ei is the overall expected number of cases for the study population, ngi is number of people at risk in age group g for study population i, and , denote the same quantities for the standard population (Waller and Gotway, 2004).

Poisson Model

The Standardized Incidence Ratio (SIR) for region i is obtained from the ratio of the observed and expected number of cases (Yi=Ei) in that region (i = 1, 2, …, m). Indeed, independently in each region i, the number of cases are supposed to follow a Poisson distribution.

The Poisson model:

$$y_i \sim Poisson\ (e_i, \theta_i) \quad (2)$$

where θ_i is the relative risk of disease in region *i*. The 95% confidence interval (CI) for SIR can be calculated as

$$[SIR_i * \exp(-1.96/\sqrt{Y_i}\ ); SIRi * \exp(1.96/\sqrt{Y_i}\ )] \quad (3)$$

This is equal to the CI [SIR_i/errfac; SIR_i * errfac] with *errfac* is an error factor defined as

$$errfac = \exp\left(z_{1-\alpha/2}\sqrt{\frac{1}{y}}\right) \quad (4)$$

(Clayton and Hills, 1995).

The conventional approach of mapping standardized disease rates based on Poisson inference gives a good illustration of the geographical distribution of the underlying rates when the disease is not rare. However, for rare disease or small areas, these maps often produce a mix of colors that are difficult to interpret. Moreover, the numbers of disease cases observed in each small area are often more variable than that implied by the standard Poisson model. Bayesian models have been developed in disease mapping in order to take into account the extra Poisson variation. One way is to shrink the most unreliable standardized rates towards the overall mean rate.

Bayesian Methods

Modern approaches to relative risk θi estimation rely on smoothing methods. These methods often involve additional assumptions or model components. Here, a bayesian modeling approach was used. The bayesian method was assumed that the relative risk estimator has a distribution. In the Bayesian terms, this is called a prior distribution. In the Poisson count, the most common prior distribution is to assume that θi has Gamma distribution.

Here, three models were applied, i.e., Poisson-Gamma, Poisson-Lognormal and Conditional Autoregressive (CAR) model, to investigate spatial distribution of appendectomy and incidental appendectomy for each district in the study area for the 6 years period from 2001-2006. Model fitting was carried out using MCMC simulation methods implemented in the WinBUGS software. Two separate chains which starting from different initial values were run for each model. Convergence was checked by visual examination of "time series" style plots of the samples for each chain, and by computing the Gelman-Rubin convergence statistic (Gelman and Rubin, 1992).

Poisson-Gamma model

When the disease is rare, the numbers of diseases in each area are assumed to be mutually independent and follow Poisson distributions

$$y_i = \text{Poisson}(e_i, \theta_i) \tag{5}$$

for $\forall_i, \theta_i \sim Gamma(a,b)$ with mean $m_{\theta_i} = a/b$ and variance $v_{\theta_i} = a/b^2$. As the result, the relative risk has the following distribution (posterior)

$$\theta_i \sim Gamma(a + y_i, b + e_i) \tag{6}$$

Poisson-Lognormal Model

The log-normal model for the relative risk is defined as

$$y_i = Poisson(e_i, \theta_i) \tag{7}$$

$$\log(\theta_i) = \alpha + v_i \tag{8}$$

where $v_i \sim N(0, \sigma_v^2)$ is the heterogeneity random effect, capturing extra-Poisson variability in the log-relative risks. The Lognormal model for the relative risk is more flexible (Lawson, et al., 2008). A major drawback with gamma prior is this method does not take into account the geographical location of the region. The models do not cope the spatial correlation. It is possible to account for the spatial pattern in disease by using Conditional Autoregressive (CAR) model.

Conditional Auto Regressive model (CAR)

The conditional autoregressive (CAR) model proposed by Besag et al. (1991) is used. In this model for relative risks, area specific random effects are decomposed into a component that takes into account the effects that vary in a structured manner in space

(clustering or correlated heterogeneity) and a component that models the effects that vary in an unstructured way between areas (uncorrelated heterogeneity).

The model of CAR can be represented as:

$$y_i = Poisson(e_i, \theta_i), \tag{9}$$

$$\log(\theta_i) = \alpha + u_i + v_i, \tag{10}$$

where α is an overall level of the relative risk, ui is the correlated heterogeneity, and $v_i \sim N(0, \sigma_v^2)$ is the uncorrelated heterogeneity. The spatial correlation structure is used then the estimation of the risk in any area depends on neighboring areas $[u_i \mid u_j, i \neq j, \tau_u^2] \sim N(\bar{u}_i, \tau_i^2)$. The u_i is smoothed towards the mean risk in the set of neighboring areas, with variance inversely proportional to the number of neighbors (Lawson, et al., 2008).

The three models above are then compared using overall goodness of fit measures, such as Deviance Information Criteria (DIC) that has been proposed by Spiegelhalter et al. (2002)

$$DIC = 2E_{\theta|x}(D) - D(E_{\theta|x}(\theta)) \tag{11}$$

with D(.) the deviance *(-2*log(likelihood))* of the model and x the observed data. The model with a smallest DIC is the best model to predict a replicate data set of the same structure as that currently observed.

Bernardinelli Model

The most common way to consider the analysis of disease maps that have a temporal dimension is to count number of cases of diseases within small areas that are available for a sequence of T time periods. In this section we are going to analyze the space-time distribution of appendectomy cases and incidental appendectomy cases in Belgium over a period of six years (2001-2006) using Bernardinelli model.

Bernardinelli et al. (1995) suggests a model in which both area-specific intercept and temporal trend are modeled as random effects. This formulation allows for spatiotemporal interactions where temporal trend in risk may be different for different spatial locations and may even have spatial structure. All temporal trends are assumed to be linear. The Bernardinelli model is defined as

$$y_{ik} \sim Poison(e_{ik}\theta_{ik}) \tag{12}$$

$$\log(\theta_i) = \alpha + u_i + v_i + \beta t_k + \delta_i t_k \tag{13}$$

where α is an intercept (overall rate), u_i and v_i are area random effects (as defined in the CAR model), βt_k is a linear trend term in time t_k, δ_i is an interaction random effect between space and time. Prior distribution must be assumed for the parameters in this model.

3. RESULTS

During 2001 to 2006, the total cases of appendectomy and incidental appendectomy are 83981 and 1907, respectively. The incidence of appendectomy and incidental appendectomy was the highest in patients of 9-15 years and 69-75 years old, respectively. This is the same for both male and female patients. In general, there is a decreasing incidence of incidental appendectomy with age. Incidental appendectomy and appendectomy diseases affect people of certain age disproportionally.

We observed that the number of cases and the population decreases from 2001 to 2005. They then increase slightly in 2006. In the other hand, the number of Incidental Appendectomy cases increases from 2001 to 2005 before it decreases slightly in 2006.

Spatial Analysis

Next, the disease mapping technique is used as a way of presenting the results and demonstrating the geographical variation of appendectomy and incidental appendectomy incidence in each district. When comparing the incidence of both cases between two areas, or when investigating the pattern of appendectomy and incidental appendectomy cases for the same areas, it is important to adjust for differences in the age and gender of those populations. In this study, this was accomplished by gender-stratified and age standardization.

Appendectomy Cases

In the appendectomy cases (Figure 1), we observe the geographical variation SIR of appendectomy cases in district in Belgium, 2001-2006. The significance of SIR also was calculated using the 95% confidence interval of SIR, which is calculated by equation (3). In Figure 1, the map shows that the three areas with the highest SIRs are shaded in dark green. These areas are district Diksmuide, Tielt and Dinant (SIR > 1.2). Other districts that have significant increased incidence as compared to the whole study region are Mechelen, Nijvem, Roeselare, Eeklo, Moeskroen, Verviers, Borgworm, Bastenaken and Namen. Otherwise, district Leuven, Aalst, Gent, Charleroi, Zinnik, Thuin, Luik, Hasselt, Maaseik, Aarlen, Marche-en-Famenne and Virton have decrease incidence significantly or fewer cases occurred than expected.

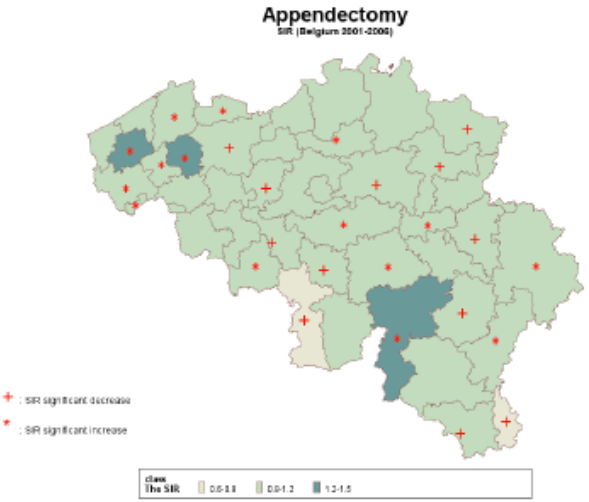


Figure 1: SIR map of the appendectomy cases 2001-2006

Incidental Appendectomy

Figure 2 shows that the value of SIR for incidental appendectomy cases have more variability of SIR as compared to the SIR in appendectomy cases. Districts with SIR > 2 are Oudenaarde, Sint-Niklaas, Hoei and Bastenaken. Some districts with significant increase on incidence are Antwerpen, Oostende, Veurnee, Bergen, Luik, Aarlen and Neufchatean. These districts indicate that more cases of incidental appendectomy occurred than expected based on the age specific incidence proportions from the standard population. On the other hand, district Turnhout, Leuven, Brugge, Roeselare, Eeklo, Charleroi, Zinnik, Hasselt, Maaseik, Tongeren, Namen and Philippeville indicate fewer cases occurred than expected.

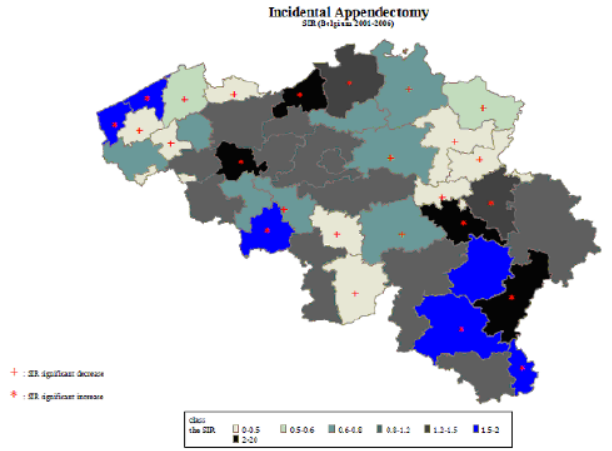


Figure 2: SIR map of Incidental Appendectomy cases 2001-2006

The result from ESDA and the non-smoothed SIRs (not shown) indicated that there is no much different for SIR of appendectomy and incidental appendectomy by gender. Hence, for the next analysis, we only consider analysis of appendectomy and incidental appendectomy that already corrected by age group and gender, not presented the SIR of these cases by gender separately.

Bayesian Methods

Appendectomy Cases

Three models were compared using Deviance Information Criterion (DIC) to determine which model gives the good estimates. They were Poisson-Gamma model (DIC=474.907), Poisson-Lognormal model (DIC=474.773) and Conditional Auto Regressive model (DIC=474.805). Poisson-LognormaL model is better than Poisson Gamma and CAR model, according to the DIC value.

The comparison map of non-smoothed SIR and smoothed SIR for the period 1996-2000 are presented in Figure 3. There is no much different between non-smoothed SIR and mean Relative Risk (smoothed SIR), but the value of SIR were smoothed by using Bayesian. The range of non-smoothed SIR is 0.78-1.36, then after smoothing method using Poisson-Lognormal model, the range of relative risk is 0.81-1.32.

District which has significant increased or decreased risk of this disease is based on the 95% Confidence Interval for SIR (CISIR) and 95% Credible Interval for mean relative risk (CI Bayes). It can be seen in Figure 3.

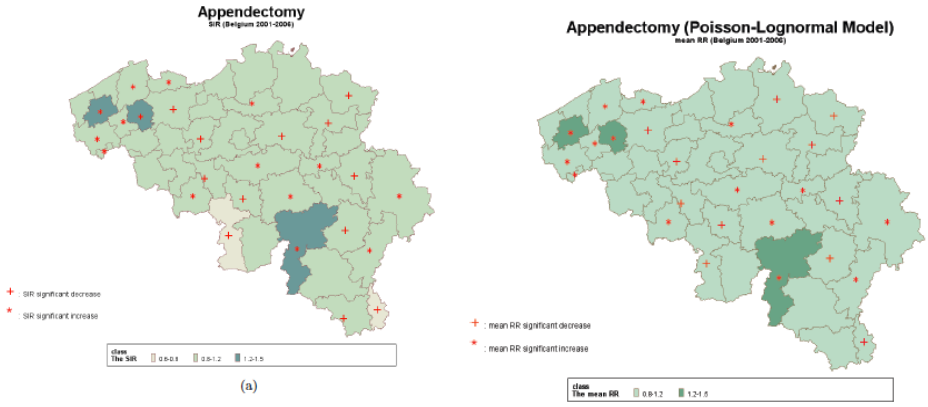


Figure 3: The comparison map of non-smoothed SIR and smoothed SIR for the period 2001-2006

Incidental Appendectomy Cases

The DIC for Poisson-Gamma model is 303.660, DIC Poisson-Lognormal model = 304.142, DIC CAR model =303.786. Based on this, Poisson-Gamma model is the best model for incidental appendectomy. Table 1 shows the posterior values for the parameters of the model after 50000 iterations.

Table 1
Posterior statistics for the parameters in the Poisson-Gamma model

	Mean	SD	Credible Interval
Alpha	3.318	0.8054	(0.01485; 5.174)
Beta	3.188	0.8305	(0.01526; 5.098)

There is few difference of map distribution (Figure 4) after using Poisson-Gamma model, the mean relative risk were smoothed. District Aarlen is not significance anymore. The range of non-smoothed SIR is between 0.18 and 3.18, and Poisson Gamma-smoothed SIR (mean relative risk) is between 0.32 and 2.67.

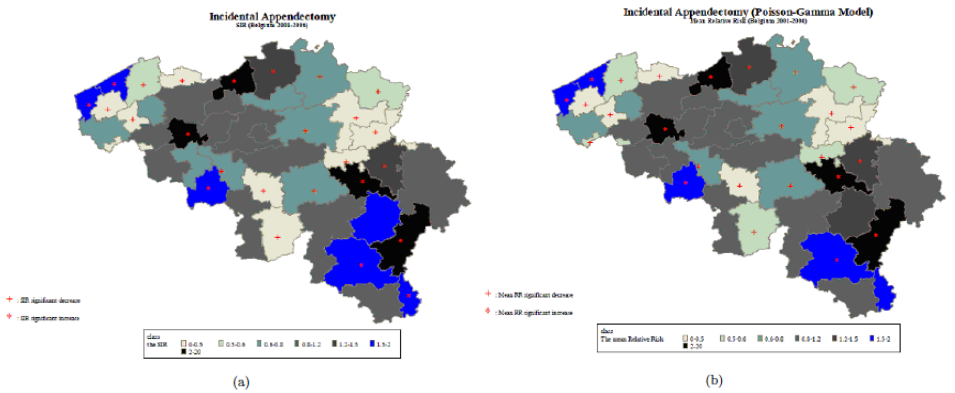


Figure 4 (a) Non-Smoothed SIRs and (b) Bayesian Mean Relative Risk (RR) of Incidental Appendectomy Cases in Belgium, 2001-2006; using Poisson-Gamma Model.

Bernardinelli Model

In this part, we considered the analysis of appendectomy and incidental appendectomy cases that have temporal dimension (from 2001-2006). Using Bernardinelli model, we obtained the estimation of the time effect for appendectomy cases as shown in Table 2.

Table 2
Mean Estimates of Time effect from Bernadinelli Model for Appendectomy

	Mean	Lower	Upper
Alpha	-0.5523	-4.378	0.0886
Beta	-0.0112	-0.0156	-0.00688
σ_u^2	0.0193	2.607E-4	0.0648
σ_v^2	1.882	9.397E-4	19.87
σ_{delta}^2	0.0016	6.766E-4	0.0033

Table 2 shows that the risk of appendectomy cases small significant decrease over time. This can be observed from the obtained trend term in time t_k which equal -0.0112. From this value, we can also obtain the ratio between two consecutive years which equal

0.9888 (*exp* [-0.0112]). Hence, the risk was multiplied by approximately 0.9888 every year. Maps of spatially smoothed relative risk of appendectomy cases at different time points are shown in Figures 5. A map of spatially smoothed time trends as shown in Figure 6 provides a visual impression of the small decreased in incidence occurring.

The map of temporal trend as derived from Bernardinelli model showed slightly decrease in risk of appendectomy cases (see Figure 6).

In contrast with the appendectomy cases, the risk for incidental appendectomy shows a small significant increase from 2001-2006. Based on Table 3, the ratio between two consecutive years was 1.039 (*exp* [0.0385]), thus the risk was multiplied by approximately 1.039 every year. A map of spatially smoothed relative risk of incidental appendectomy over time is presented in Figures 7.

Table 3
Mean Estimates of Time Effect from Bernadinelli Model for Incidental Appendectomy

	Mean	Lower	Upper
Alpha	-563.7	-1097.0	-26.75
Beta	0.0385	0.0052	0.0713
σ_u^2	33330	18.22	1.07E+5
σ_v^2	443200	685.6	1.405E+6
σ_{delta}^2	0.2504	0.1399	0.4244

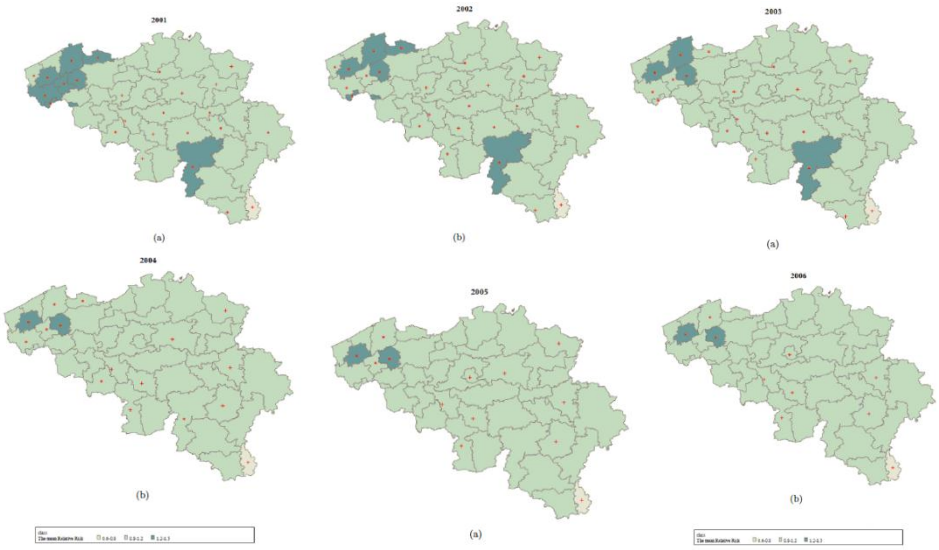


Figure 5: Smoothed Mean Relative Risk for Appendectomy Cases in Belgium 2001-2006 using Bernardinelli Model

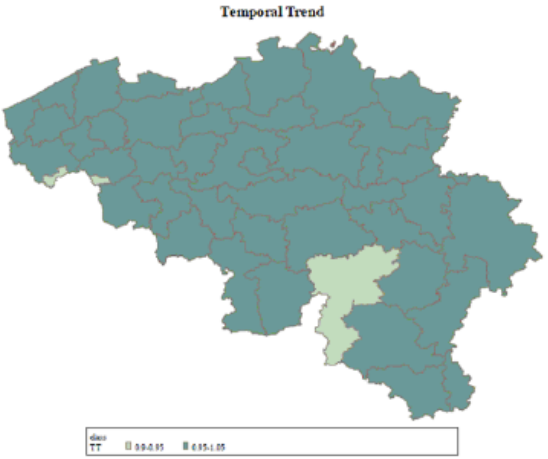


Figure 6: Temporal Trend for Appendectomy cases in Belgium 2001-2006 using Bernardinelli model.

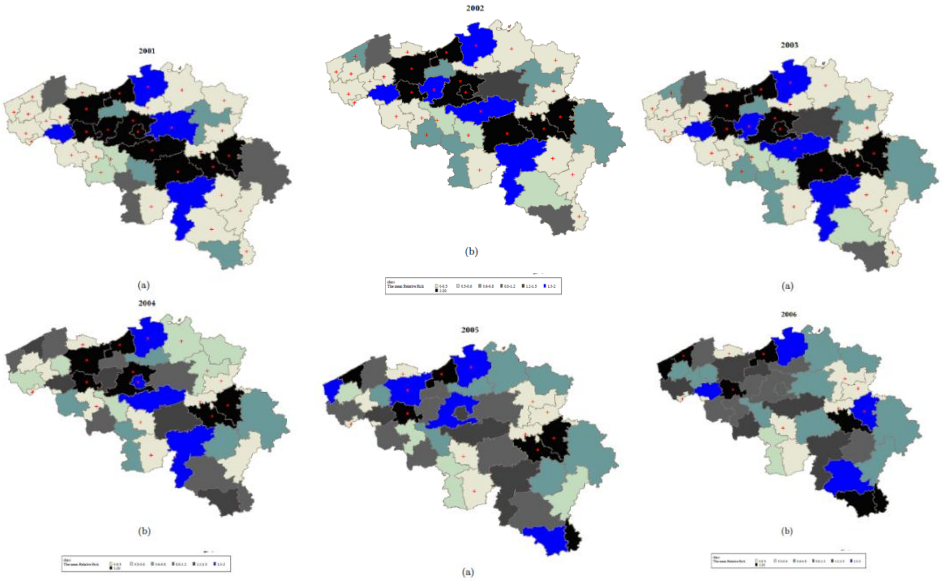


Figure 7: Smoothed Mean Relative Risk for Incidental Appendectomy Cases in Belgium, 2001-2006 using Bernardinelli Model

The map of temporal trend as derived from Bernardinelli model is shown in Figure 8.

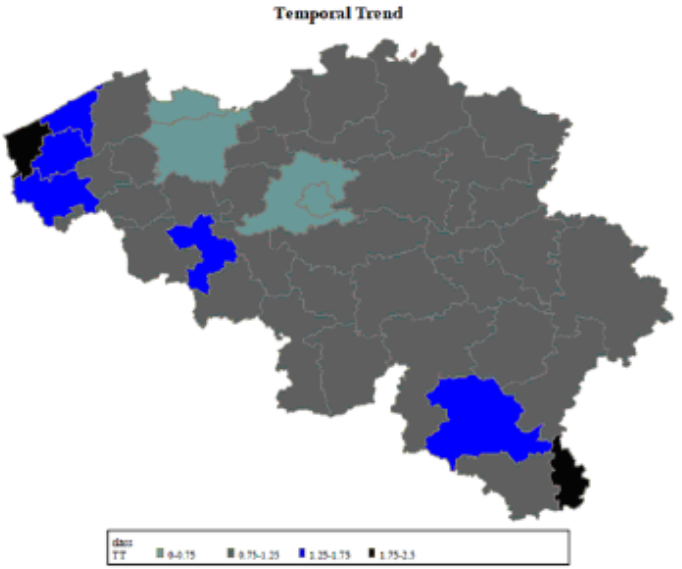


Figure 8: Temporal Trend for Incidental Appendectomy Cases in Belgium 2001-2006 using Bernardinelli Model

4. DISCUSSIONS AND CONCLUSION

The result shows that for the appendectomy cases, using non-smoothed SIRs, district Diksmuide, Tielt and Dinant have the higher SIR (SIR>1.2) than the other districts. This is due to the fact that these three districts have a rather small population and thus more prone to have more extreme values. These three districts have significant increased incidence of appendectomy disease as compare to the whole study region. For example, district Dinant has SIR=1.27, it is interpreted as 27% more cases observed than expected number.

The SIR for incidental appendectomy cases has more variability of SIR as compared to the SIR in appendectomy cases. This is because of the incidence of incidental appendectomy cases is rare. Districts with SIR> 2 are Oudenaarde, Sint-Niklaas, Hoei and Bastenaken. Other districts with significant increase incidence are Antwerpen, Oostende, Veurnee, Bergen, Luik, Aarlen and Neufchatean. A different case for district Turnhout, Leuven, Brugge, Roeselare, Eeklo, Charleroi, Zinnik, Hasselt, Maaseik, Tongeren, Namen and Philippeville. The increased or decreased incidences of incidental appendectomy cases in these districts need to be investigated further. It might also be the expression of differences in medical practice. Observing the appendectomy and incidental appendectomy cases by gender, the maps showed quite similar result. Furthermore, the difference in incidence between males and females does not seem too essential for both cases.

Using the non-smoothing SIRs, having risks when dealing with relatively small districts or for districts with relatively low numbers of cases or disease, the disease incidence rates tends to differ largely due to random error and may have misleadingly high or low values. To prevent the misleading result from non-smoothed SIRs the Bayesian smoothing with three different models was carried out. Result shows that the DIC differences between the 3 models for appendectomy and incidental appendectomy seem very small (< 0:2%), for example, in case of appendectomy: Poisson-Gamma model (DIC=474.907), Poisson-Lognormal model (DIC=474.773) and Conditional Auto Regressive model (DIC=474.805). We agreed to choose the smallest DIC, because it is estimated to be the model that would best predict a replicate data set of the same structure as that currently observed, hence the Poisson log-normal is the best model for appendectomy cases. Using this method, the value of mean relative risks were smoothed, the range of non-smoothed SIRs is 0.78 and 1.36, then after smoothing method is 0.81 and 1.32.

We repeated the analysis in incidental appendectomy. Based on the value of DIC, Poisson-Gamma model is the best model for incidental appendectomy. The value of mean relative risks were smoothed, districts Aarlen were not significant anymore compare with value of non-smoothed SIR before. The range of SIR is 0.18 - 3.18 and the range of mean RR is 0.32 - 2.67.

Finally, we can conclude that the model of smoothed SIRs (mean relative risks) of appendectomy and incidental appendectomy cases among districts in Belgium for 2001-2006 periods are not related with the environment. The best model for both cases are Poisson Lognormal and Poisson Gamma, respectively.

When we considered the analysis of appendectomy and incidental appendectomy cases which have temporal dimension using the Bernardinelli model, the risk of appendectomy cases significantly decreased in time where the risk was multiplied by approximately 0.9888 every year. In the other hand, the incidental appendectomy cases increased from 2001-2006 and the increase was significant for incidental appendectomy cases over time. The risk was multiplied by approximately 1.0392 every year.

REFERENCES

1. Bernardinelli, L., Clayton, D., Pascutto, C. Montomoli, C., Ghislandi, M. and Songini, M. (1995). Bayesian Analysis of Space-Time Variation in Disease Risk. *Statistics in Medicine*, 14, 2433-2443.
2. Besag, J., J. York, and A. Mollie (1991). Bayesian Image Restoration with Two Applications in Spatial Statistics. *Annals of the Institute of Statistical Mathematics*, 43, 1-59.
3. Clayton, D. and Hills, M. (1995). *Statistical Model in Epidemiology*. Oxford University Press, Oxford.
4. Clayton, D. and Bernardinelli L. (1992). *Bayesian Methods for Mapping Disease Risk. In Geographical and Environmental Epidemiology: methods for small-area studies*. (P. Elliot, J. Cuzick, D. English, and R. Stern ed.), 205-220. Oxford University Press.

5. Faes, C. and Abrahantes, J.C. (2009). Lecture Notes for Disease Mapping Course. Censtat - University of Hasselt. Belgium.
6. Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using Multiple Sequences. *Statistical Science,* 7, 457-72.
7. Lawson, A., Browne, W.J, and Vidal Rodeiro, C. (2003). *Disease Mapping with WinBUGS and MLwiN* John Willey & Sons Inc. USA.
8. Spiegelhalter, D.J., Best N.G., Carlin B.P. and Van der Linde, A. (2002). Bayesian measures of model complexity and fit. J. *Roy. Stat. Soc. Ser. B-Stat. Met*. 64: 583-616, Part 4.
9. Waller L.A. and Gotway C.A. (2004). *Applied Spatial Statistics for Public Health Data*. John Willey & Sons, Inc. New Jersey.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 431-438

COMPARISON OF BINARY, UNIFORM AND KERNEL GAUSSIAN WEIGHT MATRIX IN SPATIAL AUTOREGRESSIVE (SAR) PANEL DATA MODEL

Tuti Purwaningsih, Dian Kusumaningrum, Erfiani and **Anik Djuraidah**
Department of Statistics, Bogor Agricultural University,
Jl. Meranti Wing 22 level 4-5 Kampus IPB Darmaga Bogor, West Java-Indonesia
Email: purwaningsiht@yahoo.com

ABSTRACT

One of field in statistics like econometrics is discussing about spatial influence in many economics data. This research try to analyze spatial in panel data model. Panel data is combining cross-section data and time series data. If the cross-section is locations, need to check the correlation between locations. ρ is parameter in spatial autoregressive model to cover effect of data correlation between location. Value of ρ will influence the goodness of fit model, so it is important to make parameter estimation. The effect of another location is covered by make contiguity matrix until get spatial weighted matrix (W). There are some type of W, it is Binary W, Uniform W, Kernel Gaussian W and some W from real case of economics condition or transportation condition from locations. This study is aim to compare Binary W, Uniform W and Kernel Gaussian W in spatial autoregressive panel data model (SAR panel data) using RMSE value. The result of analysis showed that Uniform Weight has RMSE value less than Binary and Kernel Gaussian Weight in SAR panel data.

KEYWORDS

Latent variables, structural equation modeling, partial least square, satisfaction level, green audits programs

1. INTRODUCTION

Panel data analysis is combining cross-section data and time series data, in sampling when the data is taken from different location, it's commonly found that the observation value at the location depend on observation value in another location. In the other name, there is spatial correlation between the observation, it is spatial dependence. Spatial dependence in this study is covered by Generalized spatial model which is focussed on dependence between locations and error [1]. If there is spatial influence but not involved in model so error assumption that between observation must be independent will not fulfilled. So the model will be in bad condition, for that need a model that involve spatial influence in the analysis panel data that will be mentioned as Spatial Panel Data Model.

Some recent literature of Spatial cross-section data is Spatial Ordinal Logistic Regression by Aidi and Purwaningsih [2], Geographically Weighted Regression [3]. Some of the recent literature of Spatial Panel Data is forecasting with spatial panel data [3] and spatial panel models [4]. For accomodate spatial dependence in the model, there is Spatial weighted matrix (W) that is important component to calculate the spatial correlation between location. Spatial parameter in Spatial autoregressive panel data model, known as ρ. There are some type of W, it is Uniform W, Binary W, Invers distance W and some W from real cases of economics condition or transportation condition from the area. This research is aim to compare Binary W, Uniform W and Kernel Gaussian W in SAR panel data model using RMSE value which is obtained from simulation.

2. LITERATURE REVIEW

2.1 Data Panel Analysis

Data used in the panel data model is a combination of crosssection and time-series data. crossection data is data collected at one time of many units of observation, then time-series data is data collected over time to an observation. If each unit has a number of observations across individuals in the same period of time series, it is called a balanced panel data. Conversely, if each individual unit has a number of observations across different period of time series, it is called an unbalanced panel data (unbalanced panel data).

In general, panel data regression model is expressed as follows:

$$y_{it} = \alpha + \mathbf{x'}_{it}\boldsymbol{\beta} + u_{it} \quad (1)$$

$i = 1,2, \ldots, N$; $t = 1,2, \ldots, T$

with i is an index for crossection data and t is index of time series. α is a constant value, β is a vector of size K × 1, with K specifies the number of explanatory variables. Then y_{it} is the response to the individual cross-i for all time periods t and x_{it} are sized K × 1 vector for observation i-th individual cross and all time periods t and u_{it} is the residual / error [5].

Residual components of the direction of the regression model in equation (1) can be defined as follows:

$$u_{it} = \mu_i + \varepsilon_{it} \quad (2)$$

where μ_i is an individual-specific effect that is not observed, and ε_{it} is a remnant of crossection-i and time series-t [5].

2.2 Spatial Weighted Matrix (W)

Spatial weighted matrix is basically a matrix that describes the relationship between regions and obtained by distance or neighbourhood information. Diagonal of the matrix is generally filled with zero value. Since the weighting matrix shows the relationship between the overall observation, the dimension of this matrix is NxN [6]. There are several approaches that can be done to show the spatial relationship between the location, including the concept of intersection (Contiguity). There are three types of intersection, namely Rook Contiguity, Bishinop Contiguity and Queen Contiguity [6].

After determining the spatial weighting matrix to be used, further normalization in the spatial weighting matrix. In general, the matrix used for normalization normalization row (row - normalize). This means that the matrix is transformed so that the sum of each row of the matrix becomes equal to one. There are other alternatives in the normalization of this matrix is to normalize the columns of the matrix so that the sum of each column in the weighting matrix be equal to one. Also, it can also perform normalization by dividing the elements of the weighting matrix with the largest characteristic root of the matrix ([6]; [7]).

There are several types of Spatial Weight (W): binary W, uniform W, invers distance W (non-uniform weight) and and some W from real case of economics condition or transportation condition from the area. Binary weight matrix has values 0 and 1 in off-diagonal entries; uniform weight is determined by the number of sites surrounding a certain site in ℓ-th spatial order; and non-uniform weight gives unequal weight for different sites. The element of the uniform weight matrix is formulated as,

$$W_{ij} = \begin{cases} \frac{1}{n_i^{(l)}}, j\ is\ neighbor\ of\ i\ in\ l-th\ order \\ 0, others \end{cases} \quad (3)$$

$n_i^{(l)}$ is the number of neighbor locations with site-i in ℓ-th order. The non-uniform weight may become uniform weight when some conditions are met. One method in building non-uniform weight is based on inverse distance. The weight matrix of spatial lag k is based on the inverse weights $1/(1 + d\_ij)$ for sites i and j whose Euclidean distance dij lies within a fixed distance range, and otherwise is weight zero. Kernel Gaussian Weight follow this formulla:

$$w_j(i) = \exp\left[-\frac{1}{2}\left(d_{ij}/_b\right)^2\right] \quad (4)$$

with d is distance between location i and j, then b is *bandwith* which is a parameter for smoothing function.

2.3 Spatial Autoregressive Panel Data Model (SAR-Panel Data)

Autoregressive spatial models expressed in the following equation:

$$y_{it} = \rho \sum_{j=1}^{N} w_{ij} y_{jt} + \mathbf{x}'_{it}\boldsymbol{\beta} + \mu_i + \varepsilon_{it} \quad (5)$$

ρ where is the spatial autoregressive coefficient and w_{ij} is elements of the spatial weighted matrix which has been normalized (W). Estimation of parameters in this model use Maximum Likelihood Estimator [7].

3. METHODOLOGY

3.1 Data

Data used in this study was gotten from simulation using SAR panel data model as equation (5) with initiation of some parameter. Simulation was done use R program. The following step in methods is used to generate the spatial data panel which is consist of index n and t. Index n indicates the number of locations and index t indicates the number of period in each locations, the result can be look at Table 1.

3.2 Methods

1. Determine the number of locations to be simulated is N = 3, N = 9 and N = 25
2. Makes 3 types of Map Location on step 1
3. Creating a Binary Spatial weighted matrix based on the concept of Queen Contiguity of each type of map locations. In this step, to map the 3 locations it will form a 3x3 matrix, 9 locations will form a 9x9 matrix and 25 locations form a 25x25 matrix.
4. Creating Spatial Uniform weighted matrix based on the concept of Queen Contiguity of each type of map locations.
5. Making weighted matrix kernel gaussian based on the concept of distance. To make this matrix, previouly researchers randomize the centroid points of each location. After setting centroid points, then measure the distance between centroids and used it as a reference to build Kernel Gaussian W.

   Gaussian kernel W as follows:

   $$w_j(i) = exp\left[-\frac{1}{2}\left({}^{d_{ij}}/_{b}\right)^2\right] \quad (3)$$

6. Specifies the number of time periods to be simulated is T = 3, T = 6, T = 12 and T = 24
7. Generating the data Y and X based on generalized spatial panel data models follows equation (5).
8. Cronecker multiplication between matrix Identtity of time periods and W, then get new matrix named IW.

9. Multiply matrix IW and Y to obtain vector WY.
10. Build a spatial panel data models and get the value of RMSE
11. Repeat steps 7-9 until 1000 replications for each combination on types of W, N, T, ρ and λ.

    Description:
    Types of W: W Binary, W Uniform and Gaussian kernel W
    Types of N: 3, 9 and 25 locations
    Types of T: 3, 6, 12 and 36 Series
    Types of $\rho = 0.3, 0.5, 0.8$ and $\lambda = 0.3, 0.5, 0.8$
12. Get the RMSE value for all of 1000 replications oh each combination between W, N, T and ρ.
13. Determine the best W based on the smallest RMSE for all combinations.

4. RESULT AND DISCUSSION

Simulation generate data for vector Y as dependent variable and X matrix as independent variable. Y and X is generate with parameter initiation. After doing simulation, we can get RMSE for each combinations and proccessing it, then we can calculate RMSE for each W, N, T, ρ and λ. Here is the result. With the result in Table 1 then continued to figure it into graphs in order to look the comparison easily.

Table 1
Value of RMSE resulted from Simulation for all the combinations (W, N, T, ρ)

W Types	Location Types	Periods Types	RMSE	Average RMSE	Average RMSE
Binary W	N=3	T=3	1.562	1.930	1.606
		T=6	3.757		
		T=12	1.188		
		T=36	1.212		
		Average	1.93		
	N=9	T=3	1.324	1.385	
		T=6	1.389		
		T=12	1.406		
		T=36	1.422		
		Average	1.385		
	N=25	T=3	1.48	1.505	
		T=6	1.501		
		T=12	1.513		
		T=36	1.524		
		Average	1.505		
Uniform W	N=3	T=3	1.086	1.163	1.287
		T=6	1.163		
		T=12	1.188		
		T=36	1.213		
		Average	1.163		
	N=9	T=3	1.3	1.320	
		T=6	1.316		
		T=12	1.332		
		T=36	1.333		
		Average	1.32		
	N=25	T=3	1.363	1.379	
		T=6	1.38		
		T=12	1.389		
		T=36	1.385		
		Average	1,379		
Kernel Gaussian W	N=3	T=3	1.052	1.150	1.559
		T=6	1.133		
		T=12	1.191		
		T=36	1.224		
		Average	1.15		
	N=9	T=3	1.353	1.431	
		T=6	1.425		
		T=12	1.461		
		T=36	1.484		
		Average	1.431		
	N=25	T=3	1.922	2.099	
		T=6	2.076		
		T=12	2.166		
		T=36	2.232		
		Average	2.099		

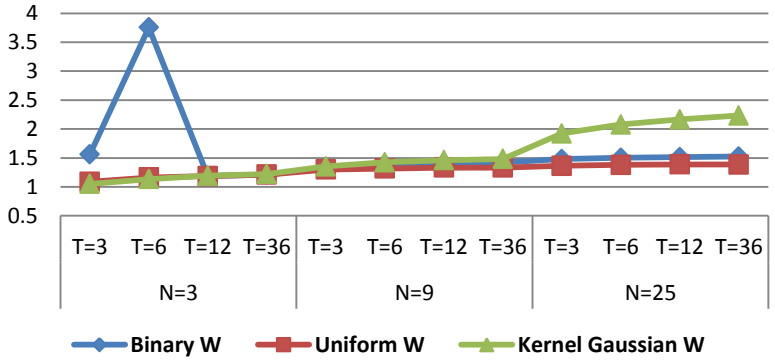


Figure 1: RMSE between Binary, Uniform and Kernel Gaussian Weight for Combinations N and T

Based on figure 1 can be said that Uniform W has smaller RMSE than Binary and Kernel Gaussian W for almost combinations of N types and T types. If we look the level of stabilization, Uniform W is better than Binary and Kernel Gaussian W. We can look at the graph in red line as Uniform W, it has value only in range 1 until 1.5 then Kernel Gaussian W has range from 1-2.5 and Binary from 1-4. So can be concluded that Uniform W is better than Binary and Kernel Gaussian W in SAR panel data model.

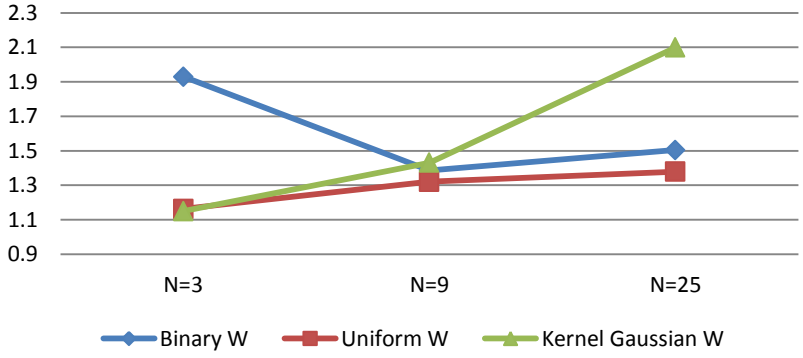


Figure 2: Comparison RMSE of W based on N types

Figure 2 try to analyze diferencies between the W based on N types (3 locations, 9 locations and 25 locations). With graph above can be concluded that Uniform W has smallest RMSE IN all N types.

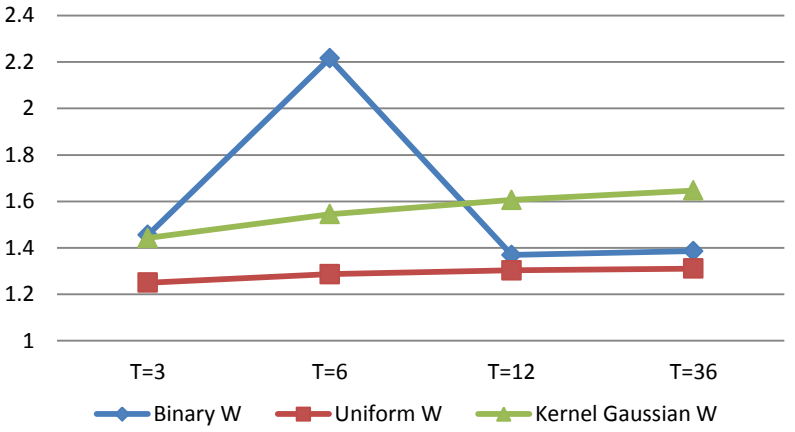


Figure 3 Comparison RMSE of W based on T types

Figure 3 try to analyze diferencies between the W based on T types (3 periods, 6 periods, 12 periods and 36 periods). With graph above can be concluded that Uniform W has smallest RMSE in all types of T.

5. CONCLUSION

Based on simulations result and after explorating the RMSE, can be concluded that Uniform W is the best W in SAR panel data model.

ACKNOWLEDGEMENTS

The first, authors would like to thankful to Allah SWT, my parents, lecturer and all of friends.

REFERENCES

1. Anselin L, Gallo Julie and Jayet Hubbert. (2008). *The Econometrics of Panel Data*. Berlin: Springer.
2. Aidi, M.N. and Purwaningsih T. (2012). Modelling Spatial Ordinal Logistic Regression and the Principal Component to Predict Poverty Status of Districts in Java Island. *International Journal of Statistics and Application*.
3. Fotheringham A.S., Brunsdon C. and Chartlon M. (2002). *Geographically Weighted Regression, the analysis of spatially varying relationships*, John Wiley and Sons, LTD.
4. Elhorst. (2011). Spatial panel models. Regional Science and Urban Econometric.
5. Baltagi, B.H. (2005). Econometrics Analysis of Panel Data. Ed ke-3. England: John Wiley and Sons, Ltd.
6. Dubin R. (2009). Spatial Weights. Fotheringham AS, PA Rogerson, editor, *Handbook of Spatial Analysis*. London: Sage Publications.
7. Elhorst, J.P. (2010). Spatial Panel Data Models. Fischer MM, A Getis, editor, *Handbook of Applied Spatial Analysis*. New York: Springer.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 439-452

EMPLOYEE INNOVATION: MANAGEMENT PRACTICES AFFECTING THE INNOVATIVE BEHAVIOR AT WORKPLACE

Samiah Ahmed, Munir Ahmad and **Suleman Aziz Lodhi**
National College of Business Administration and Economics
Lahore, Pakistan
Email: samiahahmed21@gmail.com
drmunir@brain.net.pk
sulemanlodhi@gmail.com

ABSTRACT

With globalization and increasing competing demands, higher education institutions are becoming responsive to the pressures of competition. These pressures are forcing educational sector to become efficient, effective and innovative simultaneously, adding potential to the M.Phil. and Ph.D. Scholars of Pakistan. The primary data was collected through self-administered questionnaire, at a conference, held in Pakistan. All the M.Phil. and Ph.D. scholars from all over Pakistan comprised the population for the study. Linear and Multiple regressions revealed that higher climate for innovation, lower inter role conflict and higher work life imbalance increase the research scholars' innovation at their universities or research institutions. This study seeks to be a unique attempt to look at the employee innovative behavior at work in the universities situated in the three provinces, Punjab, Sindh and Khyber Pakhtunkhwa, for the first time in Pakistan.

KEYWORDS

Climate for innovation, inter role conflict, work life imbalance, employee innovation.

INTRODUCTION

Due to a rapidly changing economic environment, globalization, and challenges emerging with the increasing societal needs, it has become of uttermost importance to generate the profitability and productivity of the organizations (Ghamin and Rasheed, 2006). Being inflexible and not adaptable to change is a step towards disaster, whether we are dealing within any sector, be it an academic institution. The educational pressures and challenges are forcing educational sector to become efficient, effective and innovative simultaneously, adding potential to the M.Phil. and Ph.D. Scholars of the educational and research institutions all around the world.

Research institutions and Universities are knowledge organizations where individuals can change their environment, where there is commitment to constant and lifelong learning and growth. Universities that want to be innovative have to flourish knowledge mechanisms, so that they do not stagnate in the long run (Friedman and Pollack, 2005).

HR practices are a bundle consisting of management practices for employees, which give them the knowledge, expertise, skills, confidence and motivation to behave in an

innovative behavior. Baker and Sinkula (2002) explored that there are different sets of HR practices needed for different innovation. Laursen and Foss (2003) argued that the impact of adopting of package of complementary HR practices could affect innovative performance much more strongly. Also, De Leede and Looise (2005) have indicated that there is a further need to research the most appropriate HRM practices in the various innovative stages. There is a further elaborate research needed on engagement and involvement in creative work tasks (Carmeli and Schaubroeck, 2007).

Studies and evidence about various practices which leads to creative and innovative behaviors has only begun (Carmeli and Spreitzer, 2009). So, the aim of our research is to explore the right bundle of HR practices conducive to the employees' innovative behavior, and which is also flexible and fit within the organization. In this study, we aim to determine the impact of three factors influencing employee innovation in the research institutions in Pakistan. The three independent variables are climate for innovation, inter role conflict and work life imbalance.

RESEARCH QUESTIONS OF THE STUDY:

i) What is the impact of climate for innovation on employee innovation?
ii) What is the impact of inter role conflict on employee innovation?
iii) What is the impact of wok life imbalance on employee innovation?

REVIEW OF THE LITERATURE AND HYPOTHESIS DEVELOPMENT

There is still no universally accepted definition of innovation. Creativity, innovativeness, and innovation are all related concepts that are frequently used interchangeably. Schumpeter (1934) defined innovation as the introduction of a new product, process, method, or system. Scientific problems and hunches and existing knowledge give birth to new ideas which then feed the beginnings of other innovative explorations (Machlup, 1962).

Increasing number of scholars and academicians have focused on determinants of individual innovation in the organizations and tried to answer the questions like, what drives employees to be creative?

Creativity is a novel/complex process which requires the structuring of the required jobs, processing tasks as a useful means to enhance creative or innovative behaviors (Binyamin and Carmeli, 2010). The work based learning strategies cultivates new ideas circling in the organization and facilitating employee learning and innovation (Bond and Flaxman, 2006). This was supported by Holman et al., (2012) who concentrated on the relationship between job design characteristics and innovation which was strengthened by the mediating factor of work based learning strategies. Another research studied empowerment practices (sharing authority, resources, information, and rewards) generating innovative proposals by the employees in the US Federal government (Fernandez and Moldogaziev, 2013).

There are several other factors as antecedents of individual innovations. The leadership factors like loyalty, affect and professional respect had been proved to predict an innovative climate in Estonian enterprises (Alas, Ubius and Vanhala, 2011). In another

study, Expected image outcomes and expected positive performance gains are indeed directly positively related to innovative behavior (Cingoz and Akdogan, 2011; Yuan and Woodman, 2010).

Climate and leadership are the factors which support creative performance of individuals (Shin and Zhou, 2003). The climate for innovation refers to the working environment, which includes the formal organization, which includes the organizational communication and the pattern communication and which also includes the catering to the special needs of the employees (Zoubi, 2006).

Positive climate can influence creativity. A study conducted by Khaja (2006) revealed that the impact of organizational climate on the empowering the employees in the Federal Government of UAE was positively high. Also, DiLiello and Houghton (2006) identified that individuals who receive organizational support in the form of self-leadership, were more capable of creativity and innovation in the German Industrial organizations. Managers can direct conversations discussing positive outcomes, they can celebrate small wins, pointing out employees' strengths and weakness and not dwelling over their mistakes or problems and telling the success stories of the employees (Cabrera, 2012).

A recent research was conducted by (Nusair, 2013) where he highlighted the importance of positive climate in relation to the Job performance in the commercial banks of Jordan. Recently, another research sheds light over the relationship between innovative work behavior and organizational climate among the knowledge workers in Malaysia (Kheng and Mahmood, 2013). When the leaders maintain a supportive behavior to employees, they circulate the knowledge around the organizations, whereby giving the employees a chance to address problems at hand creatively and to increase their creativity performance in terms of originality and fluency (Carmeli et al., 2013). Ethical leadership had been a predictor of individual innovation through the mediating factor of intrinsic motivation at the individual level (Yidong and Xinxin, 2013).

Past research studies show that the influence of the working climate on the innovation strategy remains rather limited (Nybakk and Jenssen, 2012). Climate for innovation is also a striking factor for universities and research institutions. Perceptions of Employees regarding the work environment at Jordian Private Universities were explored and their impact on the organizational creativity (Arabiyat, Balqaa and Al Saleem, 2011). This area of research was further carried out when a study was conducted to find out the impact of organizational climate on the innovative behavior at Jordanian private Universities which turned out to be significantly high (Al-Saudi, 2012).

The frequency of stress and burnout is rising (e.g., Kahn and Langlieb, 2003). This increase in stress and burnout is due to the inter role conflict, rising between work and family demands, experienced by most working individuals at work places. Role stress is consisted of role ambiguity and role conflict (Leung et al., 2011). Inter role-conflict has two important types. Work family conflict states that involvement in the work disturbs the involvement in the family whereas the family work conflict (FWC) states that involvement in the family hinders the progress in the work of the organisation (Greenhaus et al., 1985).

Keeping in line with these concepts, recently two years ago, a research paper was written on the U-shaped relationship between role stress and innovative performance. Inter role conflict had also been examined with the three dimensions of burnout out i.e., high emotional exhaustion, high depersonalization with others, and lower levels of feeling of accomplishment (Jawahar et al., 2012). Another study concentrated on the curvilinear effects of role stress, a type of hindrance stress, on the innovative performance of employees, mediated with the low perceived support for innovation in Taiwan and mainland China (Leung et al., 2011). In another recent study, the researchers examined role conflict as a mediator of the leader membership exchange and stress relationship and the mediating role Job Involvement and Role Conflict (Lawrence and Michele, 2012).

Work role resilience also increases the work satisfaction, whereas, work-family conflict is positively correlated to family-work conflict and negatively to work satisfaction. This is supported in a study where role salient reduces the inter role conflict and increases the satisfaction of dual earner couples with their work and family (Bhowon, 2013).

Another study also revealed that the higher the degree of role resiliency, the more reduced the impact of intra role conflict on job stress among the salespeople, leading to positive attitudes and increasing performance in terms of productivity and employee innovation (Krush et al., 2013).

Also, Sun, Wu and Wang (2011) revealed that role overload is a predictor of occupational stress among Chinese university teachers. This had been consistent with Kebelo (2012) who stated that 24.6% of total strain stem from role overload, role ambiguity and role conflict. In another research, the stress predictors including role conflict, role overload and role ambiguity were all related to psychological strain among university lecturers in the University of Dammam, Saudi Arabia (Jdaitawi et al., 2014). This is consistent with the previous study of Idris (2011) who contented that role stressors are related to psychological strain.

Inter role conflict also leads to job dissatisfaction and intention to leave. In another study, teacher assistants faced role ambiguity and role conflict leading to job dissatisfaction and intention to leave (Fatima and Rehman, 2012).

Prior decades have witnessed a small body of research pointed to role conflict and stress and psychological strain but studies of this caliber are still few (Achour and Boerhannaoeddin, 2011). While many antecedents of innovativeness have been studied (Zhou and Shalley, 2003), the impact of role stress has rarely been examined.

Previously, the Research has investigated a surge of definitions for work family imbalance. The work and life imbalance was first introduced by Lewin (1951) who proposed that work and "non-work" are distinct domains which is separated by a boundary. Kanter (1977) also stated that work life imbalance is a ''myth of two separate worlds. Work Life Imbalance is the inability to balance the family life/activities with the work life/activities (Netemeyer, et al., 1996). When an individual fully focuses on the work domain or is a workaholic, he fails to consider the other aspects of life including

family life, family duties and obligations and other household responsibilities and his personal interests.

Research predicted that Work life imbalance occurred when school teachers experienced long working hours with disproportionate number of working days (Madipelli, Sarma and Chinnappaiah, 2013) thereby displaying creativity in their education institution. People may work long hours for a strong desire for career development (Schaufeli, Shimazu and Taris, 2009).

Enthusiastic and well engaged have three properties of workaholism. These properties include work involvement, drive (a feeling of being compelled to work), and work enjoyment. These are the Properties found in the researcher as well. Researchers are well engrossed in their research work and are worakaholics or enthusiastic people with an opportunity cost to their success. Highly Educated individuals have increased responsibilities, heavier workloads, and high work drive with compulsive behavior (Flowers and Robinson, 2002), When an individual does more work than is expected, strictly for the fun of it, then it is treated as an enjoyment of work (Tabassum and Rahman, 2012).

There are evidences that workaholic and well engaged individuals suffer in diminished sleep quality (Kubota, Shimazu, Kawakami, Takahashi, Nakata and Schaufeli, 2010) increased marital problems (Robinson, Flowers and Caroll, 2001), lower relationship satisfaction (Bakker et al., 2009) thereby leading to work life imbalances (Aziz and Zickar, 2006; Aziz et al., 2013).

Therefore, it is important to understand these factors which predict the employee innovation. In the next section, we draw from the broad literature three hypotheses about the factors that might encourage the research scholars in the research universities of Pakistan to innovate.

STUDY HYPOTHESES

In this section, we identify and discuss three factors that are likely to encourage innovative behaviour. Therefore, based on the aforementioned study background, the objectives, prior theoretical and empirical research rationales; this research proposes the following list of hypotheses below

$H_{(11)}$: All else being equal, organizations with higher Climate for Innovation has higher employee innovation.

$H_{(12)}$: All else being equal, organizations having lower inter role-conflict has higher employee innovation

$H_{(13)}$: All else being equal, organizations having work-life imbalance have higher employee innovation

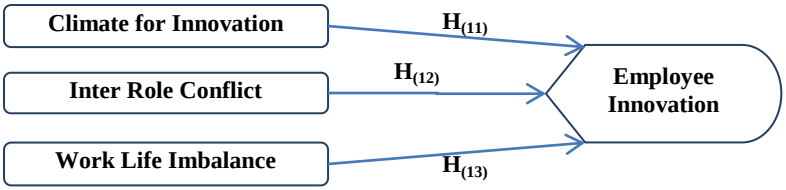


Fig. 1: Theoretical Framework

METHODS

Sample and Data

We began this research by asking how research scholars stimulate creativity at their university and in research environments. A pilot survey was conducted before, at the National College of Business and Economics, Lahore, Pakistan. From the feedback it was recognized that there was no ambiguity in understanding the terminologies or items.

The primary data was collected through self-administered questionnaire. We collected the data over the time horizon of three days at an International Conference on Statistical Sciences that was held at an university, in Karachi, Pakistan. The researcher first approached the Director of Research of the university, and explained the objectives of the survey research and asked for Director Research's approval to commence the distribution of the questionnaire after the inauguration of the conference ceremony.

Since the frame (list) of scholars was not available, All the M.Phil. and Ph.D. scholars from all over Pakistan comprised the population for the study. The researcher considered the scholars attending the conference as a purposive sample from Pakistan and administered 80 questionnaires to the sampled scholars during the conference sessions, skimming all the M.Phil. and Ph.D. scholars and excluding all the other participants present at the Karachi conference. The total complete and useable questionnaires returned back were 59, i.e. the response rate was quite high.

VARIABLE DEFINITIONS AND MEASURES

Employee innovation (EI) is measured as the creative behaviour, which is defined as behaviour consisting of activities that generate ideas that are novel and useful (Amabile, 1988; Oldham and Cummings, 1996). Furthermore, we wanted to research over the innovative behaviour of the M.Phil. or Ph.D. scholars in their respective universities or research institutions of Pakistan, so we constructed the items as per our requirement. There are 11 items for employee innovation. Sample items include 'How many research papers have you published so far?' The Cronbach Alpha of 0.839 showed the reliability of these items which was successful in capturing the consistence in the degree of innovative behaviour of the scholars.

Climate for innovation (CI) is measured as the degree of support and encouragement an organization provides to its employees to take initiative and explore innovative approaches, which predicts the actual innovation in the organization (Sarros et al., 2008). Sample items include 'Does your institution allow flexible working hours?' The internal consistency of these items was measured through Cronbach Alpha (0.703).

Inter-role-conflict (IRC) is observed as a situation where the role pressures of the two different domains are not compatible to each other (Greenhaus and Beutell, 1985). We selected the inter role conflict scale from past research study (Carlson, Kacmar and Williams, 2000), which combined work–family conflict (WFC) and family–work Conflict (FWC) scales into one measure. Sample items included 'My work keeps me from my family activities more than I would like.' All the questions for the inter role conflict construct use a five point Likert scale ranging from 'strongly disagree' (1) to 'strongly agree' (5).

Work Life imbalance (WL.IMB) occurs when the employees are captivated by work and this excessive work hinders one or more life functions (Porter, 2001). We also created the items for work life imbalances to measure the imbalance created in the scholars' lives due to their work load and creative behaviour which drive these individuals away from their homes and family duties and obligations. Sample items include 'on average, how many hours/day do you dedicate to research?' All the five questions in this section are based on five point Likert scale, ranging from (1= 1-6 hours; to 5= above 9 hours). The Cronbach Alpha appeared to be 0.550.

DEMOGRAPHICS

Some salient features of demographic characteristics of the respondents are in place in this section. Of the 59 respondents, 61% are males and 39% are females; 36% are single and 64% are married. 66% of the scholars belong to the public sector universities and 34% belong to the private sector. Since these persons come from all over Pakistan, 32 are from Sindh, 19 from Punjab and 9 are from Khyber Pakhtunkhwa and none from Baluchistan. Of 59 respondents, 61% have M.Phil. degrees and 39% are holding Ph.D. degrees. Moreover, 25% of males hold Ph.D. degrees and 36% males have M.Phil. whereas only 04% of females are having Ph.D. degrees and 25% of females have M.Phil. degrees.

The average age of respondents is about 36 years whereas the average age of males is about 39 years and of female is 35 years. Among all respondents, about 47.5% are Lecturers, 52.5% are of professorial ranks and among the professorial ranks, about 35.6% belong to Assistant Professorship, 3.4 % are Associate Professors and 13.6% are full Professors. The average experience by designation is that for all designations, the average experience is 11.2 years whereas for Lecturers the experience is 3.3 years, for Assistant Professors is 8.9 years where as for Associate and Full Professors are respectively 27.5 years and 34.8 years.

RESULTS

Linear and multiple regression analysis were carried out to test the proposed hypotheses. All hypothesized associations are shown to be true with the support of data.

LINEAR REGRESSION

Climate for innovation and employ innovation There is a strong positive relationship between climate for innovation and outcome variable employee innovation, statistically

significant at the p <0.05 level. The regression model with the co efficient beta =0.282 at p=0.039<0.05 significantly predicts the outcome variable.

Inter role conflict and employee innovation There is a negative relationship between the predictor variable, inter role conflict and the employee innovation, which is insignificant one; The regression model with the standardized co efficient beta = –0.77 at p= 0.576 >0.05 predicts that there is an inverse relationship between the inter role conflict and Employee innovation. The negative beta weight indicates that if employee innovation needs to be increased, it is necessary to reduce the Inter role conflict. This means the higher the conflict between the family domain and work domain, the lower the employee innovative behavior and the lower the inter role conflict, the higher the employee innovative behavior.

Work life imbalance and employee innovation Lastly there is a positive relationship between the third predictor variable i.e. work life imbalance and employee innovative behavior at workplace. The above regression model reveals Co efficient Beta =0.314 at significance level p=0.019<0.05 which significantly predicts the outcome variable i.e. employee innovation in M.Phil. and Ph.D. professionals.

MULTIPLE REGRESSION

The multiple regressions showed the three individual relationships with the dependent variable i.e. employee innovation.

The fitted standardized multiple regression model is

EI = 1.766 + 1.473WL.IMB + 0.420 IRC + 0.995CI,

Regression Analysis by Fitting Multiple Linear Regression between EI and WL.IMB, IRC and CI

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	1.766	17.773		0.099	0.921
Work life Imbalance	1.473	0.646	0.298	2.280	0.027
Inter Role conflict	0.420	3.361	0.017	0.125	0.901
Climate For Innovation	0.995	0.453	0.287	2.198	0.033
a. Dependent Variable: Employee Innovation					

The impact of work life Imbalance on employee innovation is at the significant level of 0.027 with the co efficient beta 0.0298, whereas the impact of Climate for innovation is significant at p=0.033<0.05 with the co efficient beta 0.287. Lastly, in the multiple regression, again the relationship of inter role conflict with employee innovation is insignificant one. Synergic effect or multiple regression reported increased employee innovation due to the combination of climate for innovation and work life imbalance. This further depicts that the combined efforts of ‘climate for innovation’ and ‘work life

imbalance' elicit employee innovation at work, suppressing the impact of inter role conflict on employee innovation.

Finally, By adopting linear and multiple regression, both $\mathbf{H_{(11)}}$ and $\mathbf{H_{(13)}}$ got supported and alternative hypothesis $\mathbf{H_{(12)}}$ did not receive any support from the data.

DISCUSSION

The alternative hypothesis, H_1 predicted that climate for innovation is positively related to employee innovation. This proved out to be true in our data analysis, thus replicating the finding from past research (Margianti et al., 2004) and also consistent with the findings of (Groenveld, 2010).

The practical explanation of this increasing scholars' innovation is that the faculty members are indeed surrounded by a positive Climate for innovation. Another recent study also illustrated the positive relation between innovative climate and innovative behavior at workplace among the Knowledge workers in Malaysia (Kheng and Mahmood, 2013).

Hypothesis H_2 predicted that inter role conflict is negatively associated with employee innovation. The hypothesis did not get any support from the data. The result shows an insignificant negative relationship with employee innovation. Inter role conflict is extremely important variable in carrying out research, as research for scholars demand no hindrance stressors. High stress from the family domain could hinder the research progress and create natural delays in research work and lead to emotional distress, burnout (Cinamon et al., 2007).Writing a research paper demands hard work of many days and nights, even months, which mean an individual would devote less time and energy to Family responsibilities and have to miss family activities due to the time he/she spends on his/her work activities (Bolino and Turnley, 2005).

Hypothesis H_3 posited that the more the imbalance between the Researcher's life and work, the higher the employee innovation. A researcher can only be innovative by devoting more time and energy to his research work, lectures, and lesser time to sleep and family. The results revealed that the average number of hours per day spent on institution is 7 hours, on carrying out research projects is 6 hours, average number hours spent with family is 5 hours and average number of hours per day sleep is 6 hours approximately by 59 respondents.

This appeared out to be normal as a researcher can only spend a very limited amount of time and energy with his family while he is loaded with his research responsibilities. These findings are consistent with (Aziz et al., 2010) where workaholism was significantly correlated with high levels of work–life imbalance. If an individual gains the fun and loves his work, he would keep engrossed in his work and would eliminate the division between work and life to attain the sweet success (Premuzic, 2013).

LIMITATIONS AND FUTURE DIRECTIONS

The present study has few limitations. First, the results of the current study are grounded on cross-sectional data Thus, future researchers may benefit from more longitudinal data (Kelly et al., 2008). All the research scholars from the three Provinces

participated in the Conference, except Baluchistan, so we could not identify to what extent climate for innovation, inter role conflict and work life imbalance affect the innovative behavior of the Scholars working in the research institutions of Balochistan.

Secondly, we could not make a comparison between international research scholars and the national scholars' innovative work behavior in terms of the difference in their varying number of research publications and research projects and their participation in the number of conferences, research or academic meetings and workshops which help them build their research display.

Thirdly, we did not measure the impact of monetary or extrinsic rewards on the innovative behavior of the employees. We recommend that leaders in the universities, whose agenda is to be research oriented institution, should express praise and offer extrinsic rewards to group members in R&D settings as a way to promote creativity.

As high education standards are revolving and becoming challenging, a considerable increase in the administrative burden was also experienced by academic staff members. It is noticed that educator's work is generally becoming more complex and demanding (Jackson and Ruthann, 2006) and variables are coming into play with more complexity. To resolve these complexities, Universities or research institutions in Pakistan could reprioritize work objectives and modify work schedules.

Furthermore, this study urges university leaders to facilitate better working environment in order to minimize the conflict in lecturer's or scholars' work. Management should employ steps to control stress and strain in the university environment, which would lead to improving the conditions of lecturers and improving their depleted performance (Jdaitawi et al., 2014).

CONCLUSION

With Globalization and increasing competing demands, Higher education institutions too are becoming responsive to the pressures of competition (Mathew, 2010). These pressures are forcing educational sector to become efficient, effective and innovative simultaneously (Herbs and Comrade, 2011). Education and research work done by the Universities must be of competitive edge satisfying employers' demands and add potential to the M.Phil. and Ph.D. Scholars.

This research is a unique attempt to look at the employee innovative behavior at work among the research professionals in the universities/educational sector situated in the three provinces which were Punjab, Sindh and Khyber Pakhtunkhwa, for the first time in Pakistan. The study contributes to the literature of employee innovation at work by empirically supporting the independent relationships climate for innovation and work life balances with the dependent variable i.e. employee innovation in the universities of Pakistan.

REFERENCES

1. Achour, M. and Boerhannoeddin, A.B. (2011). The role of religiosity as A Coping Strategy fin Coping with Work-Family Conflict and Achieving Employees' Well-

Being. *Paper Presented in International Conference on Social Science and Humanity.* Singapore: IACSIT Press, Singapore.
2. Alas, R., Übius, Ü. and Vanhala, S. (2011). Connections between organisational culture, leadership and the innovation climate in Estonian enterprises. In *E-Leader Conference* (pp. 3-5).
3. Al-Saudi, M.A. (2012). The Impact of Organizational Climate upon the Innovative Behavior at Jordanian Private Universities as Perceived by Employees: A Field Study. *International Business and Management*, 5(2), 14-27.
4. Amabile, T.M. (1988). A model of creativity and innovation in organizations. *Research in Organizational Behavior*, 10, 123-167.
5. Arabiyat, B., Balqaa, A. and Al-Saleem, B.T.I. (2011). The Extent of Application of the Principles of the Organizational Justice and Its Relationship to the Organizational Commitment of the Faculty Members at the University of Jordan. *International Journal of Human Resource Studies*, 1(2), Pages-52.
6. Aziz, S. and Zickar, M.J. (2006). A cluster analysis investigation of workaholism as a syndrome. *Journal of Occupational Health Psychology*, 11, 52-62.
7. Aziz, S., Uhrich, B., Wuensch, K.L. and Swords, B. (2013). The workaholism analysis questionnaire: emphasizing work-life imbalance and addiction in the measurement of workaholism. *Institute of Behavioral and Applied Management*, 71-86.
8. Baker, W.E. and Sinkula, J.M. (2002) Market orientation, learning orientation and product innovation: delving into the organization's black box. *Journal of Market Focused Management*, 5(1), 5-23.
9. Bhowon, U. (2013). Role Salience, Work-Family Conflict and Satisfaction of Dual-Earner Couples. *Journal of Business Studies Quarterly*, 5(2), p78.
10. Binyamin, G. and Carmeli, A. (2010). Does structuring of human resource management processes enhance employee creativity? The mediating role of psychological availability. *Human Resource Management*, 49(6), 999-1024.
11. Bolino, M.C. and Turnley, W.H. (2005). The personal costs of citizenship behavior: The relationship between individual initiative and role overload, job stress, and work-family conflict. *Journal of Applied Psychology*, 90, 740-748.
12. Bond, F.W. and Flaxman, P.E. (2006). The ability of psychological flexibility and job control to predict learning, job performance, and mental health. *Journal of Organizational Behavior Management*, 26, 113-130.
13. Cabrera, E.F. (2012). The Six Essentials of Workplace Positivity. *People and Strategy*, 35(1), 50.
14. Carmeli, A. and Schaubroeck, J. (2007). The influence of leaders' and other referents' normative expectations on individual involvement in creative work. *The Leadership Quarterly*, 18(1), 35-48.
15. Carmeli, A. and Spreitzer, G.M. (2009). Trust, connectivity, and thriving: Implications for innovative behaviors at work. *The Journal of Creative Behavior*, 43(3), 169-191.
16. Cinamon, R.G., Rich, Y. and Westman, M. (2007). Teachers' occupation-specific work-family conflict. *Career Development Quarterly*, 55, 249-261.

17. Cingöz, A. and Akdoğan, A.A. (2011). An empirical examination of performance and image outcome expectation as determinants of innovative behavior in the workplace. *Procedia-Social and Behavioral Sciences*, 24, 847-853.
18. De Leede, J. and Looise, J.K. (2005). Innovation and HRM: towards an integrated framework. *Creativity and Innovation Management*, 14(2), 108-117.
19. DiLiello, T.C. and Houghton, J.D. (2006). Maximizing organizational leadership capacity for the future: Toward a model of self-leadership, innovation and creativity. *Journal of Managerial Psychology*, 21(4), 319-337.
20. Fatima, G. and Rehman, W. (2012). Impact of Role (Ambiguity and Conflict) on Teaching Assistants' Satisfaction and Intention to Leave: Pakistani HEIs. *International Journal of Business and Management*, 7(16), p56.
21. Fernandez, S. and Moldogaziev, T. (2013). Employee empowerment, employee attitudes, and performance: testing a causal model. *Public Administration Review*, 73(3), 490-506.
22. Flowers, C.P. and Robinson, B. (2002). A structural and discriminant analysis of the Work Addiction Risk Test. *Educational and Psychological Measurement*, 62, 517-526.
23. Friedman, H.H., Friedman, L.W. and Pollack, S. (2005). Transforming a University from a Teaching Organization to a Learning Organization. *Review of Business*, 26(3), 31-35.
24. Ghamin, S. and Rasheed, Z. (2006). *Promoting Innovation in the Higher Education: A Study on the Faculty at the University of Mosul*. A study submitted to the Conference on Creativity and the Administrative and Economic Transformation. Yarmouk University, Jordan.
25. Greenhaus, J.H. and Beutell, H.J. (1985). Sources of Conflict between Work and Family Roles. *Academy of Management Review*, 10(1), 76-88.
26. Herbst, T. and Conradie, P. (2011). Leadership effectiveness in Higher Education: Managerial self-perceptions versus perceptions of others. *Journal of Industrial Psychology*, 37(1), 1-14.
27. Holman, D., Totterdell, P., Axtell, C., Stride, C., Port, R., Svensson, R. and Zibarras, L. (2012). Job design and the employee innovation process: The mediating role of learning strategies. *Journal of Business and Psychology*, 27(2), 177-191.
28. Idris, M. (2011). Over Time Effects of Role Stress on Psychological Strain among Malaysian Public University Academics. *International Journal of Business and Social Science*, 2*(*9), 154-161.
29. Jawahar, I.M., Kisamore, J.L., Stone, T.H. and Rahn, D.L. (2012). Differential effect of inter-role conflict on proactive individual's experience of burnout. *Journal of Business and Psychology*, 27(2), 243-254.
30. Jdaitawi, M.T., Mutawa, A.A., Musallam, F. and Talafha, F. (2014). Stress and Psychological Strain among University Lecturers in Saudi Arabia. *Global Conference on Business and Finance Proceedings*, *9*(1), 361-369.
31. Kahn, J. and Langlieb, A.M. (2003). Mental health and productivity in the workplace: A handbook for organizations and clinicians. Hoboken, NJ.
32. Kanter R.M. (1977). *Men and Women of the Corporation*. New York: Basic Books.

33. Kebelo, K.K. (2012). Occupational Role Stressors as Predictors of Psychological Strain among Academic Officers of Higher Educational Institutions. *Pakistan Journal of Psychological Research*, 27(2).
34. Kelly, E.L., Kossek, E.E., Hammer, L.B., Durham, M., Bray, J., Chermack, K., Murphy, L.A. and Kaskubar, D. (2008). Getting there from here. Research on the effects of work-family initiatives on work-family conflict and business outcomes. *Academy of Management Annals*, 2, 305-349.
35. Kheng, Y.K. and Mahmood, R. (2013). The Relationship between Pro-Innovation Organizational Climate, Leader-member Exchange and Innovative Work Behavior: A Study among the Knowledge Workers of the Knowledge Intensive Business Services in Malaysia. *Business Management Dynamics*, 2(8), p15.
36. Krush, M.T., Agnihotri, R., Trainor, K.J. and Krishnakumar, S. (2013). The salesperson's ability to bounce back: examining the moderating role of resiliency on forms of intra-role job conflict and job attitudes, behaviors and performance. *Marketing Management Journal*, 23(1), 42-56.
37. Kubota, K., Shimazu, A., Kawakami, N., Takahashi, M., Nakata, A. and Schaufeli, W.B. (2010). Association between workaholism and sleep problems among hospital nurses. *Industrial Health,* 48, 864-871.
38. Laursen, K. and Foss, N.I. (2003). New human resource management practices, complementarities and impact on innovation performance. *Cambridge Journal of Economics*, 27, 243-283.
39. Lawrence, E.R. and Michele, K.K. (2012). Leader-Member Exchange and Stress: The Mediating Role of Job Involvement and Role Conflict. Journal of Behavioral & Applied Management, 14(1), p39.
40. Leung, K., Huang, K.L., Su, C.H. and Lu, L. (2011). Curvilinear relationships between role stress and innovative performance: Moderating effects of perceived support for innovation. *Journal of Occupational and Organizational Psychology*, 84(4), 741-758.
41. Lewin, K. (1951). *Field theory in social science*. New York: McGraw-Hill.
42. Machlup, F. (1962). *The production and distribution of knowledge in the United States* (Vol. 278). Princeton University Press.
43. Madipelli, S., Sarma, V.S. and Chinnappaiah, Y. (2013). Factors Causing Work Life Imbalance among Working Women-A Study on School Teachers. *Indian Journal of Industrial Relations*, 48(4), p621.
44. Margianti, E.S., Aldridge, J.M. and Fraser, B.J. (2004). Learning environment perceptions, attitudes and achievement among private Indonesian university students. International Journal of Private Higher Education [online]. Retrieved from: www.xaiu.com/xaiujournal
45. Mathew, V. (2010). Service delivery through knowledge management in higher education. *Journal of Knowledge Management Practice*, 11(3), 1-14.
46. Netemeyer, R.G., Boles, J.S. and McMurrian, R. (1996). Development and Validation of Work-Family Conflict and Family-Work Conflict Scales. *Journal of Applied Psychology*, 81(4), 400-410.
47. Nusair, T.T. (2013). The role of climate for innovation in job performance: Empirical evidence from commercial banks in Jordan. *International Journal of Business and Social Science*, 4(3), 208-217.

48. Nybakk, E. and Jenssen, J.I. (2012). Innovation strategy, working climate, and financial performance in traditional manufacturing firms: An empirical analysis. *International Journal of Innovation Management*, 16(2), 1250008-1-1250008-26.
49. Oldham, G.R. and Cummings, A. (1996). Employee creativity: Personal and contextual factors at work. *Academy of Management Journal*, 39(3), 607-634.
50. Porter, G. (2001). Workaholic tendencies and the high potential for stress among co-workers. *International Journal of Stress Management*, 8, 147-164.
51. Premuzic, T.C. (2013). *Embrace Work-Life Imbalance, Finweek*. 40-41.
52. Robinson, B.E., Flowers, C. and Carroll, J. (2001). Work stress and marriage: A theoretical model examining the relationships between workaholism and marital cohesion. *International Journal of Stress Management*, 8, 165-175.
53. Sanders, K.S., Moorkamp, M., Torka, N., Groenveld, S. and Groenveld, C. (2010). How to Support Innovative work behavior? The Role of LMX and Satisfaction with HR Practice. *Technology and Investment*, 1, 59-68.
54. Sarros, J.C., Cooper, B.K. and Santora, J.C. (2008). Building a climate for innovation through transformational leadership and organizational culture. *Journal of Leadership and Organizational Studies*, 15(2), 145-15.
55. Schaufeli, W.B., Bakker, A.B., Van der Heijden, F.M. and Prins, J.T. (2009). Workaholism, burnout and well-being among junior doctors: The mediating role of role conflict. *Work & Stress*, 23(2), 155-172.
56. Schaufeli, W.B., Shimazu, A. and Taris, T.W. (2009). Being driven to work excessively hard: the evaluation of a two-factor measure of workaholism in the Netherlands and Japan. *Cross-Cultural Research*, 43, 320-348.
57. Schumpeter J.A. (1934). *The theory of economic development*. Cambridge, Mass.: Harvard University Press.
58. Shin, S.J. and Zhou, J. (2003). Transformational leadership, conservation, and creativity: Evidence from Korea. *Academy of Management Journal*, 46(6), 703-714.
59. Sun, W., Wu, H. and Wang, L. (2011), Occupational Stress and its Related Factors among University Teachers in China. *J. Occup. Health*, 53, 280-286.
60. Tabassum, A. and Rahman, T. (2012). Gaining the insight of workaholism, its nature and its outcome: A literature review. *International Journal of Research Studies in Psychology*, 2(2), 81-92.
61. Yidong, T. and Xinxin, L. (2013). How ethical leadership influence employees' innovative work behavior: A perspective of intrinsic motivation. *Journal of Business Ethics*, 116(2), 441-455.
62. Yuan, F. and Woodman, R.W. (2010). Innovative behavior in the workplace: The role of performance and image outcome expectations. *Academy of Management Journal*, 53(2), 323-342.
63. Zhou, J. and Shalley, C.E. (2003). Research on employee creativity: A critical review and directions for future research. *Research in Personnel and Human Resources Management*, 22, 165-217.
64. Zoubi, J. (2006). *Organizational Climate Factors Influencing the Creative Behavior of Managers in the Ministries of Jordan: A Field Study* (Unpublished Master Thesis). University of Jordan, Amman, Jordan.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 453-462

SMALL AREA ESTIMATION FOR NON-SAMPLED AREA USING CLUSTER INFORMATION AND WINSORIZATION WITH APPLICATION TO BPS DATA

Rahma Anisa, Khairil A. Notodiputro and **Anang Kurnia**
Department of Statistics, Bogor Agricultural University
Jl. Meranti Wing 22 level 4-5 Kampus IPB Darmaga Bogor, West Java-Indonesia
Email: anangk@apps.ipb.ac.id

ABSTRACT

Empirical Best Linear Unbiased Predictor (EBLUP) is an indirect method that used to estimate parameters of small area. The standard EBLUP usually predicts parameters of non-sampled area using a synthetic model ignoring the area random effects due to lack of non-sampled area information. Hence, this prediction is distorted. The idea of modifying the prediction model by incorporating cluster information to acquire local model has been developed in the previous study. One of the approaches was modifying both intercept and slope of the prediction model assuming that auxiliary variables were random. Since the model requires that the auxiliary variables were fixed then this paper proposed a modification assuming that the auxiliary variables were fixed. A simulation study has been carried out to evaluate the performance of the proposed model compared with standard EBLUP and the existing modified EBLUP models. The criteria used to evaluate the models were the value of Relative Bias (RB) and Relative Root Mean Squares Error (RRMSE). It was shown, by mean of simulations, that the proposed model provided better prediction of non-sampled area compared with other models. Real data from the Centre of Statistic in Indonesia (BPS) was utilized to predict average per capita expenditures per month at sub-district levels in regency and municipality of Bogor using the standard EBLUP as well as the proposed model. The result showed that proposed model has been capable to predict parameters in non-sampled area with smaller RMSE compared with standard EBLUP model. In addition, handling of outliers in the real data using one-sided wisorization has improved the estimation of non-sampled area using the proposed model.

KEYWORDS

EBLUP, Clustering, Linear Mixed Models, Small Area Estimation.

1. INTRODUCTION

Small area estimation (SAE) is used as an alternative approach to estimate the parameters of the areas with a very small or even zero sample size, known as non-sampled area. This demand appears when we need an estimator with good accuracy in a smaller subpopulation level, such as at the level of regency/municipality, sub-district, or even village (Kurnia, 2009). Indirect estimator is preferred to estimate parameters in a small area because it is able to overcome the weakness of direct estimator which can produce a large standard error (Rao, 2003).

Empirical Best Linear Unbiased Predictor (EBLUP) is an indirect method to estimate the parameter of small areas. It has been noted that there is a problem when this model will be used to predict the parameters of non-sampled area. Standard EBLUP model predict the parameters using synthetic model which will ignore the area random effects because of the lack of information of non-sampled area (Saei and Chambers, 2005). As a consequence, the resulting predictive values will be distorted into a single line of the synthetic model which is global model and may cause considerable bias.

The ideas that have been developed in the previous study to acquire local prediction model were incorporating cluster information into standard EBLUP prediction models to modify intercept and slope of the model. One of the approaches was incorporating the average of random effects estimates of areas and auxiliary variables within each cluster by assuming that auxiliary variables are random (Anisa *et al.*, 2014). This assumption was found as disadvantages of the models because auxiliary variables basically were assumed to be fixed. Therefore, in this paper we assume that the auxiliary variables were fixed. This approach was conducted by adding the effects of cluster information as dummy variables and its multiplication with the auxiliary variables.

A simulation study is carried out to evaluate the performance of the proposed model compared with standard EBLUP and the existing modified EBLUP models. This paper also presents application of the proposed models using data from the Centre of Statistic in Indonesia (BPS) to predict average per capita expenditures at sub-district level in regency and municipality of Bogor. However, economic survey data with highly skewed distribution usually contained outlier observations (Chambers and Ren, 2004). It can affect the outcome of the estimation (Kurnia *et al*., 2013). Therefore, we also applied one-sided winsorization method into the proposed model to handle outliers in the data.

2. EMPIRICAL BEST LINEAR UNBIASED PREDICTOR (EBLUP)

Consider special case of linear mixed model for i -th area and j -th unit:

$$y_{ij} = x'_{ij}\beta + \nu_i + e_{ij} \tag{1}$$

with y_{ij} denotes sample observation, x_{ij} denotes auxiliary variable whose values is known for all units in population, ν_i denotes area random effects which distributed $\nu_i \sim iid \ \left(0, \sigma_\nu^2\right)$ and e_{ij} is error term that $e_{ij} \sim iid \ \left(0, \sigma_e^2\right)$, which depends on parameter $\sigma = \left(\sigma_e^2, \sigma_\nu^2\right)$ called variance component.

EBLUP is a two-stage estimator of parameter $t(\sigma)$ which depends on unknown parameter σ (Das *et al*, 2004). This approach replace the parameter σ with its estimator, $\hat{\sigma}$, so estimation will be carried out on parameter $t(\hat{\sigma})$. Note that the variance components σ have to be estimated before we can estimate the parameter of interest.

If it is defined that parameter of interest is the i -th small area mean, EBLUP estimator for the sampled area mean can be written as:

$$\bar{Y}_i = \frac{1}{N_i}\left(\sum_{j \in s_i} y_{ij} + \sum_{j^* \in r_i} \hat{y}_{ij^*} \right) \tag{2}$$

with s_i denotes sampled units and r_i denotes non-sampled units in the i -th area. Thus, $\hat{y}_{ij^*}$ is estimated value for non-sampled units which calculated with following formula:

$$\begin{aligned} \hat{y}_{ij^*} &= x'_{ij^*}\tilde{\beta} + \gamma_i \left(\bar{y}_{is} - \bar{x}_{is}\tilde{\beta} \right) \\ &= x'_{ij^*}\tilde{\beta} + \hat{\nu}_i \end{aligned}$$

where $\tilde{\beta} = \tilde{\beta}(\hat{\sigma}) = \left(X'\Sigma^{-1}X \right)^{-1} X'\Sigma^{-1}y$ is generalized least squares estimator of β, $\Sigma = \mathrm{cov}(y)$, and $\hat{\gamma}_{ij} = \hat{\sigma}_\nu^2 \Big/ \left(\hat{\sigma}_\nu^2 + n_i^{-1}\hat{\sigma}_e^2 \right)$. EBLUP estimator for non-sampled area mean can be written as follow:

$$\bar{Y}_{i^*} = \frac{1}{N_{i^*}}\left(\sum_{j^* \in r_{i^*}} \hat{y}_{i^* j^*} \right) \tag{3}$$

with $\hat{y}_{ij^*}$ is an estimated value which calculated by the following formula:

$$\hat{y}_{i^* j^*} = x'_{i^* j^*}\tilde{\beta}.$$

3. CLUSTER ANALYSIS

Cluster analysis is a multivariate technique to classify objects based on its similarities. In other words, the basic idea of cluster analysis is grouping similar observations into one cluster. There are two approaches in clustering method, hierarchical and non-hierarchical (Johnson and Wichern, 2007). Hierarchical clustering method is used when the number of clusters is unknown, while non-hierarchical clustering method is used when the number of clusters is known.

4. DEVELOPMENT OF THE MODELS

Kurnia (2009) showed that the application of small area estimation on BPS data to estimate poverty rate which reflected by per capita expenditures required logarithmic transformation. This call for the development of small area estimation under log-scale linear mixed model. One approach to calculate Root Mean Squares Error (RMSE) of the estimator is using the following equation:

$$RMSE\left(\hat{\bar{y}} \right) = \sqrt{MSE\left(\hat{\bar{y}} \right)} \approx \sqrt{A_1\left(\hat{\sigma}_\nu^2, \hat{\sigma}_e^2 \right) + A_2\left(\hat{\sigma}_\nu^2, \hat{\sigma}_e^2 \right)} \tag{4}$$

with

$$A_1\left(\hat{\sigma}_\nu^2, \hat{\sigma}_e^2 \right) = N_i^{-2}\left(\exp\left\{ \hat{\sigma}_\nu^2 + \hat{\sigma}_e^2 \right\}\left[\exp\left\{ \hat{\sigma}_\nu^2 + \hat{\sigma}_e^2 \right\} - 1 \right] \sum_{r_i} \exp\left\{ 2x'_{ij}\tilde{\beta} \right\} \right)$$

$$A_2\left(\hat{\sigma}_\nu^2, \hat{\sigma}_e^2 \right) = N_i^{-2}\left(\sum_{r_i} \exp\left\{ x'_{ij}\tilde{\beta} + \frac{1}{2}\left(\hat{\sigma}_\nu^2 + \hat{\sigma}_e^2 \right) \right\}^2 \left[x_{ij} Var\left(\tilde{\beta} \right) \right] \right).$$

4.1 The Basic Model

The basic model which is used in this study is a standard EBLUP model under log-scale linear mixed model, hereinafter referred to as Model-0. This study use unit level small area modeling, with i and j respectively denotes area and unit of sampled area, while $i*$ and $j*$ respectively denotes area and unit of non-sampled area. If we have one auxiliary variable in Model-0, the model can be written as follow:

a) Model for population:
$$\log\left(y_{ij}\right) = \beta_0 + \beta_1 x_{ij} + \nu_i + e_{ij}$$
b) Prediction model for sampled area:
$$\log\left(\hat{y}_{ij*}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_{ij*} + \hat{\nu}_i$$
c) Prediction model for non-sampled area:
$$\log\left(y_{i*j*}\right) = \tilde{\beta}_0 + \tilde{\beta}_1 x_{i*j*}$$
with y_{ij} denotes observed variable, x_{ij} denotes auxiliary variable, ν_i denotes area random effects, and e_{ij} is sampling error for sampled area. Prediction of non-sampled area $\left(\hat{y}_{i*j*}\right)$ is obtained by utilizing information from auxiliary variables of non-sampled area $\left(x_{i*j*}\right)$. This model assume that $\nu_i \sim iid \ \left(0, \sigma_\nu^2\right)$ and $e_{ij} \sim iid \ \left(0, \sigma_e^2\right)$.

4.2 Existing Modified EBLUP Models

The existing model developed by Anisa *et al.* (2014) are as follow:

a) Model for population:
$$\log\left(y_{ijk}\right) = \beta_0 + \beta_1 x_{ijk} + \gamma_{0i} + \gamma_{1i} x_{ijk} + e_{ijk}$$
$$\beta_0 + \gamma_{0i} + \left(\beta_1 + \gamma_{1i}\right) x_{ijk} + e_{ijk}$$
b) Prediction model for sampled area:
$$\log\left(\hat{y}_{ij*k}\right) = \tilde{\beta}_0 + \hat{\gamma}_{0i} + \left(\tilde{\beta}_1 + \hat{\gamma}_{1i}\right) x_{ij*k}$$
c) Prediction model for non-sampled area:
$$\log\left(\hat{y}_{i*j*k}\right) = \tilde{\beta}_0 + \overline{\hat{\gamma}}_{0(k)} + \left(\tilde{\beta}_1 + \overline{\hat{\gamma}}_{1(k)}\right) x_{i*j*k}$$
with k denotes cluster, i and j respectively denote sampled area and unit while $i*$ and $j*$ respectively denote non-sampled area and unit. This model hereinafter referred to as Model-1. This model assumes that X is random variable so the i-th area random effects and random effects of auxiliary variable X in the i-th area ($\hat{\gamma}_i$) can be obtained. Defined that $\overline{\hat{\gamma}}_{0(k)} = \frac{1}{m_k}\sum_{i=1}^{m_k}\hat{\gamma}_0$ is average of area random effects estimator and $\overline{\hat{\gamma}}_{1(k)} = \frac{1}{m_k}\sum_{i=1}^{m_k}\hat{\gamma}_1$ is random effects estimator of auxiliary variable X for k-th cluster which containing m_k sampled areas.

Another approach is by incorporating the effects of dummy variables from k-th cluster $d_1, d_2, \ldots, d_{k-1}$ into the previous model (Model-1).

a) Model for population:

$$\log\left(y_{ijk}\right) = \beta_0 + \gamma_{0i} + \left(\beta_1 + \gamma_{1i}\right) x_{ijk} + \sum_{l=1}^{k-1} \alpha_l d_l + e_{ijk}$$

b) Prediction model for sampled area:

$$\log\left(\hat{y}_{ij*k}\right) = \tilde{\beta}_0 + \hat{\gamma}_{0i} + \left(\tilde{\beta}_1 + \hat{\gamma}_{1i}\right) x_{ij*k} + \sum_{l=1}^{k-1} \tilde{\alpha}_l d_l$$

c) Prediction model for non-sampled-area:

$$\log\left(\hat{y}_{i*j*k}\right) = \tilde{\beta}_0 + \bar{\hat{\gamma}}_{0(k)} + \left(\tilde{\beta}_1 + \bar{\hat{\gamma}}_{1(k)}\right) x_{i*j*k} + \sum_{l=1}^{k-1} \tilde{\alpha}_l d_l \ .$$

The model above hereinafter referred to as Model-2.

4.3 The Proposed Model

The proposed model is a modification of standard EBLUP model by incorporating the effects of dummy variables from each k-th cluster $d_1, d_2, \ldots, d_{k-1}$ and the effects of multiplication between the dummy variables with auxiliary variable X into the basic model. This model also use the average of area random effects estimator within each k-th cluster in prediction of non-sampled area which can be written as $\bar{\hat{\nu}}_{(k)} = \frac{1}{m_k} \sum_{i=1}^{m_k} \hat{\nu}_i$, with m_k is the number of sampled area in k-th cluster. The proposed model can be written as follow:

a) model for population:

$$\log\left(y_{ijk}\right) = \beta_0 + \beta_1 x_{ijk} + \sum_{l=1}^{k-1} \alpha_l d_l + \sum_{l=1}^{k-1} \delta_l d_l x_{ijk} + \nu_i + e_{ijk}$$

b) prediction model for sampled area:

$$\log\left(\hat{y}_{ij*k}\right) = \tilde{\beta}_0 + \tilde{\beta}_1 x_{ij*k} + \sum_{l=1}^{k-1} \tilde{\alpha}_l d_l + \sum_{l=1}^{k-1} \tilde{\delta}_l d_l x_{ij*k} + \hat{\nu}_i$$

c) prediction model for non-sampled area:

$$\begin{aligned}
\log\left(\hat{y}_{i*j*k}\right) &= \tilde{\beta}_0 + \tilde{\beta}_1 x_{i*j*k} + \sum_{l=1}^{k-1} \tilde{\alpha}_l d_l + \sum_{l=1}^{k-1} \tilde{\delta}_l d_l x_{i*j*k} + \bar{\hat{\nu}}_{(k)} \\
&= \left(\tilde{\beta}_0 + \sum_{l=1}^{k-1} \tilde{\alpha}_l d_l + \bar{\hat{\nu}}_{(k)}\right) + \left(\tilde{\beta}_1 + \sum_{l=1}^{k-1} \tilde{\delta}_l d_l\right) x_{i*j*k} \\
&= \tilde{\beta}_0^* + \tilde{\beta}_1^* x_{i*j*k}
\end{aligned}$$

with k denotes cluster, i and j respectively denote sampled area and unit while $i*$ and $j*$ respectively denote non-sampled area and unit.

5. ONE-SIDED WINSORIZATION

According to Chambers *et al.* (2000), one of the basic ways in winsorization method to handle outliers is adjusting the outlying (positive) observed value Y_i. Thus, we will get the adjusted value Y_i^* which is defined as follow:

$$Y_i = \begin{cases} K & if \quad Y_i > K \\ Y_i & Otherwise \end{cases} \tag{5}$$

with $K > 0$ is a pre-specified constant, the cut-off value for the winsorization procedure. In this study, the specified cut-off value is defined as three times standard deviation of the observed variable from its mean, it can be written as: $K = \bar{y} + 3\hat{\sigma}_y$.

6. SIMULATION STUDY

A simulation study was carried out consisting of 496 units, 46 areas, and are grouped into 8 cluster. Without loss of generality, one auxiliary variable was used in this study. Fixed coefficients were predetermined as $\beta_0 = 13.0797$ and $\beta_1 = 0.0608$. Our study assumed that there were single intercept (α_k) and slope (δ_k) for each cluster. These intercept (α_k) and slope (δ_k) were added into the coefficients and yielded β_k^* coefficients. The response variable y_{ijk} was calculated using equation below:

$$\log\left(y_{ijk}\right) = x'_{ijk}\beta_k^* + \nu_i + e_{ij}$$

with different β_k^* coefficients for each cluster (

Table 1) assuming that auxiliary variable x_{ijk} was fixed. Area random effects (ν_i) and sampling error (e_{ij}) for i-th area and j-th unit respectively were generated from normal distribution $\nu_i \sim iid \quad N\left(0, \sigma_\nu^2 = 0.0745\right)$ and $e_{ijk} \sim iid \quad N\left(0, \sigma_e^2 = 0.1356\right)$.

Table 1
Predetermined Intercept and Slope of each Cluster

Cluster	α_k	δ_k	β_{0k}^*	β_{1k}^*
1	-0.199	0.151	12.881	0.212
2	-0.400	-0.042	12.680	0.019
3	0.311	0.008	13.390	0.068
4	0.215	-0.063	13.295	-0.003
5	0.402	-0.055	13.482	0.006
6	-0.167	0.148	12.913	0.208
7	0.062	-0.371	13.142	-0.310
8	0.208	-0.007	13.288	0.054

The data was divided into two parts, sampled area and non-sampled area. The number of sampled area was 44 areas while the two others were defined as non-sampled area. All models were used to predict small area mean calculated using equation (2) and (3). This process was repeated 1000 times and the resulting Relative Bias (RB) and Relative Root Mean Squares Error (RRMSE) of each area have been shown in Table 2.

Table 2
Median of RB and RRMSE for Non-Sampled Area (%)

	Model-0	Model-1	Model-2	Proposed Model
Median of RB	84.3268	74.7944	7.5564	-1.1512
Median of RRMSE	99.6123	98.1212	40.5751	38.9159

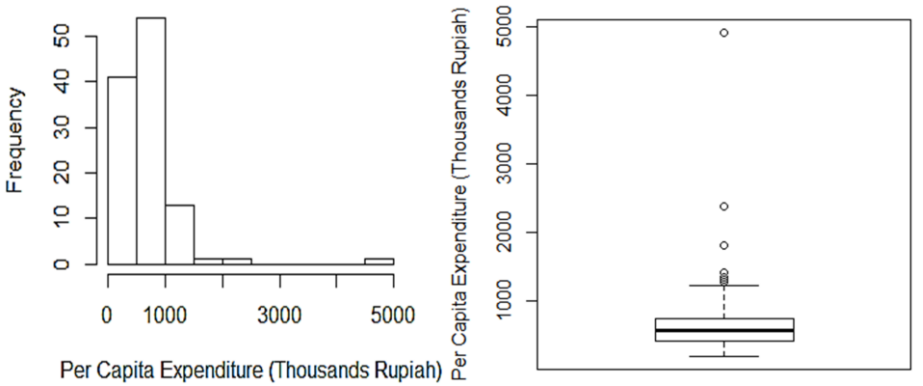
The proposed model demonstrated that the prediction of non-sampled area has provided smallest RB and RRMSE among all models (Table 2). This has indicated that proposed model showed the best performance compared with the others in predicting parameters of non-sampled area. It should be noted that this would work with proper clustering technique and variable selection which are most capable to describe variations of the observed variables of interest.

7. APPLICATION TO BPS DATA

The Centre of Statistic in Indonesia (BPS) provides data of all regions in Indonesia. SUSENAS is a national survey which is conducted annually by BPS to gather socio-economic data in Indonesia. PODES is an administrative record of village data in Indonesia which is published by BPS in every three years.

As an illustration of applying the proposed model into real data, in this study we used average per capita expenditures per month in small area (sub-district) obtained from SUSENAS 2010 as the observed variable and data from PODES 2011 as the auxiliary variable. Sub-districts in regency and municipality of Bogor were defined as area levels and the villages were defined as unit level.

The observed variable have a high positively skewed distribution and contain outliers or the so-called "extreme values" (Figure 1). Therefore, we applied one-sided winsorization method to handle these outliers using equation (5) with the specified constant value $K = 1,806,753.80$ Rupiah into the proposed model.



(a) (b)

Figure 1: Histogram and Boxplot of Average Per Capita Expenditure at Village Level in Regency and Municipality of Bogor

Normality test was conducted and the result showed that average expenditure per capita could not be assumed normally distributed (Table 3). Hence, the logarithmic transformation was applied to the data.

Table 3
Anderson-Darling Normality Test

	Y	$Log(Y)$
AD	9.367	0.506
P-value	< 0.005	0.197

In our previous study, data from PODES 2011 were used to cluster sub-districts which yielded one cluster having only one sub-district. This caused nuisance when we are using more than one auxiliary variable in the proposed model. Therefore, in this study we used area random effects estimator of each area from initial model to classify the sub-districts. This approach has the same basic idea with clustering that is classifying the similar observation into one group or cluster. This approach yielded three clusters (Table 4). First cluster and third cluster consecutively contains sub-districts with lower and higher area random effects estimator, while the second cluster contains sub-districts with area random effects estimator is around zero.

Table 4
Cluster of Sub-Districts in Regency and Municipality of Bogor

Cluster	The number of cluster members	Cluster members (Name of sub-districts)
1	7	Nanggung, Pamijahan, Cibungbulang, Kelapa Nunggal, Ranca Bungur, Parung Panjang,Tanah Sereal
2	31	Leuwiliang, Leuwisadeng, Ciampea, Tenjolaya, Dramaga, Tamansari, Cijeruk, Cigombong, Ciawi, Cisarua, Megamendung, Sukaraja, Sukamakmur, Cariu, Tanjungsari, Jonggol, Cileungsi, Citeureup, Cibinong, Bojong Gede, Parung, Ciseeng, Gunung Sindur, Rumpin, Cigudeg, Sukajaya, Jasinga, Tenjo, Bogor Selatan, Bogor Utara, Bogor Barat
3	8	Ciomas, Caringin, Babakan Madang, Gunung Putri, Tajur Halang, Kemang, Bogor Timur, Bogor Tengah

The variable of interest in this study is average per capita expenditures per month in non-sampled sub-districts in Bogor. Data from PODES 2011 which is utilized as auxiliary variable are:

1. Distance from the observed village to municipality/regency office
2. The number of minimarket in the observed village
3. The number of clinic in the observed village
4. The common job field of the observed village is farming (yes/no)

The proposed model is used to produce the prediction as it is considered as the best model in the simulation study. Standard EBLUP model is used as a comparison to the

proposed model. All models showed that Leuwisadeng sub-district has a higher average per capita expenditures compared to Tenjolaya sub-district (
Table 5).

Table 5
Estimates of Per Capita Expenditures Per Month Prediction of Non-Sampled Area (Thousands Rupiah)

Sub-districts	Standard EBLUP Model	Proposed Model	Winsorized Proposed Model
LEUWISADENG	511,764.60	511,830.00	525,760.60
TENJOLAYA	509,815.80	515,607.30	531,999.20

Table 6
Estimates of Root Mean Squares Error (RMSE) Over Non-Sampled Area (Thousands Rupiah)

Sub-districts	Standard EBLUP Model	Proposed Model	Winsorized Proposed Model
LEUWISADENG	151,129.89	144,101.94	129,943.16
TENJOLAYA	128,523.87	121,999.86	109,014.95

The models were evaluated based on RMSE value calculated using equation (4). Based on the smaller RMSE value, prediction of non-sampled area yielded by the proposed model was better than the standard EBLUP model (
Table 6). This result indicated that cluster information could improve predictive ability on non-sampled area by modifying the global synthetic model into local prediction model using the proposed model. Applying winsorization method on the proposed model was yielded even better estimation. This result shows us that handling of outliers can improve the estimation of non-sampled area using the proposed model.

8. CONCLUSION

Regarding both simulation study and the application to BPS data, cluster information can improve predictive ability on non-sampled area by modifying the global synthetic model into local prediction model using the proposed model. It should be noted that this would work with proper clustering technique and variable selection which are most capable to describe variations of the observed variables of interest. In addition, handling of outliers can improve the estimation of non-sampled area using the proposed model.

REFERENCES

1. Anisa, R., Kurnia, A. and Indahwati. (2014). Cluster Information of Non-Sampled Area in Small Area Estimation. *IOSR Journal of Mathematics*, 10, 15-19.
2. Chambers, R.L, Kokic, P., Smith, P. and Cruddas, M. (2000). Winsorization for Identifying and Treating Outliers in Business Surveys". *Proceedings of the Second International Conference on Establishment Surveys. Statistics Canada*, 717-726.

3. Chambers, R.L. and Ren, R. (2004). Outlier Robust Imputation of Survey Data. *Proceedings of the Section on Survey Research Method of the American Statistical Association*, 3336-3345.
4. Das, K., Jiang, J. and Rao, J.N.K. (2004). Mean Square Error of Empirical Predictor. *The Annals of Statistics*, 32, 818-840.
5. Johnson, R.A. and Wichern, D.W. (2007). *Applied Multivariate Statistical Analysis 6th Edition*. London, Prentice-Hall.
6. Kurnia, A. (2009). An empirical best prediction method for logarithmic transformation model in small area estimation with particular application to susenas data [doctoral dissertation]. Bogor (ID): Graduate Program, Bogor Agricultural University.
7. Kurnia, A., Kusumaningrum, D., Silvianti, P. and Handayani, D. (2013). Winsorization on Small Area Inference With Positively Skewed Distributions. *Proceedings of The International Conference on Applied Statistics of Padjadjaran University and Statistical Forum for Higher Education (FORSTAT)*, 210-216.
8. Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons.
9. Saei, A. and Chambers, R. 2005. Empirical Best Linear Unbiased Prediction for Out of Sample Area, *Working Paper M05/03*, Southampton Statistical Sciences Research Institute.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 463-476

GROUPING OF PUBLIC WELFARE IN PROVINSI ACEH PRICIPAL COMPONENT ANALYSIS

Winny Dian Safitri, Erfiani and **Bagus Sartono**
Department of Statistics, Bogor Agricultural University
Jl. Meranti Wing 22 level 4-5 Kampus IPB Darmaga Bogor, West Java-Indonesia
Email: winny.safitri@gmail.com

ABSTRACT

This research is conducted to see the changes of prosperity rate of Aceh people each year and clustering district/city in province of Aceh based of prosperity rate from 2007 until 2010. The variables that used are variables that comes from prosperity indicator of Aceh people, includes population indicator, health and nutrition indicators, education indicators, employment indicators, poverty indicators, housing and environmental indicators. Method used for analyzing the variables of people prosperity is principal component analysis. In general, the prosperity rate of districts/cities in province of Aceh is in medium prosperity rate. The result of this research found that the cluster changes each year based of prosperity rate, the districts/cities that change are Simeulue, Aceh Singkil, Aceh Timur, Aceh Tengah, Aceh Besar, Pidie, Aceh Utara, Aceh Barat Daya, Aceh Jaya, Bener Meriah, Pidie Jaya, Kota Sabang, Langsa and Subulussalam, while Banda Aceh is in the cluster of high prosperity rate.

KEYWORDS

Prosperity Rate, Cluster Analysis, Principal Component Analysis.

1. INTRODUCTION

Background

Regression analysis is a method that is widely used to examine the relationship between the response variable (Y) with the explanatory variables (X). There are various regression analysis, including linear regression analysis, nonlinear regression analysis, parametric and non-parametric. In regression analysis, in order to avoid a mismatch model of the least squares method is used based on the following assumptions:

1. ε_i spread independent follow a normal distribution ($\varepsilon_i \sim NID\ (0, \sigma^2)$);
2. Variety of ε_i constant or $V\ (\varepsilon_i) = \sigma^2$ for i = 1,2, ..., n (homoscedastisitas);
3. There is no residual autocorrelation or $Cov\ (\varepsilon_i,\varepsilon_j) = 0$ where $i \neq j$;
4. There is no multicollinearity among the explanatory variables (X) or $Cov\ (X_i, X_j) = 0$ where $i \neq j$.

Multicollinearity occurs when there is a correlation between independent variables can affect a variety of least squares estimators and estimation models are made (Wetherhill, 1986). These assumptions lead to the alleged irregularities for the regression coefficient has variance estimates are suspect more (overestimate), although still takbias and explanatory variables that should have significant influence on the response variable will be declared otherwise (not statistically significant).

Pricipal Component Analysis was first introduced by Pearson in 1901, further developed by Loeve in 1963. Pricipal Component Analysis is a pretty good method to obtain estimates of the regression coefficients that have a multicollinearity problem by reducing the data dimension becomes large dimensions small, which is expected to smaller dimensions, the correlation between the variables will be resolved. The perl note that by reducing the variation in the data is maintained at least 80% (Johson and Winchern, 1992: 359).

In the economic and social fields, the explanatory variables are more likely correlated (multicollinearity), so that the case raised in this research associated with indicators of welfare of the people with the title "Grouping of Public Welfare in Provinsi aceh with the pricipal Component Analysis".

Research Purposes

The purpose of this study is:

1) To see the change in the level of welfare of the people of Aceh annually.
2) To classify districts / cities in Provinsi Aceh based on the level of people's welfare changes from year 2007-2010.

2. LITERATURE REVIEW

People's Welfare

Welfare of the people according to the 1974 Act 6 years old is a living system and social, material and spiritual, filled with a sense of safety, decency and peace outwardly and inwardly, that allows every citizen to hold a business meeting physical needs, spiritual and social as well as for themselves, their families and communities to uphold the rights or obligations of the people according to the Pancasila (Adi, 2005). Indonesian Social Welfare (2010) defines that the welfare of the people shows how much money is paid by the government to the people who need financial assistance, but cannot work or the state earned income to meet basic needs is not sufficient. The amount paid is usually far below the poverty line and also have special conditions, such as proof of looking for a job or other conditions, such as the inability or obligation to keep the child, which prevented him from working. In some cases even grant recipients are required to work and is known as workfare. Meanwhile, according to Edi Suharto (2010) the welfare of the people is a state of fulfillment of all forms of subsistence, particularly fundamental as food, clothing, housing, education and health care.

Badan Pusat Statistik (BPS, 2010) publish the welfare of the people through several aspects, including:

1) Indikators of population
2) Indicators of health and nutition
3) Indicatorrs of Education
4) Indicators of labor
5) Indicators of poverty
6) Indicators of housing and environment

Principal Component Analysis

Principal component analysis (PCA) is a good method to obtain estimates of the regression coefficients that have a multicollinearity problem. Independent variables in the regression of the main components in the form of a linear combination of the results of

the origin of the Z variables, called principal components (PC). The coefficient estimator obtained from this method of depreciation dimensional principal component, wherein the selected subset of the main components must maintain maximum diversity. Where Z is a standard normal result of the variable X. The result is the standard normal is by subtracting each original independent variables Xi with average and divided by the standard deviation, denoted:

$$Z_i = \frac{X_i - \bar{X}}{\sigma}$$

Elimination of the major components of the selection procedure begins characteristicroots of an equation: $| AX - \lambda I | = 0$.

If the characteristic roots λ_j values are sorted from largest to smallest value, then the influence of the main components W_j corresponds to the influence of λ_j. This means that these components are explained proportion of the diversity of the dependent variable Y that is smaller. The main components W_j mutually orthogonal each other and formed through a relationship:

$$W_j = v_{1j} Z_1 + v_{2j} Z_2 + v_{3j} Z_3 + \ldots + v_{pj} Z_p$$

P is the number of variables used.

Vector vj obtained from each root traits that λj satisfy a system of homogeneous equations:

$$| AX - \lambda_j I | v_j = 0$$

where Is $v_j = (v_{1j}, v_{2j}, v_{3j}, \ldots, v_{pj})$.

There are three commonly used methods for determining the number of major components, namely:

1) If the resulting number is the main component q where q ≤ p, then that has been transformed (the main component score data) have as many variables q. Suppose proportion to root traits to - i is $\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}$ the determination of the amount of the main component (q) based on the cumulative proportion of the roots of the character. If the cumulative proportion of root traits q first reaches 80% or more, then the number of principal components is q (Johnson and Wichern, 1992).
2) Selection of eigen value
3) The third method is the use of a graph called scree plot. Scree plot is a plot of the characteristic roots λ_k with k. By using this plot, the number of principal components chosen is k, if the point of the plot k steep but not steep left to right.

3. DATA AND METHOD

Data

The data used in this research is secondary data, namely the people's welfare indicator data sourced from the Badan Pusat Statistik (BPS) Provinsi Aceh. The period of data used are annual, from 2007 to 2010.

Definition of Variables

In this study variables observed indicators of welfare of the people there are 10 variables as in the table below:

Table 3.1
Variable Used

Indicators of welfare of the people	**Variable**
Population growth rate (year)	X_1
Population density (km^2)	X_2
Young people 0 – 14 years (%)	X_3
Elderly population ≥ 65 years (%)	X_4
Residents who have health complaints (%)	X_5
Life expectancy (%)	X_6
Population 10 years and over with a junior high education or more (%)	X_7
Average expenditure per capita a month (USD)	X_8
Households with piped drinking water source (%)	X_9
Households with a floor area <10 m2 (%)	X_{10}

Research Procedures

Steps being taken in this study as follows:

1) Looking for an average of all the variables from the data of 2007 to 2010.
2) Perform principal component analysis of the average data.
3) Determine the number k principal components (PC) of average data.
4) Looking principal component scores (W) of the pricipal component (PC) are elected.
5) Perform principal component analysis of the data per year, then use the feature vector or a_{ij} coefficients obtained through principal component analysis to obtain k principal component scores (W) 2007, 2008, 2009 and 2010.
6) Calculate the change scores major components (W) from the year 2007 to year 2008, 2008 and 2009 to year 2009 to year 2010.
7) Classify districts / cities by category size changes.
8) Category group welfare of the people of each district / city to-h can be denoted as follows
   - high, if $y_{h1} > \bar{y}_1 + s_{y1}$
   - being, if $\bar{y}_1 - s_{y1} \leq y_{h1} \leq \bar{y}_1 + s_{y1}$
   - low, if $y_{h1} < \bar{y}_1 - s_{y1}$ (Vincent, 1992).
9) Interpret the results obtained from the grouping level of prosperity of each county/city in the province of Aceh.

4. RESULTS AND DISCUSSION

Principal Component Analysis (PCA)

Indicators of welfare of the people of Aceh from year to year is a significant change. Along with the changes made to the data analysis of the average variable indicators of welfare of the people of Aceh, and classify districts / cities in Provinsi Aceh based on common characteristics possessed by the average data.

Results of correlation analysis the average data show that there are some high correlation between independent variables that indicate the presence of multicollinearity. Multicollinearity can be overcome by principal component analysis (PCA), by first standardizing the variables X into variable Z. There are 4 PC elected to the proportion of

0.422 PC_1 diversity means that PC_1 can explain the origin of the data variability 4.22% of the total variability. In the same way for PC_2, PC_3 and PC_4 of 0.92%. PC is the cumulative 4 a proportion of total variability of approximately 0.855 which means all four PC able to explain the origin of the diversity of data 85.5% of the variability in total, the rest is explained by other PC. It is claimed that the process of grouping districts / cities in Provinsi Aceh by the average data of the people's welfare indicators used 4 pieces PC enough that PC_1, PC_2, PC_3 and PC_4, because four PC has been able to explain the diversity of data on the average indicator of people's welfare by 85.5%, a high level of diversity that can be explained by the 4 PC. From PC_1, PC_2, PC_3 and principal component scores obtained PC_4 (W) is W_1, W_2, W_3 and W_4 are listed in Table 4.1.

Table 4.1
Pricipal component scores (W) to the data 23 districts / cities

Districts/Cities	W_1	W_2	W_3	W_4
Simeulue	-2.23667	1.59975	-0.79883	0.42683
Aceh Singkil	-2.06023	1.89281	0.77868	-0.79807
Aceh Selatan	0.00978	-0.87433	0.71929	1.28712
Aceh Tenggara	-1.34807	-0.13593	-0.79292	0.19223
Aceh Timur	-1.98402	0.65192	1.06416	-1.32667
Aceh Tengah	-0.16409	0.18396	1.18227	-0.22918
Aceh Barat	0.49656	-0.12861	0.11454	1.22634
Aceh Besar	1.55743	-1.27249	0.07905	0.13795
Pidie	-0.54316	-2.80091	0.56869	-1.42817
Bireuen	0.44168	-1.11409	-0.66007	-0.00911
Aceh Utara	-1.15383	-0.08144	-0.05467	-0.51284
Aceh Barat Daya	-0.55562	-0.98826	-0.39830	1.63416
Gayo Lues	-2.04308	2.41856	-0.03730	1.06345
Aceh Tamiang	0.01966	-0.38028	-0.80164	-0.41300
Nagan Raya	-0.44683	-0.85085	-0.49680	1.60281
Aceh Jaya	-0.49174	-0.94997	2.80074	1.09083
Bener Meriah	-0.57984	-0.85410	-0.92949	-0.06005
Pidie Jaya	0.16425	-3.13972	0.06442	-1.57131
Banda Aceh	6.99177	2.48662	1.32610	-0.27367
Sabang	2.84193	0.28209	-1.28644	-0.20658
Langsa	1.71999	0.21849	-1.09404	-0.59674
Lhokseumawe	1.66279	0.85271	-1.28820	-0.01270
Subulussalam	-2.29865	2.98406	-0.05925	-1.22363

Furthermore made a major component score distribution plot (W) district / city, so it looks layout districts / cities in Aceh province. The plot between the main component score (W), are presented in the following figure:

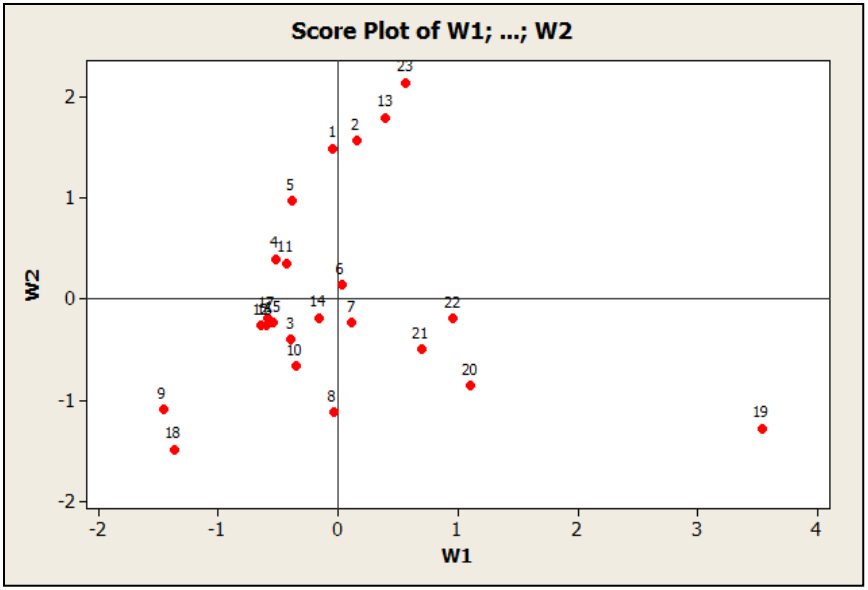


Figure 4.1: Plot W_1 and W_2 to the Data 23 Districts/Cities

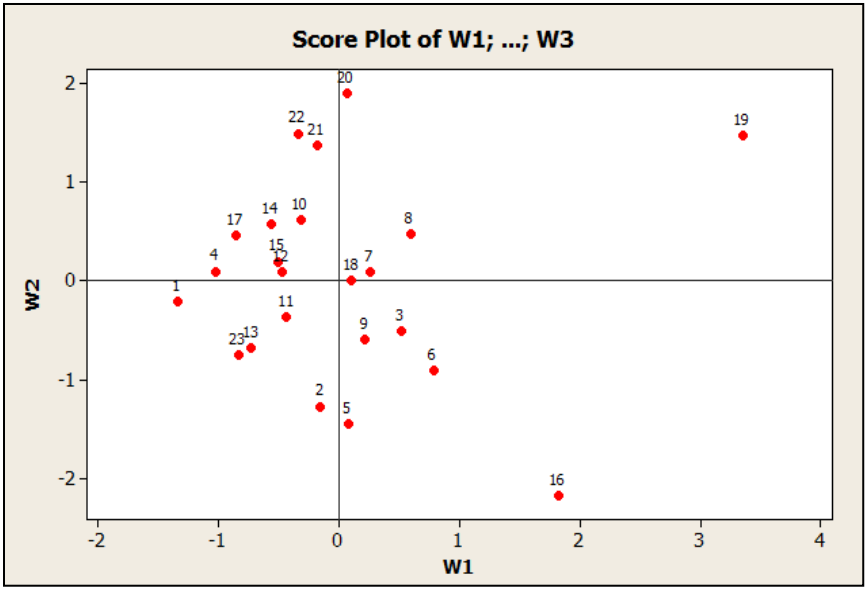


Figure 4.2: Plot W1 and W3 to the data 23 districts / cities

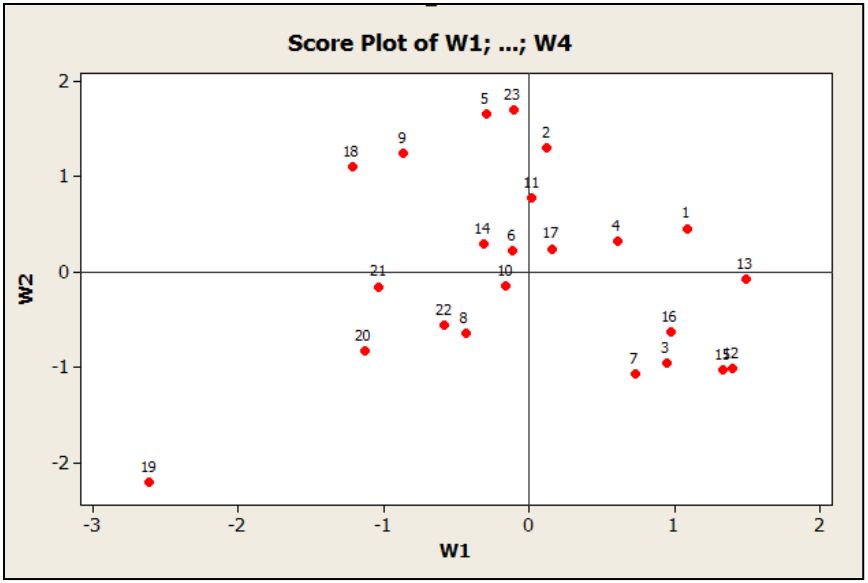


Figure 4.3: Plot W1 and W3 to the Data 23 Districts / Cities

The main component of the score plot (W) data is an average of 23 districts / municipalities is that Banda Aceh to form groups of outliers, so it is necessary to analyze data on the average 22 districts / cities without Banda Aceh. PCA to the average data 22 districts / cities together with PCA to the average data 23 districts / cities, From the correlation matrix R can be lowered 10 pricipal components (PC), there are 4 PC who meet the selection criteria are viewed from the characteristic root more than 0.7 and the first characteristic roots cumulative proportion reaches 80% or more 4 PC was shown in

Table 4.2
The Pricipal Components (PC) for the Data 22 Districts / Cities

Variables	PC_1	PC_2	PC_3	PC_4
Z_1	0.219	0.241	-0.440	-0.496
Z_2	0.282	-0.291	0.462	0.070
Z_3	-0.455	-0.216	0.216	-0.196
Z_4	0.233	0.536	0.025	-0.084
Z_5	0.297	0.072	0.364	-0.620
Z_6	-0.361	0.125	0.436	-0.368
Z_7	0.198	0.411	0.189	0.109
Z_8	0.210	-0.376	-0.413	-0.378
Z_9	0.358	-0.401	0.045	-0.151
Z_{10}	-0.426	-0.179	0.122	0.082
Eigenvalue	3.5625	2.2855	1.2686	0.9905
Proportion of Variance (%)	0.356	0.229	0.127	0.099
Cumulative Variance (%)	0.356	0.585	0.712	0.811

From table 4.2 proportion PC_1 diversity by 0.356 means that PC_1 can explain the origin of the diversity of data 35.6% of the total variability. In the same way for PC_2, PC_3 and PC_4 of 0.99%. PC is the cumulative 4 a proportion of total variability of approximately 0811 which means all four KU able to explain the origin of the diversity of data 81.1% of the variability in total, the rest is explained by other PC. It is claimed that the process of grouping districts / cities in Provinsi aceh by the average data of the people's welfare indicators used 4 pieces PC enough that PC_1, PC_2, PC_3 and PC_4, due to the PC-4 has been able to explain the diversity of data on the average indicator of people's welfare amounted to 81.1%, a high level of diversity that can be explained by the 4PC.

Table 4.3
Pricipal Component Score (W) to the Data 22 Districts / Cities

Districts/Cities	W_1	W_2	W_3	W_4
Simeulue	-2.45365	-0.79243	-0.01018	0.76066
Aceh Singkil	-2.83787	-0.89206	0.08381	-1.13028
Aceh Selatan	0.35469	0.77483	-1.45149	0.42872
Aceh Tenggara	-0.86701	0.68765	0.53884	0.85412
Aceh Timur	-2.03618	0.41225	0.54866	-1.48407
Aceh Tengah	-0.02081	-0.36636	-1.06429	-1.49581
Aceh Barat	0.74433	-0.43035	-1.64789	0.36261
Aceh Besar	2.34395	-0.07430	-1.09395	-0.60553
Pidie	0.75902	2.85901	1.31489	-0.92844
Bireuen	1.10825	0.72580	0.51705	0.68937
Aceh Utara	-0.87756	0.60309	0.94772	0.09588
Aceh Barat Daya	0.01330	1.29514	-0.56828	1.82760
Gayo Lues	-3.02969	-1.13152	-0.53824	1.13280
Aceh Tamiang	0.40568	0.05871	0.64746	0.20698
Nagan Raya	0.11049	0.97000	-0.91209	1.59439
Aceh Jaya	0.02854	1.28145	-2.43087	-1.22309
Bener Meriah	0.15797	0.87848	0.85374	0.71187
Pidie Jaya	1.82753	2.55745	1.54823	-0.92740
Sabang	3.33836	-2.83551	-0.79091	-0.71336
Langsa	2.16209	-1.91810	1.06403	-0.04409
Lhokseumawe	2.36693	-2.82735	1.64193	0.70474
Subulussalam	-3.59837	-1.83590	0.80182	-0.81766

Table 4.4
Correlation Original Variables (X) with 4 Main Component Score (W) First to the Data 22 Districts / Cities

Variables	W_1	W_2	W_3	W_4
X_1	-0.414	0.364	-0.495	-0.494
X_2	0.532	-0.440	0.521	0.069
X_3	-0.859	-0.327	0.243	-0.195
X_4	0.440	-0.810	0.028	-0.084
X_5	0.560	0.109	0.410	-0.617
X_6	-0.511	0.189	0.792	-0.366
X_7	0.374	0.622	0.213	0.109
X_8	0.396	-0.568	-0.465	-0.376
X_9	0.675	-0.606	0.051	-0.151
X_{10}	-0.803	-0.270	0.137	0.081

The correlation between the variables of origin (X) with a score of major components W are shown in Table 4.4 that the variables X_9 has a positive correlation is large enough to W_1, while the variable X_3 and X_{10} has a negative correlation is large enough to W_1. This means that the higher the value W_1 if the value of the percentage of households with piped drinking water sources (X_9) and a high percentage of the population aged 0-14 years (X_3), the percentage of households with a floor area <10 m^2 (X_{10}) is low. Conversely, a low value W_1 if the value percentage of the population aged 0-14 years (X3), the percentage of households with a floor area <10 m^2 (X_{10}) high and a source of drinking water tap (X_9) is low.

Variables X_7 has a positive correlation large enough to W_2, while negatively correlated variables X_4 large enough to W_2, the meaning of the correlation value that the value of W_2 will be high if the value of the percentage of the population 10 years and older with a high school junior or more (X_7) high and percentage of the population $\geq$ 65 years (X_4) is low. X_6 variables positively correlated well to the W3, meaning that if a district / city in the province has a high W_3, mean percentage survival (X_6) in the district / city is high. X_5 variables are positively correlated with W4 large enough, the result of this value explains that if the value W_4 of the district / city high, then the percentage of people who have health complaints (X_5) districts / cities in Provinsi aceh low. Correlation above shows that the variables of each indicator welfare of the people in Provinsi aceh berkosistensi, so it can be done grouping districts / cities in Provinsi aceh with a plot between W_1 and W_2 as follows.

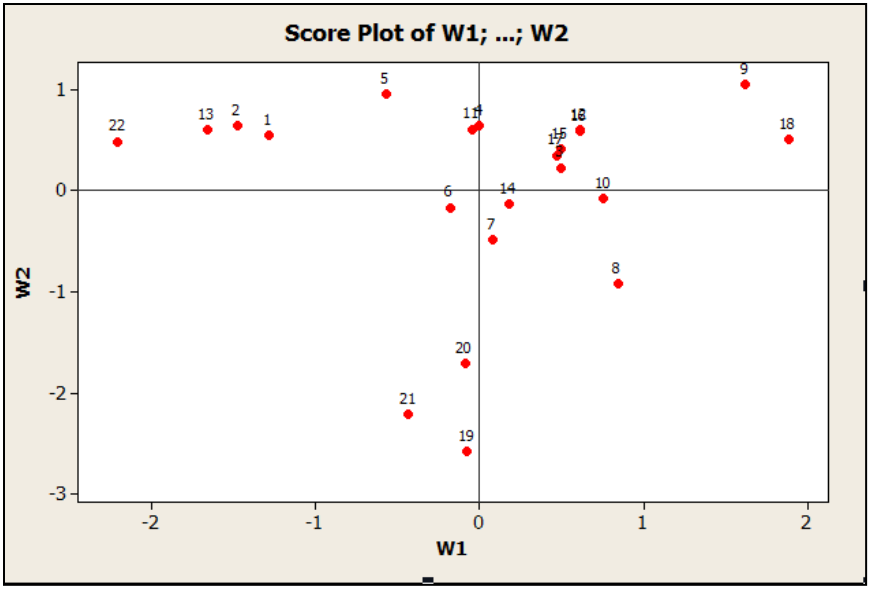


Figure 4.4: Plot of W1 and W2 to the Data 22 Districts / Cities

Results Figure 4.4 shows that the absence of outliers districts / cities in Provinsi aceh shown on the location of the plot distribution of district / city. Grouping districts / cities Based on Level of Public Welfare Stages for the same year as the stage of data average data. From the data analysis stages mean assuming that PC_1 already represent a diversity of other PC because it has the highest value of root traits that none other than the value of diversity in Table 4.5.

Table 4.5
The Pricipal Components (PC_1) Data from 2007 to 2010

Variables	PC_{2007}	PC_{2008}	PC_{2009}	PC_{2010}
Z_1	0.278	-0.260	-0.207	0.316
Z_2	0.378	0.168	0.371	0.054
Z_3	-0.417	-0.471	-0.390	0.419
Z_4	0.135	0.044	0.083	0.048
Z_5	0.165	0.202	0.254	0.462
Z_6	-0.420	-0.494	-0.386	-0.462
Z_7	0.045	-0.059	-0.160	0.212
Z_8	0.314	0.351	0.423	-0.448
Z_9	0.353	0.453	0.385	0.047
Z_{10}	-0.398	-0.253	0.306	-0.210
Eigenvalue	4.1378	3.3608	4.4363	2.6786

From Table 4.5 the process of grouping the welfare of the people of each county / city in the province of Aceh are used the fruit KU from 2007 to 2010 are PC_1. Dengan Thus the process of grouping districts / cities in Provinsi aceh based on the rate of change in

people's welfare indicators can be measured by large-small difference in scores first principal component (W_1) high means the district / city has a higher level of welfare than other regions that have a score difference of the first pricipal component (W_1) is low. Difference scores first principal component (W_1) of each county / city in the Provinsi Aceh is shown in Table 4.6 below:

Table 4.6
Difference Scores First Principal Component (W_1)

No.	Districts/Cities	W_A	W_B	W_C
1	Simeulue	-1.103	0.616	-2.667
2	Aceh Singkil	-0.035	-0.091	-4.572
3	Aceh Selatan	-0.480	0.142	-1.550
4	Aceh Tenggara	-0.608	0.650	3.258
5	Aceh Timur	0.002	1.026	0.036
6	Aceh Tengah	1.768	2.157	0.528
7	Aceh Barat	-0.247	-0.811	2.607
8	Aceh Besar	-0.823	0.083	0.755
9	Pidie	1.013	0.513	-0.770
10	Bireuen	0.092	-0.084	-1.754
11	Aceh Utara	-0.733	0.764	-3.352
12	Aceh Barat Daya	0.111	0.034	4.708
13	Gayo Lues	-0.791	-0.097	0.763
14	Aceh Tamiang	-0.399	0.234	0.035
15	Nagan Raya	0.770	-0.369	1.390
16	Aceh Jaya	1.406	0.218	-1.369
17	Bener Meriah	0.503	0.882	-1.550
18	Pidie Jaya	-0.110	0.526	-2.295
19	Banda Aceh	1.431	1.806	5.102
20	Sabang	-1.148	-0.097	7.668
21	Langsa	-0.900	-0.093	3.701
22	Lhokseumawe	-0.041	0.694	1.202
23	Subulussalam	0.911	-0.775	-1.355
	Average	0.026	0.000	0.457
	Standard Deviation	0.850	0.790	2.976

Specification:
W_A = Difference scores first principal component (W1) of 2007 to 2008
W_B = Difference scores first principal component (W1) of 2008 to 2009
W_C = Difference scores first principal component (W1) of 2009 to 2010

By using the first principal component scores (W_1) in Table 4.6, the grouping of districts / cities in Provinsi Aceh based on the rate of change in people's welfare can be done through a group of people's welfare level category. Allocation district / city to-h in the province of Aceh in a group rate change people's welfare are as follows:

A. 2007 to 2008

- high, if $y_{h1} > 0.026 + 0.850$ or $y_{h1} > 0.876$
- being, if $0.026 - 0.850 \leq y_{h1} \leq 0.026 + 0.850$ atau $-0.824 \leq y_{h1} \leq 0.876$
- low, if $y_{h1} < 0.026 - 0.850$ atau $y_{h1} < -0.824$

Note: y_{h1} is the first principal component scores (W_1) of the district to-h (h = 1, 2, 3,…,23).

Results grouping districts / cities in Provinsi aceh by the rate of change of people's welfare in 2007 to 2008, namely:

1) Groups of districts / cities in Provinsi Aceh with a high level of well-being that Central Aceh, Pidie, Aceh Jaya, Banda Aceh and Subulussalam.
2) Groups of districts / cities in Provinsi Aceh with a moderate level of prosperity that Singkil, South Aceh, Southeast Aceh, East Aceh, West Aceh, Aceh Besar, Bireuen, North Aceh, Southwest Aceh, Gayo Lues, Aceh Tamiang, Nagan Raya, Bener Meriah, Pidie Jaya and Lhokseumawe.
3) Groups of districts / cities in Provinsi Aceh with a low level of welfare that Simeulue, Sabang and Langsa.

B. 2008 to 2009

- high, if $y_{h1} > 0.000 + 0.790$ or $y_{h1} > 0.790$
- being, if $0.000 - 0.790 \leq y_{h1} \leq 0.000 + 0.790$ or $-0.790 \leq y_{h1} \leq 0.790$
- low, if $y_{h1} < 0.000 - 0.790$ or $y_{h1} < -0.790$

Note: y_{h1} is the first principal component scores (W1) of the district to-h (h = 1, 2, 3,…,23).

Groups of districts / cities in Provinsi aceh with a Results grouping districts / cities in Provinsi Aceh based on the level of people's welfare changes in 2008 to the year 2009, namely:

1. Groups of districts / cities in Provinsi Aceh with a high level of well-being that East Aceh, Central Aceh, Pidie Jaya, Bener Meriah, Banda Aceh.
2. Groups of districts / cities moderate level of well-being that is Simeulue, Singkil, South Aceh, East Aceh, West Aceh, Pidie, Bireuen, North Aceh, West Aceh, Gayo Lues, Aceh Tamiang, Nagan Raya, Aceh Jaya, city of Sabang, Langsa, Lhokseumawe and Subulussalam.
3. Groups of districts / cities in Provinsi Aceh with a low level of welfare that Aceh Besar.

C. 2009 to 2010

- high, if $y_{h1} > 0.457 + 2.976$ or $y_{h1} > 3.433$
- being, if $0.457 - 2.976 \leq y_{h1} \leq 0.457 + 3.433$ or $-2.519 \leq y_{h1} \leq 3.433$
- low, if $y_{h1} < 0.457 - 2.976$ or $y_{h1} < -2.519$

Note : y_{h1} is the first principal component scores (W1) of the district to-h (h = 1, 2, 3,…,23).

Results grouping districts / cities in Provinsi Aceh based on the level of people's welfare changes in 2009 to the year 2010 are:

1. Groups of districts / cities in Provinsi Aceh with a high level of welfare that Southwest Aceh, Banda Aceh, Sabang and Langsa.
2. Groups of districts / cities in Provinsi aceh with a moderate level of well-being that South Aceh, East Aceh, East Aceh, Central Aceh, Aceh Barat, Aceh Besar, Pidie, Bireuen, Gayo Lues, Aceh Tamiang, Nagan Raya, Aceh Jaya, Bener Meriah, Pidie Jaya, Lhokseumawe and Subulussalam.
3. Groups of districts / cities in Provinsi aceh with a low level of welfare that Simeulue, Singkil and North Aceh.

5. CONCLUSION

Banda Aceh is always in the position of the high-level group welfare, due to Banda Aceh is the capital of Provinsi aceh at the center of all activities of the Acehnese people's welfare indicators that have higher or more prosperous than other districts / cities that exist in the province of Aceh. South Aceh, East Aceh, Aceh Barat, Bireun, Gayo Lues, Aceh Tamiang, Nagan Raya and Lhokseumawe always include groups districts / cities with a moderate level of well-being.

REFERENCES

1. Adi, I.R. (2005). *Ilmu Kesejahteraan Sosial dan Pekerjaan Sosial*. FISIP UI Press. Jakarta.
2. BPS Provinsi Aceh (2007). *Indikator Kesejahteraan Rakyat 2007*. Aceh: BPS Provinsi Aceh.
3. BPS Provinsi Aceh (2008). *Indikator Kesejahteraan Rakyat 2008*. Aceh: BPS Provinsi Aceh.
4. BPS Provinsi Aceh (2009). *Indikator Kesejahteraan Rakyat 2009*. Aceh: BPS Provinsi Aceh.
5. BPS Provinsi Aceh (2010). *Indikator Kesejahteraan Rakyat 2010*. Aceh: BPS Provinsi Aceh.
6. Dillon, W. dan Goldstein, M. (1984). *Multivariate Analysis*, Methods and Applications. John Wiley and Sons, Singapore.
7. Johnson, R.A dan Wichern, D.W. (1992). *Applied Multivariate Statistical Analisys*, Second edition, Prentice – Hall International, Inc., New Jersey.
8. Jolliffe, I.T. (1986). *Principal Component Analysis.* New York: Springer-Verlag.
9. Myers, R.H. (1990). Classical and Modern Regression with Application. Second Edition. PWS-Kent Publishing Company, Boston
10. Vincent (1992). Teknik *Analisis Dalam Penelitian Percobaan*. Tarsito, Bandung.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 477-486

COMPARISON OF METHOD CLASSIFICATION ARTIFICIAL NEURAL NETWORK BACK PROPAGATION, LOGISTIC REGRESSION, AND MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS) (CASE STUDY DATA OF UNSECURED LOAN)

Siti Hadijah Hasanah[1], Kusman Sadik[2] and **Farit Mochamad Afendi[2]**
[1] Graduate School, Bogor Agricultural University, Bogor, Indonesia
Email: dije_math04@yahoo.co.id
[2] Department Statistics, Bogor Agricultural University, Bogor, Indonesia.

ABSTRACT

Unsecured loan is one of the credit product which given bank to customers in the form of a loan facility without a guarantee, because there is no collateral for the loan, so the bank must carefully examine the credit rating of the customer in order to avoid the risk of losses in the future. Customer assessment of unsecured loan conducted based classification techniques. The problem of classification technique is the amount of data and the number of explanatory variables are quite large, so the logistic regression and *multivariate adaptive* regression *splines* (MARS) was not able to overcome it. Method of artificial neural network (ANN) Backpropagation is one method that is able to overcome the problems of classification unsecured loan well. Based on the results of the classification accuracy of the confusion matrix and the receiver operating characteristic curve (ROC) artificial neural network of Backpropagation has a classification accuracy of 78.40% and AUC value 84.10% on the training data, while in the testing data has an accuracy of 72.94% and AUC value 74.80%.

KEYWORDS

Unsecured Loan, Logistic Regression, Multivariate Adaptive Regression Splines, Artificial Neural Network of Backpropagation.

1. INTRODUCTION

Quite rapid banking growth in Indonesia gave rise to competition between banks to one another provide the best service to its customers. One of the best banking service that is preferred by customers is the provision of credit. Unsecured loan is one of the credit product which given bank to customers in the form of a loan facility without a guarantee, because there is no collateral for the loan, the bank must be carefully examine the credit rating of the customer in order to avoid the risk of losses in the future. Submission of application unsecured loan by the customer to the bank will be assessed based on classification techniques. Classification technique is one of the statistical methods in grouping the data compiled systematically.

Classification techniques that have been developed to date using statistical approaches such as discriminant analysis, logistic regression (Agesti, 1990; Hosmer & Lemeshow, 2000), multivariate adaptive regression splines (MARS) (Cox & Snell, 1989; Friedman, 1991; Friedman & Silverman, 1989; Sutikno, 2008), classification and regression tree

(CART), k-nearest neighbors (KNN), genetic algorithm, support vector machines (SVM), and artificial neural network (ANN) (Kusumadewi, 2004; Mirtalaei *et al.,* 2012; Puspitaningrum, 2006; Shi *et al.,* 2012). Logistic regression is a parametric method that is used to estimate the relationship between the response variable with one or multi categories of explanatory variables are continue or category. Logistic regression was not required a the assumptions must be met when analyzing data using linear regression. Another method that can be used to a classification technique that is flexible and innovative is MARS and artificial neural network (ANN). MARS method is a nonparametric method that combines spline method with recursive partitioning regression (RPR) which has been modified. MARS is focused to overcome the problem of high dimension and not continue at the knots. This method is able analyzes 3-20 predictor and 50-1000 amount of data, can be explain by both the patterns and their dynamic nonlinear interaction owned. The weakness of this method is difficult to interpretation because has considerable explanatory variables in the models.

Problems encountered in the classification technique is the amount of data and the number of explanatory variables are quite large, logistic regression and MARS are not able to solve that problem, it is necessary an information processing system which one of them is a method of ANN. ANN method is one of the nonparametric method that mimics the workings of the human brain in solving a problem is to make the process of learning by changing the weights on the links connecting (Mirtalaei *et al.,* 2012). Excess of ANN is able to model the nonlinear relationship, has the ability to learn (adaptive) so that it can present a flexible knowledge of the existence of errors (Puspitaningrum, 2006). ANN Backpropagation is an unsupervised learning algorithm and is usually used by the perceptron with many layers to change the weights are connected with existing neurons in the hidden layer (Kusumadewi, 2004). Backpropagation algorithm using residual output to change the weights in the backward direction (backward), to get this remnant propagation step forward (forward propagation) must be done first.

This study will be designed classification techniques on customer application submission unsecured loan by comparing the ANN Backpropagation method, logistic regression, and MARS based on classification accuracy with the confusion matrix and the receiver operating characteristic curve (ROC) (Gorunescu, 2011).The aim is to conduct comparative research on the method of ANN Backpropagation, logistic regression, and multivariate adaptive regression splines (MARS) in data classification unsecured loan.

2. METHODOLOGY

The data of this study is the client application submission unsecured loan in Bank Internasional Indonesia (BII) in May 2014 consists of 1700 observations, 9 input variables and 1 output variable. Software used in this study is the Matlab 7.0.1, R 2.14.0, Minitab 14, and MedCalc.

1) Input variable:
   X_1 = Age
   X_2 = Gender
   X_3 = Marital status
   X_4 = Number of dependents
   X_5 = Education
   X_6 = Type of work

X_7 = Costumers history
X_8 = Number of installments
X_9 = Salary

2) Output variable:
Y = the decision (1: approve; 0: reject)

Steps of data analysis to ANN Backpropagation is as follows:

i) Separation of data
Credit data is separated into two parts, namely the training data and testing data. Distribution of this data is done randomly. Training data by 70% and 30% of testing data.

ii) Determine the number of layers
This stage to determine the number of input layer, hidden layer and output layer to the training data.

iii) Determination of weights
Determine the weight of the hidden layer to the input layer and the hidden layer to the output layer.

iv) Specify the activation function, learning rate

v) Learning ANN Backpropagation
Forwardpropagation

a. Initialization of weights coating with small random numbers, -0.5 up to 0.5

b. Each unit of input (x_i) sends a signal to the hidden units (z_j). Calculate all values of hidden units (z_j).
$z_{\_j} = v_{oj} + \sum_{i=1}^{n} x_i v_{ij}$ with $i = 1, \dots, n$ and $j = 1, \dots, p$
v_{oj} = Bias weights of hidden units z_j
v_{ij} = The weight of the line from the input units x_i to the hidden units z_j

c. Calculate the activation function (binary sigmoid function) on all output values of hidden units (z_j)
$z_j = f(z_{\_j}) = \frac{1}{1+e^{-x}}$

d. Each hidden unit (z_j) sends a signal to the output unit (y_k). Calculate all values output units (y_k)
$y_k = w_{0k} + \sum_{j=1}^{p} z_j w_{jk}$ with $j = 1, \dots, p$ and $k = 1, \dots, m$
w_{0k} = Weight of bias on the output unit y_k
w_{jk} =Line weights of hidden units z_j to the output unit y_k

e. Calculate the activation function (binary sigmoid function) all values output units y_k.
$y_k = f(y_{\_k}) = \frac{1}{1+e^{-x}}$
Backward propagation

f. Each unit of output y_k receiving target pattern t_k furthermore, count the output layer error information δ_k.
$\delta_k = (t_k - y_k) f'(y\_k)$
$= (t_k - y_k) y_k (1 - y_k)$, with $k = 1, \dots, m$

g. Calculate large bias correction and weighting Δw_{0k} and Δw_{jk} between the hidden layer to the output layer
$\Delta w_{0k} = \alpha \delta_k$
$\Delta w_{jk} = \alpha \delta_k z_j$, with $k = 1, \dots, m$ and $j = 1, \dots, p$

h. Each unit in the hidden layer calculation error information hidden layer (δ_j)
Calculate the sum of the hidden layer error (δ_j)
$\delta_{\_j} = \sum_{k=1}^{m} \delta_k w_{jk}$ with $j = 1, \dots, p$ and $k = 1, \dots, m$
Calculate the error information hidden layer (δ_j)
$\delta_j = \delta_{\_j} f'(z_{\_j})$
$= \delta_{\_j} z_j (1 - z_j)$ with $j = 1, \dots, p$

i. Calculate large bias correction $\Delta v_{0j}$and weight Δv_{ij} between the input layer and the hidden layer:
$\Delta v_{0j} = \alpha \delta_j$
$\Delta v_{ij} = \alpha \delta_j x_i$ with $j = 1, \dots, p$ and $i = 0, \dots, n$
Stage setting weight

j. Calculate all bias and weight changes:
Changes in weight bias and output units:
$w_{jk}(new) = w_{jk}(old) + \Delta w_{jk}$ with $k = 1, \dots, m$ and $j = 0, \dots, p$
Changes in weight bias hidden units:
$v_{ij}(new) = v_{ij}(old) + \Delta v_{ij}$ with $j = 1, \dots, p$ and $i = 0, \dots, n$

k. Test conditions stop:
If large mean squared error $\sum_{k=1}^{n}(t_k - y_k)^2$is smaller than the specified tolerance or the number of training epoch have reached the maximum, then finished, if not then go back to step 1. Tolerance value (ε) used is $0 \leq \varepsilon \leq 1$
Backpropagation algorithm to generate the final weights.

vi) Determine the classification accuracy optimization with confusion matrix and ROC curves.

vii) Validation testing data with final weights that have been done in the learning process step (v).

Steps of data analysis to logistic regression is as follows:

i) Separation of data
Credit data is separated into two parts, namely the training data and testing data. Distribution of this data is done randomly. Training data by 70% and 30% of testing data.

ii) Regression coefficient significance test simultaneously with the test statistic *G* and partially with wald test.

iii) Determine optimization classification accuracy with confusion matrix and ROC curves.

iv) Validation testing data.

Steps of data analysis to MARS is as follows:

i) Separation of data
Credit data is separated into two parts, namely the training data and testing data. Distribution of this data is done randomly. Training data by 70% and 30% of testing data.

ii) Determine the number of basis functions, boundary of basis functions between 2-4 times the number of explanatory variables.

iii) Determine the maximum interaction (MI), MI used were 1, 2, and 3.

iv) Specify a minimum of observations in each subregion, a minimum of observations used were 5,10, and 20.

v) Determine the value of a penalty (γ), γ value used 2, 3, and 4.

vi) Do a trial and error by combining the functions of the base, MI, and a minimum of observation of each subregion to obtain the minimum GCV value.
vii) Minimum GCV is determining the best model MARS.
viii) Significance test of regression coefficients simultaneously with statistics and partial *F* test statistics *t* test.
ix) Determine optimization classification accuracy with confusion matrix and ROC curves.
Validation testing data.

3. RESULT AND DISCUSSION

3.1 Prediction of ANN Backpropagation

Formation ANN Backpropagation model on the training data with the number of input layer is 9 units, output layer is one unit, whereas the determination of hidden layers, activation function, and learning rate (α) will be simulated based on the value of the minimum MSE. The results of the simulation generates hidden layer 9 units, tangent sigmoid activation function in the hidden layer to the input and the linear activation function in the hidden layer to the output, α is 0.8. So ANN Backpropagation models obtained are as follows:

$$y_k = f_k(\sum_{j=1}^{9} w_{jk} f_j(v_{0j} + \sum_{i=1}^{9} x_i v_{ij}) + w_{0k})$$

3.2 Prediction of Logistic Regression

Model obtained are as follows:

$$Logit\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -1.4167 + 0.4799\,\beta_1 + 1.1512\,\beta_2 + 0.9564\,\beta_3 + 0.8139\,\beta_4$$

$$\hat{\pi}(x) = \frac{exp(-1.4167 + 0.4799\,\beta_1 + 1.1512\,\beta_2 + 0.9564\,\beta_3 + 0.8139\,\beta_4)}{1 + exp(-1.4167 + 0.4799\,\beta_1 + 1.1512\,\beta_2 + 0.9564\,\beta_3 + 0.8139\,\beta_4)}$$

where:
β_1 =Gender
β_2 =Number of installments_1
β_3 =Number of installments_2
β_4 =Salary

Table 3.1
ANOVA Logistic Regression Model

Predictor	Estimate	Standard error	*z*	P-value	*G*	P-value
Constant	-1.4167	0.3698	-3.83	0.00	83.064	0.00
Gender	0.4799	0.1313	3.66	0.00		
Number of installments_1	1.1512	0.1976	5.83	0.00		
Number of installments_2	0.9564	0.1318	7.25	0.00		
Salary	0.8139	0.3509	2.32	0.02		

Table 3.1 shows that the partial test on any basis function with a significance level of 95% is gender, number of installments_1, number of installments_2, and salary with the value of the P-value < 0.05. The simultaneously test, there is the influence of variable

gender, number of installments_1, number of installments_2, and salary to variable Y simultaneously, so that the model fit for use.

3.3 Prediction of MARS

Formation process MARS model starts with the simplest model (without interaction) to complex models (interaction of two and interaction of three). The selection of the best model simulation by combining the number of basis functions, the number of interactions, minimal observation in each sub-region. The number of basis functions used are 18, 27, and 36. The number of interactions is 1, 2, and 3, the minimum observation in each sub-region 5, 10, and 20, a penalty (γ) is 2, 3, and 4. Model obtained are as follows:

$$Y = 3.078\text{e-}17 + 0.7057\ FB_1 - 0.1517\ FB_2 - 0.0141\ FB_3 - 0.1079\ FB_4 + 0.1069\ FB_5 - 0.2848\ FB_6 - 0.1876\ FB_7$$

where:

FB_1 = max(0, Costumerhistory)
FB_2 = max(0, Gender - 1) FB_1
FB_3 = max(0, 36 - Age) FB_1
FB_4 = max(0, Number of dependents - 2) FB_1
FB_5 = max(0, 5 - Education) max(2 - Type of work) FB_1
FB_6 = max(0, Type of work - 2) FB_1 max(Number of installments - 2)
FB_7 = max(0, 2 - Type of work) FB_1 max(Number of installments - 2)

The model above is composed of one constant and 7 basis functions, the number of observations of each region is 5 knots, γ is 4, and GCV is 0.1866. Relatively important explanatory variables are customer history, education, type of work, number of installments, gender, age, and number of dependents.

Table 3.2 ANOVA MARS model

Basis Function	Estimation	Standard Error	*t*	P-value	*F*	P-value
Constant	3.08E-17	0.0411	0	1	59.65	< 2.2e-16
FB_1	0.7057	0.0514	13.72	< 2e-16		
FB_2	-0.1517	0.0284	-5.352	1.04E-07		
FB_3	-0.0141	0.0028	-4.963	7.95E-07		
FB_4	-0.1079	0.0249	-4.335	1.58E-05		
FB_5	0.1069	0.0137	7.789	1.47E-14		
FB_6	-0.2848	0.0689	-4.134	3.82E-05		
FB_7	-0.1876	0.0290	-6.475	1.39E-10		

Table 3.2 shows that the partial test on any basis function with a significance level of 95% is FB_1, FB_2, FB_3, FB_4, FB_5, FB_6, and FB_7 with the value of the P-value < 0.05. The simultaneously test, there is the influence of variable FB_1, FB_2, FB_3, FB_4, FB_5, FB_6, and FB_7 to variable Y simultaneously, so that the model fit for use.

3.4 Accuracy Optimization of Classification

Accuracy optimization of classification based on confusion matrix and ROC on ANN Backpropagation, logistic regression, and MARS are as follows :

Table 3.3
Confusion Matrix of Training Data

Actual	MARS		ANN Backpropagation		Logistic regression	
	Prediction (%)					
	Reject	Approve	Reject	Approve	Reject	Approve
Reject	72.84	27.16	64.99	35.01	66.00	34.00
Approve	29.29	70.71	11.98	88.02	42.57	57.43

Table 3.3 shows that the model MARS has classification rate approve 70.71% and 72.84% reject, ANN Backpropagation with approve 88.02% and 64.99% reject, the results of logistic regression with approve 57.43% and 66.00% reject.

Table 3.4 Confusion matrix of testing data

Actual	MARS		ANN Backpropagation		Logistic regression	
	Prediction (%)					
	Reject	Approve	Reject	Approve	Reject	Approve
Reject	71.63	28.37	56.28	43.72	66.51	33.49
Approve	28.14	70.85	14.92	85.08	35.25	64.75

Table 3.4 shows that the model MARS has classification rate approve 70.85% and 71.63% reject, ANN Backpropagation with approve 85.08% and 56.28% reject, the results of logistic regression with approve 64.75% and 66.51% reject.

Table 3.5
Classification of Training Data Models

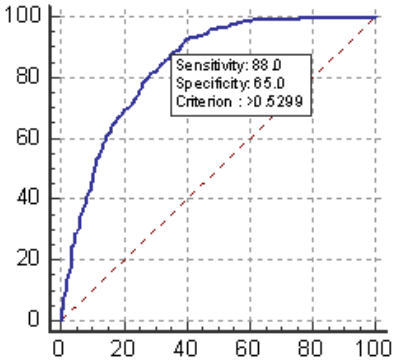
Model	Sensitivity (%)	Specificity (%)	Accuracy (%)
ANN Backpropagation	88.02	64.99	78.40
Logistic regression	57.43	66.00	61.01
MARS	70.71	72.84	71.60

Table 3.5 shows that the model which has the best accuracy value in the training data is ANN Backpropagation 78.40%, followed by the MARS 71.60%, logistic regression 61.01%.

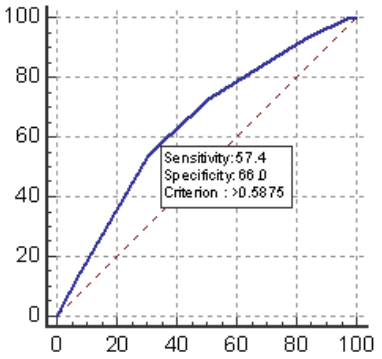
Table 3.6
Classification of Testing Data Models

Model	Sensitivity (%)	Specificity (%)	Accuracy (%)
ANN Backpropagation	85.08	56.28	72.94
Logistic regression	64.75	66.51	65.49
MARS	71.86	71.63	71.76

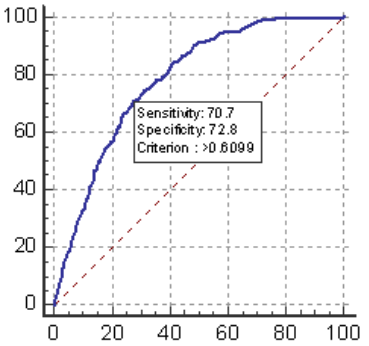
Table 3.6 shows that the model which has the best accuracy value in the testing data is ANN Backpropagation 72.94%, followed by the MARS 71.76%, logistic regression 65.49%.



ANN Backpropagation



Logistic regression



MARS

Figure 3.1: ROC of training data on ANN Backpropagation, Logistic Regression, and MARS

Figure 3.1 shows that the cutoff point were generated on ANN Backpropagation, logistic regression and MARS on training data respectively 0.5299, 0.5875, and 0.6099. Area under the curve (AUC) ANN Backpropagation 0.8410, logistic regression 0.6450, and MARS 0.7830. Based on AUC values of the three models to ANN Backpropagation model is the best model, so it can be concluded that the diagnostic accuracy of classification approve and reject the unsecured loan can be classified with either as big as 84.10% when using ANN Backpropagation.

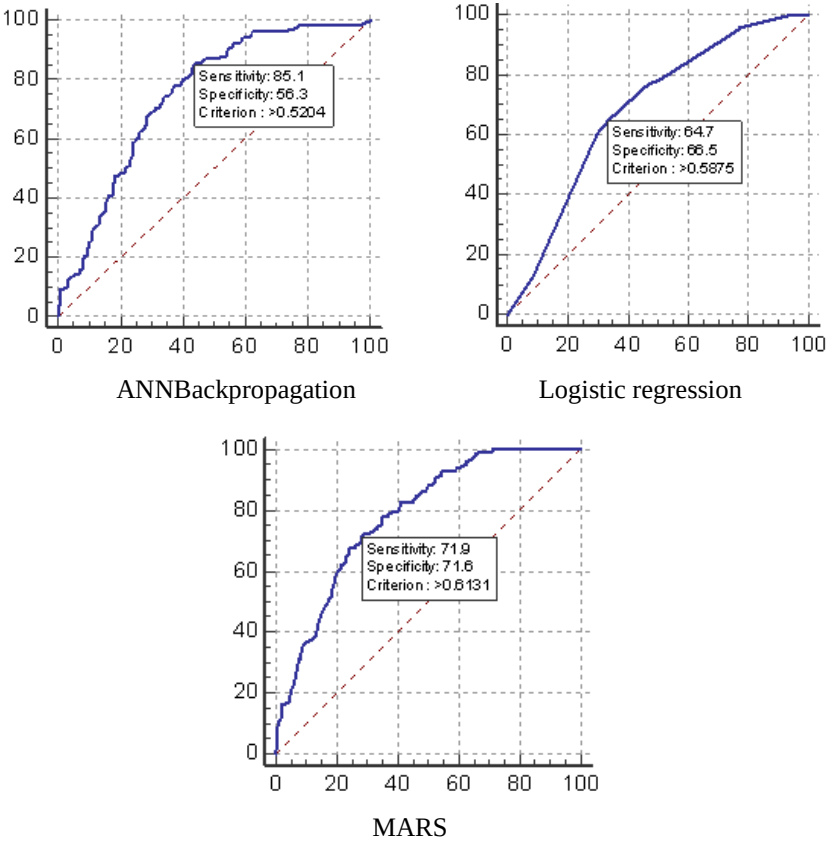


Figure 3.2: ROC of Testing Data on ANN Backpropagation, Logistic Regression, and MARS

Figure 3.2 shows that the cutoff point were generated on ANN Backpropagation, logistic regression and MARS on training data respectively 0.5204, 0.5875, and 0.6131. Area under the curve (AUC) ANN Backpropagation 0.7480, logistic regression 0.6850, and MARS 0.7830. Based on AUC values of the three models to ANN Backpropagation model is the best model, so it can be concluded that the diagnostic accuracy of classification approve and reject the unsecured loan can be classified quite well as big as 74.80% when using ANN Backpropagation.

3. CONCLUSION

Based on the value of accuracy and AUC value of the best models in a classification method for the case of an unsecured loan is ANN Backpropagation, followed by MARS and logistic regression.

REFERENCES

1. Agresti, A. (1990). *Categorical Data Analysis*. New York: John Willey and Sons.
2. Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logistik Regression*. John Wiley and Sons, New York.
3. Bank Internasional Indonesia (2014). Data Aplikasi Nasabah Kredit Tanpa Agunan Mei 2014. Jakarta: Bank Internasional Indonesia.
4. Cox, D.R. and Snell, E.J. (1989). *Analysis of Binary Data Second Edition*. London: Chapman & Hall.
5. Friedman, J.H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1), 1-141.
6. Friedman, J.H. and Silverman, B.W. (1989). Flexible Parsimony Smoothing and Additive Modelling. *Technometrics*, 31, 3-39.
7. Sutikno (2008). Statistical Downscaling Luaran GCM dan Pemanfaatannyauntuk Peramalan Produksi Padi. [Doctoral Disertation]. Bogor: Sekolah Pascasarjana, Institut Pertanian Bogor. mars
8. Gorunescu, F. (2011). Data Mining: Concepts, Models, and Techniques. Verlag Berlin Heidelberg: Springer, Berlin.
9. Kusumadewi, S. (2004). *Membangun Jaringan Syaraf Tiruan Menggunakan Matlab & Excel Link*. Yogyakarta: Graha. Ilmu.
10. Mirtalaei, M.S., Saberi, M., Hussain, O.K., Ashjari, B., Hussain, F.K. (2012). A Trust-based Bio-inspired Approach for Credit Lending Decisions. *Computing*, 94, 541-577.
11. Puspitaningrum, D. (2006). *Pengantar Jaringan Syaraf Tiruan*. Yogyakarta: Andi.
12. Shi, H.Y., Lee, K.T., Lee, H.H., Ho, W.H., Sun, D.P., Wang, J.J. and Chiu, C.C. (2012). Comparison of Artificial Neural Network and Logistic Regression Models for Predicting In-Hospital Mortality after Primary Liver Cancer Surgery. *Plos One*, 7, e35781.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 487-492

IMPACT OF CORPORATE GOVERNANCE ON FIRM FINANCIAL PERFORMANCE IN FINANCIAL SECTOR OF PAKISTAN

Nadeem Iqbal
Faculty of Management Sciences Ghazi University
DG Khan, Pakistan

Rashda Qazi
Faculty of Social Sciences Ghazi University
DG Khan, Pakistan

and

Nabila Khan
Faculty of Management Sciences Indus International Institute
DG Khan, Pakistan

ABSTRACT

Corporate governance is methods and techniques to manage and control the organization. Corporate governance (CG) focuses on transparency and accountability. This paper explores the relationship between financial performance of institutions of Pakistan and corporate governance. Corporate governance covers the three indicators including CEO duality, Board Independence and Independent Audit Committee. Return on assets and return on equity are two indicators of financial performance. Secondary data from annual reports covering period from 2009-12 is used to evaluate the result. Data show the positive relationship between financial performance and corporate governance in financial institutions of Pakistan.

INTRODUCTION

Corporate Governance (CG) is a way of being directed, organized, managed and controlled. It comprises on set of rules and regulations for all stakeholders as well as for shareholders. It defines the responsibilities and duties of the managers, board of directors, labor and shareholders. It provides proper guide to the management how to attain goals and monitor performance. A bundle of predefined standards, rules & regulations and properly stated division of labor for determining job duties and responsibilities explicitly and its implantation should be effective (Mcconomy et al. 2002). Implementation ensures the proper division of power among shareholders, management and board of directors through corporate Governance).

Corporate governance encompasses various strategies, plans and policies in order to maintain as well as improve the financial position of the firms. Corporate governance

includes corporate social responsibility, social welfare, employees career plans performance appraisal, attractive dividend policies and different strategic plans and actions but all are in accordance with business goals and objectives. Good corporate governance means directing, controlling and planning should be transparent, goal oriented and address all stakeholders stakes. Corporate governance includes CEO, Board of Directors, Chief executive officers and executive manager. Effective corporate governance always works as a Bridge between the shareholders/ stakeholders and management. That's the reasons behind this fact that they serve as Watch Dogs over each and every action or activity of management in order to secure the motives & interest of the shareholders. Hence, these shareholders are considered prime concern of corporate governance because plethora of corporate shares market value is associated with their consent and it also strengthens importance of effective governance

Therefore, the importance of effective corporate governance is going to be increasing with the passage of time because investors consider it before going to invest in a specific industry or even in an organization. Recent financial scandals like Enron, Cirio and world com where investors were more conscious about CG in order to get an insight about its business, to tap favorable impacts of CG on financial position and fore most important concern was to invest with greater awareness with minimum risk. Several studies (Recent & Past) have proved the positive link between the corporate governance and firm's value and its financial position (Drobetz et al., 2006; Beiner et al., 2006 & Brown et al., 2006).

The peculiar of firm value is massively affected by financial crises. And after the Asian financial crisis of the '90s' and late global financial crisis of the last two decades, clearly state the emergence of Governance in banking sector as well as insurance companies. Especially banks are important factor in the elaboration of the systematic risk and other risk where they tried their level best to reduce or control the risk at optimal level in order to get maximum return with minimum risk. Basel committee of banks explicitly states the relevancy of effective CG with its financial position and as this authority body issue a written document in 1990, to urge all banks to adopt the modern CG structures in order to have an effective management and satisfied customers.

Moreover, during the financial crisis of 2007-2008 many financial institutions broke down and were near to collapsed due poor governance which leads to the inactive global credit markets and financial markets. Taylor (2009) studied that the cause of their financial crisis on national economic level was due to bad government governance and ineffective monetary policies. Some recent studies found that firm's risk management & financing policies along with effective CG has a greater impact on firm's sustainability (Brunnermier, 2009). Risk management and financing policies linked with cost benefit trade off designed by board of governance (Kashyap et al., 2008).

Therefore, as a concern with financial institutions they are obliged to follow the rules and regulations in order in order to compliance with predefined standards and to satisfy their shareholders as well as their policy holders(customer). For all these, they required

effective CG to meet standards and to get clean reports from the auditors. On the basis of its importance a verdict can be extract from it that good CG plays a vital role in each organization as well as industry either it may be financial or non-financial.

Financial institutions are pillar of economy. This paper evaluates the impact of corporate governance on profitability of financial institutions in Pakistan. Good corporate governance force and attract different investors and corporate clients to invest in financial institutions with full confidence as due to this organizations compete and survive in dynamic and competitive environment. Corporate governance lead to meaningful and effectual command that facilitate to collaborative working environment in the organizations. For financial institutions, corporate governance becomes critical and essential for the stability of economy of some country.

LITERATURE REVIEW

Corporate Governance having all the standards of enterprise to support the economic agents to participate in the productive procedures, to produce excess beyond what is needed within the organization and maintain a good contribution among the partners, capturing into attention what they bring for organization. There were many researches about the link between Corporate Governance and firm performance in financial sector. Different studies concluded that corporate governance and financial performance of the company positively correlated with each other (Shleifer et al. 1997), Davis et al. 2002), Cheema, et al. 2005), Khan et al. 2011),Kumar 2012. It is also investigated that shareholders, posses condensation in firms , having a vital role to restrain and guide the management to show interest in favor of the condensation group. While corporate governance having authority to give permission to shareholders for the guidance of management for achieving better desirable position of their investment. Many studies investigated the link among four corporate governance processes (board size, board composition ,CEO/chairman duality and audit committee independence) and firm financial position measures (return on equity and profit margin) . The result have the positive interlink among the Corporate Governance procedures (board composition , board size and audit committee and performance measures , return on equity and profit margin) . The impact of study is that , the size of the board should be small within measurable limit and executive and non-executive directors must be present in the board. The research did not find out the affectionate link among the firm financial standards and CEO duality (Yasser et al. 2011).

METHODOLOGY

The research engaged mixture of primary and secondary facts and figures to find out the results. These facts gathered by the use of financial statements of the companies for the period 2009-2012. In this study various corporate governance factors have an impact on the linkage among corporate governance and firm's financial position. While in our research we just have following variables that are stated below.

INDEPENDENT VARIABLE

- CEO Duality: It prefers toward the two officers as one person was like CEO and other was Chairman.
- Board Independent: If board of the company depends one third or more upon the non executives' directors so we can say that board is independent , in case of less than one third , it is not independent.
- Audit Committee Independence: It included non-executives in the audit committee.

DEPENDENT VARIABLE

- ROA: We calculate this by dividing, net income divided by total assets of the company.
- ROE: We calculate this by Net Income divided by shareholders equity of the company.
- For exploring the facts the study has taken a four listed companies of financial institutes from KSE from Annual reports as secondary data. Annual reports of these institutions are studied from 2009- for the collection of data.

Analysis of Financial Institutions

Company Name	Year	CEO Duality	Audit Committee	Board Independent	ROA %	ROE %
Meezan Bank Ltd	2009	Yes	Yes	Yes	13.9	16
	2010	Yes	Yes	Yes	9	11
	2011	Yes	Yes	Yes	14	12
	2012	Yes	Yes	Yes	13	11.56
Bank Islami	2009	No	Yes	Yes	14.57	13.69
	2010	No	Yes	Yes	12.5	13.01
	2011	No	Yes	Yes	11.278	14
	2012	No	Yes	Yes	13.32	14.56
State Life Insurance Pakistan	2009	No	Yes	No	9.89	9.46
	2010	No	Yes	No	9.09	9
	2011	No	Yes	No	8.67	9.78
	2012	No	Yes	No	8.82	9.97
New Jubliee Life Insurance	2009	No	Yes	Yes	0	0
	2010	No	Yes	Yes	2.3	4.5
	2011	No	Yes	Yes	5.2	5.98
	2012	No	Yes	Yes	5.8	6.2

It is noted from the above table that Meezan Bank observe the two indicators of corporate governance and Bank Islami fulfill the three indicators of corporate governance. It is also evaluated from the table that there is fluctuation in financial performance of Meezan Bank and Bank Islami over the time as ROA indicator is adopted. It is observed from the table that Meezan Bank performance in ROE has much fluctuation as compare to Bank Islami. Bank Islami shows stable and increasing trend of ROE. It is also observed from the table that State Life Insurance Pakistan adopted two indicators of corporate governance and violate the board independence indicator whereas New Jubliee Life Insurance adopted three indicators of corporate governance. It is also noted that state life insurance of Pakistan is facing decreasing trend of financial performance and New Jubliee Life Insurance is earning at increasing trend by ROA and ROE indicators. It is also evaluated that New Jubliee Life Insurance adopted the all parameters of corporate governance and enjoying increasing trend of earning and financial performance.

DIRECTIONS FOR FUTURE RESEARCH

The study of financial institution has proved that there is positive relationship between the corporate governance and financial performance. However this study has focused on only three determinants of corporate governance (CEO Duality, board independent and audit committee independence). There exist other internal and external factors and determinants of corporate governance that need to be investigated. Further investigation also can be carried assuming other factors to examine the relationship between the corporate governance and firm financial performance. It can also be extended by considering other measures of performance.

REFERENCES

1. Beiner, S., Drobetz, W., Schmid, M.M. and Zimmermann, H. (2006). An integrated framework of corporate governance and firm valuation. *European Financial Management*, 12(2), 249-283.
2. Brown, L.D. and Caylor, M.L. (2006). Corporate governance and firm valuation. *Journal of Accounting and Public Policy*, 25(4), 409-434.
3. Brunnermeier, M.K. and Pedersen, L.H. (2009). Market liquidity and funding liquidity. *Review of Financial Studies*, 22(6), 2201-2238.
4. Bujaki, M. and McConomy, B.J. (2002). Corporate Governance: Factors Influencing Voluntary Disclosure by Publicly Traded Canadian Firms*. *Canadian Accounting Perspectives*, 1(2), 105-139.
5. Cheema, A., Khwaja, A.I. and Qadir, A. (2005). *Decentralization in Pakistan: Context, content and causes. Research Programs*, John F. Kennedy School of Government, Harvard University.
6. Davis, E.P. (2002). Institutional investors, corporate governance and the performance of the corporate sector. *Economic Systems*, 26(3), 203-229.

7. Hadani, M., Goranova, M. and Khan, R. (2011). Institutional investors, shareholder activism, and earnings management. Journal of Business Research, 64(12), 1352-1360.
8. Iqbal. N, Ahmad and Hammad, N. (2014) Corporate Social Responsibility and its Possible Impact on Firm's Financial Performance. *Arabian Journal of Business and Management Review* (OMAN Chapter) 3(12), July. 2014.
9. Kashyap, A., Rajan, R. and Stein, J. (2008). Rethinking Capital Regulation. *Maintaining Stability in a Changing Financial System*, 431-471.
10. Kumar, N. and Singh, J.P. (2012). Outside directors, corporate governance and firm performance: Empirical evidence from India. *Asian Journal of Finance & Accounting*, 4(2), 39-55.
11. Shleifer, A. and Vishny, R.W. (1997). A survey of corporate governance, *J. Finan. Econ.*, 20, 431-460.
12. Taylor, J.B. (2009). *Getting Off Track: How Government Actions and Interventions Caused, Prolonged, and Worsened the Financial Crisis*. Hoover Press.
13. Yasser, Q.R., Entebang, H. and Mansor, S.A. (2011). Corporate governance and firm performance in Pakistan: The case of Karachi Stock Exchange (KSE)-30. *Journal of Economics and International Finance*, 3(8), 482-491.

Proc. ICCS-13, Bogor, Indonesia
December 18-21, 2014, Vol. 27, pp. 493-502

THE DYNAMICS APPROACH FOR REGIONAL DISPARITY IN CENTRAL SULAWESI AFTER DECENTRALIZATION POLICY

Krismanti Tri Wahyuni
Department of Statistics, Institute of Statistics (Sekolah Tinggi Ilmu Statistik) Indonesia. Email: krismanti@stis.ac.id

ABSTRACT

Central Sulawesi is a region that is rarely studied in the national scope, but it is the North Sulawesi and South Sulawesi. However, Central Sulawesi is not classified as advanced, in term of income and welfare. Range area of Central Sulawesi is quite spacious and has different characteristics across the regions. It is influenced by the small kingdoms in ancient.

This study aimed to analyze the dynamics of disparity in economic development among regions in Central Sulawesi after decentralization policy. The study employs the data of 11 districts/cities from 2001 up to 2013. It estimates by using dynamic panel data. The results showed that the regencies/cities inequality in Central Sulawesi is relatively small when viewed from the Williamson Index since the implementation of fiscal decentralization and tending to convergent. The same result happens when the data were analysed using the oil and gas GRDP.

KEYWORDS

Convergence, inequality, dynamics panel data.

1. INTRODUCTION

Central Sulawesi is a region that is rarely studied in the national scope, but it is important to the economy of the island of Sulawesi, that become the main route connecting the North Sulawesi and South Sulawesi. However, Central Sulawesi is not classified as advanced, in term of income and welfare. Range area of Central Sulawesi is quite spacious and has different characteristics across the regions. Central Sulawesi is the largest province in Sulawesi Island, covering a land area of 61,841.29 square kilometers (based on Home Affairs Ministerial Regulation Number 18/2013) which include the eastern peninsula and a part of the long northern peninsula, including the island group of Togean in the Tomini Bay and Banggai Island in Tolo Bay.

With a wide area and is located on separate islands, there are diverse cultures that affect the economic life of Central Sulawesi. It is also influenced by the small kingdoms in ancient. Territorial Boundaries of Central Sulawesi Province in northern area bordered by Sulawesi Sea and Gorontalo Province, eastern area border on Maluku Province, southern area border on West Sulawesi Province and South East Sulawesi, and western area bordered by Makassar strait.

Central Sulawesi is a relatively distant area from the growth center in eastern Indonesia, that is Makassar. However, the role of local revenue is reflected in 16.04 percent of GRDP at current market prices, including oil and gas compared to the overall provinces on the Sulawesi island or about 17.03 percent to GRDP at constant 2000 market prices of oil and gas in 2013. Excluding oil and gas, Central Sulawesi's GRDP is about 15.85 percent to GRDP at current market prices or about 16.89 percent to GRDP at constant 2000 market prices in 2013. The percentages are second order after South Sulawesi. Based on these data, Central Sulawesi has increased ratings higher than North Sulawesi's GRDP in the last three years. The Central Sulawesi's contribution on Sulawesi Island's GRDP including oil and gas have increased slightly. Observed by source, value-added oil and gas resources in Central Sulawesi comes from one region only, ie Morowali.

Based on regency/city's GRDP in Central Sulawesi province, the highest gross value-added comes from Palu, a difference of 167 billion dollars with Parigi Moutong's GRDP in 2013. Palu's GRDP ranks first with a contribution of 16.60 percent at current market prices or about 15.87 percent at constant market prices in the same year. The smallest Gross value-added in Central Sulawesi derived from Tojo Una-una, about 3.13 percent at current market prices or about 2.75 percent at constant market price in 2013.

The difference in per capita income by the region in 2013 is quite large, reaching nearly 20 million per year for GRDP per capita, including oil and gas and nearly 16 million per capita per year for non-oil GRDP. The highest oil and gas GRDP per capita in 2013 achieved by Morowali (32.01 million per capita) and the lowest is Banggai Islands (12.31 million per capita) in the same year. Though based on Williamson's coefficient of variation, which showed inequality region, the average inequality of regencies/city in Central Sulawesi is relatively small (in the range of 0.2), with the fluctuated numbers from 2000 to 2013, as shown in Table 1.

Table 1
Williamson's Coefficient of Variation in Central Sulawesi 2000 - 2013
(calculated using both oil and gas and non-oil GRDP)

Year	Williamson Index of Oil and Gas GRDP	Williamson Index of Non-Oil GRDP
2000	0,2624	0,2624
2001	0,2343	0,2343
2002	0,2119	0,2119
2003	0,2060	0,2060
2004	0,2139	0,2139
2005	0,2091	0,2089
2006	0,2156	0,2131
2007	0,2178	0,2111
2008	0,2228	0,2125
2009	0,2157	0,2091
2010	0,2231	0,2084
2011	0,2374	0,2116
2012	0,2533	0,2243
2013	0,2606	0,2372

Inequality between regions at the beginning of economic development is a natural thing in the concept of national development. Williamson (1965) in Tambunan (2001) found that in the early stages of economic development, income inequality will be enlarged and concentrated in certain areas that are relatively advanced, for example in industrial development, infrastructure and human resources. Later in the stage of rapid economic growth, convergence and the inequality in the distribution of income will decline. Is it true that areas in Central Sulawesi has reached a stage of development in such level of convergence thus decreasing in inequality and led to the point of common progress in terms of economic growth? This is very important question to answer since the government's focus in development are the regencies/city, as from fiscal decentralization implemented in 2001. The decentralization policy goal is that regencies/city can maintain production levels and develop the economic sectors that are important according to the potential and diversity of resources that are available. Reviewing the economic growth, especially in the area of fiscal policy implementation, that is the level of the regencies/city occupies an important position in the development of the wider region and the constellations surrounding regions. Structural economic conditions in each region is also associated with various factors endowment of every area, which lead to inequalities between regions.

Based on this background, this study essentially, aims to:

1) Getting the best models in dynamic panel data analysis to calculate the convergence of the regencies/city in Central Sulawesi and the level of convergence using the data of oil and non-oil GRDP.
2) Analyze the occurrence or non-occurrence of regencies/city's convergence in Central Sulawesi.

The scope of this research is whole regencies/city in Central Sulawesi (10 regencies and one city), by separating the data for the expansion area to ensure the consistency of data so that the research's analysis conducted on 11 areas. The time period of the study was 13 years, started the implementation of fiscal decentralization in 2001 to 2013, which is the recent collected data.

2. LITERATURE

The economic growth's theory begins from Keynesian growth model applied to many developing countries, which emphasizes the dual role of investment through capital accumulation process that increase the capital stock and production capacity to work on aggregate supply (Jhingan, 2008). According to the Solow model, economic growth is a series of activities that are rooted in human capital accumulation, the use of modern technology and output, in order to achieve sustainable economic growth and constantly converging towards balanced growth (Blanchard, 2006). Furthermore, endogenous growth model is based on the importance of human resources, the knowledge stock as a source of increased economic productivity and the importance of learning by doing and human capital (Capello, 2007).

Convergence theory is based on the Solow-Swan model. Long-term growth rate is determined by exogenous variables in steady state, where k, y and c per capita does not grow and aggregate variables K, Y and C grew at the rate of population growth rate n

(Barro and Sala-i-Martin, 1995). The smaller the value of k, then the value of $\dot{k}/k$ greater, ceteris paribus. This shows that the economy with lower individual capital will grow faster (convergence trend). An area / country that began with a low ratio of worker capital will have a higher growth rate $\dot{k}/k$ per capita.

The region inequality analysis is based on the differences in the ability of potential resources in the region and the relative growth between regions and the possibility of convergence in growth rates or average income (Capello, 2007). Inequality between regions has implications for the level of social welfare among regions and also to the formulation of regional development policies carried out by the Local Government. The government has an important role in the context of the inter-regional inequality prevention through the deployment of communications facilities and labor, the development of growth centers, and the implementation of decentalization (Sjafrizal, 2008).

3. METHODS

This study uses data of regencies/city in Central Sulawesi from 2001 to 2013, which consists of: GRDP at constant 2000 market prices, both oil and non-oil, the population to calculate GRDP per capita, the amount of labor and investment are obtained from the data GFCF (GRDP components according to usage, which is based on the price of the base year 2000). GRDP at current prices is used for calculating the Williamson Index.

Measurements made with the region inequality coefficient of variation Williamson with the formula:

$$CV_w = \frac{\sqrt{\frac{f_i}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}{\bar{y}}$$

That y_i is the GRDP per capita of the region, $\bar{y}$ is GRDP per capita across the region, f_i is the population of the region and n is population of the entire region. CV_w value is between 0 and 1. If the approach to zero means very evenly, otherwise if close to 1 means very lame.

Estimation of convergence is done by using panel data. Panel data can be defined as repeated observations on each unit of the same cross section, which has a characteristic in which $N > 1$ and $T > 1$. Let y_{it} is dependent varabel value for unit i-th cross section at a time to-t dengan $i = 1, 2,\ldots, N$ and $t = 1, 2,\ldots,T$. Standard linear panel data models can be expressed as $y = X'\beta + \varepsilon$.

Panel data used in this research is a dynamic panel data. Relationships between economic variables in fact many dynamic. Panel data analysis can be used in the dynamic model related to the analysis of dynamic adjustment. This dynamic relationship is characterized by the presence of lagged dependent variable among the regressors variables. As an illustration, a dynamic panel data model is as follows:

$$y_{it} = \delta y_{i,t-1} + x'_{it}\beta + u_{it}; i = 1, \ldots, N; t = 1, .., T$$

with δ denotes s scalar, x'_{it} states 1xK sized matrix and β is Kx1 sized matrix. In the static panel data model, it can be shown the consistency and efficiency of both the FEM

and REM-related treatment of μ_i. In the dynamic model, this situation is very different in substance, that is y_{it} fungtion from μ_i then $y_{i,t-1}$ is also a fungtion of μ_i. Because μ_i is a fungtion of u_{it}, there will be a correlation between regressors variables $y_{i,t-1}$ with u_{it}. This will cause the least square estimator (as used in the static panel data model) to be biased and inconsistent, even if v_{it} uncorrelated series though.

To overcome this problem, the method of moments approach can be used. Arrelano and Bond suggests an approach to the generalized method of moments (GMM). There are two types of GMM estimation procedures are commonly used:

(i) *First-difference* GMM (FD-GMM atau AB-GMM)
Transformation in addressing inconsistencies that occur by using the first difference to approach the instrument variables.
(ii) *System* GMM (SYS-GMM)
The basic idea of using the system GMM method is to estimate the system of equations both in first-differences as well as at the level where the instruments are used at the level of first-differences is the lag of the series.

Specifications Model

The study was carried out by assuming the function of the Cobb-Douglas constant return to scale the output (Y) and three inputs, that are capital (K), labor (L) and labor augmenting technological progress (A):

$$Y(t) = K(t)^{\alpha}(A(t)L(t))^{1-\alpha},\ 0 < \alpha < 1$$

With the decline of the formula, the resulting panel data literature equation:

$$\ln y_{it} = \gamma \ln y_{i,t-1} + \beta_1 \ln s_{i,t-1} + \beta_2 \ln(n + g + \delta)_{i,t} + \eta_i + \upsilon_{i,t}$$

where $x_{it} = (\ln(s_{it}), \ln(n_{it} + g + \delta), \theta = ((1-\varsigma)\frac{\alpha}{1-\alpha}, -(1-\varsigma)\frac{\alpha}{1-\alpha})$ and $\gamma = 1 + \beta = \varsigma$. And the convergence of the model, can be written as follows:

$$\Delta z_{it} = (1-\alpha)\,\Delta\, z_{i,t-1} + \beta'\,\Delta\, x_{it} + D_i + \Delta u,$$ with $i = 1, 2,\ldots, N$ and $t = 1, 2,\ldots, T$.

Referring to these theories, the research model used in this study are:

$$\ln pdrb_{it} = (1-\alpha)\ln pdrb_{i,t-1} + \beta_1 \ln pmtb_{it} + \beta_2 \ln labour_{it} + \upsilon_{it}$$

The dependent variable y is GRDP per capita. With the purpose of performing a comparison between the oil and non-oil and gas GRDP data, there are two models used, with the dependent variable as a distinguishing, that are oil and gas GRDP in the first model and the non-oil GRDP in the second model. The process of convergence occurs when the coefficients of (1 - α) is less than one, the convergence rate is expressed as - ln lag of the dependent variable coefficients. While independent data on both models is the same, that are investment per capita and labor. Data investments approximated using data GFCF (Gross domestic Fixed Capital Formation), which is calculated based on constant market prices. GFCF is defined as the procurement, manufacture, purchase of new capital goods from domestic and new and used capital goods from abroad, reduced net sales of second-hand capital goods. Inclusion of second-hand goods from abroad as the new capital goods in the country, because of its value in the economy has not been taken into account. Capital goods can also be interpreted as the goods or equipment used in the

production process repeatedly and have one year lifetime or more. In its application, the calculation includes the addition of reduced reduction GFCF assets (property) remains both new and secondhand broken down by assets, such as residential buildings, non-residential buildings, other buildings, machinery and equipment, vehicles and livestock; major repairs of tangible assets and the cost of transfer of ownership of assets.

4. RESULT AND DISCUSSION

Estimation with a dynamic panel data model is done by two procedures, that are First-difference GMM (FD-GMM or AB-GMM) and System GMM (Sys-GMM). But apparently the result of running the data indicate that the processing of data with both procedures yield the same model, both the level of significance and the coefficient. Independent variables in the estimated coefficients are all positive and in accordance with the theory. Likewise with all the independent variables in the model have a significant effect on the dependent variable, including the lag dependent variable. Based Sargan test statistic, the null hypothesis that the variable is valid instrument rejected, with p-value 1.0000 for FD-GMM and Sys-GMM in both models. This indicates that the instrument variables used in both dynamic panel data model is valid. Test of model consistency is done by looking at the level of significance of AB m_1 which is significant at 5 percent level and AB m_2 which is not significant at the 5 percent level, meaning there is no serial correlation in the Sys-GMM models for both data oil and non-oil or both models are consistent.

The process of convergence can be seen from the coefficients of autoregressive parameters of the dependent variable. The value of the coefficient of y_{t-1} are less than 1 indicate a convergence process, while a value greater than 1 indicates that the dependent variable persistent. Based on the model of dynamic data obtained by the panel, the coefficient of y_{t-1} in both models the same, there are model with oil-gas GRDP as dependent variable and models with non-oil GRDP as dependent variable. However, the results for each different method are different. FD-GMM method produces coefficient of y_{t-1} which amounted to 0.755, while the Sys-GMM method amounted to 0,806 with the rate of convergence of each of 28.086 percent and 21.525 percent. Estimation obtained by dynamic model with Sys-GMM method that meets the test instruments and test the consistency of the model is the best model generated in this study. Henceforth, the model of the research below refers to the Sys-GMM models were produced, according to the following table.

Table 2
Results Running Dynamic Panel Data by Using Gas Dependent Variables and Non-Oil, both FD-GMM and Sys-GMM Method

Comment	MIGAS		NON MIGAS	
	FD-GMM	**Sys-GMM**	FD-GMM	**Sys-GMM**
ln $PDRB_{t-1}$	0,755 (0,000)	**0,806** **(0,000)**	0,755 (0,000)	**0,806** **(0,000)**
ln PMTB	0,058 (0,030)	**0,048** **(0,029)**	0,058 (0,030)	**0,048** **(0,029)**
ln LABOUR	0,157 (0,021)	**0,124** **(0,014)**	0,157 (0,021)	**0,124** **(0,014)**
Implied λ	28,086	**21,525**	28,086	**21,525**
Wald-Test	839,330 (0,000)	**1201,910** **(0,000)**	839,330 (0,000)	**1201,910** **(0,000)**
Sargan Test	9,335 (1,0000)	**9,286** **(1,0000)**	9,335 (1,0000)	**9,286** **(1,0000)**
AB m_1	-2,800 (0,0051)	**-2,816** **(0,0049)**	-2,800 (0,0051)	**-2,816** **(0,0049)**
AB m_2	-0,012	**0,065**	-0,012	**0,065**

Note: the numbers in the brackets indicate the level of significance

Based on the research results, the variable coefficient GFCF (investment) per capita in the model is equal to 0.048, meaning that changes in the percent of per capita investment will increase 4.80 percent of GRDP per capita. While the labor coefficient of 0.124 means that changes in the percent of the workforce will increase by 12.40 percent of GRDP. It turned out that in the economy of Central Sulawesi, the effect of labor is more dominant than the investment. If we compare our model using oil and non-oil data is no different, it can be concluded that the inequality regions not affected by oil and gas production in the province of Central Sulawesi. It is as if contrary to depictions Williamson index, as shown in Figure 1 (depiction prepared from Table 1).

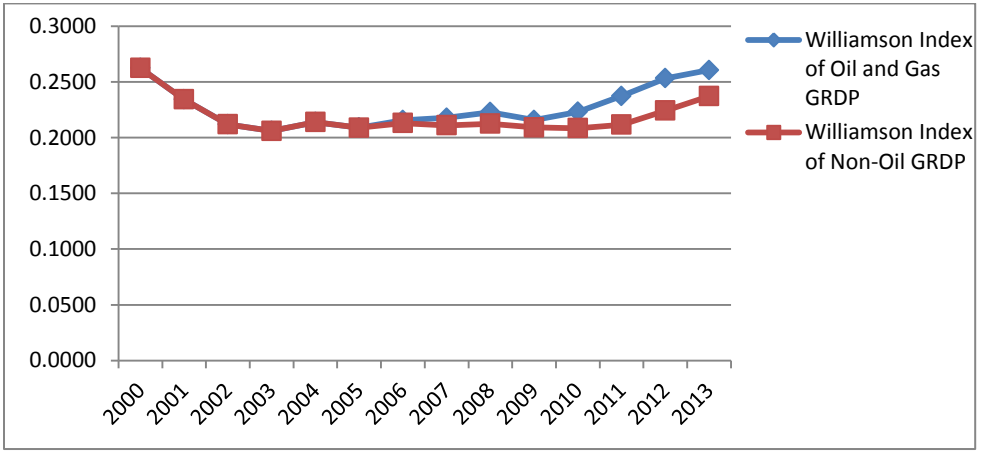


Fig. 1: Coefficient of Variation Williamson Central Sulawesi Year 2000-2013 (calculated with the use of oil and gas and non-oil GRDP)

Oil and gas data did not affect the estimation of GRDP convergence in Central Sulawesi, because the estimation for oil and non-oil produces exactly same equation data model. Therefore, it is quite interesting to analyze the data of oil and gas produced in Central Sulawesi, the newly calculated in GRDP beginning in 2005, due to new oil and gas production process implemented in that year. Oil and gas production in this region is only produced in Morowali, which is an expansion area of Poso Regency in 1999.

Table 3
Oil and Gas Data, Contributions, GDP Deflator, Growth of GDP Deflator, and Growth of Oil and Gas Data of Central Sulawesi Year 2005 – 2013

Year	Oil and Gas Data (Million Rupiahs)		Contribution to GRDP of Morowali		Contribution to GRDP of total regencies/city in Central Sulawesi		GDP Deflator	Growth of GDP Deflator	Growth of Oil and Gas Data
	At current price	At constant price	At current price	At constant price	At current price	At constant price			
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
2005	63 254	41 385	4.57	3.91	0.38	0.36	152.84	-	-
2006	241 984	115 098	13.98	9.55	1.26	0.92	210.24	37.55	178.12
2007	463 534	216 633	21.59	15.60	2.08	1.59	213.97	1.77	88.22
2008	661 960	280 027	24.55	18.16	2.46	1.91	236.39	10.48	29.26
2009	644 362	272 204	21.61	16.43	2.09	1.72	236.72	0.14	-2.79
2010	766 156	289 816	20.57	15.54	2.12	1.68	264.36	11.68	6.47
2011	946 160	307 556	19.86	14.39	2.23	1.62	307.64	16.37	6.12
2012	1 029 590	301 447	17.49	12.51	2.07	1.45	341.55	11.02	-1.99
2013	907 316	260 528	12.81	9.58	1.55	1.14	348.26	1.96	-13.57

Based on the data in Table 3, the role of oil and gas in the economy in Central Sulawesi is very small, less than 3 percent annually to GRDP at current market prices. Oil and gas contribution to GRDP at constant 2000 prices, Central Sulawesi smaller, less than 2 percent annually. The oil and gas production has ever experienced a decrease compared to the previous year, indicated by a negative growth of value added of oil and gas in 2009, 2012 and 2013. These small figures and this fluctuating production did not affect the estimation of convergence regencies/city in Central Sulawesi.

The percentage ratio between GDP at current prices and GDP at constant 2000 prices produce a measure called the GDP deflator, which reflects the price of oil and gas that occurred during the year. The growth of GDP deflator shows the change in gas prices in general, and the prices at the level of these producers have fluctuated during the extraction of oil and gas in Morowali. With the fluctuation of oil and gas production and the price of oil and gas produced causes no significant effect on the economy of Central Sulawesi. In other words, economic development can ignore the existence of oil and gas in this region.

How is the increase in value-added economy between regencies/city in Central Sulawesi? Value added in the forward areas in Central Sulawesi was experiencing a

lower rise than in the less developed regions, is evidenced by the level of income convergence regions reached 21.525. Is it true that areas "relatively advanced" has been declining their economic growth? With the Local Autonomy Law that gives freedom to the region to develop local strengths, areas already considered advanced tend to bloomed in the development of infrastructure and local potential exploration. On the other hand, true regions "less advanced" in Central Sulawesi has been more rapid in comparison with areas that have been developed so as to achieve convergence? The new expansion areas usually tend to get more attention from the government's investment in terms of public development and improvement of human resources, which would increase the workforce in the process of formation of value-added economy.

5. CONCLUSION

The conclusions of this research are:

1) Inequality of regencies/city in Central Sulawesi is relatively small when viewed from the Williamson Index since the implementation of fiscal decentralization.
2) The process of income convergence (based on GRDP) in Central Sulawesi when analyzed using dynamic panel data approach, evidenced by the lag of the dependent variable which is less than one.
3) The best model was chosen in research is a dynamic panel model with Sys-GMM method, the model was valid in test instruments and consistence in the test that there is no serial correlation (consistent).
4) The rate of convergence generated in this study indicated the implied value of λ is equal to 21.525.
5) Convergence is happening in Central Sulawesi supported by the presence of per capita investment and labor where the influence of labor is more dominant than the investment per capita. That is, the increase in investment has yet to reach the optimum value in the economy and can be upgraded to achieve a greater revenue growth rate.

Based on the conclusions obtained from this study, suggested an increase in economic activity in manufacturing by increasing investment, especially in areas less developed, in order to avoid concentration of investments leading to higher values of capital goods in the course of growth centers, in order to improve the distribution of income and the speed of convergence and reduce the depletion of resources in the vicinity. In terms of methodology, future studies should incorporate spatial effects (spatial filtering) in dynamic panel data model as the locus of the regencies/city, because the interaction of economic and spatial dependence between regions must be very high (Badinger, et al., 2002). Furthermore, future research needs to be done by comparing the Central Sulawesi with a wider area, such as the convergence of the regencies/city on Sulawesi Island, so it can see the "position" of Central Sulawesi more objectively, internally and externally. Does not rule out the possibility that the imbalance in the already small province in fact undergone a process of convergence with a relatively high speed, but it turns out if positioned with the other regions, the economic progress of Central Sulawesi still not significant.

REFERENCES

1. Badinger, Harald, Werner, Muller, Gabriele Tondl. (2002). Regional Convergence in the European Union (1985–1999). *IEF Working Papers*, 47, 7-17.
2. Baltagi, Badi Hani (2005). *Econometric Analysis of Panel Data*. Ed ke-3. Chicester: John Wiley & Sons. Ltd.
3. Barro, Robert dan Xavier Sala-i-Martin. (1995). *Economic Growth*. New York: McGraw-Hill.
4. Blanchard, Olivier. (2006). *Macroeconomics*. New York: Prentice Hall Business Publishing.
5. Capello, Robert. (2007). *Regional Economics*. New York: Routledge.
6. Firdaus, Muhammad (2006). *Impact of Investment Inflows on Regional Disparity in Indonesia* [disertasi]. Malaysia: Universiti Putra Malaysia.
7. Jhingan (2008). *Ekonomi Pembangunan dan Perencanaan*. Jakarta: PT. RajaGrafindo Persada.
8. Kuncoro, Mudrajad (2002). *Analisis Spasial dan Regional: Studi Aglomerasi dan Kluster Industri Indonesia*. Yogyakarta: UPP AMP YKPN.
9. Pontoh, Nia dan Iwan Kustiawan. (2009). *Pengantar Perencanaan Perkotaan*. Bandung: Penerbit ITB.
10. Rumayya, Wirya Wardana dan Erlangga Agustino Landiyanto. (2005). Growth in East Java: Convergence or Divergence? *EconWPA* 0508, 15-16.
11. Sala-i-Martin, Xavier. (1994). Regional Cohesion: Evidence and Theories of Regional Growth and Convergence. *Economic Working Paper*, 104, 31-33.
12. Sjafrizal. (2008). *Ekonomi Regional Teori dan Aplikasi*. Sumatera Barat: Pranita Offset.
13. Wahyuni, Krismanti Tri. (2011). Konvergensi dan Faktor-faktor yang Memengaruhi Ketimpangan Wilayah Kabupaten/Kota di Sulawesi Tengah [tesis]. Bogor: Institut Pertanian Bogor.