Department of Statistics, Universitas Islam Indonesia

Advances on Statistical Methods in Engineering, Science, Economy, Education and Disaster Management

Proceedings of
The First International Conference on
Statistical Methods in Engineering, Science,
Economy and Education (SESEE-2015)
Universitas Islam Indonesia,
Yogyakarta, Indonesia
September 19th-20th, 2015.

ISBN : 978-602-99849-2-7

Advances on Statistical Methods in Engineering, Science, Economy, Education and Disaster Management

Proceedings of
The First International Conference on Statistical Methods in Engineering, Science, Economy and Education (SESEE-2015)
Universitas Islam Indonesia,Yogyakarta, Indonesia
September 19th-20th, 2015.

Editor
Dr. RB. Fajriya Hakim, S.Si., M.Si.
Department of Statistics,
Universitas Islam Indonesia
e-mail: hakimf@uii.ac.id

ISBN : 978-602-99849-2-7

Advances on Statistical Methods in Engineering, Science, Economy, Education and Disaster Management

Proceedings of
The First International Conference on Statistical Methods in Engineering, Science, Economy and Education (SESEE-2015)
Universitas Islam Indonesia,Yogyakarta, Indonesia
September 19th-20th, 2015.

The First International Conference on Statistical Methods in Engineering, Science, Economy and Education (SESEE-2015)

Organized by:
Department of Statistics
Faculty of Mathematics and Natural Sciences.
Universitas Islam Indonesia

Published by
Department of Statistics
Faculty of Mathematics and Natural Sciences.
Universitas Islam Indonesia
Sleman, Yogyakarta
Indonesia.

Conference Organizer

Patron

Dr. Ir. Harsoyo, M.Sc.	The Rector of Universitas Islam Indonesia

Advisor

Prof. Drs. Suryo Guritno, M. Stat., Ph.D.	Universitas Islam Indonesia
Prof. Akhmad Fauzy, S.Si., M.Si., Ph.D.	Universitas Islam Indonesia
Drs. Allwar, M.Sc., Ph.D.	Universitas Islam Indonesia

General Chair

Dr. RB. Fajriya Hakim, S.Si., M.Si.	Universitas Islam Indonesia

Organizing Committee

Arum Handini Primandari, S.Pd.Si., M.Sc.	Universitas Islam Indonesia
Asmadini Handayani, S.Si., M.M.	Universitas Islam Indonesia
Atina Ahdika, S.Si., M.Si.	Universitas Islam Indonesia
Ayundyah Kesumawati, S.Si., M.Si.	Universitas Islam Indonesia
Muhammad Hasan Sidiq, S.Si., M.Sc.	Universitas Islam Indonesia
Muhammad Muhajir, S.Si., M.Si.	Universitas Islam Indonesia
Rahmadi Yotenka, S.Si., M.Si.	Universitas Islam Indonesia
Tuti Purwaningsih, S.Stat., M.Si.	Universitas Islam Indonesia

Reviewer

Prof. Dr.rer-nat., Dedi Rosadi, S.Si., M.Sc.	Gadjah Mada University, Indonesia
Dr. Edy Widodo, S.Si., M.Si.	Universitas Islam Indonesia
Dr. Fatekurrohman, S.Si., M.Si.	UNEJ, Indonesia
Iwan Tri Riyadi Yanto, M.Sc.	Ahmad Dahlan University, Indonesia
Dr. Jaka Nugraha, S.Si., M.Si.	Universitas Islam Indonesia
Kariyam, M.Si.	Universitas Islam Indonesia
Dr. Techn. Rohmatul Fajriyah, S.Si., M.Si.	Universitas Islam Indonesia
Sugiyarto, S.Si., M.Si., Ph.D.	Ahmad Dahlan University, Indonesia
Dr. Suhartono, S.Si., M.Sc.	Institut Teknologi Surabaya, Indonesia
Assoc. Prof. Dr. Tutut Herawan	Universiti Malaya, Malaysia
Dr. Yosza Dasril	Universiti Teknikal Malaysia, Malaysia

Student Committee

M Nasihun Ulwan	Universitas Islam Indonesia
Rudy Prietno	Universitas Islam Indonesia
Septianusa	Universitas Islam Indonesia

ISBN : 978-602-99849-2-7

Preface

We are honored to be part of this special event in The First International Conference on Statistical Methods in Engineering, Science, Economy, and Education (SESEE-2015) together with Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia.

This first edition of the conference is a continuation of the First Workshop on Big Data and Hadoop Implementation which has been successfully held in the Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia during the period of September 16-18, 2015 which has attracted more than 100 participants.

This first conference event we accept 26 submissions after rigorous review processes by program committee members of SESEE-2015. All registered and presented papers will be published by online proceedings and indexed by Google Scholar. The best selected papers will be invited by Scopus indexed journal.

On behalf of SESEE-2015, we would like to express our highest appreciation to be given the chance to do cooperation with KOMPAS Media for their support. Our special thanks go to the Rector of Universitas Islam Indonesia, the Dean of Faculty of Mathematics and Natural Sciences, Steering Committee, Program Committee Chairs, Organizing Chairs, all Program and Reviewer Committee members and all the additional reviewers for their valuable efforts in the review process that helped us to guarantee the highest quality of the selected papers for the conference.

We also would like to express our thanks to the three keynote speakers, Prof Dr Jemal H. Abawajy from Deakin University, Australia, Prof Dr Mustafa Mat Deris from Universiti Tun Hussein Onn Malaysia and Mr. Younes Sa'adi, MT., from Batna Universite, Algeria. We are also very pleased to have with us our best special coordinator in conducting this Conference, Dr. Tutut Herawan from Universiti Malaya, Malaysia.

We cordially thank all the authors for their valuable contributions and other participants of this conference. The conference would not have been possible without them. We wish to thank the members of the student committee for their very substansial work, especially whose who played essential roles, Septianusa, Rudy and Ulwan.

We hope that the papers become very useful references to enrich the ideas that can support the advancement of statistical method in various application.

SESEE-2015 Chair

RB Fajriya Hakim

Table of Contents

The Performance of Radial Basis Function Neural Network Model in Forecasting Foreign Tourist Flows To Yogyakarta

Dhoriva Urwatul Wutsqa[1], and Syalsabila Mei Yasfi[2]

[1]Department of Mathematics and Science, Yogyakarta State University
[2]Department of Mathematics and Science, Yogyakarta State University

dhoriva@yahoo.com and syalsabilay@gmail.com

Abstract: The purpose of this research is to compare between the use of global ridge-regression and local ridge-regression methodsfor radial basis function neural network (RBFNN) modelingof foreign tourist flows to Yogyakarta. Many activation functions are studied here, those are gaussian, cauchy, and multiquadratic functions. The K-means clustering method is used to obtain centeroids and standart deviations of activation functions of RBFNN. The weight estimation is proceed only from hidden layer to output layer by using global ridge-regression and local ridge-regression method. The result gives evidence in favor of the RBFNN model with gaussian activation function whose weight estimation is the global ridge-regression method.
Keywords: RBFNN; Global ridge-regression; Local ridge-regression; K-Means Clustering.

1. Introduction

Yogyakarta becomes a magnet for foreign tourists. Frequently, Yogyakarta becomes the second destination after Bali for foreign tourists. Yogyakarta is an attractive place because of the two famous temples Borobudur and Prambanan, and the uniqueness of Javanese culture. Additionally, Yogyakarta is very easy to be accessed by land and air transportations.The foreign tourism to Yogyakarta tends to grow up and it can be observed from the international arrivals to airport in Yogyakarta Adisucipto. The forecast of those numbers is important information for the government and the tourism business to have effective planing to provide satisfied services and facilities. This problem deals with time series forecasting.

Various methods have implemented for time series forecasting. Recently, the soft computing method such as neural network has extensively utilized in time series forecasting. One attractive neural network model is radial basis function neural network (RBFNN), since the algorithm of the RBFNN model is very reliable for solving forecasting problem, as well it is simple and fast [9]. The architecture of RBFNN consists of one input layer, one hidden layer, and one output layer. The characteristics of the RBFNN model is each hidden node represents one nonlinear activation function known as basis function, while the output is linear function of the basis functions. The activation functions commonly used are Gaussian, Cauchy and multiquadratic. Theoretically, there is no explanation which function is more superior then others. So, it is still an interesting issue to investigate empirically which function is preferable.

The specific of the RBFNN model is the learning process or the parameter estimation, it involves two steps. The first step is to estimate the center and width parameterfor constructing basis function. We use the typical method K-Means clustering. The second step is to estimate the weights between hidden layer and output layer. Two familiar methods are global ridge-regression and local ridge-regression methods. The difference between that two methods is on the generated regulation parameter where the global ridge-regression method produces single parameter whereas the local ridge-regression method produces m parameters, where m is the number of hidden nodes.This research aims to compare the performance of RBFNN models for forecasting the number of foreign tourists to Yogyakarta that come through Adisucipto airport, regarding different activation functions and different weight estimations as explain above.

ISBN : 978-602-99849-2-7

2. Literature Review

The RBFNN model has been widely applied in time series forecasting. Many researchers have given a great attention to theoritical and emprical studies on this issue. Theoritical studies have been reported by [8] and [14], who combine RBFNN modelfor time series and statistical modeling. They propose forward selection to gain the optimal RBFNN model.They also provide empirical result to explain the propose procedure to inflation forecasting problem [8] and to exchange rate of US dollar against Rupiah forecasting problem [14]. Their works focus onthe global ridge-regression method for the weight estimation and the gaussian function for the activation function.

The empirical studies which support the advantage of the RBFNN model for time series forecasting have done by [5] and [6]. They apply the RBFNN model in dengue fever case at Yogyakarta and in stock index ofIndonesianSyariah, respectively.Simmulation study by [1] demonstrate the superiority of gaussian function to cauchy, multiquadratic and invers multiquadratic functions in predictive subset selection using regression tress and RBF Neural Network Hybridized with the genetic algorithm. Those researches also use global ridge-regressionas weight estimation method.The dominance of RBFNN model have also been reported by [3] and [12]. They are use the RBFNN with gaussian function for forcesting the performance of biofilter treating toluene [3] and the weather eficiency [12].

Based on that previous researches, we can perceive that their works focus on the global ridge-regression method for the weight estimation. Theoritically, the weight estimation can be employed by using local ridge-local regression method [10]. However, research on this area has not attracted the researcher, yet. The performance of local ridge-local regressioncompared with global ridge-regression methodstill becomes an interesting issue.Therefore, this research aims to compare those two methods, each with the activation function gaussian, cacuchy and multiquadratic. The data used here is the number of foreign tourist arrivals inYogyakarta through Adisucipto international airport in Yogyakarta.

3. Material & Methodology

3.1. Data

The data set used in this research is the total number of monthly arrival foreign touristto Yogyakarta through the entrance of Adisucipto Airport in 2010-2014. This data is drawn from Central Bureau of Statistics (BPS) Yogyakarta. Figure 1 depicts the time series plot of the arrival foreign tourist flows to Yogyakartaand figure 2 depicts its autocorrelation functions (ACF) plot.

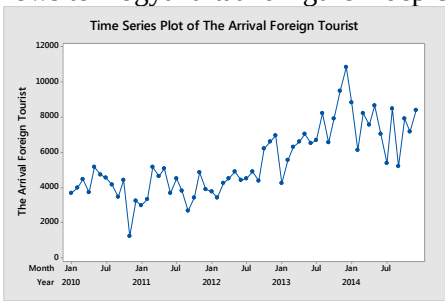


Figure 1. The time series plot of the arrival foreign tourist flows to Yogyakarta

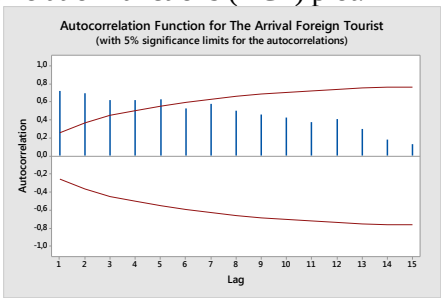


Figure 2. The ACF plot of the arrival foreign tourist flows to Yogyakarta

Figure 1 exibits the nonlinear and nonstationary time series pattern. The later pattern is strongly supported by Figure 2 that demonstrates sample autocorrelations remain significantly different from zero for five time periods. They are the autocorrelations at lag 1, 2, 3, 4, and 5.

3.2 Method

The procedure of modeling RBFNN generally consists of four steps, those areinputsidentification, data split, parameterestimation, and error diagnostic checking. The inputs of RBFNNmodel are determine by considering the autocorrelation of the variable in each lag. It can done by observing the ACF plot as describe in Figure 2. The inputs are the lags of variable whose autocorrelations are significantly different from zero, they are pass the red lines in the ACF plot.Thenext step isdividingdata into training and testing sets.The training set intends to generatemodel and testing setintend to validate the model, means how well the genarated model can be generalized.

The RBFNN modeling requires two steps parameter estimation, those are the parameters of activation functions in the hidden layer and the weights to produce the forecast on output layer. The activation functions in the hidden layer are nonlinear functions represented by basis functions. The output unit is a linear combination of the basis functions calculated in hidden layers. Then,RBFNN model can be expressed as

$$y = f(\omega, x) = \sum_{j=1}^{q} \omega_j h_j(x) + b \qquad (1)$$

whereω_j is the weightfrom the *j*th hidden neuron to the output layer, h_j is the basis function in the *j*th hidden neuron and b is bias. The following expressions are basis functions in the hidden layer addressed from [2] and [10]:

$$\text{Gaussian RBF (GS)}: \quad \boldsymbol{h}_j(x) = \exp\left(-\sum_{l=1}^{p} \frac{(x_l - c_{jl})^2}{r_{jl}^2}\right) \qquad (2)$$

$$\text{Cauchy RBF (CH)}: \boldsymbol{h}_j(x) = \frac{1}{1+\exp\left(-\sum_{l=1}^{p} \frac{(x_l - c_{jl})^2}{r_{jl}^2}\right)} \qquad (3)$$

$$\text{Multiquadratic RBG (MQ)} \boldsymbol{h}_j(x) = \sqrt{1 + \exp\left(-\sum_{l=1}^{p} \frac{(x_l - c_{jl})^2}{r_{jl}^2}\right)} \qquad (4)$$

wherec_{jl} and r_{jl}^2 are the *j*th center and the *j*th width parameter of *l*th variables, respectively.

In this research, the center and width parameters of basis functions have to be estimated are mean and standard deviation, respectively. They are estimated by using K-Means clustering method. The clustering process relied on euclidean distance[8]. The weight estimation can be performed by using global ridge-regression and local ridge-regression methods. Global ridge-regression is known as ridge-regression or weight decay with single parameter. It is employs by adding the single positive regulatory parameter λ to the Sum Square Error (SSE), so the function has to be minimized is written in the following equation

$$\text{C} = \sum_{n=1}^{m} \left(\boldsymbol{y_n} - \mathbf{f}(\boldsymbol{x_n})\right)^2 + \lambda \sum_{j=1}^{q} \boldsymbol{w}_j^2. \qquad (5)$$

The optimal weight vector yielded by global ridge-regression method is

$$\widehat{\boldsymbol{w}} = (\boldsymbol{H}^T \boldsymbol{H} + \lambda \boldsymbol{I_q})^{-1} \boldsymbol{H}^T y \qquad (6)$$

where H is the design matrix, with $H_{ij} = \boldsymbol{h}_j(x_i)$ and $\boldsymbol{A}$is the inverse of$\boldsymbol{H^T H}$matrix, $\boldsymbol{I_q}$is the q dimensional identity matrix, and $y = (y_1, y_2, \cdots, y_m)^T$ is the m-dimensional output vector.
Different from the global ridge-regression method, the local ridge-regression method add m positive regulatory parametersassociated with m basis functions to the Sum Square Error (SSE), so the function has to be minimized is

$$\text{C} = \sum_{n=1}^{m} \left(\boldsymbol{y_n} - \mathbf{f}(\boldsymbol{x_n})\right)^2 + \sum_{j=1}^{q} \boldsymbol{\lambda}_j \boldsymbol{w}_j^2 \qquad (7)$$

Theoptimal weight vectoris given by

$$\widehat{\boldsymbol{w}} = (\boldsymbol{H}^T \boldsymbol{H} + \Lambda)^{-1} \boldsymbol{H}^T y \qquad (8)$$

where $\Lambda = diag\{\boldsymbol{\lambda}_j\}_{j=1}^{q}$is a diagonal regularization parameter matrix.

The last step is checking the model adequacy. The model is adequate if it produce white noise errors. That is, errors are uncorrelated random shocks with zero mean and constant variace. They can be resolve by observingthe plot of ACF and partial autocorrelation function (PACF) of residuas. The white noise errors are displayed by their residuals ACF and PACF which are identically equal to zero (Figure 2).

4. Results and Discussion

4.1.Result

To obtain the Radial Basis Function Neural Network (RBFNN) model for forecasting foreign touristflows to Yogyakarta, the procedure explained above has been performed. That variable is represented by the total number of monthly arrival foreign touristto Yogyakarta through the entrance of Adisucipto Airport in 2010-2014. The inputsare determined by observing the ACF plotin Figure 2. Based on the ACF plot in Figure 2, the inputs are the variable at time lags 1, 2, 3, 4, and 5. Those 5 input variables are denoted as x_{t-1},

ISBN : 978-602-99849-2-7

x_{t-2}, x_{t-3} , $x_{t-4}$and x_{t-5}. To find the best model, we divide the data into four different proportions training and testing sets. They are 80 % and 20 %, 75 % and 25 %, 70 % and 30 %, and 60 % and 40 %.

Then, we process K-Means clusteringalgorithm for all the different data splitting. Once again, to find the best model, we examine the model with the number of cluster are 2 until 20 clusters. The number of hidden neuronsare the same as the number of cluster. This algorithm produce the centeroid and the standard deviation, the parameter estimator that forming basis function. Those values are used as inputs on the RBF design matrix forming.

We consider three basis functions, those are Gaussian, Cauchy, and Multiquadratic functions. We compare the weights estimation global ridge-regression and local ridge-regression methods. The best RBFNN model is assign by regarding the *Mean Absolute Percentage Error*(MAPE)values on training and testing data, which leads model with the leastMAPE value.

After doing all the steps, the best models resulted by Global and Local Ridge Regression Methods of each basis function are reported in Table 1 and Table 2, respectively.

Table 1. The performance of the RBFNN model by the Global Ridge-regression Method

The proportion of Training Data (%)	Gaussian			Cauchy			Multiquadratic		
	q	**MAPE Training (%)**	**MAPE Testing (%)**	***q***	**MAPE Training (%)**	**MAPE Testing (%)**	***q***	**MAPE Training (%)**	**MAPE Testing (%)**
80	**14**	23,15	24.82	13	18.58	16.89	2	20.91	23.24
75 *	**19**	**13.69**	**14.88**	20	14.28	16.56	7	18.19	18.10
70	**20**	13.89	16.87	20	12.81	22.76	9*	**17.85**	**18.15**
60	**20**	15.61	14.92	**20***	**15.01**	**15.43**	18	17.38	24.32

q :the number of cluster

* : the best model

Tabel 2. The performance of the RBFNN model by the Local Ridge-regression Method

The proportion of Training Data (%)	Gaussian			Cauchy			Multiquadratic		
	q	**MAPE Training (%)**	**MAPE Testing (%)**	***q***	**MAPE Training (%)**	**MAPE Testing (%)**	***q***	**MAPE Training (%)**	**MAPE Testing (%)**
80	**8**	25.38	25.13	13	18.43	17.00	**7**	19.92	25.08
75*	**18**	**14.80**	**15.13**	18*	**15.14**	**15.83**	**7***	**18.37**	**18.82**
70	**18**	15.13	16.82	13	15.74	17.22	6	18.57	18.84
60	**20**	15.10	15.54	**17**	17.37	16.67	9	19.41	19.12

q : the number of cluster

* : the best model

Table 1 and 2 shows that both global ridge-regressionand local ridge-regression methodslead to the superiority of the Gaussian function. However, the global ridge-regression method is better then local ridge-regression method.

5. Discussion

These results indicate that the accuracy of the RBFNN model is strongly influenced by the basis function, the estimation method, the cluster number, and the training data testing proportion. The Gaussian function is dominant compared to the other activation functions, both by the global ridge-regression and the local ridge-regression estimation methods. The global ridge-regression method has better performance than the local ridge-regression, that is, it produce less MAPE values for each basis functions.However, when viewed from the perspective of training testing data proportion there is no consistency that global ridge-regression method is always more accurate than the of local ridge regression method.

To obtain the best model, the number of cluster in the model with Gaussian and Cauchy function are far more than the model with Multiquadratic, which are about 18-19 and 7-9, respectively. So, even though the model with Mutiquadratic function has the lowest accuracy, but it is the most efficient. The ideal data setting (proportion of training testing data) is 75% -25%. This is shown by the highest accuracy in the local ridge-regression method on that data setting. While the accuracy of global ridge-regression method on that data setting is the highest only in the model with Gaussian function. However, in the models with Cauchy and Multiquadratic functions their accuracy on that data setting is slightly smaller than on the data testing with the highest accuracy. The results can be seen clearly in Figure 3.

ISBN : 978-602-99849-2-7

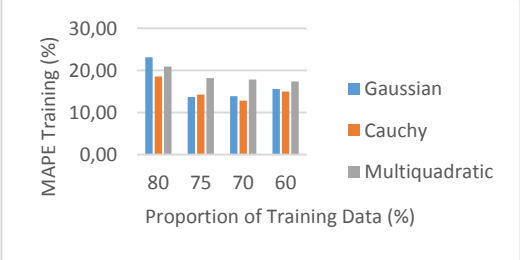


Figure 3(a).MAPE training data of the RBFNN model by the Global Ridge-regression Method

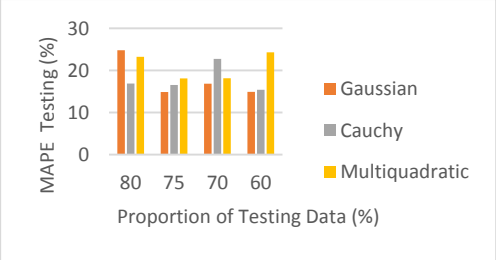


Figure 3(b).MAPE testing data of the RBFNN model by the Global Ridge-regression Method

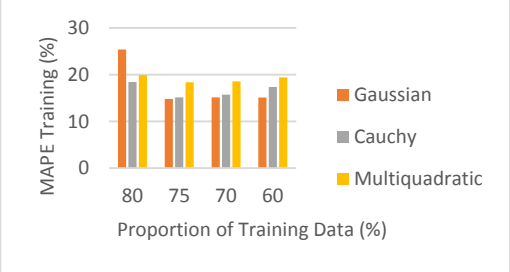


Figure 3(c).MAPE training data of the RBFNN model by the Local Ridge-regression Method

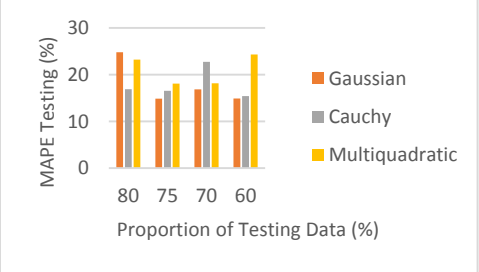


Figure 3(d).MAPE testing data of the RBFNN model by the Local Ridge-regression Method

In general, the preferred basis functions, weights estimation methods, and the proportion of training testing data for generating RBFNN model are Gaussian function, global - ridge regression method, and the proportion of training testing data 75 % -25 % . While the number of clusters ideal depends on those three other aspects.So, the best RBFNN model is model with gaussian activation function, 19 clusters, and proportion training testing data 75%-25%, which is estimated by global ridge-regression method. The MAPE values are 13,69% on the training data and 14,88% on the testing data. The architecture of the best RBFNN model is model with inputs x_{t-1}, x_{t-2}, x_{t-3} , $x_{t-4}$and x_{t-5}, 19 neurons and 1 bias on the hidden layer, single output neuron. The activation function is the Gaussian function in the hidden layer and linear function in the output layer. Figure 4 is presents the best architecture of RBFNN.

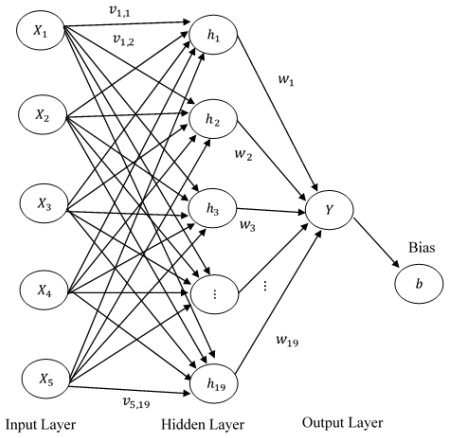


Figure 4. The Best Architecture of RBFNN Model

The architecture leads the following RBFNN model

$$y = f(\omega, x) = \sum_{j=1}^{19} \omega_j h_j(x) + b \quad (9)$$

The RBFNN model including the center and width parameters, and the weights is expressed as

$$y = f(\omega, x)$$
$$= 617 \times \exp\left(-\left\{\frac{(x_{t-1} - 5219)^2}{701{,}066^2} + \frac{(x_{t-2} - 4342{,}75)^2}{433{,}362^2} + \frac{(x_{t-3} - 4700)^2}{415{,}971^2} + \frac{(x_{t-4} - 3926{,}5)^2}{456{,}773^2} + \frac{(x_{t-5} - 4094{,}5)^2}{419{,}965^2}\right\}\right)$$
$$+ 2702 \times \exp\left(-\left\{\frac{(x_{t-1} - 5671{,}5)^2}{1308{,}85^2} + \frac{(x_{t-2} - 5687{,}5)^2}{731{,}86^2} + \frac{(x_{t-3} - 4059{,}5)^2}{461{,}74^2} + \frac{(x_{t-4} - 4688{,}5)^2}{313{,}25^2} + \frac{(x_{t-5} - 4263)^2}{363{,}45^2}\right\}\right) + \ldots$$
$$- 1056 \exp\left(-\left\{\frac{(x_{t-1}-4851)^2}{4851^2} + \frac{(x_{t-2}-3417)^2}{3417^2} + \frac{(x_{t-3}-2672)^2}{2672} + \frac{(x_{t-4}-3826)^2}{3826^2} + \frac{(x_{t-5}-4519)^2}{4519^2}\right\}\right) + 7418 \quad (10)$$

The white noise residual yielded by the model (9) is prven by the residual ACF plot in the Figure 5 and the residual PACF plot in Figure 6. They display no autocorrelation and partial autocorrelation that significantly different from zero. So, the RBFNN model (9) is adequate model.

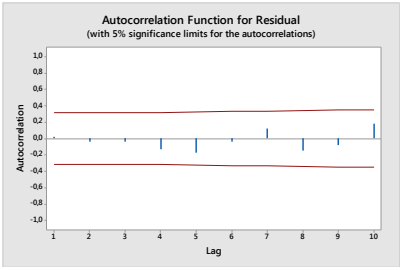


Figure 5. Residual of autocorrelation functions (ACF) plot

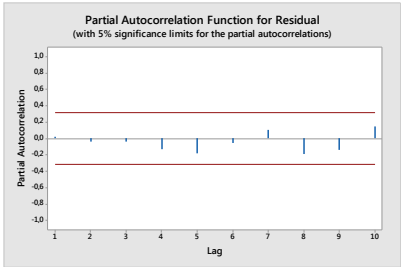


Figure 6. Residual of partial autocorrelation functions (PACF) plot

6. Conclusion

In this paper we utilize gaussian, cauchy, and multiquadratic as activation functions of RBFNN model. We also examine two weight estimation methods : the global ridge-regression and local ridge-regression methods. The result show that the gaussian function is outperform of other basis functions when applied to the data of foreign arrival tourist flows to Yogyakarat, for both estimation. In addition, the result shows that the global ridge-regression method provides more accurate forecasts than the local ridge-regression method for all the three basis functions considered.The research results give empirical evidences that the gaussian function and the global ridge-regression method is appropriate to forecast the foreign arrival tourist flows to Yogyakarat. Accordingly, the researcher can consider the gaussian function and the global ridge-regression method in RBFNN modeling for other problem. The proportion training-testing data 75%-25% is more preffered than the others. However, the result also shows that the forecast accuracy of the best model still unsatiesfied. Future research is addressed to improve the performance of RBFNN. We plan to achieve that by examining other clustering method, since it possibly has significant contribution to increase the accuracy of the model.

References

[1] Akbilgic, O.,& H. Bozdogan.,"Predictive Subset Selection using Regression Trees and RBF Neural Networks Hybridized with the Genetic Algorithm," European Journal Of Pure And Applied Mathematics 4 (4), 467-485, ISSN 1307-5543(2011).

[2] Bishop, C. M., *"Neural Networks for Pattern Recognition,"* London : Clarendon Press Oxford, (1995).

[3] Deshmukh, S. C, "Comparison of Radial Basis Function Neural Network and Response Surface Methodology for Predicting Performance of Biofilter Treating Toluene," National Environmental Engineering and Research Institute. India. Journal of Software Engineering and Applications5, 595-603, (2012).

[4] Fausett, L., *"Fundamentals of Neural Networks (Architectures,Algorithms, and Applications)," Upper* Saddle River, New Jersey: Prentice,(1994).

[5] Irman, A.,*"Forecasting of Indek Saham Syariah Indonesia (ISSI) use backpropagation and radial basis function neural network model,"* thesis, Department of Mathematicand Science, Yogayakarta State University, (2014).

[6] Johnson, R.A & Wichern, D.W, *Applied Multivariate Statistical Analysis*, Sixth Edition, Prentice Hall : New Jersey, (2007).

[7] Juliaristi, F., *"Forecasting the number of dengue fever in D. I. Yogyakarta used to Radial Basis FunctionNeural Ntework,"*thesis, Department of Mathematicand Science, Yogayakarta State University, (2014).

[8] Khasanah, U., *"Forward Select to determinate of Radial Basis Function Neural Network Model (RBFNN) on time series data,"* thesis, Department of Mathematic, Gajah Mada University, (2008).

[9] Kusumadewi, S.,*"Artificial Intelligence"*, First Edition. Yogyakarta : Graha Ilmu Press,(2003).

[10] Orr, Mark. J. L., *"Introduction to Radial Basis Function Neural Networks,"* Edinburgh: University of Edinburgh, (1996).

[11] Paul, A., et al.,"Eigen Value and It's Comparison with Gaussian RBF, Multi-Quadratic RBF and Inverse Multi-Quadratic RBF Methods," Information Sciences Letters An Foreign Journal 3 (2), 69-75(2014).

[12] Santhanam,T. & A.C. Subhajini, "An Efficient Weather Forecasting System using Radial Basis Function Neural Network," Journal of Computer Science 7 (7): 962-966, ISSN 1549-3636, (2011).

[13] Wei, W. W. S., *Time Series Analysis Univariate and Multivariate Method*.Second Edition. New York: Pearson Education, (2006).

[14] Zuliana, S. U. (2008). *"The Study on forming of RBFNN through a method of constructive learning to time series data in the financial field (case study: modeling number of exchange US dollar to rupiah),"* thesis, Department of Mathematic, Gajah Mada University, (2008).

Estimating The Break-Point of Segmented Simple Linear Regression using Empirical Likelihood Method

Muhammad Hasan Sidiq Kurniawan[1], Zulaela[2]
[1]Department of Statistics, Universitas Islam Indonesia
[2]Department of Statistics, Universitas Gadjah Mada

shiddiqkurniawan@yahoo.com

Abstract: Regression linear analysis is a statistical tool to study the linearity between two variables or more. Therefore one of the variables can be predicted using the other variables. If there is only two variables (predictor and response), then the analysis is called simple regression analysis. Segmented simple linear regression is developed from simple linear regression. Segmented simple linear regression is used if the predictor and response have more than one pattern relation. Also, it is used in a case which the predictor and response have fixed relation but when the predictor on the specific value we can see that the regression equation is different. If general regression linear analysis is used for that kind of data, the model we get will less representative. To overcome that situation we use segmented regression. In the segmented regression construction model, we have to estimate the break-point. This study is talking about break-pointestimation for a segmented simple linear regression using empiricallikelihoodconcept.
Keywords: Segmented Linear Regression; Break-point; Empirical Likelihood

1. Introduction

Regression analysis is one of the most popular statistical analysis methods and widely used in many fields. The aim of the analysis is to find the relation's pattern between the predictor and response variable so we can predict the value of response variable based on the predictor we have. The first step to see whether there is linearity between the predictor and response is constructing the scatter plot. In the reality, the predictor and response variable may have more than one pattern relation. Let $(X_1, X_2, \ldots, X_k, X_{k+1}, \ldots, X_n)$ be the predictor and be $(Y_1, Y_2, \ldots, Y_k, Y_{k+1}, \ldots, Y_n)$ the response. For $i = 1,2,\ldots,k$ the regression equation estimation is, for example $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. But, for $i = k+1,\ldots,n$ the estimated regression equation is $\hat{Y}_i = \hat{\beta}_0^* + \hat{\beta}_1^* X_i$. That example shows that at some points, the pattern relation between the predictor and response may be different. If the usual simple linear regression is used to analyze this kind of data, the estimated model will be less accurate and less precise. One of the methods is using the segmented regression analysis. To construct the segmented regression model, the first step is estimating the break-point. Break-point is threshold value. The estimated regression model before and after the break-point is different, like the example mentioned above. If the break-point estimation is wrong then the regression model will be less accurate. This study is talking about break-point estimation using the empirical likelihood concept so we can construct the segmented linear regression. The empirical likelihood is chosen because the data is not need to be assumed following some parametric distribution. We limited our study to estimate the break-point when there is only one predictor. In section 2, the segmented simple linear regression and the empirical likelihood are discussed. In section 3, the main topic of this study is discussed. That is estimating the break-point using the empirical likelihood concept. In section 4, we applied the method on the real data and construct the segmented simple linear regression. The comparison result between the usual simple linear regression and segmented simple linear regression also discussed in this section. Section 5 concludes the conclusion of this study.

2. Literature Review

2.1 The Segmented Simple Linear Regression

The difference between the usual simple regression model and the segmented simple linear regression model is in the model construction. On the segmented simple linear regression, the estimated model may be change due to the value of the predictor variable. It is highly possible in the construction of the segmented linear regression that more than one break-point are occurred. Let X and Y be the predictor and response variable, respectively. Let $\{(X_i, Y_i)\}_{i=1}^{n}$ be the n observations from the population (X, Y). Then we construct the simple linear regression model which divided into two parts at $X_i = \tau$.

$$Y_i = \begin{cases} \beta_0 + \beta_1 X_i + e_i; X_i \le \tau \\ \beta_0^* + \beta_1^* X_i + e_i; X_i > \tau \end{cases} \tag{1}$$

The expected value for Y_i is

$$E(Y_i) = \begin{cases} \beta_0 + \beta_1 X_i; X_i \le \tau \\ \beta_0^* + \beta_1^* X_i; X_i > \tau \end{cases} \tag{2}$$

From (2) we know that $X_i = \tau$ is a threshold value for the predictor variable X where we will have different regression model before and after the τ. In this study, we called τ as break-point. A regression model is called the linear segmented regression if it continuous on its break-point. But, it is rare in occasion that the model we construct automatically continuous on $X_i = \tau$. We can say that (2) is continuous on τ if the following condition is fulfilled:

$$\beta_0 + \beta_1 \tau = \beta_0^* + \beta_1^* \tau \tag{3}$$

We can see that $\beta_0^* = \beta_0 + (\beta_1 - \beta_1^*)\tau$. Therefore, the Eq. (2) become

$$E(Y_i) = \begin{cases} \beta_0 + \beta_1 X_i; X_i \le \tau \\ (\beta_0 + (\beta_1 - \beta_1^*)\tau) + \beta_1^* X_i; X_i > \tau \end{cases} \tag{4}$$

Sometimes more than one break-point occurred in the model construction, for example N break-point. Then we need to divide a simple linear regression model into $N+1$ regression models based on its break-point. Therefore we have the expected value of Y_i is

$$E(Y_i) = \begin{cases} \beta_0 + \beta_1 X_i; X_i \le \tau_1 \\ \beta_0^* + \beta_1^* X_i; \tau_1 < X_i \le \tau_2 \\ \vdots \\ \beta_0^{**} + \beta_1^{**} X_i; \tau_{N-1} < X_i \le \tau_N \\ \beta_0^{***} + \beta_1^{***} X_i; X_i > \tau_N \end{cases} \tag{5}$$

In order to make (5) continuous on $\tau_1, \tau_2, \ldots, \tau_n$ we can applied the method as in one break-point case. But usually we will get the less representative segmented regression model. Therefore we have to use another method. The following algorithm can be used to make the regression model continuous on its break-point:

Step 1. Construct a scatter plot between the predictor and response variable.

Step 2. Draw the regression line on the scatter plot based on the regression equation that not continuous on its break-points.

Step 3. Set the value of $E(Y_i)$ when $X_i = \tau_1, \tau_2, \ldots, \tau_n$, then adjust the regression model in each segment so that it continuous on the break-points.

2.2 The Empirical Likelihood

The construction of empirical likelihood functions basically same as the construction of likelihood function. We don't need the data to follow some parametric distributions in order to construct the empirical likelihood function. Let $X = (X_1, X_2, \ldots, X_n)$ i.i.d. with some probability function p. If $(x_1, x_2, \ldots, x_n)$ is a random sample from X then the empirical likelihood function is

$$L = \prod_{i=1}^{n} p_i \tag{6}$$

where $p_i = P(X = x_i)$. We can see that $p_i \geq 0; \forall i$ and $\sum_{i=1}^{n} p_i = 1$ because p is probability function. The probability function p is unknown. Therefore it needs to be estimated. Let $l = \log(L)$. Using Lagrange method, we will find the value of p_i that maximize l. The Lagrange function is

$$G = l - \lambda\left(\sum_{i=1}^{n} p_i - 1\right)$$

$$\Leftrightarrow G = \sum_{i=1}^{n} \log(p_i) - \lambda\left(\sum_{i=1}^{n} p_i - 1\right) \tag{7}$$

l will optimum if the first derivative of (7) to p_i is equal to zero (0), for each i.

$$\frac{\partial G}{\partial p_i} = 0 \Leftrightarrow p_i = \frac{1}{\lambda}; i = 1,2,\ldots,n \tag{8}$$

The equation above have the Lagrange multiplier constant, λ, which the value is unknown. The value of λ that maximize l can be estimated if the first derivative of (7) to λ is equal to zero (0).

$$\frac{\partial G}{\partial \lambda} = 0 \Leftrightarrow 1 - \sum_{i=1}^{n} p_i = 0 \tag{9}$$

Substitute the Eq. (8) into (9) we will get $\hat{\lambda} = n$. Therefore

$$\hat{p}_i = \frac{1}{n} \tag{10}$$

ISBN : 978-602-99849-2-7

Because $\frac{\partial^2 G}{\partial p_i^2} = -\frac{1}{p_i^2} < 0$ we can say that $\hat{p}_i$ is maximizing the function l. Automatically it is also maximizing the empirical likelihood function, L. The Eq. (10) is called empirical probability function that can be estimated from the sample $(x_1, x_2, \ldots, x_n)$.

2.3 The Empirical Likelihood Ratio

As in the original likelihood, the empirical likelihood also had ratio. The empirical likelihood ratio (ELR) is defined by

$$R(F) = \frac{L}{L(F_n)} \tag{11}$$

where $L(F_n)$ is an empirical likelihood function which the value of p is estimated by $\hat{p}$. Therefore we have

$$L(F_n) = \prod_{i=1}^{n} \hat{p}_i = \left(\frac{1}{n}\right)^n \tag{12}$$

Substitute (6) and (12) into (11) we will get

$$R(F) = \frac{\prod_{i=1}^{n} p_i}{(1/n)^n} = \prod_{i=1}^{n} np_i \tag{13}$$

If a random variable X have an expected value $E(X) = \mu$, then the ELR is defined as

$$R(F) = \sup\left\{ \prod_{i=1}^{n} np_i \,\middle|\, p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i x_i = \mu \right\} \tag{14}$$

The value of p_i can be estimated so we will get the estimated value of $R(F)$. If the value of $R(F)$ is small enough so that the value is less than the critical value of the distribution of X, then the ratio that X have an expected value $E(X) = \mu$ is small.

3. Estimating the Break-point of Segmented Simple Linear Regression

3.1 Estimating one Break-point

From a set of data $\{(X_i, Y_i)\}_{i=1}^{n}$, the simple regression model which divided into two parts on X_k is written as

$$Y_i = \begin{cases} \beta_0 + \beta_1 X_i + e_i; i = 1,2,\ldots,k \\ \beta_0^* + \beta_1^* X_i + e_i; i = k+1,\ldots,n \end{cases} \tag{15}$$

Let $\beta = (\beta_0, \beta_1)^T$ and $\beta^* = (\beta_0^*, \beta_1^*)^T$. To detect if there is a break-point based on the data, we construct the null hypothesis $H_0 : \beta = \beta^*$. Let $\{X_i\}_{i=1}^{n}$ is the value of predictor variable which sorted in ascending order. For a fixed k that fulfill $1 < k < n$, the data-set is divided into two cluster, those are $\{(X_i, Y_i)\}_{i=1}^{k}$ and $\{(X_i, Y_i)\}_{i=k+1}^{n}$. For each cluster, the regression parameter is estimated using the

ISBN : 978-602-99849-2-7

ordinary least square method or maximum likelihood. Let $\hat{\beta}(k)=\left(\hat{\beta}_0(k),\hat{\beta}_1(k)\right)^T$ and $\hat{\beta}^*(k)=\left(\hat{\beta}_0^*(k),\hat{\beta}_1^*(k)\right)^T$ are the estimator for β and β^*, respectively. Then the residuals are

$$\hat{e}_i(k)=\begin{cases} Y_i-\left[\hat{\beta}_0(k)+\hat{\beta}_1(k)X_i\right], i=1,\ldots,k \\ Y_i-\left[\hat{\beta}_0^*(k)+\hat{\beta}_1^*(k)X_i\right], i=k+1,\ldots,n \end{cases} \tag{16}$$

Under the null hypothesis is true, then the value of $\hat{\beta}(k)$ and $\hat{\beta}^*(k)$ will be close enough. Therefore, the residuals can be written as

$$\tilde{e}_i(k)=\begin{cases} Y_i-\left[\hat{\beta}_0^*(k)+\hat{\beta}_1^*(k)X_i\right], i=1,\ldots,k \\ Y_i-\left[\hat{\beta}_0(k)+\hat{\beta}_1(k)X_i\right], i=k+1,\ldots,n \end{cases} \tag{17}$$

If the null hypothesis is not rejected then $E\left(\tilde{e}_i(k)\right)=E\left(\hat{e}_i(k)\right)=0$. The null hypothesis will be rejected if the empirical likelihood ratio (ELR)

$$\Re(k)=\sup\left\{\prod_{i=1}^{n} np_i \middle| \sum_{i=1}^{n} p_i\,\tilde{e}_i(k)=0,\ p_i\geq 0,\ \sum_{i=1}^{n} p_i=1\right\} \tag{18}$$

is small enough. The logarithm value for ELR is

$$-2\log\Re(k)=-2\sup\left\{\sum_{i=1}^{n}\log(np_i) \middle| \sum_{i=1}^{n} p_i\,\tilde{e}_i(k)=0,\ p_i\geq 0,\ \sum_{i=1}^{n} p_i=1\right\} \tag{19}$$

The probability function p_i is unknown and will be estimated with Lagrange method. We construct the Lagrange function with two constraints.

$$G=\sum_{i=1}^{n}\log(np_i)-n\lambda\sum_{i=1}^{n} p_i\,\tilde{e}_i(k)-\gamma\left(\sum_{i=1}^{n} p_i-1\right) \tag{20}$$

λ and γ are the Lagrange multiplier. $\sum_{i=1}^{n}\log(np_i)$ will optimum if the first derivative of (20) to p_i is equal to zero for all i.

$$\frac{\partial G}{\partial p_i}=0\Leftrightarrow 1-n\lambda p_i\,\tilde{e}_i(k)-\gamma p_i=0, i=1,2,\ldots,n \tag{21}$$

Taking $\gamma=n$ we will get

$$p_i=\frac{1}{n\left(\lambda\,\tilde{e}_i(k)+1\right)} \tag{22}$$

To estimate p_i we have to estimate λ. Substituting Eq. (22) into (20) we will get

$$-2\log\Re(k,\lambda) \qquad =-2\left\{\sum_{i=1}^{n}\log\left[n.\frac{1}{n\left(\lambda\,\tilde{e}_i(k)+1\right)}\right]\right\}$$

ISBN : 978-602-99849-2-7

$$= -2\left\{\sum_{i=1}^{n}\log\left[\frac{1}{\lambda\tilde{e}_i(k)+1}\right]\right\}$$

$$= 2\left\{\sum_{i=1}^{n}\log\left[\lambda\tilde{e}_i(k)+1\right]\right\} \quad (23)$$

The Eq. (23) can be seen as the logarithm of empirical likelihood function with parameter λ. Therefore, we will derive (23) to λ. Let $\delta(k,\lambda)$ is the first derivative of (23) to λ, then

$$\delta(k,\lambda) = 2\left(\sum_{i=1}^{n}\frac{\tilde{e}_i(k)}{\left[\lambda\tilde{e}_i(k)+1\right]}\right) \quad (24)$$

The Eq. (23) will optimum if $\delta(k,\lambda)=0$. The first derivative of $\delta(k,\lambda)$to λ is negative.

$$\frac{\partial(k,\lambda)}{\partial\lambda} = -2\left(\sum_{i=1}^{n}\frac{\left(\tilde{e}_i(k)\right)^2}{\left[\lambda\tilde{e}_i(k)+1\right]^2}\right) < 0 \quad (25)$$

Therefore the estimated value of λ is called MELE (Maximum Empirical Likelihood Estimator). For a fixed k the estimator for $-2\log\Re(k,\lambda)$is

$$-2\log\hat{\Re}\left(k,\hat{\lambda}\right) = 2\left\{\sum_{i=1}^{n}\log\left[\hat{\lambda}\tilde{e}_i(k)+1\right]\right\} \quad (26)$$

where $\hat{\lambda}$ can be found if $\delta(k,\lambda)=0$.

Under the null hypothesis is true, there will be no break-point. Let $\hat{\tau}$ is the estimated break-point. The null hypothesis will be rejected if the value of $\hat{\Re}\left(k,\hat{\lambda}\right)$is small enough, or the value of $-2\log\hat{\Re}\left(k,\hat{\lambda}\right)$is big enough. To detect whether there is a break-point in the regression model or not, we compute $-2\log\hat{\Re}\left(k,\hat{\lambda}\right)$for various k and choose $-2\log\hat{\Re}\left(k,\hat{\lambda}\right)$that has the biggest value or maximum. Let $MA_n$is the maximum value of $-2\log\hat{\Re}\left(k,\hat{\lambda}\right)$for various k.

$$MA_n = \max_{(L\le k\le U)}\left\{-2\log\hat{\Re}\left(k,\hat{\lambda}\right)\right\} \quad (27)$$

where $L$is the lower bound for k and U is the upper bound for k. The value of L and U can be chosen at random, for example $L=(\log n)^2$ and $U=n-L$, where n is the number of observations. Liu and Qian (2009) defined the statistical test to detect a break-point is

$$GU_n = \sqrt{MA_n} \quad (28)$$

GU_n is following gumbel extreme value distribution. The null hypothesis is rejected at significance level α if GU_n is bigger than the critical value of gumbel extreme value distribution with location and scale parameter are 0 and 1, respectively. That is G_α.

ISBN : 978-602-99849-2-7

3.2 Estimating N Break-point

The algorithm to estimate one break-point can be applied if we want to detect if there are more than break-point. Let there is a break-point from n observations $\{(X_i, Y_i)\}_{i=1}^{n}$. For example, on X_{k^*}. Then the data-set is divided into two clusters, $\{(X_i, Y_i)\}_{i=1}^{k^*}$ and $\{(X_i, Y_i)\}_{i=k^*+1}^{n}$. Then the same algorithm which used on estimating one break-point is applied to see whether there is a break-point in each cluster. The algorithm to detect a break-point for cluster $\{(X_i, Y_i)\}_{i=1}^{k^*}$ is:

Step 1. Set the value of l so that $1 < L \le l \le U < k^*$. Then the data-set $\{(X_i, Y_i)\}_{i=1}^{k^*}$ is divided into two sub-clusters, $\{(X_i, Y_i)\}_{i=1}^{l}$ and $\{(X_i, Y_i)\}_{i=l+1}^{k^*}$. The value of L and U can be set at random.

Step 2. For each sub-cluster, estimate the regression parameter using the least square method. Let $\left(\hat{\alpha}_0(l), \hat{\alpha}_1(l)\right)$ is estimated regression parameter for sub-cluster $\{(X_i, Y_i)\}_{i=1}^{l}$ and $\left(\hat{\alpha}^*_0(l), \hat{\alpha}^*_1(l)\right)$ is estimated regression parameter for sub-cluster $\{(X_i, Y_i)\}_{i=l+1}^{k^*}$.

Step 3. Compute the value of $\tilde{e}^*_i(l) = Y_i - \left[\hat{\alpha}^*_0(l) + \hat{\alpha}^*_1(l) X_i\right]$ for $i = 1,2,\ldots,l$ and $\tilde{e}^*_i(l) = Y_i - \left[\hat{\alpha}_0(l) + \hat{\alpha}_1(l) X_i\right]$ for $i = l+1,\ldots,k^*$.

Step 4. The result from Step 3. is used to compute $-2\log\hat{\Re}\left(l, \hat{\lambda}\right)$.

Step 5. Repeat Step 1. to Step 4. for various l so that we have $\left\{-2\log\hat{\Re}\left(l, \hat{\lambda}\right)\right\}_{l=L}^{U}$. Compute the value of GU_n. If GU_n is bigger than G_α, then there is a break-point on cluster $\{(X_i, Y_i)\}_{i=1}^{k^*}$. The break-point is $X_i = X_{l^*}$. l^* is the value of l that maximizing $-2\log\hat{\Re}\left(l, \hat{\lambda}\right)$.

To detect a break-point on cluster $\{(X_i, Y_i)\}_{i=k^*+1}^{n}$ the identical algorithm can be applied.

4. Case Study

4.1 Data

We use the chlorine's content's data on chemical product. The data is reported by H. Smith and S.D. Dubey in "*Some reliability problems in the chemical industry*" (Draper and Smith, 1992). Chlorine is a chemical material which can kill the bacteria. It is usually used to clean the water on swimming pool. In a certain dosage, it is being mixed with drinking water so the water will long-lasting. The research was using a certain chemical product that contained 50% of chlorine when it was produced. As the time goes on, it is normal to assume that the chlorine's content will reduce. The predictor variable is time since the production (week) and the response variable is the chlorine's content. From 44 observations, we construct the scatter plot.

ISBN : 978-602-99849-2-7

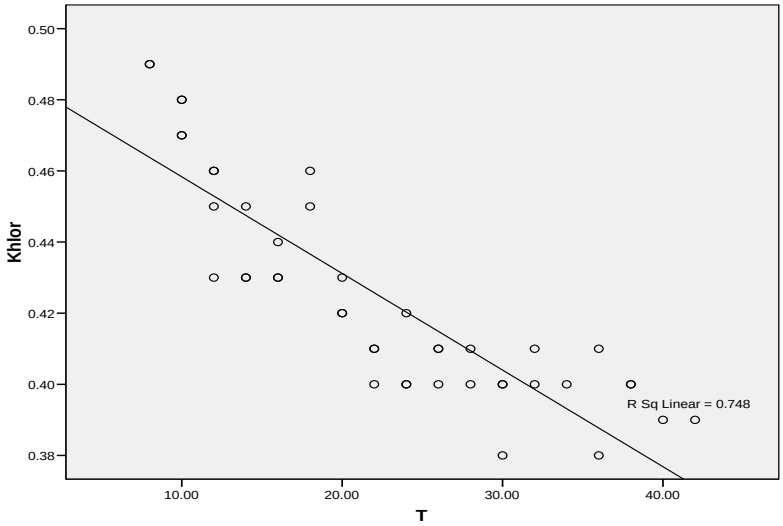


Figure 1. Scatter Plot Time vs Chlorine's Content

Using the least square method, the estimation for the regression model is

$$\hat{Chlorine} = 0.4855 - 0.00272(Time) \tag{29}$$

The coefficient of determination of the model is 0.748. So we can say that the percentage of the Chlorine variable variation that is explained by a linear model is 74.8%. We also get the Sum Squared of Error (SSE) of the model is 0.009941.

4.2 Segmented Simple Linear Regression with One Break-point

Now we want to know if there is a break-point on the model (29). We set $L = [\log n]^2 = [\log 44]^2 = 14.32 \approx 14$ and $U = n - L = 44 - 14 = 30$. The value of $-2\log \hat{\Re}\left(k, \hat{\lambda}\right)$ for $k = 14,\ldots,30$ are shown in the **Table 1.**

Table 1. The value of $-2\log \hat{\Re}\left(k, \hat{\lambda}\right)$ for $k = 14,\ldots,30$ on data-set $\{T_i, Chlorine_i\}_{i=1}^{44}$

k	$-2\log \hat{\Re}\left(k, \hat{\lambda}\right)$	k	$-2\log \hat{\Re}\left(k, \hat{\lambda}\right)$
14	174.5045	23	95.762
15	172.6441	24	119.4032
16	173.723	25	138.8887
17	202.829	26	127.7394
18	88.6485	27	111.7352
19	92.3683	28	122.7356
20	95.9104	29	106.5654
21	86.5573	30	132.7436
22	91.6802		

ISBN : 978-602-99849-2-7

Based on Table 1. We knew that $MA_n = 202.829$ and $GU_n = 16.8175$. Using library `evd` on R, we know that $G_{0.05} = 2{,}9702$. The null hypothesis is then rejected. So we can say that there is a break-point on the model. That is on $T_{17} = 18$ because the value of $-2\log\hat{\Re}\left(k, \hat{\lambda}\right)$ is maximum at $k = 17$. Now we will divide the regression line on its break-point and divide it into two parts. The result is

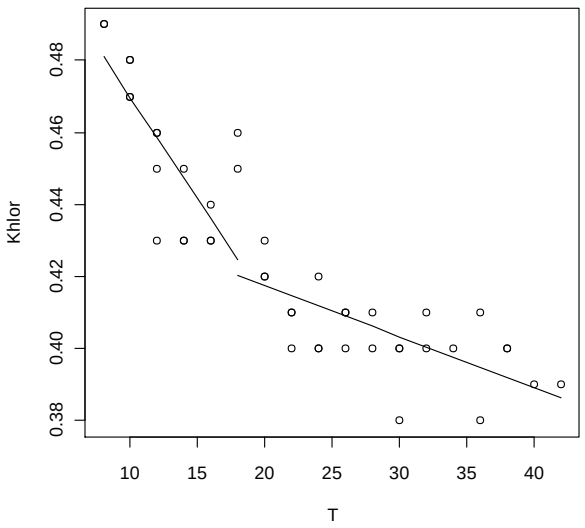


Figure 2. Regression Linear Plot which Divided into Two Parts

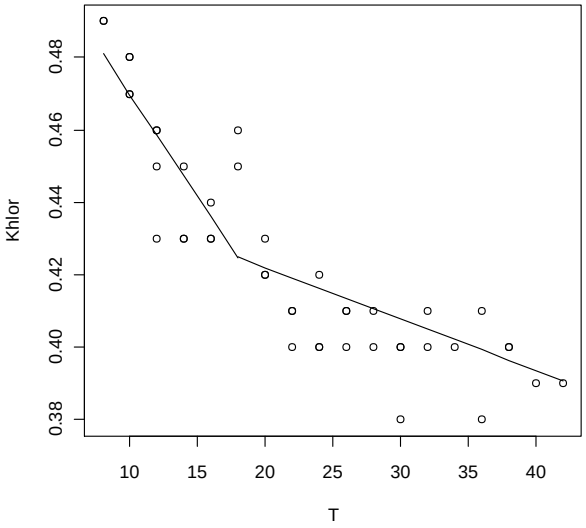


Figure 3. Segmented Simple Linear Regression with 1 Break-point

The estimation for the regression model is

$$\hat{Chlorine} = \begin{cases} 0.5259 - 0.005612(Time); Time \le 18 \\ 0.4459 - 0.001419(Time); Time \ge 18 \end{cases} \quad (30)$$

We could see that the regression line is not continuous on $T = 18$. Based on Eq. (4), we can make the regression equation continuous on its break-point.

$$\hat{Chlorine} = \begin{cases} 0.5259 - 0.005612(Time); Time \le 18 \\ 0.4504 - 0.001419(Time); Time \ge 18 \end{cases} \quad (31)$$

The regression line is shown on Figure 3.

The coefficient of determination of the new regression model is 0.8284. It is higher than the coefficient of determination on simple linear regression model. Also the SSE of the model is 0.006779. We can say that the segmented regression model with 1 break-point is better than the usual simple regression model.

4.3 Segmented Simple Linear Regression with More Than One Break-point

Based on the result from section 4.2., we can divide the data into two clusters $\{T_i, Chlorine_i\}_{i=1}^{17}$ and $\{T_i, Chlorine_i\}_{i=18}^{44}$. For each cluster we will detect if there is a break-point. We set $L = \left[\log n^*\right]^2$ and $U = n^* - L$, where n^* is number of observations for each cluster.

ISBN : 978-602-99849-2-7

For cluster $\{T_i, Chlorine_i\}_{i=1}^{17}$ we get $L=8$ and $U=9$. Then we get the value of $-2\log\hat{\Re}\left(k,\hat{\lambda}\right)$ for $k=8,9$ are 8.4283 and 15.264 respectively. We can see that $GU_n=\sqrt{15.264}=3.9527$ is bigger than $G_{0.05}=2.9702$. Therefore there is a break-point on this cluster, that is on $i=9$. Then we can divide the data into two sub-clusters $\{T_i, Chlorine_i\}_{i=1}^{9}$ and $\{T_i, Chlorine_i\}_{i=10}^{17}$.

For cluster $\{T_i, Chlorine_i\}_{i=18}^{44}$ we get $L=10$ and $U=17$. The value of $-2\log\hat{\Re}\left(k,\hat{\lambda}\right)$ for $k=10,\ldots,17$ are shown on **Table 2.**

Table 2. The value of $-2\log\hat{\Re}\left(k,\hat{\lambda}\right)$ for $k=10,\ldots,17$ on data-set $\{T_i, Chlorine_i\}_{i=18}^{44}$

k	$-2\log\hat{\Re}\left(k,\hat{\lambda}\right)$
10	63.3748
11	57.2521
12	46.8727
13	53.5123
14	33.7442
15	31.8474
16	26.0805
17	24.4227

The value of GU_n is 7.96. For $\alpha=0.05$ we rejected the null hypothesis so there is a break-point on this cluster. Then we divide the data into two sub-clusters, $\{T_i, Chlorine_i\}_{i=18}^{27}$ and $\{T_i, Chlorine_i\}_{i=28}^{44}$. Now we can make a scatter plot for the segmented regression model.

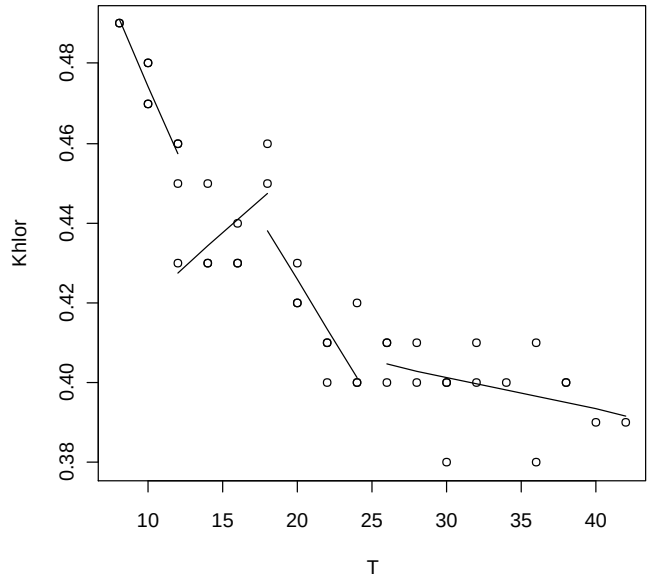


Figure 4. Regression Linear Plot for each Cluster Data

ISBN : 978-602-99849-2-7

The regression equation is

$$\hat{Chlorine} = \begin{cases} 0.5582 - 0.008409(Time); Time \le 12 \\ 0.3875 + 0.003333(Time); 12 \le Time \le 18 \\ 0.5488 - 0.006146(Time); 18 \le Time \le 24 \\ 0.4253 - 0.0007979(Time); Time \ge 26 \end{cases} \quad (32)$$

We could see that the regression line is not continuous on $Time = 12,18,24,26$. We have to modify the regression equation so that it continuous on its break-point. The modification result is

$$\hat{Chlorine} = \begin{cases} 0.5582 - 0.01089(Time); Time \le 12 \\ 0.3875 + 0.003333(Time); 12 \le Time \le 18 \\ 0.5488 - 0.005628(Time); 18 \le Time \le 26 \\ 0.4253 - 0.000878(Time); Time \ge 26 \end{cases} \quad (33)$$

Then the segmented regression plot is shown on **Figure 5.**

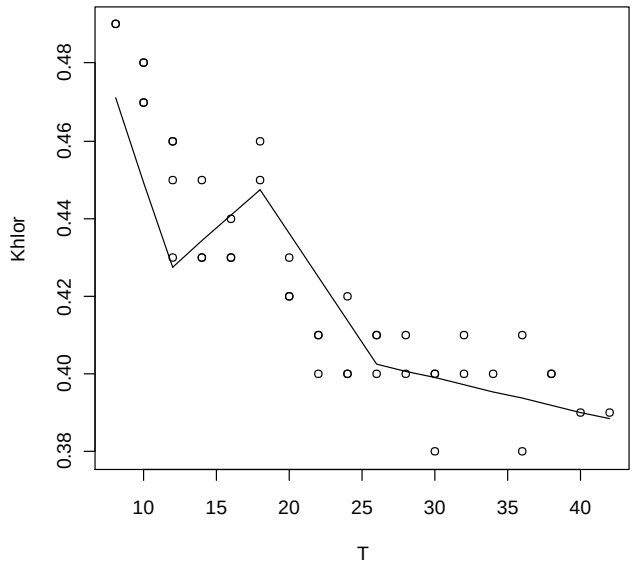


Figure 5. Segmented Simple Linear Regression with 3 Break-points

The coefficient of determination and SSE of the segmented simple linear regression with 3 Break-points are 0.7427 and 0.01016. The comparison of the three regression models is shown on **Table 3.**

Table 3. The comparison of the three regression models

Model	SSE	Coefficient of Determination
Simple Linear Regression	0.009941	0.7480
Segmented Regression (1 Break-point)	0.006779	0.8284
Segmented Regression (3 Break-points)	0.010160	0.7427

ISBN : 978-602-99849-2-7

We could see that the segmented regression model with one Break-point has the lowest SSE. Also, its coefficient of determination is the highest. Therefore we can say that the segmented regression model with one Break-point is the best model among the threes, for the Chlorine data. The best regression model is not determined by the number of break-points. Sometimes the simple linear regression model gives the best result. So, it is depend on the characteristic of the data.

4.4 Discussion

The number of estimated break-point may be more than one. But we can't directly estimate more than one break-point. First, we have to estimate whether there is a break-point on the data. Then we divided the data into two clusters. Then we can estimate the break-point for each cluster. It is better if we can detect more than one break-point at one time.

5. Conclusion

Estimating the break-point of segmented regression model can be done using empirical likelihood ratio. The number of estimated break-point may be more than one. But we can't directly estimate more than one break-point at one time. The best regression model is not determined by the number of break-points. It is depend on the characteristic of the data.

References

The list of references should only include works that are cited in the text and that have been published or accepted for publication. Personal communications and unpublished works should only be mentioned in the text. Do not use footnotes or endnotes as a substitute for a reference list.

[1] Liu, Z., and Qian, L., "Change-point Estimation via Empirical Likelihood for a Segmented Linear Regression." Department of Mathematical Science, Florida Atlantic University. (2009).

[2] Qin, J., and Lawless, J., "Empirical Likelihood and General Estimating Equations." University of Waterloo, Ontario, Canada. (1994).

[3] Rao, J.N.K., and Wu, C., "Empirical Likelihood Methods." University of Waterloo, Ontario, Canada. (2008).

[4] Chatterjee, S., and Price, B., "Regression Analysis by Example." John Wiley & Sons, inc. New York (1977).

[5] Draper, N., and Smith, H., "Analisis Regresi Terapan Edisi Kedua." PT Gramedia Pustaka Utama. Jakarta. (1992).

[6] Utami, D., "Regresi Linear Tersegmentasi." Skripsi. Department of Statistics, Universitas Gadjah Mada, Yogyakarta, 2011.

ISBN : 978-602-99849-2-7

Risk Factor of Formaldehyde Detection on Sales Location of Jambal Rotisalted Fish (*Arius Thalassinus*) in Yogyakarta

Roza Azizah Primatika[1]

[1] Education staff of veterinary public health, veterinary medicine faculty of Universitas Gadjah Mada
roza.azizah@gmail.com

Abstract: Fish is a source of animal protein needed by the human body. The excess of fish in the harvesting moment can cause economic loss for the fisherman if not directly sold out. Fish preservation is needed but it need long time and difficult to apply in the small scale fisherman, it cause the fisherman choose formaldehyde as illegal food preservation chemical to avoid spoilage process in the fish. The aim of the present study was to determine the formaldehyde content in the jambal roti salted fish using Schiff solution and risk factor on sales location of jambal roti salted fish. A total of 32 samples of jambal roti salted fish were taken from several traditional markets and supermarkets in Yogyakarta. All of samples were analyzed using Schiff solution to detect the content of formaldehyde in jambal roti salted fish. All of data were analyzed by descriptive and inferential. Descriptive analysis results showed that 64.7% (11/17) samples of jambal roti salted fish came from traditional markets were positive detected formaldehyde and 35.3% (6/17) samples were negative. While the samples came from supermarket shows that 60% (9/15) were positive detected formaldehyde and 40% (6/15) samples were negative. The Chi-Square test was performed to determine the risk factor for the sales location of jambal roti salted fish based on the descriptive analysis data. Chi-square test showed that no statistically significant association between the sales location and detection of formaldehyde in jambal roti salted fish (p-value>0.05). However, the traditional markets have risk 1,222 times to have positive detected formaldehyde than Supermarket (OR = 1.222).
Keywords : Jambal roti salted fish, Schiff solution, Descriptive analysis, Chi-square test

1. Introduction

Indonesia is anarchipelago country where have a lot of product comes from the sea. One of product is fish. Fish is a source of animal protein needed by human body. One of fish that fisherman take from the sea is manyung fish (*arius thalassinus*) and usually called jambal roti salted fish if fish did salted. The excess of fish in the harvesting moment can cause economic loss for the fisherman if not directly sold out. Selection of jambal roti salted fish as the object of research, due to the shape and texture of the fish that tend to be thicker than the salted fish generally. So it takes a long time in the preservation with the help of sunlight. But if in the rainy season, fisherman need a long time to preservation. It is indicated that fishermen use preservatives in order to save time, cost and effort. Preservation is a step for fisherman to do economic loss. Fish preservation is needed but it need long time and difficult to apply in the small scale fisherman, it cause the fisherman choose formaldehyde as illegal food preservation chemical to avoid spoilage process in the fish.According Fraizier and Westhoff (1981), the use of formalin in food is not permitted because of toxic effects, except for a small degree in wood smoke, although this compound is effective against fungi, bacteria and viruses [4].The purpose research are to determine the formaldehyde content in the jambal roti salted fish and risk factor on sales location of jambal roti salted fish.

2. Related Works/Literature Review

a. Related Works

Author	Title	Originaly Research
Riaz Uddin et al, 2011	Detection of formalin in fish samples collected from Dhaka City, Bangladesh	- Samples in this research are fish (do not salted fish) from Dhaka City, Bangladesh - Qualitative detection of formalin detection kit for fish developed by Bangladesh Council of Scientific and Industrial Research (BCSIR)
C. H. Castell and Barbara Smith, 1972	Measurement of formaldehyde in fish muscle using TCA extraction and the Nash reagent	- Samples in this research are fish muscle from atlantic cod (Gadus Morhua) - This research is calculating concentration formaldehyde in fish muscle using TCA extraction and Nash reagent
Noordiana N et al, 2011	Formaldehyde content and quality characteristic of selected fish and seafood from wet markets	- Samples in this research are fish and seafood. - The aim of this research is detection of formaldehyde in fish and seafood using nass's reagent and TCA

b. Jambal Roti Salted Fish

Jambal roti salted fish is a fish product comes from manyung fish (*Arius thalassinus*). The term is use because the character of jambal roti meat texture will be broken down after fried like toast with a special aroma [3]. Jambal roti salted fish is made through a fermentation process that changes the weight and characteristics [2].

c. Formaldehyde

Formalin is a commercial chemical solution that is commonly used as an antiseptic, germicide and preservative. Formaldehyde is a pure form (100%) is not available in the market because at normal temperature and pressure easily polymerized to solids formed [1]. Formalin can damage growth and cell division, causing structural damage to the body's tissues to trigger cancer [6].

d. Schiff solution

Schiff solution is a solution made from a mixture of 0.01% aqueous solution of fuchsin, sodium metabisulphite and 1 NHCl [4]. According to the Center for Testing and Quality Development of Fishery Cirebon Schiff solution is derived from the fuchsin solution, distilled water, sodium metabisulfite and concentrated HCl were mixed into one.

e. Chi – square test

Chi – square test is an analysis to determine association between the two factors that define the contingency table. In the chi – squared test of association in a contingency table with two columns (e.g. defining groups) and two rows (e.g. defining outcomes). A formula for calculating the test statistic, when the contingency table has only two rows and two columns, is [7].

ISBN : 978-602-99849-2-7

$$x^2 = \sum \frac{(O - E)^2}{E}$$

And a formula for calculating risk factor between groups was using odd ratio (OR) is [7]

$$OR = \frac{axd}{bxc}$$

3. Material & Methodology

3.1. Data

Sampling was performed by sampling method. In total, 32 samples of jambal roti salted fish were collected from tradisional markets (17 samples) and supermarkets (15 samples) in Yogyakarta. All of samples were tested in Veterinary Public Health Laboratory, Faculty of Veterinary Medicine, Universitas Gadjah Mada Yogyakarta.

3.2. Method

3.2.1. Schiff solution

The mixture of 0.2 grams Fuchsin and 200 ml of warm distilled water were stirred until dissolve. Sodium metabilsulfit piecemeal (2 grams) and HCl (2 ml) were slowly added and stirred until a homogenous solution was obtained. Afterwards, the solution was left for half toone hour until the color disappeared. When the color was not disappeared then the mixture was filtered with activated charcoal or added 1 ml of concentrated HCl. The solution should be stored in the refrigerator for durability.

3.2.2. Sample Detection

The average weight of each samples was 2 grams. The sample was crushed with a mortar and 2 ml of distilled water was added, then filtered to get the extract. One drop of Schiff solutionwas added in the salted fish extract and the color should be changed to purple. Two drops of concentrated HCl was added to salted fish extract.The positive samples was determined when the color remains purple, otherwise negative.

3.2.3. Data analysis

Statistical analysis was performed on the basis of the individual jambal roti salted fish as the unit. A jambal roti salted fish was considered test positive when the color of extract was change into purple color, otherwise negative. Categorical variables were analyzed as risk factors for association with the acquisition of positive formaldehyde detection using chi-square test. First, descriptive analysis was performed withpercentage of positive and negative formaldehyde detection. The associations between outcome and risk factors were estimated by calculation odds ratios with the 95% confidence interval.

Equation of chi-square test was showed below [7]:

$$x^2 = \sum \frac{(O - E)^2}{E}$$

Risk factor can be calculated with odd ratio (OR) that have equation [7]:

$$OR = \frac{axd}{bxc}$$

Statistical analysis data was conducted using SPSS 21.0 version with license from Universitas Gadjah Mada.

ISBN : 978-602-99849-2-7

4. Results and Discussion

4.1. Result

Figure 1 and 2 showed the results of formaldehyde detection on jambal roti salted fish from traditional markets and supermarkets was using Schiff solution.

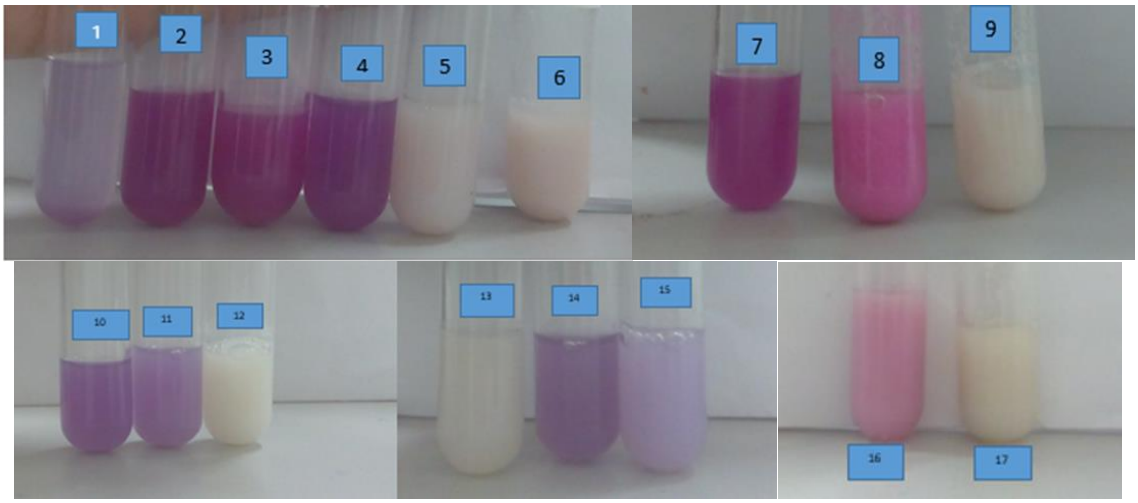


Figure 1. Results of 17 samples in formaldehyde detection test using Schiff solution come from 17 traditional markets in Yogyakarta

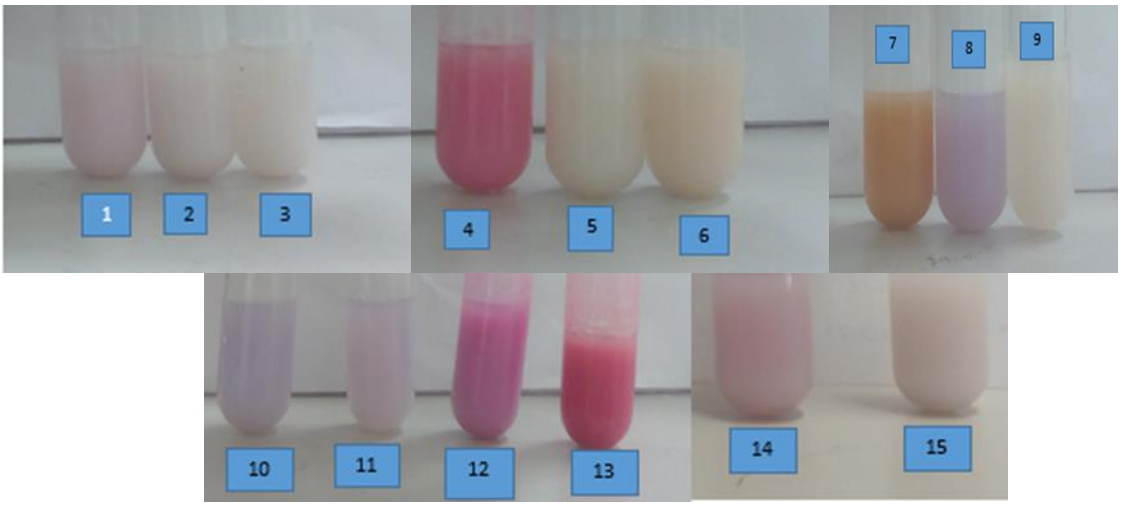


Figure 2. Results of 15 samples in formaldehyde detection test using Schiff solution come from 15 supermarkets in Yogyakarta

Table 1 showed the result of formaldehyde detection and it was determined based on the location of the sales (traditional markets and supermarkets) in Yogyakarta.

Table 1. Formaldehyde detection using Schiff solution

Sales location	+	-	**Total**
Traditional markets	11	6	17
Supermarkets	9	6	15
Total	20	12	32

Risk factor of sales location was analyzed based on descriptive result (Table 1) by chi – square test using SPSS 21.0 version. Table 2 showed the result of chi – square test.

ISBN : 978-602-99849-2-7

Table 2. Chi – square analysis

Sig. of Chi-square test	Odd Ratio (OR) of sales location (traditional markets/supermarkets)
0.784	1.222

4.2. Discussion

Descriptive analysis on figure 1, 2 and table 1 showed that 64.7% (11/17) samples of jambal roti salted fish came from traditional markets were positive detected formaldehyde and 35.3% (6/17) samples were negative. While, the samples came from supermarket shows that 60% (9/15) were positive detected formaldehyde and 40% (6/15) samples were negative.

Chi-Square test was performed to determine the risk factor for the sales location of jambal roti salted fish based on the descriptive analysis data (Table. 1). Chi-square test (Table. 2) showed that no statistically significant (p_value > 0.05) association between the sales location (traditional markets and supermarkets).Positive detected formaldehyde on jambal roti salted fish not depend on the sales location, because all of seller didn't know if the fisherman using formaldehyde in the preservation process of jambal roti salted fish.

However, the traditional markets have risk 1,222 times to have positive detected formaldehyde than Supermarket (OR = 1.222). It is because almost all of sellers in the traditional market haven't assessment system especially for formaldehyde content in the process of unloading product (jambal roti salted fish) from fisherman.

5. Conclusion

Percentage of positive detected formaldehyde in jambal roti salted fish in traditional markets was greater than supermarket, In traditional market showed that 64.7% positive detected formaldehyde and Supermarket showed that 60% positive detected formaldehyde. But no statistically significant association between the sales location and detection of formaldehyde in jambal roti salted fish (pvalue > 0.05). However, the traditional markets have risk 1,222 times to have positive detected formaldehyde than Supermarket.In order to control illegal preservation using formaldehyde in salted fish, training is the best way to give awareness to the fisherman and fish seller. In training programs, legal preservation method and impact of formaldehyde for human health should will be addressed under supervision of government in Yogyakarta.

Acknowledgement. Financial support for this research study was provided by Faculty of Veterinary Medicine, Universitas Gadjah Mada. The authors would like to thank the Department of Veterinary Public Health, Faculty of Veterinary Medicine, Universitas Gadjah Mada, Yogyakarta, Indonesia.

References

[1] Arifin, Z., *Stabilitas Formalin dalamDagingAyamselamaPenyimpanan*, in the national seminar of veterinary and livestock (2007).

[2] Burgess, G.H.O., C.L. Cutting, J.A. Lovern dan J.J. Waterman, *Fish Handling and Processing*. Her majesty's Stationary Office, Edinburg, 1965.

[3] Burhanudin, A.D., S. Martosewojo dan M. Hoetomo, *Sumber Daya Ikan Manyung di Indonesia*. LON-LIPI, Jakarta, 1987

[4] Fraizer, W.C., dan Westhoff, D.C., *Food Microbiology*.3rd Edition. Tata McGraw Hill Publishing Co., Ltd. New Delhi, 1981.

[5] Keusch, P, Schiff's Reagent, www.schiff'sreagent.com, 2013

[6] Rinto, E., Arafah, S.B. Utama, Kajian keamanan pangan (formalin, garam dan mikrobia) pada ikan sepat asin produksi Indralaya, Jurnal Pembangunan Manusia, 8 (2): 20-25 (2009).

[7] Watson, P, Avian P., *Statistics for Veterinary and Animal Science, 3rd edition*, Wiley-Blackwell, 2013.

ISBN : 978-602-99849-2-7

Cluster Analysis and Its Various Problems

Erfiani
Bogor Agricultural University

erfiani_ipb@yahoo.com

Abstract: Cluster analysis is one of exploration data techniques which is widely used and has extensive application. Due to its development, clustering analysis can be applied on many types of data. Classical Clustering Method assumes that data only posses either numeric or categoric scale. Nowadays, however, clustering method can be applied on data with mixed numeric and categoric scales. Innitially, clustering method was only used on cases with complete data. Along with its development, nevertheless, this method gives many solutions for incomplete or missing data, these days. Clustering method is even used for clustering data with time series. This paper will review many clustering method approaches for various data characteristics and case studies.
Keywords: Cluster analysis; Classic methode; alternative cluster analysis

1. Introduction

The number of cluster analysis techniques has increased over the last fifty years, and these techniques have been used in many areas of scientific fields. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). This method concern some ways to identify homogeneous groups of objects. Thus, a cluster is a set of entities that are alike, and entities from different clusters are not alike.

Cluster analysis is one of exploration data techniques which is widely used and has extensive apllication. Due to its development, clustering analysis can be applied on many types of data, but there are many problems in its applications . Firstly, mMost of clustering methods assume that the data only have numeric or categorical scale only and do not have any correlation between the variables. However, in the reality almost all of the data are mixed scales between numeric and categorical data, so that it is not effective when applied to the traditional clustering method In addition, innitially, clustering method was only used on cases with complete data. Along with its development, nevertheless, clustering method gives many solution for incomplete or missing data, these days. Thirdly, clustering method is even used for clustering data with time series, although there are problems in sensitivity to small changes, choices of distance, , expensive computation, and redundant information.

Each clustering method is applied on a specific distance matrix such as Euclidean distance, and Mahalanobis distance. The results of clustering depend on the choice of the distance measurement..

There have been many studies on general cluster analysis, however, only a few discuss cluster analysis with time series data. This paper, therefore, will discuss clustering time series data using several distance measurements.

The is one of data mining techniques developed for clustering object based on time series data. Clustering algorithm on time series data is done to change or alter the distance for static data with the appropiate distance for time series. Cluster analysis is one of multiple variable techniques that are widely used and has broad applications. The cluster analysis was also used in a clustering that includes the element of time which case the data in the form of time series. However, the concept of similarity between time series are not simple and can be viewed in various ways. On the other hand, the problem of clustering is to find similarities between the time series occurred in many fields such as economics, finance, medicine, bioinformatics, ecology, and engineering..This paper will discuss clustering with time series data, using approach of three distance and are applied in export-import data.

Liao (2005) introduces the basics of clustering time series data, the criteria to evaluate the results of the clustering, and the steps to determine similarities / dissimilarities between the two time series. Liao also shows increased interest in the clustering of time series data. Several illustrations of previous research, among others are the clustering of industrialized countries according to historical data of CO2 emissions (Alonso et al 2006)[1], a clustering of banks on the basis of share price weekly (Vilar et al, 2009)[10], and a clustering of the industrial production index (Vilar et al, 2010[11]).

One important thing in the clustering analysis is to determine the size of similarities or dissimilarities between two data objects. The similarities between objects can be viewed from the proximity of the distance between the objects (Mattjik and Sumertajaya 2011)[7]. Different kinds of distances were introduced to overcome the clustering of time series data including distance based on autocorrelation, distance based on the complexity, and distance of Dynamic Time Warping (DTW).

There are two commonly used method in the clustering ;namely, hierarchy and non-hierarchy. One method of non-hierarchy that is often used is the k-average algorithm which has ability to classify large amounts of data faster than does the hierarchy method. However, k-average method has a weakness caused by the initial determination of the center of the cluster. The results of the grouping formed may dependon the initiation of the initial value of the center of cluster given (Mattjik and Sumertajaya 2011)[7]. Therefore, the clustering method used is a merger between the methods of hierarchy and k-average algorithms.

2. Material & Methodology

a. Data

This paper used secondary data from Bank of Indonesia from http://data.go.id/dataset/nilai-ekspor-indonesia-berdasarkan-negara-tujuan. These data cover export value of Indonesia to 20 destination countries, observed from January 1999 to December 2013.

b. Methods

This study was aimed to cluster time series data using several distance measurements; namely, autoccorelation distance, complexity distance, and Dynamic Time Warping (DTW) distance for Indonesian export values by destination countries. The clustering method used the incorporation of hierarchical method and k-means algorithm. The threeapplications of distance measurements in this study were also performed using the bootstrap method to see consistency in the determining the best distance measurement. Cophenetic correlation coefficient was used to determine the best distance measurement from the hierarchical clustering method, and the criteria used to evaluate the clustering results from k-means algorithm is coefficient silhouette.

3. Literature Review

3.1. Distance Measurement

Autocorrelation Distance

Both parametric and nonparametric approachescan be done in time series clustering. Nonparametric clustering technique is less studied because of the difficulty in defining the size of the distance between the orders of the stationary time series. The distance that may be used to measure the similarity between points is the euclidean distance but this distance does not consider the correlation structure.

Autocorrelation distance is a distance based on the autocorrelation function estimation approach proposed by Galeano and Pena (2000)[5]. Autocorrelation distance is used to find a similar correlation structure in time series data. The first step in determining a direct measurement of the distance in this case is calculating the autocorrelation coefficient. This step is linked to a parametric approach of auto regression parameters. The distance between the two time series can be formed from vectors of autocorrelation by the following equation:

$$d_{ACF} = \{(\hat{\rho}_X - \hat{\rho}_Y)^{\mathrm{t}} \mathbf{\Omega} (\hat{\rho}_X - \hat{\rho}_Y)\}^{1/2}$$

with:

d_{ACF} : the distance between two time series
$\hat{\rho}_X$: estimated autocorrelation coefficients vector of time series X
$\hat{\rho}_Y$: estimated autocorrelation coefficients vector of time series Y
$\boldsymbol{\Omega}$: the weighting matrix
d_{ACF} equation can produce euclidean distance with the provision of a uniform weight on the autocorrelation function that makes the weighting matrix becomes an identity matrix (Caiado 2006)[3].

Complexity Distance

Errors in clustering time series data often occur when the nearest neighbor classification is a complex object assigned to the class that is much simpler. The problem is that most of the domains have a diverse complexity that subjectively may look very similar. Complexity in the time series is defined as the number of peaks and values that change with the time (Batista et al 2011)[2].
Complexity distance is the distance that is formed by using information about the differences between the two time series complexity as a correction factor of the distance that has been there. Euclidean distance is used as a starting point in determining this distance (Batista et al 2011)[2]. The first step taken in calculating the correction factor of complexity between the two time series is to estimate the complexity of each of the compared time series. Calculation of the correction factor of complexity between the two time series is defined by the following equation:

$$CF(Q,C) = \frac{\max\{CE(Q),CE(C)\}}{\min\{CE(Q),CE(C)\}}$$

with CE(T) is an estimation of the complexity of the time series T. The distance based on the complexity can be formulated by the following equation:

$$CID(Q,C) = ED(Q,C) \times CF(Q,C)$$

with ED(Q,C) is euclidian distance from time series Q and C.
Batista et al (2011) made an approach to estimate the complexity as an attempt to calculate difference in complexity at the compared time series. This approach was done by aligning the pattern of the time series to a straight line. The more complex the time series data, the longer the straight line generated. Complexity estimation can be calculated from the equation:

$$CE(Q) = \left\{\sum_{i=1}^{n-1}(q_i - q_{i-1})^2\right\}^{1/2}$$

withq_i is the time series value of i, i = 1, 2, ... , n.

Dynamic Time Warping (DTW) Distance

Dynamic Time Warping (DTW) is an algorithm to calculate the distance between the two time series by determining the optimal warping path. This algorithm works on the basis of dynamic programming techniques in an attempt to find an optimal warping path to test every possible path that is warping between the two time series (Niennattrakul and Ratanamahatana 2007)[8]. At the initial stage, local costs matrix will be formed with size of n x m in an effort to align the two sequences of X and Y representing all pairwise distances between them. Each element in the matrix of local costs is derived from the equation:

$$c_{i,,j=} \parallel x_i - y_j \parallel, i \in [1:n], j \in [1:m]$$

The algorithm will then look for the warping path which has a minimum cost with the order of the points $p = (p_1, p_2, \ldots, p_K)$, $p_l = (p_i, p_j) \in [1:n] \times [1:m]$ for $l \in [1:k]$ (Senin 2008)[9]. Each element (i,j) in the matrix is a cumulative cost of the points (i,j) and the minimum value of three adjacent elements (i,j) with $0 \leq i \leq n$ and $0 \leq j \leq m$, and it can be formulated with the following equation:

$$e_{i,j} = c_{i,j} + \min\{c_{(i-1)(j-1)}, c_{(i-1)j}, c_{i(j-1)}\}$$

That matrix is used to find the optimal warping path, that is the path that gives the smallest cumulative distance of all the possible paths warping (Niennattrakul and Ratanamahatana 2007)[8].
Conditions that must be met in DTW algorithm are as follows:
1. Boundary

ISBN : 978-602-99849-2-7

This condition that requires the starting point and the ending point of warping path is the starting point and the ending point of a series of data / sequence which $p_1 = (1,1)$ and $p_K = (m, n)$.

2. Monotonicity
   This condition describes the process that follows a sequence based on the time; that is,$n_1 \leq n_2 \leq \cdots \leq n_K$dan $m_1 \leq m_2 \leq \cdots \leq m_K$.
3. Continuity
   This condition requires the index i and j from warping path develops gradually with a maximum increase of 1 unit every step (Liao 2005)[6].

3.2. Cluster Analysis

Hierarchical Method

The basic principle of the hierarchical method is clustering the objects in a structure based on the similarities of the properties. The similarities of the properties can be determined from the proximity between objects. In the hierarchical method, the amount of desired cluster is unknown. In general, there are two ways of clustering using hierarchical method; that is, by merging and splitting. In the clustering method by merging, each object is ascribed from different clusters. The objects will then be merged gradually to obtain a cluster at the final clustering stage. Instead, splitting method of the hierarchical methods have the opposite process by the merging method (Everitt et all 2011)[4].

Mattjik and Sumertajaya (2011)[7] describe three kinds of algorithms to form a cluster with the hierarchical method; namely, the single linkage, the average linkage, and the complete linkage. Single linkage methodsare based on a minimum distance, while complete linkage methods are based on maximum distance. The distance between the clusters on the average linkage methods is determined from the average between the distance around the object of a cluster of and other objects in the cluster. This method is considered more stable than other hierarchical methods.

Non-Hierarchical Method

One of non-hierarchical method that is often used is k-means algorithm. The clustering results formed will depend on the determination of the number of cluster and the central election of cluster in the early stages of the clustering. In the k-means algorithm, the desired amount of cluster has been determined since the beginning. Each cluster is represented by the average value of the entire objects in the cluster. The objects are then clustered iteratively until there is no more transfer of objects between clusters (Mattjik and Sumertajaya 2011)[7].

Cophenetic Correlation Coefficient

Dendogram or the tree diagram is a visual representation of the complete clustering procedures from the analysis of hierarchical clustering. Dendogram has a point that shows the cluster and the length of the bar, and the distance between the objects that are combined into a cluster. Selection of the distance and the methods used in clustering affects the structure of dendogram formed. The resulting hierarchical structure needs to be identified for accuracy. The suitability of the structure of the data generated by the analysis of hierarchical cluster with observations proximity between objects can be determined by calculating the cophenetic correlation coefficient of the clustering (Everitt et al 2011)[4].

Cophenetic correlation matrix is the correlation between the distance of object matrix and cophenetic matrix which is a distance matrix as a result of merging all the objects into a single cluster (Everitt et al 2011)[4]. Matrices are formed into vectors that contain the corresponding elements of the upper triangular distance matrix. Cophenetic correlation coefficient is defined by the equation:

$$r_{XY} = \frac{n(\sum_{i=1}^{n} X_i Y_i) - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{\left\{\left(n(\sum_{i=1}^{n} X_i^2) - \left(\sum_{i=1}^{n} X_i\right)^2\right)\left(n(\sum_{i=1}^{n} Y_i^2) - \left(\sum_{i=1}^{n} Y_i\right)^2\right\}^{\frac{1}{2}}}$$

This correlation coefficient can be used to determine the best clustering method and the best clustering distance by comparing the accuracy measurement of the resulting cluster.

ISBN : 978-602-99849-2-7

4. Results and Discussion

The author aplied hierarchical clustering with three distance measurement; namely, Autocorrelation distance, Complexity distance and DTW distance for export value of Indonesia. Cophenetic correlation of three distance are shown on Table 1.

Table1Cophenetic correlation

Distance	Cophenetic correlation
Autocorrelation	0.7022418
Complexity	0.9207987
DTW	0.9895711

Table 1 shows that DTW distance is the largest value of Cophenetic correlation (0.9896). It means that DTW distance is the best distance among other distances. Using resampling bootstrap with 100 repeatation, all of repeatation of DTW distance have the largest Cophenetic correlation of all other distances. Figure 1 shows the cluster of 20 destination countries.There are two cluster, and countriesin cluster one are Russia, India, Argentina, and Mexiko. Members for cluster two are the USA, Vietnam, Italia, Japan, Malaysia, Hongkong, Germany, Singapure, Belgium, South Korea, Australia, Arab Emirate, China, Thailand, and Taiwan.

5. Conclusion

Clustering export destination countries of Indonesia produce two clusters. DTW distance is the best distance coefficent to cluster export value of Indonesia.

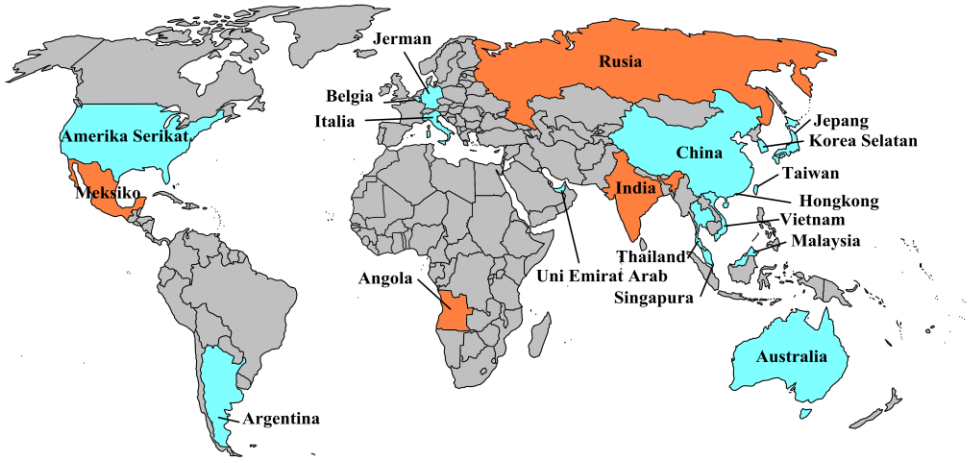


Figure 1. Mapping of Cluster of Indonesian Export Value

References

[1] Alonso AM, Berrendero JR, Hernandez A, Justel A. 2006. Time series clustering based on forecast densities. Computational Statistics & Data Analysis. 51(2006): 762-776. ISSN 0167-9473.

[2] Batista GE, Wang X, Keogh EJ. 2011. A complexity-invariant distance measure for time series. in: Liu B, Liu H, Clifton C, Washio T, Kamath C, editor. Proceedings of the 2011 SIAM International Conference on Data Mining; 2011 Apr 28-30; Arizona, USA. Arizona (VI): SIAM. hlm 699-710. doi: 10.1137/1.9781611972818.60.

[3] Caiado J, Catro N, Pena D. 2006. A periodogram-based metric for time series classification. Computational Statistics & Data Analysis. 50(2006): 2668-2684. doi: 10.1016/j.csda.2005.04.012.

ISBN : 978-602-99849-2-7

[4] Everitt BS, Sabine L, Leese M, Stahl D. 2011. Cluster Analysis, 5th ed. United Kingdom (GB): John Wiley and Sons, Inc.
[5] Galeano P, Pena D. 2000. Multivariate analysis in vector time series. Resenhas. 4(4): 383-403. doi: 10.1.1.486.9669.
[6] Liao TW. 2005. Clustering of time series data-survey. Pattern Recognition. 38(2005): 1857-1874. doi: 10.1016/j.patcog.2005.01.025.
[7] Mattjik AA, Sumertajaya IM. 2011. Sidik Peubah Ganda dengan Menggunakan SAS. Bogor (ID): IPB Press.
[8] Niennattrakul V, Ratanamahatana CA. 2007. On clustering multimedia time series data using k-means and dynamic time warping. International Conference Multimedia and Ubiquitous Engineering, 07; 2007 Apr 26-28; Seoul, Korea Selatan. Seoul: IEEE. hlm 733-738. doi: 10.1109/MUE.2007.165.
[9] Senin P. 2008. Dynamic Time Warping Algorithm Review. Honolulu (USA): University of Hawaii.
[10] Vilar JA, Alonso AM, danVilar JM. 2010. Nonlinear time series clustering based on nonparametric forecast densities. Computational Statistics & Data Analysis. 54(11):2850-2865. ISSN 0167-9473.
[11] Vilar JM, Vilar JA, danPertega S. 2009. Classifying time series data: A nonparametric approach. J Classification. 26(1): 3-28. ISSN 0176-4268.

ISBN : 978-602-99849-2-7

Volatility Modelling Using Hybrid Autoregressive Conditional Heteroskedasticity (ARCH) -Support Vector Regression (SVR)

Hasbi Yasin[1], Tarno[2], and Abdul Hoyyi[3]

[1,2,3] Department of Statistics, Faculty of Science and Mathematics, Diponegoro University
Jl. Prof. Soedharto SH, Tembalang, Semarang 50275

hasbiyasin@live.undip.ac.id, tarno@undip.ac.id, and ahy_stat@undip.ac.id

Abstract: High fluctuations in stock returns is one problem that is considered by the investors. Therefore we need a model that is able to predict accurately the volatility of stock returns. One model that can be used is a model Autoregressive Conditional Heteroskedasticity (ARCH). This model can serve as a model input in the Support Vector Regression (SVR) model, known as Hybrid ARCH-SVR. This modeling is one of the alternatives in modeling the volatility of stock returns. This method is able to show a good performance in modeling the volatility of stock returns. The purpose of this study was to determine the stock return volatility models using a Hybrid ARCH-SVR model on stock price data of PT. Indofood Sukses Makmur Tbk. The result shows that the determination of the input variables based on the ARIMA (3,0,3)-ARCH (5), so that the SVR model consists of 5 lags as input vector. Using a this model was obtained that the Mean Absolute Percentage Error (MAPE) of 1,98% and R^2 =99,99%.
Keywords:ARCH; ARIMA; SVR; Volatility

1. Introduction

The Volatility is statically a standard deviation of returning stock thatrepresents the share price returns [3],[12]. The higher the volatility, the higher the risk of profit or loss [5],[11]. The uncertainty value of the volatility in the financial markets leds to the need for a tool to foresee. Whilst the value at risk (VaR) is a concept that is used for measuring a risk in risk management. VaR can be simply defined as how much investors can lose their money during the investment period. In calculating VaR, the main problem to be solved is to determine a prediction of the volatility stock returns accurately which will be used as basis for calculating VaR.

According to Jorion [6], data stock returns have usually variances that are not constant at any point of time, called conditional heteroskedasticity. One of the financial time serie models that can accomodate heteroskedastisity is Autoagressive Conditional Heteroskedasticity (ARCH) which was introduced by Engle [4]. Whereas the more flexible model for modeling variance which is not constant is Generalized Autoregressive Conditional Heteroskedasticity (GARCH) proposed by Bollerslev [2]. GARCH structure consists oftwo equations, one is conditional mean equation which is ARCH standard model and the other is conditional variance equation that allows the variance changes anytime [13]. This model will be less optimal when used for prediction of stock return volatility. One of theforecasting method developed at this time is using Support Vector regressions (SVR). SVR is anon-linearapproachthat isbased onmachine learning. SVRisa modification of the Support Vector Machine (SVM) which is used forregression approach. The concept of SVR is maximizing hyperplane to collect data that can be support vector. One of the advantages is SVR able to overcomeoverfitting.

Therefore, this study will develop an alternative model that combines ARCH and SVR (Hybrid ARCH-SVR) for modeling the volatility shares of PT. Indofood Sukses Makmur Tbk, which later would beused to calculate Value at Risk (VaR).

2. Literature Review

2.1 Autoregressive Conditional Heteroskedasticity (ARCH)

Generally, ARCH models of order q is used to form the conditional variance models $\left(\sigma_t^2\right)$ at all time (t) based on the squared error at a time (t – 1) to (t – q). E.g.the average models are:

$$Z_t = \mu_t + e_t$$

According to Tsay [12] that μ_t is a expectation value Z_t conditional F_{t-1}, with $F_{t-1} = \left\{Z_{t-1}, Z_{t-2}, Z_{t-3}, \cdots, Z_2, Z_1\right\}$. So the models of ARMA(*r*,*m*) of Z_t are:

$$\begin{aligned}\mu_t &= E\left(Z_t \mid F_{t-1}\right)\\ &= \theta_0 + \sum_{i=1}^{r}\phi_i Z_{t-i} + \sum_{j=1}^{m}\theta_j e_{t-j}\end{aligned}$$

with:

X_t = return at a time -*t*

F_{t-1} = the entire set of information at a time -1 to- *t*-1

μ_t = expectation value X_t conditional F_{t-1}

e_t = residual ARMA at a time-*t*

Tsay[12] stated that ARCH model is a remnant e_t of the ARIMA model which is in the high order will be correlated, e_t could be describes as follows:

$$e_t = \varepsilon_t \sigma_t \qquad e_t \mid F_{t-1} \sim iidN\left(0, \sigma_t^2\right)$$
$$\varepsilon_t \sim iidN(0,1)$$

Acquired conditional variance for e_t :

$$\begin{aligned}\mathrm{Var}\left(e_t \mid F_{t-1}\right) &= E\left(e_t^2 \mid F_{t-1}\right)\\ &= E\left(\varepsilon_t^2 \sigma_t^2 \mid F_{t-1}\right)\\ &= \sigma_t^2 E\left(\varepsilon_t^2 \mid F_{t-1}\right)\\ &= \sigma_t^2\end{aligned}$$

so that the conditional variance that defines the order q ARCH models, is:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q}\alpha_i e_{t-i}^2$$

with q > 0, $\alpha_0 > 0$, and $\alpha_i \geq 0$ for *i* = 1, 2, 3, …, *q*.

2.2 Support Vector Regression (SVR)

Support Vector Regression (SVR) is a development of SVM for regression case. The goal of SVR is to find out a function $f(x)$ as a *hyperplane* in the form of regression functions which correspond to all the input data by an error ε and made ε as thin as possible[10]. Suppose there is l data training, $(\boldsymbol{x_i}, y_i)$, $i = 1, .., l$ in which $\boldsymbol{x}_i$ an input vector $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n\} \subseteq \boldsymbol{\Re}^n$ and scalar output $y = \{y_i, \ldots y_l\} \subseteq \Re$ and l is the number of training data. With SVR, will be determined a function $f(x)$ which has the biggest variation ε from the actual target y_i, for all the training data. if ε equal to 0 then obtained a perfect regression equation [9].

ISBN : 978-602-99849-2-7

The purpose of SVR ist o mapping input vector into the higher dimension [1]. For example a function below the regression line as the optimal hyperplane:

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \varphi(\boldsymbol{x}) + b$$

with :

$\boldsymbol{w}$ = dimensional weight vector*l*

$\varphi(\boldsymbol{x})$ = function that maps *x* to the space with *l* dimension

b = bias

2.3 Kernel Function

Many techniques of data mining or *machine learning* developed with the assumption of linearity, so that the resulting algorithm is limited to linear cases. With K*ernel Trick*, the data x in the *input space* mapped to the *feature space* with higher dimension through φ[9].

- Linear : $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}.\boldsymbol{y}$
- *Polynomial* : $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}.\boldsymbol{y} + \boldsymbol{c})^{\boldsymbol{d}}$
- *Radial Basis Function* (RBF) : $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{y}) = exp(-\gamma\|\boldsymbol{x} - \boldsymbol{y}\|^2)$, with $\gamma = \frac{1}{2\sigma^2}$
- *Tangent hyperbolic* (sigmoid): $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{y}) = tanh(\sigma(\boldsymbol{x}.\boldsymbol{y}) + c)$

$\boldsymbol{x}$ and $\boldsymbol{y}$ are two pairs of data from all parts ofthe training data. Parameter $\sigma, c, d > 0$, is constant. According toVapnikand Haykin, legitimate Kernel function provided by Mercer theory where these functions should be qualified continouos and positive definite [9].

2.4 Selection Parameters

According to Leidiyana [7], *cross-validation* is a standard test that is performed to predict error rate. Training data are randomly divided into several parts with the same ratio then the error rate is calculated section by section, and then calculate the overall average error rate to get the overall error rate. The rate of error can be calculated with the following formula:

$$CV = \sum_{i=1}^{n} \left(y_i - \hat{y}_{\neq i} \right)^2$$

with: $\hat{y}_{\neq i}$: fitting value y_i where the observation toi removed from the assessment process

y_i : actual value y on observation to i

in the *cross-validation*, known validation *leave-one-out* (LOO). In the LOO, data is divided into two subsets, one subset contains N-1 data for training and the rest oft he data for testing [9].

2.5 Hybrid ARCH-SVR

Hybrid ARCH-SVR is a combination model between SVR and ARCH, where ARCH models are used as an initial model for the determination of the input variables in the model SVR. Modeling a number of return data Y_t at the time $t_1, t_2, t_3, \cdots, t_n$ then used to estimate the value of the return at time t_{n+1}. One of the important things in ARCH-SVR model is determining the input variables. For example, to specify the input and the target of ARCH models (1). Suppose ARCH models (1) $\sigma_t^2 = \omega + \alpha_1 e_{t-1}^2$, then the used input is e_{t-1}^2 with the target σ_t^2. So that the model can be written $\sigma_t^2 = f\left(e_{t-1}^2\right)$.

2.6 Value atRisk(VaR)

Value at Risk (VaR) to return a single asset PT. Indofood Sukses Makmur, Tbk with a confidence level(1-α) and theholding period(hp), can be calculated using the formula:

$$VaR(1-\alpha, hp) = -Z_{1-\alpha} * S_0 * \sqrt{\sigma_t^2 * hp}$$

with:

ISBN : 978-602-99849-2-7

S_0 = initial investment

σ_t^2 = The volatility of stock returnsPT. Indofood Sukses Makmur, Tbk at the time t

3. Material & Methodology

1. Preparing daily stock return data PT Indofood Sukses Makmur Tbk.
2. Determining the independent variables based on the model of the best ARCH
3. Dividing the data into training data and testing the data to the percentage of a certain proportion.
4. Performing modeling stock returns using SVR method with kernel function, the values of kernel parameters and cost parameters and parameter optimization hyperplane epsilon for the training data.
5. Using the hyperplane with the best parameters obtained in the data testing.
6. Evaluating of regression models in testing using the coefficient of determination (R^2) and MAPE.

4. Results and Discussion

4.1 Result

In Modeling stock returns PT. Indofood Sukses Makmur, Tbk. Conducted by using GARCHmodels. Based on the resultsof data processing usingMATLABGUIprogram,it could be found that the identificationinitial modelis ARMA(3,3) ARCH(3). But toobtain thebestGARCH model, overfit process and underfit to parameter model used need to be done, and the resultsshown in Table1.

Table 1. Determination of the best ARCH model for return stocks of PT. Indofood Sukses Makmur, Tbk.

NO	MODEL	AIC
1	ARMA(3,3) ARCH(3)	-3229.8079
2	ARMA(2,2) ARCH(3)	-3220.8947
3	ARMA(3,3) ARCH(4)	-3231.8740
4	ARMA(2,2) ARCH(4)	-3221.8036
5	**ARMA(3,3) ARCH(5)**	**-3263.0156**
6	ARMA(2,2) ARCH(5)	-3261.0707
7	ARMA(1,1) ARCH(1)	-3171.7286

The best model for modeling stock returns PT. Indofood Sukses Makmur, Tbk is a model ARMA(3,3) ARCH(5) which mathematically can be written as follows:

$$Z_t = 2,4017\times10^{-4} + 0,17698Z_{t-1} - 0,17352Z_{t-2} + 0,18189Z_{t-3}$$
$$-0,23633e_{t-1} + 0,09151e_{t-2} - 0,36459e_{t-3} + e_t$$

with $e_t \sim N\left(0,\sigma_t^2\right)$ and

$$\sigma_t^2 = 9,33\times10^{-5} + 0,14318e_{t-1}^2 + 0,20259e_{t-2}^2 + 0,3671e_{t-3}^2 + 0,010765e_{t-4}^2 + 0,17057e_{t-5}^2$$

4.2 Determination of Kernel function and parameters for hyperplane

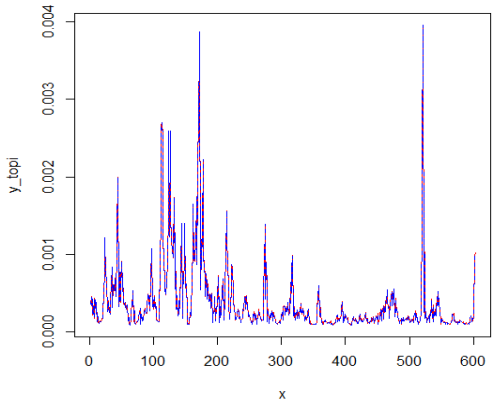


Figure 1. Plot of predicted and actual results

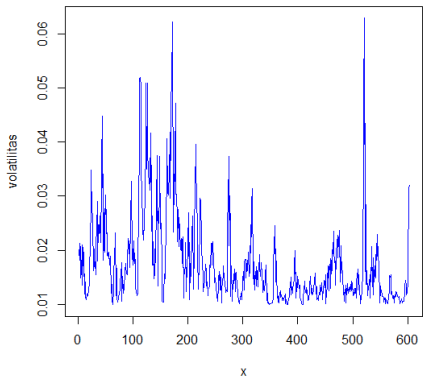


Figure 2. Plot of stock return volatility predicted results of PT. Indofood Sukses Makmur, Tbk

ISBN : 978-602-99849-2-7

The This study only used Kernel linear functions in linear at hyperplane SVR. The best parameters on the kernel function is determined by trying out some of the values in a specific range to build hyperplane. Optimized parameters is the value of C and the value of epsilon. The best parameters for the hyperplane determined by the smallest error value. From the selected parameters could be found that the best parameters for the hyperplane with linear kernel function is C = 10 and epsilon = 0.01. SVR modeling results with the parameter values obtained very high accuracy of the model, namely R^2=99.99% and MAPE = 1.98%.Visually, the results of prediction data can be seen in Figure 1. While the results of the predictive value of the stock return volatility can be seen in Figure 2. In those figures show that the data pattern has followed the same pattern so obtained SVR models used for prediction decent stock return volatility PT. Indofood Sukses Makmur Tbk.

4.3 Calculation of VaR using the best model

Value at Risk (VaR) to return a single asset of PT. Indofood Sukses Makmur Tbk with a confidence level (1-α) and the holding period (hp) can be calculated using the formula:

$$VaR(1-\alpha, hp) = -Z_{1-\alpha} * S_0 * \sqrt{\sigma_{INDF}^2 * hp}$$

with:

S_0 = the value of the initial investment

σ_{INDF}^2 = The volatility of stock returns PT. Indofood Sukses Makmur, Tbk.

VaR return value shares of PT. Indofood Sukses Makmur, Tbk with a 95% confidence level and 1 day holding period is $VaR(95\%, 1) = -1{,}645 * S_0\sqrt{\sigma_{INDF}^2}$. Volatility estimation results to the data in the sample shown in Figure 2.

5. Conclusion

Estimation of the model inputs used to predict the volatility of stock returns PT. Indofood Sukses Makmur, Tbk is ARIMA (3,0,3) -ARCH (5). So that the SVR model consists of 5 lags as input vector. This method is capable of performing well in modeling the volatility of stock returns with MAPE of 1.98% and R^2 = 99.99%.

References

[1] Abe, S. 2005. *Support Vector Machine for Pattern Classification*. Springer - Verlag. London Limited.
[2] Bollerslev, T., Generalized Autoregressive Conditional Heteroscedasticity, *Journal of Econometrics*, 1986, **31**: 307-327
[3] Engle, R.F., Autoregressive Conditional Heteroscedasticity with Estimates of Variance United Kingdom Inflation, *Econometrica*, 1982, Vol. 50, No.4: 987-1007.
[4] Engle, R.F. and S. Manganeli, *Value at Risk Models in Finance*, Working Paper Series No. 75 August 2001 European Central Bank. Germany.
[5] Holton, G., *Value at Risk, Theory and Practice*, Academic Press, Boston, 2003.
[6] Jorion, P., *Value at Risk:The New Benchmarking for Managing Financial Risk*. Mc Graw Hill, 2002.
[7] Leidiyana, H. 2013. Penerapan Algoritma K-Nearest Neighbor untuk Penentuan Risiko Kredit Kepemilikan Kendaraan Bermotor.*Jurnal Penelitian Ilmu Komputer. Sistem Embedded & Logic*. Vol. 1. No. 1: 65-76.
[8] Morgan J.P., *Risk Metrics – Technical Document*, J. P. Morgan Global Research Foueth Edition, Reuters, 1996.
[9] Santosa, B.2007.*Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis.*Yogyakarta:Graha Ilmu.

ISBN : 978-602-99849-2-7

[10] Scholkpof, B., and Smola A. 2003. *A Tutorial on Support Vector Regression*.
[11] Situngkir, H. dan Surya, Y., *Value at Risk yang Memperhatikan Sifat Statistika Distribusi Return*, Working Paper WDP 2006, Bandung Fe Institute.
[12] Tsay, *Analysis of Financial Time Series, second Edition*, Wiley Interscience, A John Wiley and Sons. Inc. Publication, USA, 2005.
[13] Wei, W.W.S., *Time Series Analysis: Univariate and Multivariate Methods*, Second Edition, Pearson Education Inc., Boston, 2006.

ISBN : 978-602-99849-2-7

Optimization of Fuzzy System Using Point Operation Intensity Adjustment for Diagnosing Breast Cancer

Kurrotul A'yun[1], Agus Maman Abadi[2]

[1]Mathematics Department of Mathematics and Science Faculty of Yogyakarta State University
[2]Mathematics Department of Mathematics and Science Faculty of Yogyakarta State University

kurrotulayun29@gmail.com, agusmaman@uny.ac.id

Abstract: Breast cancer was one of causing of woman death. Therefore, early detection and diagnosis are needed to determine the possibility of breast cancer. The aim of this research is knowing the different of diagnose accuracy using fuzzy system with point operations using 120 mammogram images that implemented by Graphical User Interface (GUI) and no using point operations. The point operation that is used was intensity adjustment point operation to increase quality of mammogram image. This research used fuzzy system with 10 feature extraction of the mammogram images as input variables. Fuzzy Mamdani Method is used in inference process and Centroid Method is used in defuzzification process. The accuracy of the fuzzy system with point operation reached 96.875% in the training data and 91.67% on the testing data. The accuracy of the fuzzy system without point operation only amounted to 94.79% on the training data and 50% on the testing data. So, the fuzzy system with point operation better than fuzzy system without point operation to diagnose breast cancer.
Keywords: breast cancer, fuzzy system, Graphical User Interface, mammogram image, point operation

1. Introduction

Breast cancer is one of the causing of woman death. A part of 30% Indonesian with cancer are breast cancer patient [1]. Therefore, early detection of breast cancer is important to do. Early detection of breast cancer can be found with two ways. There are self knowing and getting information from doctor. There are two ways, that are needed by the doctor to detection. There are mammograph and ultrasonograph (USG). The using of mammograph can produce mammograph image. This method is better than another method [2].

The researchers increase the accuracy of breast cancer diagnose incessantly. They utilize variety of methods and kind of data to develop their researches. This researches such as Schaefer, Zavisek, and Nakhasima [3] with termogram data, Al-Daoud [4] with fuzzy c-means radial basis function network method, Zadeh, et al [5] with FNN method, Keles and Keles [2] with NEFCLASS method and Mei Mutlimah [6] with Fuzzy Mamdani method. But, they classify to two outputs (benign and malignant) and with low accuracy.

The intensity adjustment of point operation is one of image correction method with linear mapping from the last histogram to the beginning histogram [7]. The intensity value of new image is better than last image. Fuzzy logic is a function that the domain is mapped to interval 0 and 1 [8]. It is mean, the value-truth of fuzzy logic are not absolute. It is different with strict logic. He assert the truth absolutely, 0 if it is false and 1 if it is true. Fuzzy logic can explain and tolerance apparent values. Therefore, fuzzy logic is appropriate for some fields included diagnose of breast cancer. Fuzzy logic applied on fuzzy system that is using the inference methods, such as Mamdani method. Mamdani method is the simple inference method because it has easy computation and comprehension [9].

The steps of building fuzzy system can be splved with Matlab program. Matlab is a software to make computation of mathematics analyse become easier, included fuzzy system. Then, the result of fuzzy system are pointed out with Graphical User Interface (GUI). GUI is one of feature on Matlab to make users easier to operate this system without know the script [10].

Base on this explanation, the author arranged this research to diagnose breast cancer with and without point operation intensity adjustment on mammogram image to see how it make the different on accuracy. Then we can build fuzzy system, determine the accuracy and show it on GUI.

2. The Modeling Process

This research used 120 mammograms from 322 mammograms of breast cancer images [11]. After the data have been extracted, the data are classified became 80% training data and 20% testing data. The extraction result became input and the output data is classified to 3 parts, there are normal, benign, and malignant. The steps of research are shown at figure 1.

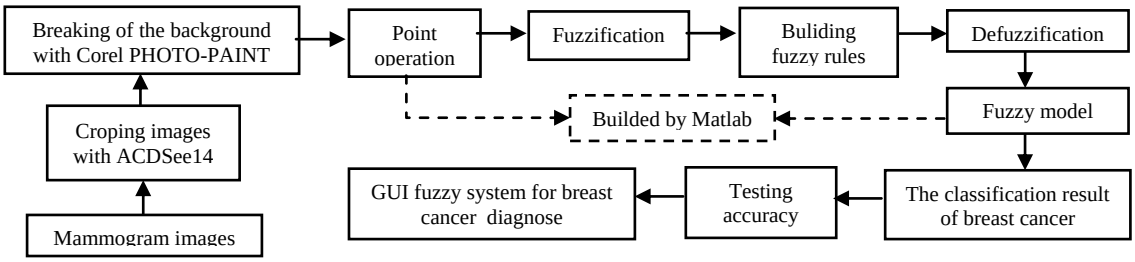


Figure 1. The steps of research

System testing is solved by determine the accuracy base on the true data and the false data. The result of fuzzy system that was build is shown at GUI. GUI can show pictures, graphs, and looked nice.

3. Results and Discussion

The first step to diagnose breast cancer is preprocessing mammogram image. There are cropping withACDSee14, breaking of background with Corel PHOTO-PAINT X7, and taking point operation intensity adjustment with Matlab R2010a. The example of the result of preprocessing image are shown at Figure 2.

Figure 2 show the changes of image on preprocessing step. The second step is extracting images. The images are extracted to 10 features using Matlab there are contrast, correlation, energy, homogeneity, mean, variance, standard deviation (SD), skewness, kurtosis, and entropy. The formula of each feature are:

$$\begin{aligned}
&Contrast\ [12] = \sum_i \sum_j (i-j)^2 p(i,j) && Correlation[13] = \sum_i \sum_j \frac{\{(ij)p(i,j)\} - \mu_x \mu_y}{\sigma_x \sigma_y} \\
&Energy[14] = \sum_i \sum_j p^2(i,j) && Homogeneity[12] = \sum_i \sum_j \frac{p(i,j)}{1+|i-j|} \\
&Mean\ (\mu)[15] = \sum_i \sum_j (i,j) p(i,j) && variance\ (\sigma^2)[16] = \sum_i \sum_j (i-\mu)^2 p(i,j) \\
&SD\ (\sigma)[16] = \sqrt{\sum_i \sum_j (i-\mu)^2 p(i,j)} && Skewness[17] = \frac{1}{\sigma^3} \sum_i \sum_j (i-\mu)^3 p(i,j) \\
&Kurtosis\ [18] = \frac{1}{\sigma^4} \sum_i \sum_j (i-\mu)^4 p(i,j) - 3 && Entropy\ [15] = -\sum_i \sum_j p(i,j) \log_2 p(i,j)
\end{aligned} \quad (1)$$

where $p(i,j)$ refer to pixel row-i column-j, μ_x is mean value of column on histogram, μ_y is mean value of row on histogram, σ_x is SD of column on histogram, and σ_y is SD of row on histogram. Table 1 show the result of image extractions `mdb004.png` with 10 features using Matlab.

The extraction result from Table 1 is used to build fuzzy system. The steps of building fuzzy system are given as follows:

Step 1. Indentifing the universal set of discourse U for input and output

The universal set is the possible value on operation of fuzzy system. The data have been covered by the universal set. There are interval base on minimum and maximum value on histogram from 96 training data. The universal set for each input variable are given as follows:

Contrast $(U_A) = [0.134\ 0.235]$, correlation$(U_B) = [0.955\ 0.989]$, energy$(U_C) = [0.123\ 0.639]$, homogeneity$(U_D) = [0.939\ 0.979]$, mean$(U_E) = [127.6\ 234]$, variance$(U_F) = [1973\ 7827]$, SD$(U_G) = [44.42\ 88.47]$, skewness$(U_H) = [-3.121\ 0.71]$, kurtosis $(U_I) = [1.36\ 13.13]$, and entropy$(U_J) = [2.995\ 7.394]$.

The universal setof discourse U for output variable is defined by $U_O = [1\ 3]$. One gives sign of normal with main value 1.5 and range diagnose in [1 1.7]. Two gives sign of benign with main value

ISBN : 978-602-99849-2-7

2 and range diagnose in (1.7 2.3]. Three gives sign of malignant with main value 2.5 and range diagnose in (2.3 3].

Step 2. Defining fuzzy set on input and output variables

Fuzzification is transform crisp set to fuzzy set using membership function. The membership function of input variables are using Gauss membership function. The formula of Gauss membership function [9] defined by:

$$G(x;k,\gamma) = e^{-\frac{(x-\gamma)^2}{2k^2}} \quad (2)$$

where k is width of curve and γ is domain value of curve center.

Each input variable defined by 9 fuzzy sets with Gauss membership function. Contrast variable is defined by 9 fuzzy sets, there are $A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8$ and A_9. The width on each fuzzy set on contrast variable is 0.005361 that acquired by observe domain value on cross point inter-fuzzy sets. The representation of Gauss curve on contrast variable base on equation (2) is given at Figure 3. The fuzzy sets in another input variables are analog. The domain value and width of curve are different.The output is defined by representation from combination of triangle curve and trapezoid curve. Figure 4 shown the representation of output curve.

Step 3. Building fuzzy rules

The extraction result is used to build fuzzy rule from training data. First, search the membership degree for each value from image extraction result and then used the highest membership degree to build fuzzy rule. The number of fuzzy rules are 96 rules accord with the number of training data. The steps to build fuzzy rules accord with image `mdb004.png` are given as follows.

The contrast value accord with Table 1 is 0.17724 called x. Base on 9 fuzzy sets on contrast variable, the value of x at the setsA_3, A_4, and A_5. Therefore, the membership degree in another sets are zero. Equation (2) used to determine membership degree.The highest value chosen that is used base union operation function [8] as shown in equation (3).

$$\mu_{A\cup B}(x) = \max[\mu_A(x), \mu_A(x)], \forall x \in U \quad (3)$$

The membership degree which acquired according to equation (3) is

$$\max(0,0,0.00306,0.60891,0.39973,0,0,0,0) = 0.60981$$

The value 0.608981 is membership degree of A_4, so the extraction of contrast from image `mdb004.png` included in A_4. Another features are analog and shown at Table 1. According to Table 1 acquired the rule " If contrast is$A_4$and correlation is$B_5$and energy isC_3andhomogeneityis$D_5$and mean is$E_7$and variance is$F_4$and SD is$G_4$and skewness$H_5$and kurtosis is$I_3$and entropy isJ_6then diagnose is normal". Another rules are analog.

Step 4. Inferenceing fuzzy Mamdani method

The Mamdani method or min-max inferenceing use min or AND implication function and use max or OR aggregation rule. The determination of fuzzy inference can be solved with Matlab. The manual computation can be solved to check the accuracy of system. The membership degree of image `mdb004.png` on Table 2 accord with rule 1, 2, and 21. Then we can determine the minimum value of this rules using equation (4) [8].

$$\mu_{A\cap B}(x) = \min[\mu_A(x), \mu_A(x)], \forall x \in U \quad (4)$$

Then, the result of determining rule 1, 2 and 21 using equation (4) in succession are $p = 0.5626$, $q = 0.50067$, and $r = 0.63795$. Figure 5 is shown the membership function of output for rule 1, 2, and 21.

The aggregation for this rules searchable with formula

$$\mu_{B^k}(y) = \max_k[\min[\mu_{A_1^k}(x_i), \mu_{A_2^k}(x_j)]] \quad (5)$$

ISBN : 978-602-99849-2-7

for $k = 1,2, \dots n, A_1^k$ and A_2^k explain fuzzy set antecedent-k pairs and B^k is fuzzy set of concecuent-k[19].

The agregation value according to equation (5) is s = max(0.5626, 0.50067, 0.63795) = 0.63795.The representation of fuzzy set base on this value is shown on Figure 5(d).

Then the next step is finding the cross point on Figure 5(d) using membership function of normal fuzzy set on output. For s = 0.6379 then

$$s = \frac{2-x}{2-1{,}5}$$
$$0.6379 = \frac{2-x}{0{,}5}$$
$$x = 2 - 0.31895 = 1.68105$$

Then the membership function for Figure 5(d) is

$$\mu(x) = \begin{cases} 0 & x \le 1 \; dan \; x \ge 2 \\ 0.6379 & 1 < x < 1.68105 \\ \frac{2-x}{0.5} & 1.68105 < x < 2 \end{cases} \quad (6)$$

Step 5. Defuzzification

The defuzzification process with Centroid method can be solved through Matlab. But then, we will be explain the analysis result. The fuzzification result is diagnosis of breast cancer that is classified by three parts. According to image extraction `mdb004.png`, membership function on equation (6) changed to crisp set using centroid method. The formula of fuzzification with centroid method [20] is given at this equation

$$D^* = \frac{\int_x x\,\mu_B(x)\,dx}{\int_x \mu_B(x)\,dx} \quad (7)$$

According to equation (6) and (7) acquired

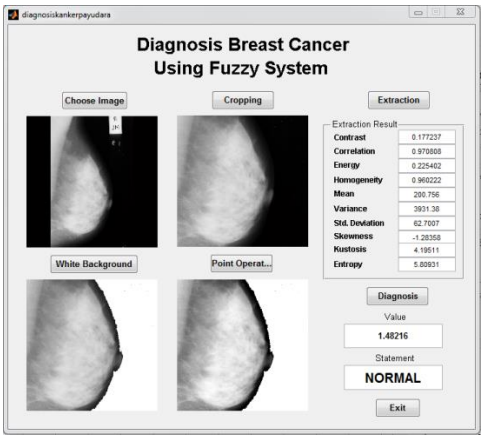
$$D^* = \frac{\int_1^{1.68105} x(0.6379)dx + \int_{1.68105}^{2} x\left(\frac{2-x}{0.5}\right)dx}{\int_1^{1.68105}(0.6379)dx + \int_{1.68105}^{2}\frac{2-x}{0.5}dx} = \frac{0.6379\left[\frac{x^2}{2}\right]_1^{1.68105} + \frac{1}{0.5}\left[x^2 - \frac{x^3}{3}\right]_{1.68105}^{2}}{0.6379[x]_1^{1.68105} + \frac{1}{0.5}\left[2x - \frac{x^2}{2}\right]_{1.68105}^{2}} = 1.42530$$

The value of $D^* = 1.42530$ include at [1 1.7). Thereby, the diagnosis of breast for image `mdb004.png` is normal. Another data are analog. Then, fuzzy system is consist of 96 images data and it is able to diagnose breast cancer for another mammogram images.

But, this fuzzy system is not good yet before trial. The examination was doing by determine accuracy and error value. The formula [21] to determine accuracy is

$$Accuracy = \frac{the\ number\ of\ correct\ data}{the\ number\ of\ all\ data}\ x100\% \quad (8)$$

Here is table of the analysis result with and without point operation.



ISBN : 978-602-99849-2-7

Figure 6. The GUI screen of fuzzy system of breast cancer diagnosis

From Table 2, the accuracy values of fuzzy system without point operation are 94.79% for training data and 50% for testing data. The accuracys of fuzzy system with point operation intensity adjustment are analog and the value are 96.875% for training data and 91.67% for testing data.

Then the last step are apply all of the system to GUI from preprocessing step until know the diagnosis of mammogram data. This aim of this step is helping the user easier to know the diagnosis of breast cancer from mammogram image. The way is using the feature of GUI guide in Matlab then enter the fuzzy system that has built to program GUI. The application of GUI is shown at Figure 6.

4. Conclusion

The result of the diagnosis of breast cancer using fuzzy system with point operation of intensity adjustment are better than the diagnosis of breast cancer using fuzzy system without point operation. It can be shown by the value of accuracy on training and testing data of system with point operation intensity adjustment is bigger than the accuracy value on training and testing data of system without point operation. However, this research did not consider the diagnosis of breast cancer for woman. But, this system can help doctor to analyze and take a decision about the patient's breast. In the future, the researchers can concentrate for another program to increase the image quality.

References

[1] Departemen Kesehatan Republik Indonesia, http://www.depkes.go.id Retrieved 27 January, 2015.

[2] Keles, A. and Keles, A.,"Extracting Fuzzy Rules for the Diagnosis of Breast Cancer,"*Turkish Journal of Electrical Engineering and Computer Sciences* 21, 1495-1503 (2013).

[3] Schaefer, G., Zavisek, M., dan Nakashima, T., "Thermography Based Breast Cancer Analysis Using Statistical Features and Fuzzy Classifications,"*Pattern Recognition* 42 (6), 1133 – 1137 (2009).

[4] Al-Daoud, E.,"Cancer Diagnosis Using Modified Fuzzy Network,"*Universal Journal of Computer Science and Engineering Technology* 1 (2), 73-78(2010).

[5] Zadeh, H.G., et al.,"Diagnosing Breast Cancer with the Aid of Fuzzy Logic Based on Data Mining of a Genetic Algorithm in Infrared Images," *Middle East Journal of Cancer 2011* 3 (4), 119-129(2011).

[6] Mutlimah, M.,"*Penerapan Sistem fuzzy Untuk Diagnosis Kanker Payudara (Breast Cancer)*," S.Si. thesis, Department of Mathematics, Yogyakarta State University, Yogyakarta, 2014.

[7] Munir, R.,"P*engolahan Citra Digital dengan pendekatan algoritmik*,"Informatika, 2004.

[8] Klir, G. J., Clair, U. S., and Yuan, B.,"*Fuzzy Set Theory Fondations and Applications*,"Prentice-Hall International, 1997.

[9] Kusumadewi, S.,"*Analisis dan Desain Sistem Fuzzy Menggunakan Toolbox Matlab*,"Graha Ilmu, 2002.

[10] Mathematics Laboratory, http://www.mathworks.com/discovery/matlab-GUI.html. Retrieved 09 March, 2015.

[11] The Pilot European Image Processing Archive, http://peipa.essex.ac.uk/pix/mias/ Retrieved 20 January, 2015.

[12] Sharma, M. and Mukherjee, S.,"Artificial Nueral Network Fuzzy Inference System (ANFIS) for Brain Tumor Detection,"*Advances in Intelligent System and Computing* 177, 329-339 (2013).

[13] Soh, L. and Tsatsoulis, C.,"Texture Analysis of SAR Sea Ice Imagery Using Gray Level Co-Occurence Matrices,"*IEEE Transactions on Geoscience and Remote Sensing* 37 (2), 780-795 (1999).

[14] Mohanainah, P., Sathyanarayana, P., and Guru Kumar, L.,"Image Texture Feature Extraction Using GLCM Approach,"*International Journal of Scientific and Research Publications* 3 (5), (2013).

[15] Haralick, R.M., Shanmugam, K., and Dinstein, I.,"Textural Features for Image Classification,"*IEEE Transaction on System, Man and Cybernetics* 3, 610-621 (1973).

[16] Wijanarto,"Image Retrieval Berdasarkan Properti Statistik Histogram,"*Jurnal Techno Science Fakultas Teknik Uiniversitas Dian Nuswantoro Semarang* 3 (2), (2009).

[17] Srivastava, M.S.,"A Measure of Skewness and Kurtosis and Graphical Method for Assessing Multivariate Normality,"*Statistics and Probability Letters* 2 (5), 263-267(1984).

[18] Pradeep, N., et al., "Feature Extraction of Mammograms,"*International Journal of Bioinformatics Research* 4 (1), 241-244(2012).

[19] Kusumadewi, S. andPurnomo, H.,"*Aplikasi Logika Fuzzy untuk Pendukung Keputusan, 2nd edition* ," Graha Ilmu, 2013.

[20] Wang, L.,"*A Course in Fuzzy Systems and Control*,"Prentice-Hall International, 1997.

ISBN : 978-602-99849-2-7

[21] Nithya, R. and Santhi, B.,"Classification of Normal and Abnormal Patterns in Digital Mammograms for Diagnosis of Breast Cancer,"*International Journal of Computer Applications* 28 (6), (2011).

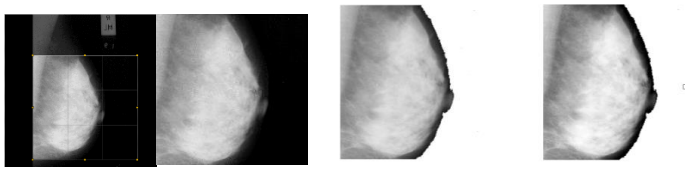
Appendix

Table 1. The extraction result and grouping fuzzy sets of image `mdb004.png`

Features	Extraction	Membership degree	Fuzzy set
Contrast	0,17724	0,60891	A_4
Correlation	0,97081	0,80467	B_5
Energy	0,2254	0,62402	C_3
Homogeneity	0,96022	0,8478	D_5
Mean	200,7564	0,50067	E_7
Variance	3931,3796	0,74827	F_4
SD	62,7007	0,753095	G_4
Skewness	-1,2836	0,92805	H_5
Kurtosis	4,1951	0,98512	I_3
Entropy	5,8093	0,96165	J_6
Diagnose			Normal

Table 2. The diagnosis result of fuzzy system

Kind		Diagnosis without point operation				Diagnosis with point operation			
Training data	Diagnosis	Normal	Benign	Malignant	Sum	Normal	Benign	Malignant	Sum
	Normal	31	1	-	32	32	-	-	32
	Benign	-	32	-	32	-	32	-	32
	Malignant	-	4	28	32	-	3	29	32
	Sum	31	37	28	96	32	35	29	96
Testing data	Diagnosis	Normal	Benign	Malignant	Sum	Normal	Benign	Malignant	Sum
	Normal	3	5	-	8	7	1	-	8
	Benign	-	7	1	8	1	7	-	8
	Malignant	2	4	2	8	-	-	8	8
	Sum	5	16	3	24	8	8	8	24



(a) (b) (c) (d)

Figure 2. Image `mdb004`(a) before (b) after cropping (c) after breaking of background (d) after point operation

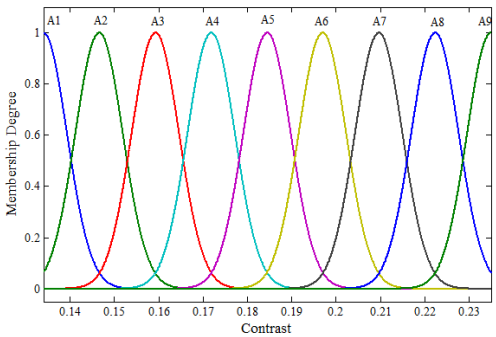


Figure 3. The representation of fuzzy set on contrast variable

ISBN : 978-602-99849-2-7

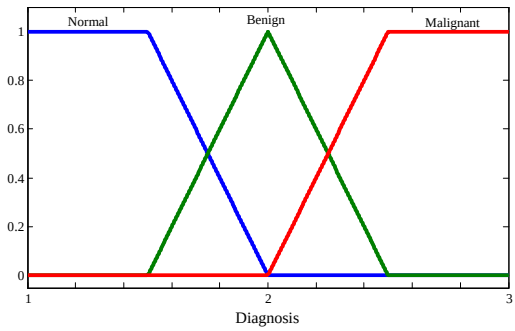


Figure 4. The representation of fuzzy set on output variable

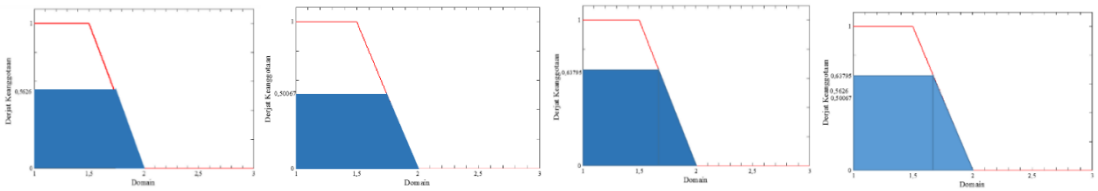


Figure 5. The curve of membership function on output for (a) rule 1; (b) rule 2; and (c) rule 21 (d) aggregation

ISBN : 978-602-99849-2-7

Thurston Method, Area Development Project Impact Evaluation in Pasaman Barat

Aam Alamudi, Kusman Sadik, and Khiril Anwar Notodiputro

Statistics Department, FMIPA IPB, Bogor

aamalamudi@gmail.com, kusmansadik@gmail.com, and kh.notodiputro@gmail.com

Abstract: A study was intended to evaluate comprehensively the impact of the ADP on the development of West Pasaman. One of problems in the study of is to determine how effective the ADP compared to other project. There are five activities considered effect Pasaman Barat development, they are (1) access road project, (2) Oil Palm Plantation Project in Ophir, (3) Rural Banking Project, (4) ADP, and (5) others. This paper presents Thurston Methods to access comparison the five activities as one assessment to be considered. The result show relative importance (RI) of the factors considered to have some contribution towards the changes in West Pasaman during ADP periods and post-ADP periods. The result show that during both periods, the Access Road Project is the most important factor affecting the observed changes in West Pasaman, followed by ADP, Oil Palm Plantation in Ophir, Rural Banking Project, and the other projects. The results are consistent for both changes in village infrastructure and village economic development.
Keywords: Keyword Area Development Project; Integrated Regional Rural Development; Relative Importance; rescaling method.

1. Introduction

A study of impact of Area Development Project (ADP) has been conducted in West Pasaman on March 2004. The study was intended to evaluate comprehensively the impact of the project on the development of West Pasaman. ADP was one of first integrated area development projects implemented by the German Agency for Technical Cooperation (GTC), since its scope includes agricultural development, social development, and private sector development. ADP undertake in 1980 to 1992.

Together with ADP, there are some other project that impact the community, they are Oil Palm Plantation in Ophir (OPPP) in 1981 – 1996, and Access Road Project in West Pasaman or ARP in 1978 – 1984. In addition to this project, other funding agencies also provided financial assistance to improve irrigation, education, health, water supply and other facilities in West Pasaman.

Basically, the study was intended to answer the following questions: "what effect do development measure have? What can we learn from the success and failure of the past so that the German Ministry for Economic Cooperation and Development (BMZ) can further improve development cooperation". The study was undertaken by the team of ten international and domestic consultants from 1 November 2003 to 29 February 2004.

One of problems in this study is to determine how effective the ADP compared to other project. There are five activities considered effect Pasaman Barat development, they are (1) access road project, (2) Oil Palm Plantation Project in Ophir, (3) Rural Banking Project, (4) ADP, and (5) others. This paper presents Thurston Methods to access comparison the five activities as one accessment to be considered.

2. Integrated Regional Rural Development of West Pasaman

In 1980, the Government of Indonesia (GOI) and the Government of Federal Republic of Germany agreed to adopt a concept of integrated regional rural development for the development of West Pasaman, which was at that time the least developed and the most isolated area compared to the other part of West Sumatera Province [5]. Under this concept the German Government provided financial assistance in the form of grant and soft loan for the implementation of the following [5]:

1. Area Development Project in West Pasaman (1980 – 1992). ADP was one of first integrated area development projects implemented by the German Agency for Technical Cooperation (GTC), since its scope includes agricultural development, social development, and private

sector development. ADP used a cross sector approach with an intended target group orientation and participation.

2. Oil Palm Plantation in Ophir or OPPP (1981 – 1996). Under this project, the German Bank for Reconstruction and Development (KfW) and GTZ provided financial assistance to GOI to (a) develop oil palm estates covering about 6,000 ha (1,200 ha nucleus estate, and 4,800 ha smallholder estates), (b) establish and strengthen viable farmers' groups and cooperatives to manage the smallholder oil palm (4,800 ha), (c) construct a network of farm-to-market roads, (d) built an oil palm mill with a capacity of 20 ton fruit fresh bunch/hour which was later expanded to 50 ton/hour. The project adopted a concept of nucleus estate with the participation of small farmers (NESP). This project served as catalyst for the rapid development of the agriculture sector in West Pasaman as it has become a model for development of oil palm in the area.
3. Access Road Project in West Pasaman or ARP (1978 – 1984). ARP is a prerequisite for the successful development of ADP and OPPP. Under this project, KfW provided financial assistance to construct a connecting road from Lubok Alung to Manggopoh in Agam District, and from Manggopoh to Simpang Empat in West Pasaman, which reduced the travel time to Padang to Simpang Empat by at list 8 hours. In addition, KfW also provided financial assistance for the construction of 56 km of feeder roads to connect areas with the connecting road, and to construct the Air Gadang Bridge crossing Batang Pasaman River. GOI, using its own resources, constructed and rehabilitated the main roads between Simpang Empat and Air Bangis, between Simpang Empat and Sasak, and between Simpang Empat and Panti.

In addition to this project, other agencies also provided financial assistance to improve irrigation, education, health, water supply and other facilities in West Pasaman.

Asurvey carried out ex-post evaluation of ADP in terms of its relevance, efficiency, effectiveness, impact, and sustainability.

3. Material & Methodology

a. Data

In the survey, 775 respondents has been chosen in Pasaman Barat and asked to value the contribution of the five activities on development of Pasaman Barat. The five activities are (1) access road project, (2) Oil Palm Plantation Project in Ophir, (3) Rural Banking Project, (4) ADP, and (5) others. The developments to be considered are (1) village infrastructure, and (2) village economic development. The development to be considered differentiate as development in the periods 1980 – 1992 (ADP Periods), and 1992 – Now (post-ADP periods). In other to value the two periods, the respondents should be in the age of 39 years at the time of study, and have been living in pasaman since 1980.

In the post-ADP periods there are also government programs to be considered.

b. Method

Thurston method is a rescaling method for comparing some likert scale measurement. For some objects put scale of 1 – k (1 – 2 – 3 – ⋯ – k). Some number of respondent (n) prepared to value the objects by the scale 1 for the highest, k for the lowest. So that for p objects can be built n×p matrix with object values as the element. By this matrix, further steps then to be done to get new scale. The steps of this scaling are:

1. For the i-th object, compare to j-th object (i,j=1,2,⋯,p), put 1 score if better, 0 if worst, 0.5 if of the same. The scoring to be done for n respondent.
   Sum the score of each object. These score are the object frequencies better than other objects.
2. Calculate the proportion; it is object frequencies divide by total frequencies.
3. Calculate Z-value; it is the standard normal score for each proportions (standard normal score is $\Phi^{-1}\left((i - \frac{3}{8})/(p + \frac{1}{4})\right)$, with i is the score of the rank, p is number of score; $\Phi^{-1}(x)$is the function of standard normal distribution.
4. Construct new scale by interpolating the range {min (Z) – max (Z)} into {1 – k}. The new scale reflects relative importance of the factor.

ISBN : 978-602-99849-2-7

4. Results and Discussion

a. Result

As the result, Relative Importance Of Factors Contributing to Changes in West Pasaman Result, is presented in Table 1. The table show relative importance (RI) of the factors considered to have some contribution towards the changes in West Pasaman during ADP periods (1980 – 1992) and post-ADP periods (1992 – now). The result show that during both periods, the Access Road Project is the most important factor affecting the observed changes in West Pasaman, followed by ADP, Oil Palm Plantation in Ophir, Rural Banking Project, and the other projects. The results are consistent for both changes in village infrastructure and village economic development.

Table 1. Relative Importance Of Factors Contributing to Changes in West Pasaman

Changes	Factors	ADP Periods		Post-ADP Periods	
		RI	Rank of RI	RI	Rank of RI
Village infrastructures	Acces Road Project	0.0196	1	0.0195	1
	Oil Palm Plantation Project in Ophir	0.0125	3	0.0131	3
	Rural Banking Project	0.0095	4	0.0096	4
	ADP	0.0156	2	0.0151	2
	Others	0.0043	5	0.0042	5
Village economic development	Acces Road Project	0.0295	1	0.0294	1
	Oil Palm Plantation Project in Ophir	0.0199	3	0.0213	3
	Rural Banking Project	0.0157	5	0.0156	5
	ADP	0.0228	2	0.0221	2
	Government programs	0.0173	4	0.0166	4
	Others	0.0057	6	0.0057	6

b. Discussion

As expected, the method presented in this paper can be used to measure relative importance of the four projects had developed in West Pasaman. This method can be applied in other field of research of that manner. Validity of the measure then depend on sample drawed.

5. Conclusion

Thurston method can be applied in accessing relative importance of numbers of project, judged by people. The method can be applied in other field of research of that manner.

Acknowledgement. This research is supported by ex-post evaluation of ADP.

References

[1] Agrar und Hyrotechnik, GMBH. 1987. Consulting Enginer, Ophir Palm Oil Project Review Study, Draft Final Report, Essen Germany.
[2] Alfred Jaeckle. 1985. Integrated Agricultural Extention for Food Crop, Final Report, ADP Project. Sukamenanti
[3] Birgit Kerstan. 1991. ADP/PDP Women and Youth Promotion and the Cooperation withNGOs: 1985-1990. Final Report. Padang.
[4] GTZ and KfW. 2000. Ex-Post Impact Analysis of Area Development Project, Rural Finance Project West Pasaman, Access Road Project in West Pasaman and Oil Palm Plantation Project NESP Ophir. Bonn. Germany.
[5] GTZ. 1993. Project Completion Report on Area Development Project/provincial Development Programme in West Pasaman. GTZ. Germany

ISBN : 978-602-99849-2-7

Simulation Study of Robust Regression in High Dimensional Data Through the LAD-LASSO

Septian Rahardiantoro[1], Anang Kurnia[1]

[1]Department of Statistics, Bogor Agricultural University

rahardiantoro.stk@gmail.com* and anangk@apps.ipb.ac.id

Abstract: The common issues in regression, there are a lot of cases in the condition number of predictor variables more than number of observations (p ≫ n) called high dimensional data. The classical problem always lies in this case, that is multicolinearity. It would be worst when the datasets subject to heavy-tailed errors or outliers that may appear in the responses and/or the predictors. As this reason, Wang *et al*in 2007 developed combined methods from Least Absolute Deviation (LAD) regression that is useful for robust regression, and also LASSO that is popular choice for shrinkage estimation and variable selection, becoming LAD-LASSO. Extensive simulation studies demonstrate satisfactory using LAD-LASSO in high dimensional datasets that lies outliers better than using LASSO.
Keywords: high dimensional data, LAD-LASSO, robust regression

1. Introduction

The classical multiple linear regression problem follows the model $\boldsymbol{y}_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i;\ i = 1,2,\dots,n$, with $\boldsymbol{x}_i = \left(x_{i1}, x_{i2}, \dots, x_{ip}\right)'p$ −dimensional regression covariates, a response $\boldsymbol{y}_i$, and $\boldsymbol{\beta} = \left(\beta_1, \beta_2 \dots, \beta_p\right)'$ the associated regression coefficients where assumes errors $\boldsymbol{\varepsilon}_i \sim N(0, \sigma^2)$. Estimation of regression parameter, $\boldsymbol{\beta}$ could be using ordinary least squre (OLS) that minimize the sum square of error. The formula follows $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, implies assume $\boldsymbol{X}'\boldsymbol{X}$ is a nonsingular matrix, with matrix covariates $\boldsymbol{X}_{n\times p}$ and response vector $\boldsymbol{y}_{n\times 1}$.

Common issues in the certain background, there are a lot of regression cases in the condition number of predictor variables more than number of observations ($p \gg n$). When $\boldsymbol{X}$ is full rank ($p \leq n$), the exploration of causal relationship could be accomplished using classical multiple regression above. But when the number of predictors is large compared to the number of observations, $\boldsymbol{X}$ is likely not full rank, that means $\boldsymbol{X}'\boldsymbol{X}$ become singular and the regression approach is no longer feasible (i.e., because of multicolinearity) [1]. LASSO regression [2], is a penalized regression method that is so popular choice for handling this conditions. It is so useful for shrinkage estimation and variable selection.

The worst condition of datasets for regression problem is when they subject to heavy-tailed errors or outliers that may appear in the responses and/or the predictors. In such a situation, it is well known that the traditional OLS may fail to produce a reliable estimator, and the least absolute deviation (LAD) estimator can be very useful. Wang et al (2007) [3] developed the combined method from LAD and LASSO regression. The basic idea is to combine the usual LAD criterion and the LASSO-type penalty together to produce the LAD-LASSO method.

Simulation study have been developed to see the LASSO and LAD-LASSO processes for handling high-dimensional data containts outliers in a lot of scenarios. The simulation using R software and some of R packages.

2. LAD-LASSO

Consider the linear regression model above, moreover assume that $\beta_j \neq 0$ for $j \leq p_0$ and $\beta_j = 0$ for $j > p_0$ for some $p_0 \geq 0$. Thus the correct model has p_0 significant and $(p - p_0)$ insignificant regression variables. Usually, the unknown prameters of classical regression model can be estimated by minimizing the OLS criterion, $\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2$. Furthermore, to shrink unnecessary coefficients to 0, Tibshirani (1996) [2] proposed the following LASSO criterion

$$LASSO = \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + n\lambda \sum_{j=1}^{p} |\beta_j|,$$

where $\lambda > 0$ is the tuning parameter. Then the LASSO formula have been modified by Fan and Li 2001 [4] for avoiding the bias,

$$LASSO^* = \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + n \sum_{j=1}^{p} \lambda_j |\beta_j|,$$

As a result, $LASSO^*$ is able to produce sparse solutions more effectively than $LASSO$. To obtain a robust $LASSO$-type estimator, the modification of $LASSO^*$ into the following LAD-$LASSO$ criterion:

$$\text{LAD-}LASSO = Q(\beta) = \textstyle\sum_{i=1}^{n} |\boldsymbol{y}_i - \boldsymbol{x}_i'\boldsymbol{\beta}| + n\lambda \sum_{j=1}^{p} |\beta_j|,$$

As can be seen, the LAD-$LASSO$ criterion combines the LAD criterion and the lasso penalty, and hence the resulting estimator is expected to be robust against outliers and also to enjoy a sparse representation.

3. Simulation Study

The simulation in this research using R software that would evaluate the standard errors performance. Using R package, flare for sparse linear regression, the simulation set in $n = 100$, and vary p from 375 to 3000 as shown in Table 1. The datasets generated independently with each row of the design matrix from a p-dimensional normal distribution $N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}_{jk} = 0.5^{|j-k|}$ [5]. Then the response vector generated follows $\boldsymbol{y}_i = 3\boldsymbol{x}_{i1} + 2\boldsymbol{x}_{i2} + 1.5\boldsymbol{x}_{i4} + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{\varepsilon}_i$ is independently generated from $N(0,1)$. The scenario before generated without effects of outliers.

Next scenarios generated using the effects of outliers by replacing the distributions of errors that generated from some heavy-tailed distributions, in this research using the standard t-distribution with 5 df (t_5). For comparison purpose, all of scenarios to be evaluated using $LASSO$ and LAD-$LASSO$.

Table 1. Average of standard errors

Scenario 1 : Dataset without outliers				
Method	$p = 375$	$p = 750$	$p = 1500$	$p = 3000$
LASSO	1.762	1.742	1.769	1.728
LAD-LASSO	1.641	1.715	1.738	1.684
Scenario 2 : Dataset with outliers				
Errors : standard t-distribution with 5 df (t_5)				
Method	$p = 375$	$p = 750$	$p = 1500$	$p = 3000$
LASSO	1.514	1.653	1.623	1.638
LAD-LASSO	1.503	1.588	1.636	1.617

From the table 1 above, it is shown that in the datasets without outliers, the batter performance is from LAD-LASSO that have standard errors minimum. It is also happen almost in the datasets with outliers, the performance of LAD-LASSO is better than LASSO.

4. Conclusion

In the conditions of high-dimensional datasets contains outlier, the LAD-LASSO result the better solution (smaller standard errors) than LASSO. The concept is from combination of LAD and LASSO. And it could be as a suggestion for some researchers when handling high-dimensional datasets with $p \gg n$, and also contains outliers.

ISBN : 978-602-99849-2-7

References

[1] Myers, RH. "*Classical and Modern Regression with Applications Second Edition"*. Boston: PWS-Kent,1990.

[2] Tibshirani, R. (1996). "Regression Shrinkage and Selection via the LASSO", *Journal of the Royal Statistics Society Series B*, 58, 267-288.

[3] Wang, H. Li,G. Jiang, G. (2007). "Robust Regression Shrinkage and Consistent Variable Selection Through the LAD-Lasso", *JBES asa v.2007.*

[4] Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihoodand Its Oracle Properties," *Journal of the American Statistical Association*,96, 1348–1360.

[5] Li, X. Zhao, T. Yuan, X. *et al*(2015). "An R Package flare for High Dimensional Linear Regression and Precision Matrix Estimation", *R publication.*

ISBN : 978-602-99849-2-7

An Implementation of Genetic Algorithm to Generate the Most Compromised Decision when Information of the Alternativesis Incomplete

Bagus Sartono[1], Septian Rahardiantoro[1]

[1]Department of Statistics, Bogor Agricultural University

bagusco@ipb.ac.id / bagusco@gmail.com* and rahardiantoro.stk@gmail.com

Abstract: The present paper discusses a method in a decision support system to order some alternatives based on multiple criteria while some information is missing or incomplete. It is important since in many applications we may not be able to have complete evaluation scores for all objects in all variables. A method which utilizes a genetic algorithm to find the best possible solution is proposed to maximize the correlation among the preference rank order and all variables involved. A small illustration to select a car out of ten alternatives was given to present how it works to find a good solution in a quite short time.
Keywords: correlation; decision support system; genetic algorithm; missing data

1. Introduction

In most circumstances, an individual or organization needs to make a rational decision to choose one out of several available alternatives. The complexity of the decision making arises when several criteria are involved to select the best. Indeed there are several approaches developed to help the decision makers determining the choices.

Some approach that can be mentioned here such as weighted sum [1] where the ranking of the alternative is based on the final score of the linear combination of all criteria using different importance weights. Another famous technique is the analytic hierarchy process [2] which decomposes the decision problem into several subproblems with a hierarchical arrangement. Using a pairwise comparison the decision maker decide which one is relatively more important than the other and use the comparison score to generate the weight. A quite new approach can be found in [3] called best-worst method.

The aforementioned approaches have been implemented in many applications and empirically assessed to be useful for the decision makers. However, they could not work when there are some missing or incomplete information on one or more alternatives in one or more variables. For example, suppose that the Government of Indonesia would like to develop a rank of the districts across the country. They use many indicators to measure the advancement of the districts. It is possible that in some districts, we could not obtain the value of some variables due to the limitation of data collection process or other reasons.

The present paper would address that problem, by providing a methodology to generate the sorted list of alternatives. Because there are several criteria used in the decision making, the order of alternatives in the list should be compromised with all criteria. We define the property of compromised by having an optimal correlation to all criteria used.

2. Basic Notation and Problem Definition

Suppose there are n alternatives which are characterized by p criteria. Each criterion has at least ordinal scale so that the alternatives could be sorted with a certain meaning. In this paper we assume that higher value is preferable for all criteria. A $n \times p$ matrix $\boldsymbol{Z} = [z_{ij}]$ represents the data for the decision making with z_{ij} is the value of the j-th criterion for the i-th alternative. If $\boldsymbol{z}_j$ is the j-th column of $\boldsymbol{Z}$ then it is the column containing all values of j-th criterion of all alternatives.

The ultimate goal of the methodology is finding a vector $\boldsymbol{s} = (s_1, s_2, \ldots, s_n)$,with $s_i \in \{1, \ldots, n\}$ represents the rank order of the i-th alternative. An alternative with $s_i = n$ is the one which is most preferable, and the one with $s_i = 1$ is the least. We put constraints that $s_i \neq s_j$. Therefore the vector $\boldsymbol{s}$ can be seen as a preference rank vector.

The basic idea of the approach is that the preference rank vector should have high agreement to the criteria. Since all criteria are measured in at least an ordinal scale, we propose to use a rank correlation coefficient to represent the degree of agreement.

Suppose that R_k is the correlation between criterion k and the preference rank vector $\boldsymbol{s}$, or R_k = corr($\boldsymbol{z}_k$, $\boldsymbol{s}$). For all k = 1, …, p, we would like to have high values of R_k's. To ensure that each of criteria has high value of R_k, we propose to the restate the problem of finding preference rank vector $\boldsymbol{s}$ as follow: "Find $\boldsymbol{s}$ so that the minimum value of R_k's is maximum" or "Find $\boldsymbol{s}$ which maximize $min\{R_k\}$".

3. Proposed Approach

As described, the problem of choosing the best alternative can be seen as an optimization problem. We could obviously view that it is a combinatorial problem since the solution basically is the permutation of n different things labeled as 1, 2, …, n. Since it is a combinatorial optimization problem, we could employ a genetic algorithm methodology [4] to find an optimal solution.

The main components of the algorithm could be summarized as follow. First, a preference rank vector could be seen as a gene consisting n chromosomes with the value as one of the element of {1, …, n} and be distinguished each other. Second, the fitness value of each gene or the objective function is the $min\{R_k\}$. Each of the genes has a single fitness value, and the gene with higher fitness value is the preferable one.

As previously mentioned, the R_k value is the rank correlation between the score of the k-th criterion of decision making and the preference rank. It is possible that one or more objects do not have a score of the criterion. Whenever it happened, the correlation was calculated by involving m ($<n$) pair observations whose available score.

The pseudocode of the proposed algorithm is as follow

1. set n = the number of alternatives
   set p = the number of criteria
   set g = the maximum number of iteration of the genetic algorithm
   set $npop$ = number of genes in a population
2. initiate a population of$npop$ preference rank vectors,S = {$\boldsymbol{s}$ | $\boldsymbol{s}$ = (s_1, s_2, …, s_n)} by randomly permute {1, 2, … n}} and n(S) = $npop$
3. for t = 1 to g
   a. select as many as $nsel$ genes from S whose higher fitness value
   b. do a cross-over procedure appropriate for combinatorial case problem and do mutation
   c. collect the new genes resulted by cross-over and mutation to a new population S
   d. if the fitness value improves, then repeat a – c, otherwise set t = g

4. Illustrative Example

As an illustration, we implemented the proposed approach to the following case. Suppose a man would like to purchase a car. At that time, he had ten alternatives of cars. To be able to find the best, he considered nine attributes that describe the quality of the alternatives: X1 Attractive , X2 Quiet , X3 Reliable, X4 Well Built , X5 Comfortable , X6 Roomy , X7 High Prestige, X8 Unique, and X9 Worth Value. Table 1 provides the score of the attributes for each car where an alternative whose a higher score means more preferred. Suppose that there are several attribute scores are not available. As a note, the data was taken from the discussion of Positioning the Nissan Infiniti G20 in[5].

We implemented the algorithm and a program written in SAS/IML language was developed to perform the genetic algorithm approach. Readers who are interested to have could have it upon request to the authors.

A result that we obtained from the algorithm yields the order of the preference of the alternatives as follow: Nissan G20, BMW 318i, SAAB 900, Honda Prelude, Toyota Supra, Audi 90, Ford T-Bird, Mercury Capri, Eagle Talon, andPontiac Firebird. This order has the fitness value of minimum correlation coefficient as large as 0.728. In other word, the preference order has high correlation to all of attributes by 0.728 or larger.

Table 1. The attribute score of the ten alternative cars

ISBN : 978-602-99849-2-7

Cars	X1	X2	X3	X4	X5	X6	X7	X8	X9
Toyota Supra	5.6	4.2	7.0	6.9	5.3	3.5	5.3	6.1	5.5
SAAB 900	5.3	4.8	5.3	6.2	6.2	5.1	5.7	7.1	6.1
Pontiac Firebird	3.9	2.8	5.1	4.6	4.7	3.3	3.8	4.7	4.7
Nissan G20	5.6	6.3	6.1	7.4	6.6	5.6	5.4	5.5	5.6
Mercury Capri	3.9	3.3	5.0	4.7	4.6	3.6	.	5.1	5.2
Honda Prelude	5.2	5.4	5.8	6.2	5.7	3.9	4.7	5.1	6.4
Ford T-Bird	4.0	3.6	.	4.8	5.0	3.9	3.5	5.4	4.7
Eagle Talon	4.0	3.5	4.7	4.7	5.0	3.6	2.8	4.7	5.4
BMW 318i	5.7	5.0	6.7	7.2	5.5	4.3	6.4	6.2	5.7
Audi 90	4.6	5.2	5.3	6.4	.	5.3	5.6	5.6	4.7

To compare how good is the result of the proposed algorithm, we ran 2000 repetitions of random permutation. Each repetition evaluated 1000 random order and selected the best order. Figure 1 presents the distribution of the best order from the random order. We could see that the best result using this procedure is around 0.713 which is worse than the one that we obtained using a genetic algorithm.

In addition that we could get a better solution, the proposed algorithm ran faster than the greedy search using random permutation. To obtain best preference order with minimum correlation of 0.713, the computer took 375 seconds, while the genetic algorithm needed 6 seconds only to obtain the best of 0.728.

5. Conclusion

The proposed method deals with a problem of ordering several alternatives and expects that the ordering optimally agree with all criteria. It is utilizing a genetic algorithm approach and proved to be able to provide a good decision to rank alternatives based on multicriteria information which some information is incomplete. Not only it results excellent solutions, the method also ran quickly. By this fact, we recommend scientists and decision makers to adopt the method for the cases that they might face to handle.

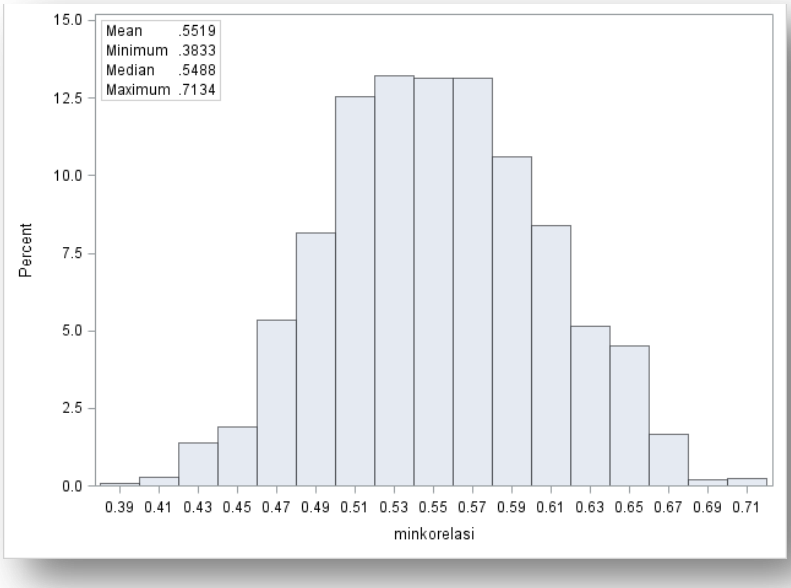


Figure 1. The distribution of minimum correlation of 1000 random preference order

ISBN : 978-602-99849-2-7

References

[1] Gass, S.; Saaty, T. "Parametric Objective Function Part II".*Operations Research* 3: 316–319. (1955).
[2] Saaty, T.L. “*The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation*”. New York: McGraw-Hill, 1980.
[3] Rezaei, J. (2015) "Best-Worst Multi-Criteria Decision-Making Method", *Omega*, 53, 49-57.
[4] Mitchell, M. “*An Introduction to Genetic Algorithms*”. Cambridge, MA: MIT Press, 1996
[5] Lilien GL, Rangaswamy A. “*Marketing Engineering*”. Create Space Independent Publishing Platform. 2004

ISBN : 978-602-99849-2-7

Estimation of MedianGrowth Charts for Children Based on Biresponse Semiparametric Regression Model by using Local Linear Estimator

Nur Chamidah [1], Marisa Rifada[1]

[1]Department of Mathematics, Faculty of Sciences and Technology,Airlangga University, Surabaya, Indonesia

nur-c@fst.unair.ac.id, marisa_rifada@yahoo.com

Abstract: Physical children growth can be measured based on anthropometric measure i.e. weight and length. There is significant of the coeficient correlation between weight and length of the children. It means that the modeling of children growth charts would be more realistic if it was modeled simultaneously by using biresponse model approach. In this study, we investigategrowth charts of childrenfrom birth up to two years old based on not only age but also sex. So, we use biresponse semiparametric model that the response variables i.e., weight (kg) and length (cm), while a parametric predictor variable is sex and a nonparametric predictor variable is age (month). The children around two years old grow rapidly,then decrease slowly along with increasing of children's age, so the modeling of children growth charts is more appropriate modeled by using locally model approach. Local linear estimator is one of estimation methods in local smoothing. The data was collected of children up to two years old in Surabaya, 2015. Based on generalized cross validation criterion, we get the optimal bandwidth, coefficient of determinationand the mean squared error, i.e., 1.94, 0.997 and 0.21 respectively. The average of weight and length growth of boys from birth up to two years old in Surabaya is higher than girls i.e., 0.24 kg and 1.78 cm, respectively.
Keywords: Median Growth Charts; Local Linear Estimator ; Biresponse Semiparametric Model

1. Introduction

Growth charts are important clinical tools to assess and monitor growth[1]. Physical children growth can be measured based on anthropometric measure i.e. weight and height/length [2]. There is significant of the coeficient correlation between weight and height of the children. It means that the modeling of children growth charts would be more realistic if it was modeled simultaneously by using biresponse model approach [3].The children around two years old grow rapidly, then decrease slowly along with increasing of children's age [4], so the modeling of children growth charts is more appropriate modeled by using locally model approach.

The WHO Multicenter Growth Reference Study Group [1] have proposed the child growth chart standards constructed based on single response by using cubic spline smoothing. Several smoothing terms can be used in generating the curves [1]. Local linear estimator is one of estimation methods in local smoothing [5]. Chamidah and Eridani [3] studied for designing of growth reference chart by using biresponse semiparametric regression approach based on P-Spline estimator. From [3], based on children data in Surabaya city 2013, and the 50^{th} percentiles estimation of weight and height, they obtained that in average the growth reference chart of boy higher than those of girl. In addition,the mean squared error value is0.5684.Therefore, the aim of this study isestimate median of weight and length for boys and girls based on biresponse semiparametric regression model by using local linear estimator and then construct the median growth chart of weight-for-age and length-for-age for boys and girls.

2. Related Works/Literature Review

Chamidah and Eridani [3], for designing of growth reference chart used biresponse semiparametric model based on P-Spline estimator. Firstly, given *n* observations data $\left(y_{i1}, y_{i2}, x_i, t_i\right)$, where y_{ij} represents i^{th}observation of j^{th} response that satisfy the following multi-response semiparametric regression model:

$$\underset{\sim}{Y}_i = \mathbf{X}_i^T \underset{\sim}{\alpha} + \underset{\sim}{g}\left(t_i\right) + \underset{\sim}{\varepsilon}_i \ , \qquad i = 1,2 \ , \ \ldots \ , \ n \tag{1}$$

where $\underset{\sim}{Y}_i = \left(y_{i1}, y_{i2}\right)^T$ and $\underset{\sim}{\varepsilon}_i$ are responses and error for i^{th}observation, respectively. Next, $\underset{\sim}{g}\left(t_i\right)$is function of population mean that is assumed smooth. $\mathbf{X}_i^T = \left(\underset{\sim}{1} \quad \underset{\sim}{x}_{i1} \quad \ldots \quad \underset{\sim}{x}_{iq}\right)^T$ represents parametric component that form of the function is assumed to be known fori^{th}observation and $\underset{\sim}{\alpha} = \left(\alpha_0, \alpha_1, ..., \alpha_q\right)^T$ represents coefficient of parametric predictor variables. The model (1) contain two parts, i.e., parametric component $\mathbf{X}_i^T \underset{\sim}{\alpha}$ and nonparametric component $\underset{\sim}{g}\left(t_i\right)$.Based on model (1), function $\underset{\sim}{g}\left(t_i\right)$ is estimated by using penalized splines estimator.Penalized Spline function with order *p* and knots $\left(\tau_1, \tau_2, ..., \tau_K\right)$ can be expressed as follows:

$$g\left(t\right) = \sum_{r=1}^{p} \beta_r t^r + \sum_{l=1}^{K} \beta_{p+l} \left(t - \tau_l\right)_+^p \tag{2}$$

Next, we estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by minimizing penalized least square (PLS) criterion as follows:

$$\sum_{j=1}^{2} \sum_{i=1}^{n} \left(y_{ij} - x_{ij}^T \underset{\sim}{\alpha} + t_{ij}^T \beta\right)^2 + \lambda \sum_{l=1}^{K} \beta_{p+l}^2 \tag{3}$$

where λ is smoothing parameter, and *K* is number of knots and *p* is order of polynomial.

3. Material & Methodology

We collected data of children growth for ages 0 to 24 months from integrated health center records which is in Indonesia is called"*POSYANDU*" in Surabaya, 2015. The data contains 629 observations for boys and 613 observations for girls. We computed the 50th percentiles value (P_{50}), for weight and length children in all age groups (age 0 to 24 months) for boys and girls. Then, we analyzed the 50th percentiles to assessed the median growth chart of children based on bi-response semiparametric regression that the response variables i.e., weight (kg) and length (cm), while a parametric predictor variable is sex and a nonparametric predictor variable is age (month).

We have pairs observations $(y_{1i}, y_{2i}, x_i, t_i)$, $i = 1,2,...,n,$ that follows biresponse semiparametric regression model:

$$\underset{\sim}{y}_i = \underset{\sim}{x}_i \underset{\sim}{b} + \underset{\sim}{f}\left(t_i\right) + \underset{\sim}{e}_i \tag{4}$$

where $\underset{\sim}{y}_i = \left(y_{1i}, y_{2i}\right)^T$ represents response variable, $\underset{\sim}{x}_i = \begin{pmatrix} 1 & x_i & \cdots & x_i^d \end{pmatrix}$is parametric predictor variable, $\underset{\sim}{b} = \begin{pmatrix} b_0 & b_1 & \cdots & b_d \end{pmatrix}^T$, $\underset{\sim}{f}\left(t_i\right) = \left(f_1(t_i), f_2(t_i)\right)^T$ is function which is

ISBN : 978-602-99849-2-7

estimated by nonparametric approach, and $\underset{\sim}{e}_i = \left(e_{1i}, e_{2i}\right)^T$ is random errorwith mean $\underset{\sim}{0}$ and variance V where $E\left(\underset{\sim}{e}\right) = \underset{\sim}{0}$, $Cov\left(e_{ri}, e_{sj}\right) = \begin{cases} rs_{ri}s_{si}, i = j \\ 0, \quad i \neq j \end{cases}$ $, r, s = 1, 2$.

Based on equation (4), we first assumed parameter $\underset{\sim}{b}$ is given, then equation (4) can be written as:

$$\underset{\sim}{y}_i - \underset{\sim}{x}_i \underset{\sim}{b} = \underset{\sim}{f}(t_i) + \underset{\sim}{e}_i$$
$$\underset{\sim}{y}_i^*\left(t_i\right) = \underset{\sim}{f}(t_i) + \underset{\sim}{e}_i \tag{5}$$

Function $\underset{\sim}{f}(t)$ in (5) is smooth functionwhich is assumed continuous and has (d+1) continuous derivatives on an interval about $t = t_0$. We can approach the function $f_1(t)$ and $f_2(t)$ by Taylor series about t_0 as follow :

$$f_1(t) = \sum_{j=0}^{d_1} \frac{f_1^{(j)}(t_0)}{j!}\left(t - t_0\right)^j = \sum_{j=0}^{d_1} a_{1j}\left(t_0\right)\left(t - t_0\right)^j \tag{6}$$

where $a_{1j}\left(t_0\right) = \dfrac{\left(f_1^{(j)}\left(t_0\right)\right)}{j!}$, and $t \in (t_0 - h, t_0 + h)$

$$f_2(t) = \sum_{j=0}^{d_2} \frac{f_2^{(j)}(t_0)}{j!}\left(t - t_0\right)^j = \sum_{j=0}^{d_2} a_{2j}\left(t_0\right)\left(t - t_0\right)^j \tag{7}$$

where $\alpha_{2j}\left(t_0\right) = \dfrac{\left(f_2^{(j)}\left(t_0\right)\right)}{j!}$.

Based on equation (6) and (7), $\underset{\sim}{f}(t)$ can be expressed as :

$$\underset{\sim}{f}(t) = Z^*\left(t_0\right)\underset{\sim}{a}^*\left(t_0\right) \tag{8}$$

where

$$\underset{\sim}{f}(t) = \left(f_1(t), f_2(t)\right)^T$$

$$Z_{t_0}^* = \begin{bmatrix} 1 & \left(t - t_0\right) & \cdots & \left(t - t_0\right)^{d_1} & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & \left(t - t_0\right) & \cdots & \left(t - t_0\right)^{d_2} \end{bmatrix}$$

$$\underset{\sim}{a}^*\left(t_0\right) = \begin{bmatrix} a_{10}\left(t_0\right) & \cdots & a_{1d_1}\left(t_0\right) & a_{20}\left(t_0\right) & \cdots & a_{2d_2}\left(t_0\right) \end{bmatrix}^T$$

Local linear estimator is obtained ifthe degree of polynomial d=1. So, equation (8) can be written as:

$$\underset{\sim}{f}(t) = Z\left(t_0\right)\underset{\sim}{a}\left(t_0\right) \tag{9}$$

where

$$Z\left(t_0\right) = \begin{bmatrix} Z_1\left(t_0\right) & \mathbf{0} \\ \mathbf{0} & Z_2\left(t_0\right) \end{bmatrix}, Z_r\left(t_0\right) = \begin{bmatrix} 1 & t_{r1} - t_o \\ 1 & t_{r2} - t_o \\ \vdots & \vdots \\ 1 & t_{rm} - t_o \end{bmatrix}; \underset{\sim}{\alpha}(t_0) = \begin{pmatrix} \alpha_{10}\left(t_0\right) & \alpha_{11}\left(t_0\right) & \alpha_{20}\left(t_0\right) & \alpha_{21}\left(t_0\right) \end{pmatrix}^T$$

Based on equation (9), equation (5) can be expressed as follows:

ISBN : 978-602-99849-2-7

$$\underset{\sim}{y}^{*} = Z\left(t_0\right)\underset{\sim}{a}\left(t_0\right) + \underset{\sim}{e} \tag{10}$$

We obtain estimator $\underset{\sim}{\alpha}\left(t_0\right)$ in equation (10) based on *Weighted Least Square* (WLS) method by minimizing function:

$$Q\left(t_0\right) = \left(\underset{\sim}{y}^{*} - Z\left(t_0\right)\underset{\sim}{\alpha}\left(t_0\right)\right)^{T} V^{-1}\mathbf{K}_h\left(t_0\right)\left(\underset{\sim}{y}^{*} - Z\left(t_0\right)\underset{\sim}{\alpha}\left(t_0\right)\right) \tag{11}$$

and we get estimator $\underset{\sim}{\alpha}\left(t_0\right)$ is :

$$\underset{\sim}{\hat{a}}\left(t_0\right) = \left[Z^{T}\left(t_0\right)V^{-1}K_h\left(t_0\right)Z\left(t_0\right)\right]^{-1} Z^{T}\left(t_0\right)V^{-1}K_h\left(t_0\right)\underset{\sim}{y}^{*} \tag{12}$$

with weighted matrix V^{-1} is invert of covariance matrix of $\underset{\sim}{e}$ and $\mathbf{K}_h\left(t_0\right)$ is the diagonal matrix of weights.

$\mathbf{K}_h\left(t_0\right) = \mathrm{diag}\left[\mathbf{K}_{h_1}\left(t_0\right), \mathbf{K}_{h_2}\left(t_0\right)\right]$

Where $\mathbf{K}_{h_r}\left(t_0\right) = \mathrm{diag}\left[K_h\left(t_1 - t_0\right), K_h\left(t_2 - t_0\right), ..., K_h\left(t_n - t_0\right)\right]$ and $K_h(.)$ is kernel function with optimum bandwidth *h*.

Based on equation (12), nonparametric hat matric, $\mathbf{A}_h$,for estimate regression function about t_0 is :

$$\mathbf{A}_h\left(t_0\right) = \mathbf{e}\left(\mathbf{Z}^{T}\left(\mathrm{t}_0\right)\mathbf{V}^{-1}\mathbf{K_h}\left(\mathrm{t}_0\right)\mathbf{Z}\left(\mathrm{t}_0\right)\right)^{-1}\mathbf{Z}^{T}\left(\mathrm{t}_0\right)\mathbf{V}^{-1}\mathbf{K_h}\left(\mathrm{t}_0\right) \tag{13}$$

where

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

So, estimation of $\underset{\sim}{f}(t)$ based on local linear estimator can be written as:

$$\underset{\sim}{\hat{f}}\left(\underset{\sim}{t}\right) = \mathbf{A}_h(\underset{\sim}{t})\underset{\sim}{y}^{*} \tag{14}$$

Based on equation (5) and (14), local linear estimator for $\underset{\sim}{\hat{f}}\left(\underset{\sim}{t}\right)$ can be expressed as :

$$\underset{\sim}{\hat{f}}\left(\underset{\sim}{t}\right) = \mathbf{A}_h(\underset{\sim}{t})\left(\underset{\sim}{\mathrm{y}} - \mathbf{X}\underset{\sim}{\beta}\right) \tag{15}$$

So, function sum of squared error in equation (4) can be written as :

$$Q = \left[\underset{\sim}{\mathrm{y}} - \mathbf{X}\underset{\sim}{\beta} - \left(\mathbf{A}_h\left(\underset{\sim}{t}\right)\left(\underset{\sim}{\mathrm{y}} - \mathbf{X}\underset{\sim}{\beta}\right)\right)\right]^{T}\left[\underset{\sim}{\mathrm{y}} - \mathbf{X}\underset{\sim}{\beta} - \left(\mathbf{A}_h\left(\underset{\sim}{t}\right)\left(\underset{\sim}{\mathrm{y}} - \mathbf{X}\underset{\sim}{\beta}\right)\right)\right] \tag{16}$$

The solution of $\underset{\sim}{\hat{\beta}}$ is gotten by minimizing of Q and we get:

$$\underset{\sim}{\hat{\beta}} = \left[\mathbf{X}^{T}\left(\mathrm{I} - \mathbf{A}_h\left(\underset{\sim}{t}\right)\right)^{T}\left(\mathrm{I} - \mathbf{A}_h\left(\underset{\sim}{t}\right)\right)\mathbf{X}\right]^{-1}\mathbf{X}^{T}\left(\mathrm{I} - \mathbf{A}_h\left(\underset{\sim}{t}\right)\right)^{T}\left(\mathrm{I} - \mathbf{A}_h\left(\underset{\sim}{t}\right)\right)\underset{\sim}{y} \tag{17}$$

By substituting estimator $\underset{\sim}{\hat{\beta}}$ in equation (17) to equation (15), so equation (15) can expressed as :

$$\begin{aligned} \underset{\sim}{\hat{f}}\left(\underset{\sim}{t}\right) &= \mathbf{A}_h(\underset{\sim}{t})\left(\underset{\sim}{y} - \mathbf{X}\left[\mathbf{X}^{T}\left(\mathrm{I} - \mathbf{A}_h\left(\underset{\sim}{t}\right)\right)^{T}\left(\mathrm{I} - \mathbf{A}_h\left(\underset{\sim}{t}\right)\right)\mathbf{X}\right]^{-1}\mathbf{X}^{T}\left(\mathrm{I} - \mathbf{A}_h\left(\underset{\sim}{t}\right)\right)^{T}\left(\mathrm{I} - \mathbf{A}_h\left(\underset{\sim}{t}\right)\right)\underset{\sim}{y}\right) \\ &= \mathbf{A}_h(\underset{\sim}{t})\left(\mathbf{I} - \mathbf{X}\left[\mathbf{X}^{T}\mathbf{S}\left(\underset{\sim}{t}\right)\mathbf{X}\right]^{-1}\mathbf{X}^{T}\mathbf{S}\left(\underset{\sim}{t}\right)\right)\underset{\sim}{y} \end{aligned} \tag{18}$$

with $S\left(\underset{\sim}{t}\right) = \left(I - A_h\left(\underset{\sim}{t}\right)\right)^{T}\left(I - A_h\left(\underset{\sim}{t}\right)\right)$

Based on the solution of estimation in equation (17) and equation (18), so we obtained

$\underset{\sim}{\hat{y}} = \mathbf{X}\hat{\beta} + \underset{\sim}{\hat{f}}\left(\underset{\sim}{t}\right)$

ISBN : 978-602-99849-2-7

$$\hat{\underset{\sim}{y}} = \left\{ \mathbf{X}\left[\mathbf{X}^T S\left(\underset{\sim}{t}\right)\mathbf{X}\right]^{-1} \mathbf{X}^T S\left(\underset{\sim}{t}\right)\right\} \underset{\sim}{y} + A_h\left(\underset{\sim}{t}\right)\left(I - \mathbf{X}\left[\mathbf{X}^T S\left(\underset{\sim}{t}\right)\mathbf{X}\right]^{-1} \mathbf{X}^T S\left(\underset{\sim}{t}\right)\right)\underset{\sim}{y}$$

$$\hat{\underset{\sim}{y}} = \left(\mathbf{A}_{par} + \mathbf{A}_{nonpar}\right)\underset{\sim}{y} = \mathbf{A}_h\left(\underset{\sim}{t}\right)\underset{\sim}{y}$$

with

$$\mathbf{A}_h = \mathbf{A}_{par} + \mathbf{A}_{nonpar}$$

$$\mathbf{A}_{par} = \mathbf{X}\left[\mathbf{X}^T S\left(\underset{\sim}{t}\right)\mathbf{X}\right]^{-1} \mathbf{X}^T S\left(\underset{\sim}{t}\right)$$

and $$\mathbf{A}_{nonpar} = A_h\left(\underset{\sim}{t}\right)\left(I - \mathbf{X}\left[\mathbf{X}^T S\left(\underset{\sim}{t}\right)\mathbf{X}\right]^{-1} \mathbf{X}^T S\left(\underset{\sim}{t}\right)\right)$$

In this study, we used local linear regression approach. So, there is a parameter i.e. bandwidth (h) that control the smoothness of the fit and also affect the bias-variance trade-off. The optimal bandwidth value is obtained based on generalized cross validation (GCV) method by minimizing the GCV function as follows:

$$\mathrm{GCV}(h) = \frac{\left[\underset{\sim}{y} - \hat{\underset{\sim}{f}}\left(\underset{\sim}{t}\right)\right]'\left[\underset{\sim}{y} - \hat{\underset{\sim}{f}}\left(\underset{\sim}{t}\right)\right]}{(np)^{-1}\left[\mathrm{tr}\left(\mathbf{I} - \mathbf{A}_h\left(\underset{\sim}{t}\right)\right)\right]^2}$$

4. Results and Discussion

Theresult of analysis P_{50} data obtained Pearson's correlation coefficient between weightand length is 0.99. We estimated median of weight and length based on bi-response semi-parametric regression by using local linear estimator. We get optimal bandwidth value, i.e., 1.94, minimum value of GCV i.e., 53.62, coefficient of determination, i.e., 0.997 andthe mean squared error value i.e., 0.21. Then, we assessed the median growth chart of weight-for-age and length-for-age for boys and girls. The median growth chart of weight-for-age represents that 50% of the children of a given age and gender haveweight higherthan the fiftieth percentile (P_{50}) value, and 50% of the children of a given age and gender have weight less than P_{50} value. The children who have weight below P_{50} value means that their weight less than the average. Hence, the children who have weight higher than P_{50} value means that their weight above the average. This does not mean thatoverweight or underweight.The median growth chart of weight-for-age and length-for-age for boys and girls are compared by observation median are showed by the following four Figures (**Figure 1 -Figure4**)

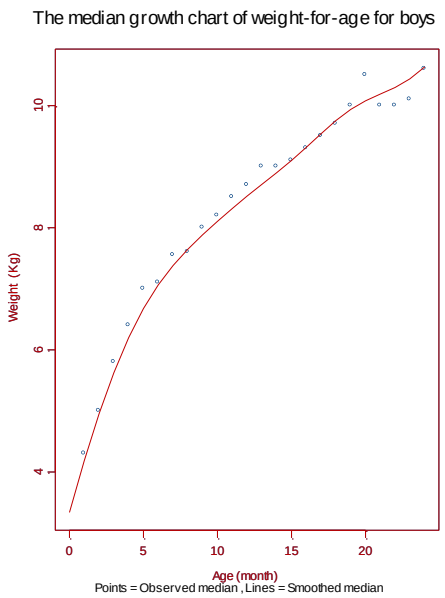


Figure 1. Comparison between smoothed median (observed median of weight-for-age for boys)

ISBN : 978-602-99849-2-7

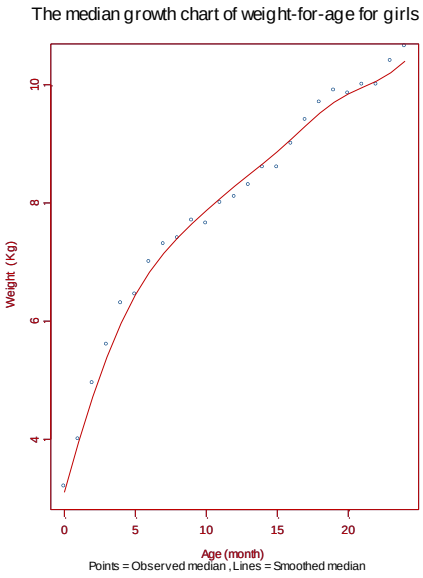


Figure 2. Comparison between smoothed median (observed median of weight-for-age for girls)

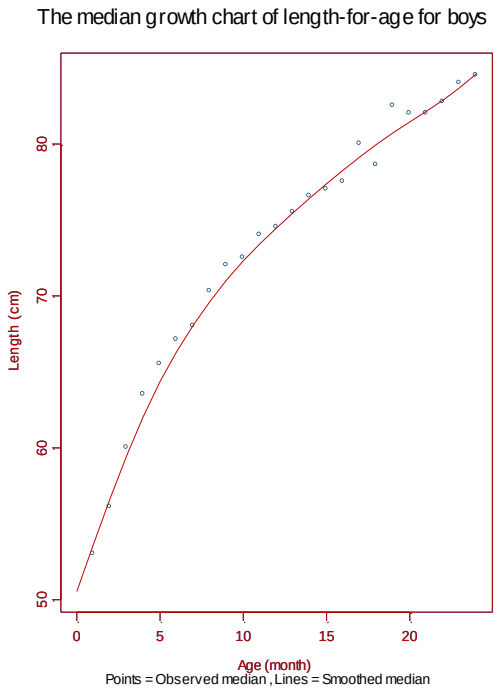


Figure 3. Comparison between smoothed median (observed median of length-for-age for boys)

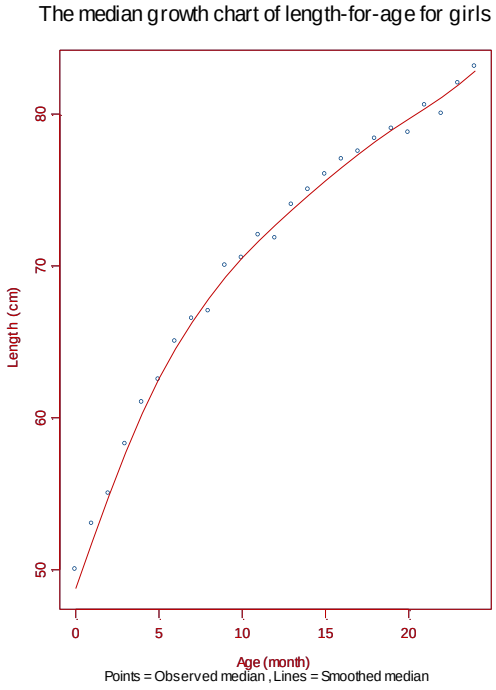


Figure 4. Comparison between smoothed median (observed median of length-for-age for girls)

Overall, the comparison between smoothed median and observed or empirical median growth chart of weight-for-age and length-for-age has been fit.The average absolute difference of smoothed and observed median was small i.e., 0.15 kg for weight-for-age for boys (**Figure 1**) and 0.14 kg for girls (**Figure 2**).Meanwhile,the average absolute difference of smoothed and observed medianfor length-for-age i.e. 0.62cm for boys (**Figure 3**) and 0.47cm for girls (**Figure 4)**. Next, plotting of comparison median growth chart between boys and girls for weight-for-age and length-for-age are showed in **Figure 5** and **Figure 6**, respectively.

ISBN : 978-602-99849-2-7

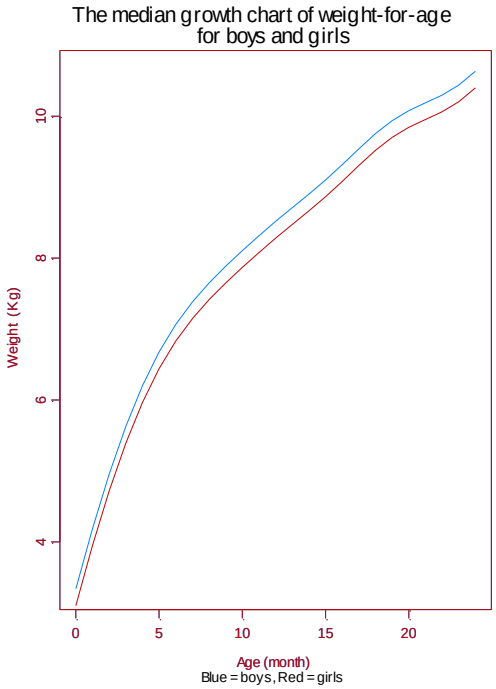


Figure 5. Comparison between median growth (chart of weight-for-age boys and girls)

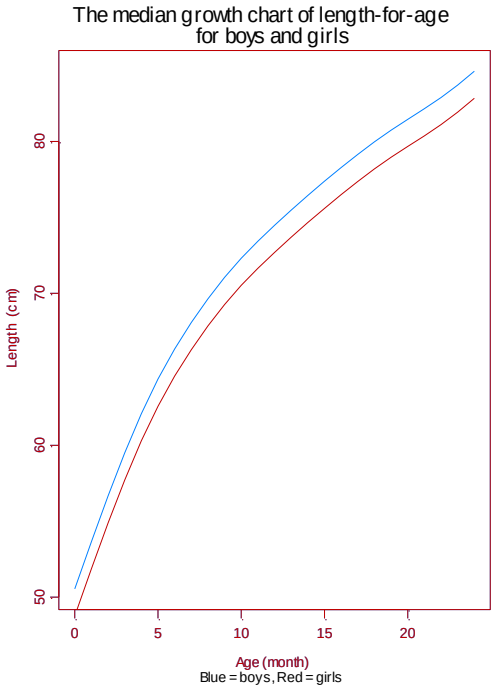


Figure 6. Comparison between median growth (chart of length-for-age boys and girls)

We can showed from **Figure 5** and **Figure 6** that the median growth chart of weight-for-age and length-for-age for boys were higher than girls.The average of weight and length growth of boys from birth up to two years old in Surabaya is higher than girls i.e., 0.24 kg and 1.78 cm, respectively.

5. Conclusion

The estimation median of weight and length children in Surabaya, Indonesia 2015, based on bi-response semiparametric regression model by using local linear estimator has been appropriate to goodness of fit criterions, i.e.,determination coefficient tend to one and mean squared error tend to zero. There is difference pattern of growth chart between boys and girls. The median growth chart of weight-for-age and length-for-age for boys were higher than girls.

Acknowledgement.

Many thanks to Directorate General of Higher Education of Indonesia for financial support of this research through FeaturedResearch University Grant 2015.

References

[1] WHO-Multicenter Growth Reference Study Group, "WHO Child Growth Standards based on length/height, weight and age", *ActaPædiatrica*, Suppl 450 : 76 - 85 (2006)
[2] M.B. Narendra, T.S. Sularyo, Soetjiningsih, H. Suyitno, and I.G.N.G Ranuh, " *TumbuhKembangAnakdanRemaja* ", Jakarta: CV. SagungSeto, 2002
[3] N.Chamidah and Eridani, "Designing of Growth Reference Chart by Using Birespon Semiparametric Regression Approach Based on P-Spline Estimator", *International Journal of Applied Mathematics and Statistics*53 (3),150-158 (2015).
[4] P.Roumeliotis, *Children's health and wellness. Growth and Development*, 2012
[5] Eubank, R., *Spline Smoothing and Nonparametric Regression*,Marcel Dekker, New York, 1998.

Feature Reduction of Wayang Golek Dance Data Using Principal Component Analysis (Pca)

Joko Sutopo[1], Adhi Susanto[2], Insap Santosa[3], Teguh Barata Adji[4]

[1,2,3,4]Departement of Electrical Engineering and Information Technology, Gadjah Mada University

[1]Departement of Electrical Engineering, Yogyakarta Technology University

jksutopo@uty.ac.id

Abstract: One of the internally acknowledged Indonesian Culture inheritance is the Wayang Golek Menak Yogyakarta, which is usually performed with wood doll characters. The objective of the study is to apply motion capture (mocap) technique with Kinect Sensor to capture the Wayang Golek Menak dance movements and apply the Principal Component Analysis (PCA) to identify the specific patterns of the data and express them so that their similarities and differences can be seen. This method is useful as well to compress the data without losing the important information. The resulting BioVision Hierarchy (BVH) motion of this Kinect sensor is to simulate the wayang Golek Menak dance of sizes 149x54 and 151x54 dimension cartesius (*x*, *y*, *z*). Then these tensor cartesius data are converted into spherical frame of $h\theta, h\varphi, hr$. Reduction of matrix dimension is to ease the process in next stage. The results show a truly acceptable wayang golek performance.

Keywords: Feature, Reduction, Kinect, PCA

1. Introduction

Wayang Golek is an internationally acknowledged Indonesian culture inheritance reflecting social, politic, economic, religious, linguistic and human relation life aspects. Wayang Golek researched in this study is Wayang Golek Menak Yogyakarta. Motion capture of Wayang Golek Menak Yogyakarta dance used motion capture (mocap) method and Kinect as input device to detect motion. Kinect has higher performance than other device, namely, it is able to capture and trace motion or action of 3D objects (human and animal), non-intrusive and work with less lighting. However, system of Kinect motion capture (mocap) needs calibration of appropriate object capture space[1][2].

Biovision Hierarchy (BVH) data processing used Principal component Analysis (PCA) methods reducing of data matrix. Objective of PCA methods is to process computation into easier way to make further processing unmeet significant constraints. The method Principal Component Analysis (PCA) to reduce the dimension of the input image for face recognition and percentage of success of face recognition process in this study was 82.81%[3].

2. Literature Review

2.1 Motion Capture (mocap) Technique

Motion capture (mocap) is digital recording technique in motion of real objects such as human or animal that can be illustrated in animation computer character[4]. Procedure of motion capture is to extract motion of an object in real world in computer using a set of input devices, furthermore actor or performer does motion with a set of input devices with motion model where pattern has been determined according to story. Strengths of motion capture are that generated images are more complex with shorter production time and lower significant production process cost because time is minimized and process is more effective, generated motion is more natural and accurate, pursuant to natural motion of taken objects. Weaknesses of motion capture are that it needs specific hardware and software, price and application of input devices becoming constraints for small industries; and it needs accuracy in synchronizing character motion when taking motion. A motion capture of data is a representation of digital data from the motion capture technique actor or character. Digital data

obtained in the data format of motion or motion in the form of a position or orientation coordinates (points) position gestures at a certain time[4]. The data format mocap consists of the skeleton which is a representation of the movement of the character as a whole, bone as the basic entity of the skeleton that became the subject of transformation, Channel or Degree Of Freedom (DOF) as a parameter for transformation of bone (translation, rotation, orientation movement) and the frame as a collection of information channels / DOF for each bone in a pose[5].

2.2 Sensor Kinect

Sensor Kinect is a control technology in game introduced by Microsoft in November 2010. Kinect develops continuously for not only games but also robotic applications, virtual reality, health and various pattern identifications without requiring additional devices. Sensor Kinect was developed by software technology from Microsoft Game Studios and camera technology from Prime Sense. Camera technology of Kinect has performance to interpret body motion or gesture movement specifically without requiring control using hands-free, utilizing infrared projector, RGB camera and microchip to trace motion of objects in 3D format. Kinect has RGB camera and depth sensor facilities. Application of motion capture technique is to use results of Sensor Kinect capture generating motion data consistent with each motion of Wayang Golek Menak dancer in motion data formats such as Biovision Hierarchy (BVH). BVH is used as motion capture data format generated by Kinect, because it has support of compatible data format to be exported and imported in some 3D software applications[6]. BVH data format consists of two parts, information on hierarchical structure of bone and information on parameters of each channel.

2.3 Principal component Analysis (PCA)

Principal component Analysis (PCA) is a way to identify patterns of data and express them so that their similarities and differences can be seen. These patterns are useful to compress data, namely, to reduce size or dimension of data without losing many kinds of information[7].

PCA is statistic technique which can be used to explain structures of variances in a group of variables through some new variables where these new variables are mutually independent, and these are linear combinations of origin variables. Furthermore, the new variables are named PCA. Generally, the objective of PCA is to reduce dimension of data and to fulfill need of interpretation.

In each multiple-variant measurement (observation), principal component is linear combination of initial variables. Main objective of PCA is to reduce dimension of changes which are interrelated and have sufficient quantity of variables so that it is easier to interpret data[8].

Mathematically, PCA as orthogonal linear transforms data into new coordinate system so that biggest variance of any data projection will exist in first coordinate, the second biggest variance exists in second coordinate, and so on[9].

PCA is one way to identify patterns in the data and express it in a way that can be seen similarities and differences. This pattern is useful to compress data, which reduces the size or dimension of data without losing much of the information contained[7]. If it be defined a matrix A, with x an eigenvector and λ is the eigenvalue, then to get the eigenvector and eigenvalue can use any general equations of PCA :

$$Ax = \lambda x \tag{1}$$

$$(A - \lambda I)x = 0 \tag{2}$$

The principal components analysis would reduce the observational data into multiple sets of data so that information from all the data we can absorb optimally. Thus the principal component analysis can be viewed as the transformation of $X_1, X_2, \ldots X_p$. For example $X_1, X_2, \ldots X_p$ has a variance-covariance matrix $\Sigma = (\sigma^2 ij)$, $i = 1,2 \ldots p$; $j = 1,2, \ldots p$ and the Σ has eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. The first Principal Component expressed by PC_1 contains the greatest amount of total variation of the data. PC_1 as linear combinations of the variables X_i. ; $i = 1,2 \ldots p$

$$PC_1 = a_{11}X_1 + a_{12}X_{12} + \cdots + a_{1p}X_p \tag{3}$$

ISBN : 978-602-99849-2-7

Where a_{1i} chosen, so as to maximize the ratio of variance PC_1 to the total variance, with a barrier that$\Sigma {a_{1i}}^2 = 1$. The principal component regression formation through principal component analysis, there are two ways.First, the establishment of the main components based on the covariance matrix. Secondly, the formation of the main components is based on the correlation matrix.

2.3.1. The Principal Component Analysis Based Formed Covariance Matrix

Through the data source to be searched X_{nxp} variance covariance matrix Σ where elements are :

$$S_{jk} = \frac{1}{n-1}\sum_{j=1}^{p}(X_{ij} - X_j)(X_{ik} - X_k) \quad (4)$$

Then from the variance covariance matrix of the sought eigenvalues λ_i with $i = 1,2, \ldots p$, obtained from the determinant equation form :

$$\left|S - \lambda_i I\right| = 0 \quad (5)$$

eigenvalues of the vector-eigenvectors calculated by an equation $S_{ei} = \lambda_i ei$, $i = 1,2, \ldots p$.

2.3.2. The Principal Component Analysis Based Formed Correlation Matrix

The main components of the i-th; W_i formed by variables that have been standardized$Z' = (Z_1, Z_2, \ldots, Z_p)$ with $cov(Z) = \rho$defined as follows:

$$W_i = e_{i1}Z_1 + e_{i2}Z_2 + \cdots + e_{ip}Z_p \qquad i = 1, 2, \ldots, p \quad (6)$$

2.4 Wayang Golek Menak Yogyakarta

Wayang Golek Menak show in Yogyakarta and surrounding reached glory in approximately 1950s, pioneered by Ki Widiprayitna[10]. Ki Widiprayitna was Dalang Wayang Golek and he was also known as wayang kulit maker. Wayang Golek show currently becomes media to present various moral messages, entertainments, advices, and announcements for people. Wayang Golek Menak has given economic, spiritual and social-political contributions to the people. As appreciation of Wayang Golek Menak story attraction, Sultan Hamengku Buwana IX as Yogyakarta Sultanate King immortalized story plots of Wayang Golek Menak in a set of drama art motions or ballets known as beksa golek Menak or golek Menak dance. Distinctive-typical characteristic of beksa golek Menak is lied in strength of dance motion including elements of self-defense, fingered hand palm, and firm dance motion. Ballet beksa Golek Menak a transformation of the story puppet show Menak into works of art and culture that has a value system of motion, meaning and philosophy are high. In terms of movement on the motion basically imitating the pupper Golek menak dance. Benchmark standard of Golek Menak dance referring to the Javanese dance style of Yogyakarta, which is modified with emphasis on the hull base motion, motion tolehan head, hands and feet[11]. Puppet Golek Menak dance is a form of transformation of the storyline puppet Golek Menak show. This dance is the creation Sultan Hamengku Buwono IX combining puppet Golek performances with classical Javanese dance which was then named Beksa Golek Menak[12]. Wayang Golek Menak motion of dance includes series of Sabetan (Tangkep Asta, Tancep, Jogetan Bapang Menak), Nyrimpet Maju, Ulap–ulap, Muryani Busana (Atrap jamang, Usap Rawis, Ngingset Udet), Lampah Sekar, Pencak Silat Gaya Minang and Peperangan[11].

3. Material & Method

3.1. Data

Motion capture (mocap) of Sensor Kinect are motion data of Biovision Hierarchy (BVH) of Wayang Golek Menak Yogyakarta dancers. Motion data of Biovision Hierarchy (BVH) generate data of motion tensor data cartesius (x, y, z) and then converted into sphere ($h\theta, h\varphi, hr$).

ISBN : 978-602-99849-2-7

3.2. Method

This study observed and interviewed directly with experts of Wayang Golek Menak dancers for basic motions of the Wayang Golek Menak dance. The following is process of motion data processing of Wayang Golek Menak Yogyakarta dance:

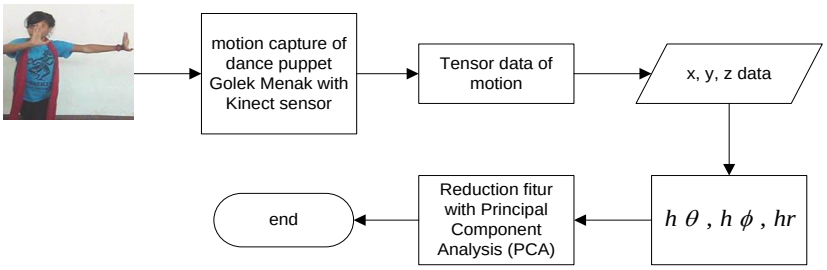


Figure 1. Flowchart of Research

4. Results and Discussion

The results of motion capture (mocap) of Sensor Kinect are motion data of Biovision Hierarchy (BVH) of Wayang Golek Menak Yogyakarta dancers.

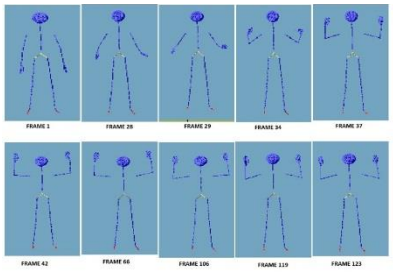


Figure 2. Frame of Body Motion Position

Wayang Golek Menak dance consists of 61 joint but just taken a major parts are 18 joint, time to capture dance ± 4 second, and count of frames between 140 – 180. The dance data consists of a skeleton name, nestdepth, parent, offset, Nchannel, order, Dxyz, Rxyz, and trans.

Table1. Tensor Data Motion X, Y, Z of Wayang Golek Menak Dance

Frame	1 HIPS X	2 HIPS Y	3 HIPS Z	4 LEFT HIP X	....	54 HEAD Z
Data *x, y, z* of Wayang Golek Menak Dance – Jogetan (149x54)						
1	-139,144	49,1302	-653,392	-132,824	.....	-655,168
2	-139,144	49,1302	-653,392	-132,824	.....	-655,168
3	-138,808	48,9704	-653,233	-132,524	.....	-654,472
4	-138,775	48,9436	-653,22	-132,496	.....	-654.41
.....up to 149 frames.....						
Data *x, y, z* of Wayang Golek Menak Dance – Sabetan (151x54)						
1	-132,285	67,4904	-688,208	-124,79	.....	-691,588
2	-132,332	67,3781	-688,121	-124,864	.....	-691,159
3	-132,332	67,3781	-688,121	-124,864	.....	-691,159
4	-132,339	67,3713	-688,113	-124,878	.....	-691,021

ISBN : 978-602-99849-2-7

.....up to 151 frames.....

Motion data of BVH generate data of motion tensor data cartesius (x, y, z) and then converted into sphere ($h\theta, h\varphi, hr$).

$$\theta = tan^{-1}(y, x) \quad (7)$$

$$\varphi = tan^{-1}(z, \sqrt{x^2 + y^2} \quad (8)$$

$$r = \sqrt{x^2 + y^2 + z^2} \quad (9)$$

Table 2. Tensor Data Motion $h\theta, h\varphi, hr$ of Wayang Golek Menak Dance

Frame	1 HIPSθ	2 HIPS φ	3 HIPS r	4 LEFT HIPθ	....	54 HEAD r
Data positions hr, $h\varphi$, $h\theta$ of Wayang Golek Menak Dance – Jogetan (149x54)						
1	2,802168	-1,34868	669,8477	2,802168	.....	669,8477
2	2,802168	-1,34868	669,8477	2,802168	.....	669,8477
3	2,802433	-1,34917	669,6112	2,802433	.....	669,6112
4	2,80253	-1,34922	669,5897	2,80253	.....	669,6112
.....up to 149 frames.....						
Data positions hr, $h\varphi$, $h\theta$ of Wayang Golek Menak Dance – Sabetan (151x54)						
1	2,669826	-1,35827	704,0486	2,669826	.....	704,0486
2	2,670643	-1,35825	703,9616	2,670643	.....	703,9616
3	2,670643	-1,35825	703,9616	2,670643	.....	703,9616
4	2,670707	-1,35825	703,9616	2,670707	.....	703,9546
.....up to 151 frames.....						

Tensor data shpere($h\theta, h\varphi, hr$)of dance motion of Wayang Golek Menak have big matrix dimension so that tensor data of dance motion are reduced. Objective of this tensor data reduction is to make computation process become simple. Method used to reduce tensor data is PCA. The result gave eigenvalue, because main objective of PCA is to obtain eigenvalue.

Table 3. Dance Motion Tensor Data Reduction Method Results PCA

Frame	Jogetan	Sabetan
1	4,011922	1,187172
2	4,011922	1,136066
3	3,674163	1,136066
4	3,54979	1,12832
5	3,54979	1,132112
6	3,54979	1,132112
7	3,54979	1,132112
8	3,54979	1,132112
9	3,54979	1,138808
10	3,54979	1,138808
.....up to 149 data.....		
149	2,51913	2,297637

ISBN : 978-602-99849-2-7

Dimension matrix of tensor data motion after reduction used PCA method from matrix dimension 149x54 and 151x54 become 149x1 matrix dimension for each one of dance data.

5. Conclusion

a. Motion capture of Wayang Golek Menak Yogyakarta dance generates motion data withBiovision Hierarchy(BVH) format compatible to be imported and exported in some 3D software forms.
b. BVH motion data are motion position matrix with 149x54 and 151x54 dimension, furthermore the matrix is reduced by Principal component Analysis (PCA) into matrix with 149x1dimension for each one of dance data.
c. Reduction of matrix dimension is to ease process of computer-based computation in next stage.
d. The results of the PCA reduction feature can be used to process motion classification.

6. Reference

[1] Gabel M., Gilad-Bachrach, R., Renshaw, E. and Schuster, A., "Full Body Gait Analysis with Kinect" Department of Computer Science, Technion – IsraelInstitute of Technology, 2012
[2] Held, Robert, Ankit Gupta, Brian Curless, and Maneesh Agrawala, "3D Puppetry: a Kinect-based Interface for 3D Animation" In Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology, 423–434, UIST '12. New York, NY, USA: ACM, 2012
[3] Pratiwi, Dian Esti, Agus Harjoko, "Implementasi Pengenalan Wajah Menggunakan PCA (*Principal Component Analysis*)", Study of Electronics and Instrumentation Program, Departement of Computer Science and Electronics, Gadjah Mada University, 2013
[4] Guerra-filho, Gutemberg B., "Optical Motion Capture: Theory and Implementation" Journal of Theoretical and Applied Informatics (RITA 12: 61–89), 2005
[5] Anggarwal, J.K., and Q. Cai, "Human Motion Analysis : a Review", IEEE Nonrigid and Articulated Motion Workshop, Proceedings, 90-102. doi:10.1109/NAMW.1997.609859, 1997
[6] Liu, Ming, Zhenjiang Miao, Jia Li, and Zhan Xu, "Design and Implementation of a Vision-Based Motion Capture System" In Fifth International Conference on Image and GraphicsICIG '09, 716–721. doi:10.1109/ICIG.2009.187, 2009
[7] Smith, L.I., "A Tutorial on Principal Component Analysis", New York : Cornell University, 2002
[8] Johnson, R.A. & Wichern, D.W., "Applied Multivariate Statistical Analysis,5th edition", Pearson Education International, 2002
[9] Jolliffe, I.T., "Principle Component Analysis", Aberdeen : Springer, 2002
[10] Sukistono, Dewanto, "Wayang Golek Menak Yogyakarta Bentuk dan Struktur Pertunjukkannya", Gadjah Mada University, 2013
[11] Supriyanto, "Bentuk Penyajian Pembakuan Tari Golek Menak Gaya Yogyakarta : Laporan Penelitian", High School Art of Indonesia (STSI), 1992
[12] Susiyanto, "Cerita Menak : Warisan Budaya Islam di Indonesia", http://susiyanto.com, accessed 2013 August 30th

ISBN : 978-602-99849-2-7

The Ability The Chi Squares StatisticsTo Rejecting The Null Hypothesis on Contingency Tables 2x2

Jaka Nugraha

Statistics Department, Islamic University of Indonesia, Yogyakarta

jnugraha@uii.ac.id

Abstract: The Chi-Square statistic is the primary statistic used for testing the statistical significance of the cross-tabulation table. The Chi-Square test is based on an approcimation that works best when the expected frequencies are fairly large. There is still a lack of consensus on the optimum method most texts recommend the use of the chi squared test and there is disagreement on the boundary between 'large' and 'small' sample sizes. So in this paper discussed the limitations and requirements for proper use, especially related to the minimum number of sample requirements for table 2x2.
The study focussed on four versions of the Chi Squared test the tests were: Pearson's Chi-Squared test, Yates's Chi-Squared test, Likelihood Ratio Chi-Square test and '*N* - 1' Chi-Squared test. From each of the test statistic equations, composed functions that can explain the relationship between the size of the sample (m), the probability of each category (k) and the amount of difference can be detected (d). Concluded that it takes a larger sample when the value of k nearest 0.5 and smaller when approaching a value of 0 or 1. The four test statistics that have the same pattern, but the statistics Yate's need a sample of the largest while the three others the sample sizes are nearly equal.
Keywords: Pearson Chi-Square; Likelihood Ratio; Normal Distribution;'N - 1' Chi-Squared

1. Introduction

TheContingency table, also known as Cross-tabulation, is a joint frequency distribution of cases based on two or more categorical variables. Contingency table analysis is a common method of analyzing the association between two categorical variables. There are two separate sampling strategies lead to the contingency table analysis. First, Test of Independence. A single random sample of observations is selected from the population of interest, and the data are categorized on the basis of the two variables of interest. Second, Test for Homogeneity [1]. Separate random samples are taken from each of two or more populations to determine whether the responses related to a single categorical variable are consistent across populations.

The Chi-Square statistic is the primary statistic used for testing the statistical significance of the cross-tabulation table. The two-way table is set up the same way regardless of the sampling strategy, and the chi-square test is conducted in exactly the same way. Chi-Square tests whether or not the two variables are independent. The chi-square test is based on an approcimation that works best when the expected frequencies are fairly large. No expected frequency should be less than 1 and no more than 20% of the expected frequencies should be less than 5. For tables larger than 2x2, the chi-square distribution with the appropriate degrees of freedom provides a good approximation to the sampling distribution of Pearson's chi-square and the Likelihood Ratio Chi-Square. However, the Chi-Square statistic is only approximated by the chi-square distribution, and that approximation worsens with small expected frequencies. When we have very small expected frequencies, the possible values of the chi-square statistic are quite discrete. The general rule is that the smallest expected frequency should be at least five. However Cochran (1952), who is generally considered the source of this rule, acknowledged that the number "5" seems to be chosen arbitrarily [3].

Statistical test of 2 × 2 tables have been under discussion for a hundred years and dozens of research papers have been devoted to them. However, there is still a lack of consensus on the optimum method most texts recommend the use of the chi squared test for large sample sizes and the Fisher-Irwin test for small sample sizes, but there is disagreement on the boundary between 'large' and 'small' sample sizes [4]. Because the Pearson Chi-Square statistics and Likelihood RatioChi-Square statistics are based on a normal distribution approach, so in this paper discussed the limitations and requirements for proper use, especially related to the minimum number of sample requirements for table 2x2. Two distinct research designs can give rise to 2×2 tables : Model I (Row totals fixed, column totals free to vary), Model II (Both row & column totals free to vary). Model 1

where we wish to test the hypothesis that two proportions are equal. Model II where we wish to test the hypothesis of independence of two variables of classification.

The validity of the chi-square test depends on both the sample size and the number of cells. Several rules of thumb have been suggested to indicate whether the chi-square approximation is satisfactory. One such rule suggested by Cochran (1954) says that the approximation is adequate if no expected cell frequencies are less than one and no more than 20% are less than five. Because of the expected cell frequency criterion in the second sampling strategy, it may be necessary to combine similar categories to lessen the number of categories in your table or to examine the data by subcategories

2. Contingency Table 2x2 Analysis

A number of experiments involve binary outcomes (i.e., 1 and 0, yes and no). Typically, these occur when you are observing the presence or absence of a characteristic such as a disease, flaw, mechanical breakdown, death, failure, and so on. The analysis of the relationship between two bivariate categorical variables results in a 2 × 2 crosstabulation table of counts. There are *2* x *2* possible combinations of responses for these two variables. The *2* × *2* crosstabulation or contingency table has *2* rows and *2* columns consisting of *2* × *2* cells containing the observed counts (frequencies) for each of the *2* × *2* combinations. The observed frequencies are presented in Table 1.

Tabel 1. 2x2 Contingency table

	B	B^c	Total
A	a	b	m_1
A^c	b	d	m_2
Total	n_1	n_2	N

This type of analysis is called a contingency table analysis and is usually accomplished using a chi-square statistic that compares the observed counts with those that would be expected if there were no association between the two variables. Two separate sampling strategies lead to the chi-square contingency table analysis :

1. *Test for Homogeneity (Model I).* Separate random samples are taken from each of two or more populations to determine whether the responses related to a single categorical variable are consistent across populations. In this setting, you have a categorical variable collected separately from two or more populations. The hypotheses are as follows:
   H_0: The distribution of the categorical variable is the same across the populations.
   H_a: The distribution of the categorical variable differs across the populations.
2. *Test of Independence (Model II).* A single random sample of observations is selected from the population of interest, and the data are categorized on the basis of the two variables of interest. In this case, you have two variables and are interested in testing whether there is an association between the two variables. Specifically, the hypotheses to be tested are the following:
   H_0: There is no association between the two variables.
   H_a: The two variables are associated.

The two-way table is set up the same way regardless of the sampling strategy, and the chi-square test is conducted in exactly the same way. The only real difference in the analysis is in the statement of the hypotheses and conclusions. The chi squared test were (1) Pearson's Chi-Squared test (2) Yates's Chi-Squared test (3) Likelihood Ratio Chi-Square test (4) The '*N* - 1' Chi-Squared test.

The original chi-square test, often known as Pearson's chi-square, dates from papers by Karl Pearson in the earlier 1900s. The test serves both as a "goodness-of-fitt" test, where the data are categorized along one dimension, and as a test for the more common "contingency table", in which categorization is across two or more dimensions. The standard Pearson chi-square statistic for 2x2 Contingency table is defined as

$$\chi^2=(ad - bc)^2N / m_1m_2n_1n_2 \qquad (1)$$

Pearson's chi-square statistic is not the only chi-square test that we have.

The likelihood ratio chi-square builds on the likelihood of the data under the null hypothesis relative to the maximum likelihood. It is defined as

$$\chi^2 = 2*(a*\log(aN/m_1n_1)+b*\log(bN/m_1n_2)+c*\log(cN/m_2n_1)+d*\mathrm{Log}(dN/m_2n_2)) \qquad (2)$$

ISBN : 978-602-99849-2-7

If an expected frequency is lower than five, you have alternatives: Yates correction (Yates's Chi-Squared test) and the *N* - 1 chi-square test. [2]

Yates' correction (Yates, 1934) is equivalent to Pearson's chi-square but with a continuity correction.In cases where an expected frequency is below 5, Yates' correction brings the result more in line with the true probability.

$$\chi^2 = (|ad - bc| - \tfrac{1}{2}N)^2 N / m_1 m_2 n_1 n_2 \quad (3)$$

The *N - 1 chi-square test* is another option. Campbell (2007) carried out a very large sampling study on 2x2 tables comparing different chi-square statistics under different sample sizes and different underlying designs. He found that across all sampling designs, a statistic suggested by Karl Pearson's worked best in most situations. The statistic is defined as

$$\chi^2 = (ad - bc)^2 (N - 1) / m_1 m_2 n_1 n_2 \quad (4)$$

3. Calculation ofMinimumSamples Size

Assuming asample size of2mbyeachpopulation ofm, observational datacan be presented inTable2.P(A | B) =khandP(Ac |B) =xwith0<k <1 and0<x <1.

Table2. Tableof contingencyfor thesampleP(A | B) =k

	B	B^c	Total
A	Km	(1-k)m	m
A^c	Xm	(1-x)m	m
Total	(k+x)m	(2-k-x)m	2m

Based onchi squarestatisticformulaas inequation(1) s/d(4), going tolook forthe minimumnumber of samplesthat candetect the differencebetween theproportion ofpopulationI (A) anda population of2(A^c). The hypothesis is

$$H_0: P(A|B) = P(A^c|B) \text{ vs } H_0: P(A|B) = \neq P(A^c|B)$$

The minimumsample sizeto be able to$\text{reject}H_0$withsignificance levelαforeachtest statisticsareasfollows:

1. Pearson's Statistics

$$\frac{(km(1-x)m - xm(1-k)m)^2 2m}{m.m(k+x)m(2-k-x)m.} \geq \chi_\alpha^2$$

$$\Leftrightarrow \quad m \geq \frac{\chi_\alpha^2 (k+x)(2-k-x)}{2(k-x)^2}$$

$$F(k;\alpha;x) = \frac{\chi_\alpha^2 (k+x)(2-k-x)}{2(k-x)^2} \quad (5)$$

$F(k;\alpha;x)$ is a function thatshowsthe minimumsamplerequired tobe able to$\text{reject}H_0$.

2. Yate's Statistics

$$\frac{(|km(1-x)m - xm(1-k)m| - m)^2 2m}{m.m(k+x)m(2-k-x)m.} \geq \chi_\alpha^2$$

$$\Leftrightarrow \frac{(|k-x|m-1)^2 2}{(k+x)(2-k-x)m.} \geq \chi_\alpha^2$$

$$\Leftrightarrow \quad 2(k-x)^2 m^2 - (4|k-x| + (k+x)(2-k-x)\chi_\alpha^2)m + 2 \geq 0$$

$$\Leftrightarrow \quad m \geq \frac{4|k-x|(k+x)(2-k-x)\chi_\alpha^2 + \left(\left(4|k-x| + (k+x)(2-k-x)\chi_\alpha^2\right)^2 - 16(k-x)^2\right)^{1/2}}{2(k-x)^2}$$

$$F(k;\alpha;x) = \frac{4|k-x|(k+x)(2-k-x)\chi_\alpha^2 + \left(\left(4|k-x| + (k+x)(2-k-x)\chi_\alpha^2\right)^2 - 16(k-x)^2\right)^{1/2}}{2(k-x)^2} \quad (6)$$

ISBN : 978-602-99849-2-7

3. Campbell's Statistics

$$\frac{(km(1-x)m - xm(1-k)m)^2(2m-1)}{m.m(k+x)m(2-k-x)m.} \geq \chi_\alpha^2$$

$$\Leftrightarrow \quad m \geq \frac{\chi_\alpha^2(k+x)(2-k-x))}{2(k-x))^2} + \frac{1}{2}$$

$$F(k;\alpha;x) = \frac{\chi_\alpha^2(k+x)(2-k-x))}{2(k-x))^2} + \frac{1}{2} \qquad (7)$$

4. Likelihood's Statistics

$$ln\left(\left(\frac{2km}{(k+x)m}\right)^{2km}\left(\frac{2(1-k)m}{(2-k-x)m}\right)^{2(1-k)m}\left(\frac{2xm}{(k+x)m}\right)^{2xm}\left(\frac{2(1-x)m}{(2-k-x)m}\right)^{2(1-x)m}\right) \geq \chi_\alpha^2$$

$$\Leftrightarrow \quad 2m.ln\left(\left(\frac{2k}{(k+x)}\right)^{k}\left(\frac{2(1-k)}{(2-k-x)}\right)^{(1-k)}\left(\frac{2x}{(k+x)}\right)^{x}\left(\frac{2(1-x)}{(2-k-x)}\right)^{(1-x)}\right) \geq \chi_\alpha^2$$

$$\Leftrightarrow \quad m \geq \frac{\chi_\alpha^2}{2.ln\left(\left(\frac{2k}{(k+x)}\right)^{k}\left(\frac{2(1-k)}{(2-k-x)}\right)^{(1-k)}\left(\frac{2x}{(k+x)}\right)^{x}\left(\frac{2(1-x)}{(2-k-x)}\right)^{(1-x)}\right)}$$

$$F(k;\alpha;x) = \chi_\alpha^2\left(2.ln\left(\left(\frac{2k}{(k+x)}\right)^{k}\left(\frac{2(1-k)}{(2-k-x)}\right)^{(1-k)}\left(\frac{2x}{(k+x)}\right)^{x}\left(\frac{2(1-x)}{(2-k-x)}\right)^{(1-x)}\right)\right)^{-1} \qquad (8)$$

4. Properties of Functions F(*K;α;X*)

The properties ofthe function F(k; α; x) illustrated using Figure 1 and Figure 2 by taking the value of α=0.05. Figure 1 and Figure 2 describes the effects of changes in the value ofk, namely P(A∩B) and the difference between the value of P(A|B) and P(A^c|B).

$\chi 2$= 3,8415 and d=|k-x|

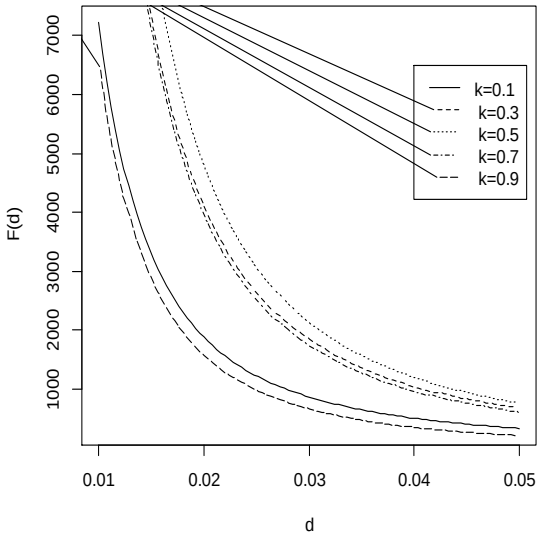


Figure 1. GraphF(k; α; x) on somevaluek

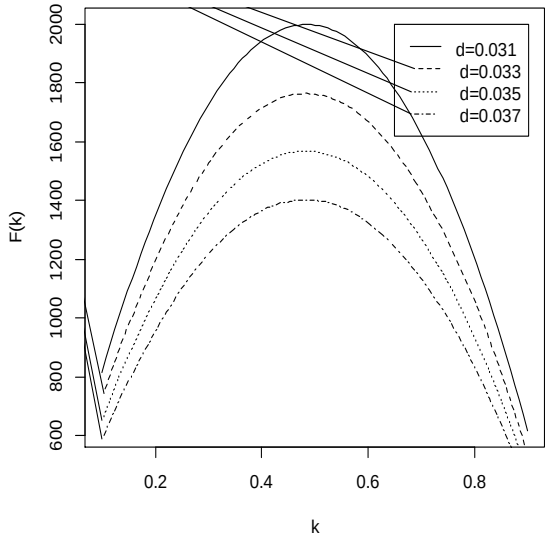


Figure 2. GraphF(k; α; x) on somevalued

As in Figure 1, the greater the ability to detect a difference (d smaller) the number of samples required is also greater. Figure 2 illustrates that the sample also needs greater when k approaches 0.5.
Furthermore, a comparison of several test statistics can be explained in Table 3 and Figure 3 to Figure 6. Table 3 is an example in the case of k = 0.5 and k = 0.4 which shows that (a) The number of

ISBN : 978-602-99849-2-7

samples required at Pearson's statistics, Likelihood's statistics and Champbell's statistics relatively similar (b) Yate's statistics require larger samples than the three other statistics. (c) the greater the ability to detect a difference (d smaller) the number of samples required also getting bigger.

Table 3. Tablecontingencyatk=0.4 andk=0.5

Nilai d	k=0,4				k=0,5			
	Pearson	Yate	Likelihood	Champbell	Pearson	Yate	Likelihood	Champbell
0,01	18514	19454	18515	18514	19206	20173	19206	19206
0,02	4647	5117	4647	4646	4800	5283	4801	4800
0,03	2073	2387	2013	2073	2133	2454	2133	2132
0,04	1170	1406	1171	1170	1199	1440	1200	1199
0,05	752	941	752	751	767	959	767	767

In general, to be able to detect the differences of the particular, Yate's statistics require larger samples than the likelihood's statistics, Campbell's statistics and Pearson's statistics. As in Figure 3 and Figure 4, likelihood's statistics, Campbell's statistics and Pearson's statisticshaverelatively similarproperties.. The four statistic has the same pattern (as a function of d or k).

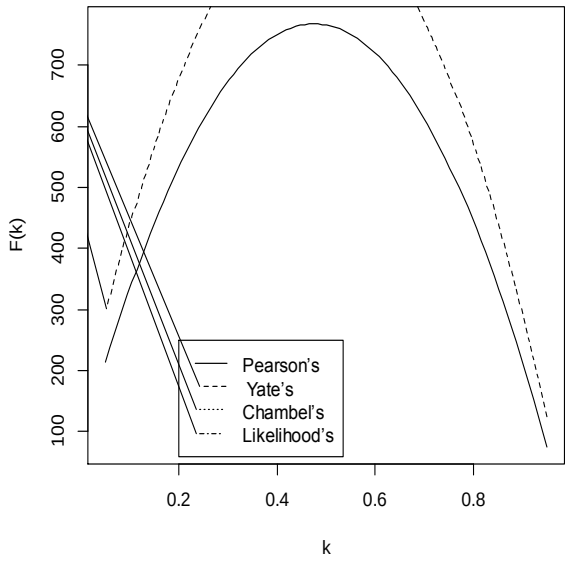


Figure 3.The sample sizeatd=0.05

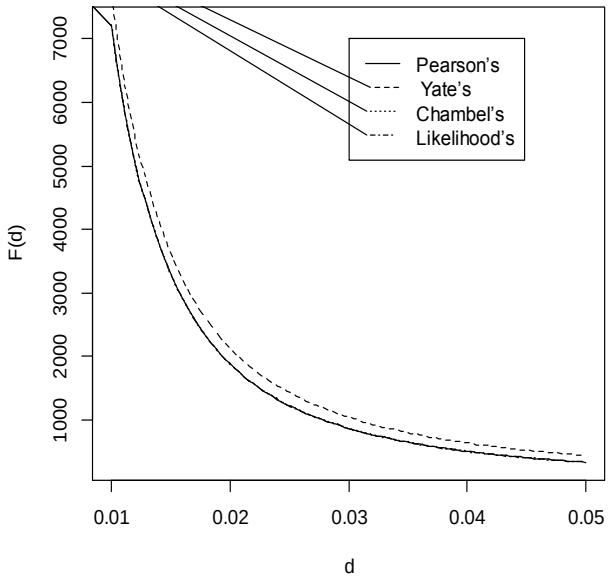


Figure 4.The sample sizeatk=0.1

In moredetail,to determine thestatistical differenceslikelihood's statistics, Campbell's statistics and Pearson's statistics taken short intervals. From Figure 5 and Figure 6, it appears that the order from the smallest sample sizer espectively : likelihood's statistics, and Pearson's statistics and Campbell's statistics.

ISBN : 978-602-99849-2-7

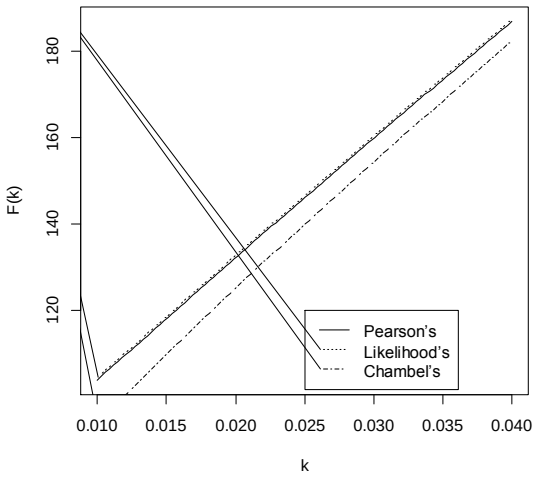


Figure 5.The sample size at k∈ (0.01,0.04)

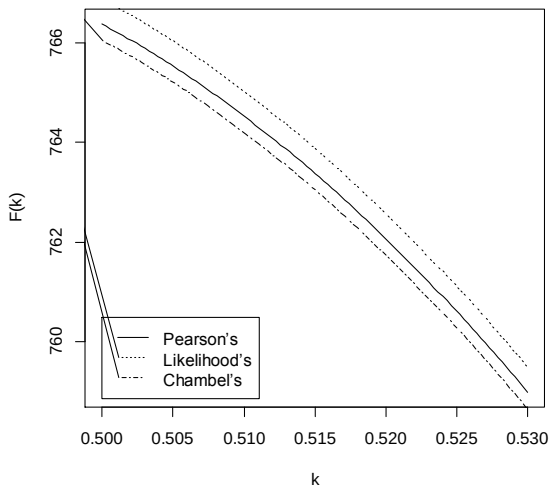


Figure 6.The sample size at k∈ (0.5,0.53)

5. Conclusion

Fromthe above discussionit can be concluded that (a) The higher thesample size, the ability todetect differences (rejecting H_0), the better. (b) The ability ofYate's statisticslowerbe compared toPearson's statistics, Likelihood's statisticsandChampbell's statistics. (c) Ability Likelihood's statistics relatively similar to the Pearson's statistics and better than the Champbell's statistics. should give a summary of:

Acknowledgement. The author would like to thank to Islamic University of Indonesia for supporting this research..

References

[1] Agresti A (2002), *Categorical Data Analysis*, John Wiley and Son
[2] Campbell, I. (2007), Chi-squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations. *Statistics in Medicine*, 26, 3661-3675,.
[3] Cochran, W. G. (1952), The χ^2 test of goodness of fit. Annals of Mathematical Statistics, 25, 315-345.
[4] Yates F. (1984), Tests of significance for 2 × 2 contingency tables (with discussion) *Journal of the Royal Statistical Society Series A*; 147: 426-463.

ISBN : 978-602-99849-2-7

Predictive Simulation of Amount of Claims with Zero Truncated Negative Binomial Distribution

Heri Kurniawan

[1] Lecture at Institut Teknologi Harapan Bangsa
email : heri_kurniawan@ithb.ac.id

Abstract: Modeling frequency and amount of claims are usually separate, and ultimately the compound will be obtained using the model predictions of amount of claims. Unlike the case with the Annual Expenditure Model, modeling of the claims made in the model using linear mixed models with negative binomial distribution in the frequency of claims. In fact, amount of claim can not be calculated if the frequency of claims is not the case, so Distribution of the frequency of claims should be modified in the event of a claim doesn't in the case using zero truncated Negative Binomial distribution. Calculation of the claims that do not conform to the characteristics of the data will produce imprecise predictions that would be detrimental to the health care institutions eg hospitals in filing a claim with the insurer. From the simulation results with only about 3% observational data, the result prediction relatively closed with real data for one year (one period of observation).
Keywords: Zero Truncated Negative Binomial; Annual Expenditure Model; Two Part Model.

1. Introduction

The health care expenditures in hospitals becomes very important with some health insurance program, such as Jamkesmas, Jamkesda, Kartu Jakarta Sehatandsocial security programs organized by theSocial Security Organizing Body(BPJS).

However, there are problems in implementation, especially in the process of claims against service providers. One possible cause is not exactly a prediction of the costs to be borne than the claims file dby the end of the program implementation, orthere is abig differencethe fundswill be reimbursed by the government with the calculations performedby the hospital, so it should be the process of adjustment (adjustment) in allocationof funds per person perservice.

Therefore, modeling andprediction the health care expenditures becomes very important to estimate how much the cost that should be provided to implement the program.One of the model that can be usedin modeling the health care expenditures, is Two Part Model (TPM). In this model there are two components to be modeled is the frequency of the (many) users of health services, and a large health care costs. The first part of the TP Misto model users and non-users of the service through probit or logistic regression models. Furthermore, the second partis to amount of claims model [2].

Toimprove TPM, Frees, et.al [3] propose theAnnual Expenditure Model (AEM) as special case on the TPM using data frequencies negative binomial distribution. The assumptions used are observation data obtained is complete for one year of observation, therefore, required a simulation to make predictions with observational datais not complete untilthedesired periodin this case isan annual. However, in practicethe claimscan not be predictedforthe datafrequency ofwhichis zero, so that the necessary predictive simulation of the claims with the data frequency ZeroTruncated negative binomial distribution (ZTNB).

This paper is organized as follows. In Section 2, we introduce some model in amount of claims modeling, and simulation. Section 3 describes data that used in this research, assumption testing and fitting distribution. Section 4 desribed result and discussion, and Finally, we describe conclusion and future work in Section 5.

2. Literature Review

2.1 Modeling of the Claims

To determine the amount of the claim needs to be done modeling that involves elements of the frequency rnumber of visitsutilization of healthservices, and a large element of the cost for each time of the visit. So the modeling of the claimsare multiplying frequency models withcost model (amount of claims model).Modeling development of the claims such as one part model (OPM), two partmodel (TPM), three part model (TRPM), and four part model (FPM) as well as a comparison of these models can be seen in Duan, et.all[2]. In this paper will be focused on Annual Expenditures Model(AEM) proposed by Frees, et.all[3]which is aspecial case of Two partsModel(TPM) whoseimplementation is still to beimproved, especially if the modeling is done not involve frequency data=0.Thus, in this paper described a solution to overcome itis to perform predictive simulation of the claimsusing Zero Truncated Negative Binomial distribution.

2.1.1 Modeling of Frequency

For Frequency Model (model first part) in the AEM used negative binomial regression with the frequency of occurrence N_i as the dependent variable and $\boldsymbol{x_{1i}}$ as an independent variable vector with the regression coefficients $\boldsymbol{\beta_i}$. This modelcan be describedas follows:

Random variables N_i which is the negative binomial distribution with parameters rand p, have the probability density function

$$Pr(N = k|r,p) = \binom{k+r-1}{r-1} p^r (1-p)^k \quad (1)$$

in Generalized Linear Model (GLM) , negative binomial probability distribution can be expressed within the parameters parameter μ and κ, with $\mu = \frac{r(1-p)}{p}$, and $\kappa = \frac{1}{r}$, so that equation (1) can be expressed as follows [1]:

$$Pr(N = k|\mu,\kappa) = \frac{\Gamma\left(k+\frac{1}{\kappa}\right)}{k!\Gamma\left(\frac{1}{\kappa}\right)} \left(\frac{1}{1+\kappa\mu}\right)^{\frac{1}{\kappa}} \left(\frac{\kappa\mu}{1+\kappa\mu}\right)^k \quad (2)$$

From the equation (2) with $E[N] = \mu$, $Var(N) = \mu(1+\kappa\mu)$, so that the equation model from negative binomial regression is :

$$ln(\mu_i) = ln(n) + \boldsymbol{x_{1i}}'\boldsymbol{\beta_1} \quad (3)$$

2.1.2 Modeling of Amount of Claims

Models for amount of claims follow OPM, but conditional on the event r_i=1, so the response variable being modeled is $(Y_{ij}|N_i \geq 1)$. Vector of independent variables is $\boldsymbol{x_2}$, the $\boldsymbol{x_2}$ is asubset of $\boldsymbol{x_1}$. The amount of claims models using mixed linear regression model developed byMcCulloch[5], which is expressedby the following equation[3].

$$\ln(Y_{ij}) = \alpha_i + \boldsymbol{x_{2i}'}\boldsymbol{\beta_2} + N_i\beta_N + \varepsilon_{ij} \quad (4)$$

with α_i = *intrafamily correlation* atorau *Intraclass correlation* (ICC); $\varepsilon \sim N(0, \sigma_\varepsilon^2)$*i.i.d* between observations; and $\beta_N \sim N(0, \sigma_{\beta_N}^2)$*i.i.d* between *family*

2.2 Amount of Claims Prediction with ZTNB

To predict of the health care expenditures or amount of claims, can be derived from equation (4), so that would be obtained:

$$\hat{Y}_{ij} = \exp(\hat{\alpha}_i + \boldsymbol{x_{2i}'}\hat{\boldsymbol{\beta}}_{\boldsymbol{2}} + N_i\hat{\beta}_N) \quad (5)$$

the health care expenditures Predictionfor one year(one period) using AEM is a function of N_i

$$\hat{S}_i(N_i) = \hat{Y}_{ij}N_i \quad (6)$$

However, this prediction can not be done at the beginning of the year because N_i is not unknown for one year of observation. To overcome these problemsit must find the expected value of N_i from amount of claims for one year (a period) showed the following results

$$\hat{S}_i = \exp\left(\hat{\alpha}_i + \boldsymbol{x_{2i}'}\hat{\boldsymbol{\beta}}_{\boldsymbol{2}}\right) . M_{N_i}'\left(\hat{\beta}_N\right) \quad (8)$$

ISBN : 978-602-99849-2-7

$M'_{N_i}(\hat{\beta}_N)$ known as the first derivative of the moment generating function$N_i$at the point $\hat{\beta}_N$. As the random variableN_i is adiscrete random variablewitha negative binomial distribution which classes (a, b, 1), So thatthe first derivative ofthe moment generating functionaboveis:

$$M'_{N_i}(\hat{\beta}_N) = \sum_k N_i . e^{Ni\hat{\beta}_N}\, \widehat{Pr}_i(N_i = k),\ k = 1,2,\ldots \quad (9)$$

$\widehat{Pr}_i(N_i = k)$ is the probability obtained from the zero truncated negative binomial distribution[4]. There are problems in calculating probability with zero truncated negative binomial distribution because observational data that should be complete within one year of observation, is precisely not fulfilled. This happens because the new data collected in the current year, so it is unknown how many times a patient will use health care services for one year or the period of observation. Therefore we need the simulation of frequency during the period of observation for zero truncated negative binomial so the claims can be predicted. The process of generating random numbers distributed ZTNB can be seen in the appendix.

3. Material & Methodology

The data used in this research is data claims paid by the insurance companyXYZ in the period of filing and payment of claims in 2012. Data claims paid by the insurance company XYZ is divided into three groups of data. Due to the first data group has fewer exogenous variables then used the data group II and group III.

Tabel 3.1 Observation Data for Each Group

	Group I	Group II	Group III
Number of Participants	16377	7881	858
Participants filing a claim	1418	865	26
Many claims were filed	1844	1305	53

Then performed the cleaning process the data with the following criteria which the participant group I excluded due to differences in field with group II and III, the participants make a claim one or more but not approvedby the company, the participants wer enot recorded the data of age, participants were not recorded the type of illness. Frequency of Claims has a value of0,1,2,3. the composition of the variable frequency ofclaims(N) ieN=0(89.9%) andN≥1(10.1%).

Tabel 3.2 Frequency of Claims

N	0	1	2	3	4	5	6	7	8	24	25	28	Total
Frequency	7854	642	158	43	22	8	5	2	2	1	1	1	8739
Percentage	89.9	7.3	1.8	0.5	0.3	0.1	0.1	0	0	0	0	0	100

Modeling of the claims is modeling based on the frequency of claims, assuming a negative binomial distribution. So that the data collected should be pass of Goodness of fit test and parameter estimation for negative binomial distribution. By using Matlab for the negative binomial distribution with parameters r and π obtained estimation results r = π = 0.148438 and 0.489505. in Generalized Linear Model (GLMs), the parameters of the binomial distribution negative is μ and κ, where E (N) = μ = (r (1-π)) / π and κ = 1 / r with Var (N) = μ (1 + κμ). So it can be calculated value E [N] = 0.154709 and Var [N] = 0.316051. by looking at the value of E [N] <Var [N], then the initial assumption that the data Frequency of Claims was proven as negative binomial distribution. But to test this assumption, then tested the Goodness of Fit (GoF) of the negative binomial distribution. GoF test performed using Pearson chi-square test. From the results of the test by removing extreme data obtained p-value = 0.072 which states that the data frequency of claims following the negative binomial distribution.

In addition tothe frequency of claims, large claims also have to meet the criteria of normality, so can meet the model assumption ofa linear mixed modeling. However, because the distribution of the claims are always skewed to the right, then the transformation process logarithm so data

ISBN : 978-602-99849-2-7

transformation results meet the criteria of normality(p-value = 0.066usingthe Kolmogorov-Smirnov testtest). After all the data fulfill to the modeling assumptions, modeling both frequency and large claims, the next step isto predict the frequency and the prediction of the claims.

4. Results and Discussion

Predictive simulation begins by generating random numbers for the frequency of claims with Zero truncated negative binomial distribution (ZTNB) withk=1,2,3, ... This is because the claims can only be modeled if the claim frequencyof at least once for a predetermined type of disease is stroke, high blood pressure, diabetes and kidney failure.

Tabel 3.3 The Comparative of Amount of Claims between Simulation Result and Real Data

Obs	N	hasil simulasi untuk satu tahun kedepan	Diprediksi dari data Real		
		Si_hat	Ln_y	Y	Si
1	1	6,301,100.38	15.62192	6088566	6,088,566.00
2	1	3,686,969.84	14.8437	2796000	2,796,000.00
3	2	17,782,796.29	15.7994	7271000	14,542,000.00
….	….	….	….	….	….
22	1	4,162,755.40	15.35453	4660000	4,660,000.00
23	1	7,535,073.98	16.03074	9163500	9,163,500.00
24	1	2,400,855.12	13.44445	690000	690,000.00
25	1	5,317,410.86	15.39922	4873000	4,873,000.00
26	2	18,052,021.48	14.4226	1835080	3,670,160.00

Simulations conducted to determine the prediction of amount of the claims for the next year, if the data is observed not until less than one year, therefore the data observed frequency of claims is random. Here are the results of a simulation conducted on the number of participants of the third group of participants who filed at least claims assumed that the data has not been up to one year of observation. From the table above can be seen the results of simulations carried out by generating 10,000 random numbers distributed Z T N Bandreplication processes performed10 times(r =10). The simulation processwill produce better results if done generating random numbers>10000 and replication more reproduced so that it will get better predictionresults.

5. Conclusion

From the above explanation can be concluded as follows:

a. Annual expenditure Model is used if the data is not completewithin one year of observation, so as to determine the prediction of the claims at the end of years of observation required a simulation predictions.
b. Predictive simulation conducted on data with a frequency ≥ 1 claims so that the necessary modifications to the data frequency of claims negative binomial distribution becomes Zero Truncated NegativeBinomial.

References

[1] De Jong. P., Heller, G. Z. *Generalized Linier Models for Insurance Data*. Cambridge University Press. 2008.
[2] Duan, N. H. , Manning, W. G. Jr., Morris, C. N., Newhouse, J. P. A Comparison of Alternative Models for Demand for Medical Care. *Journal of Business & Economic Statistics* 1(2): 115–126 (1983).
[3] Frees, E.W., Jie. G., Rosenberg, M. A. Predicting The Frequency and Amount of Health Care Expenditures. *North American Actuarial Journal*. Vol 15. No3: 377-392 (2011).

ISBN : 978-602-99849-2-7

[4] Klugman, S. A., Panjer, H. H., Wilmot, G. E. *Loss Models from Data to Decision*. John Willey & Sons. 2004.
[5] McCulloch, Charles. E, Shayle R., Searle. *Generalized, Linier, and Mixed Models*. Jhon Willey and Sons. 2001.
[6] Tse, YiuKuen. *Nonlife Actuarial Models: Theory, Methods and Evaluation*. Cambridge University Press. 2009.

Appendix

Generating random numbers distributed ZTNB

1. Determine the distribution parameter base (base distribution) of the distribution that will be raised is the negative binomial distribution with parameters (r, β).
2. Determine the CDF fromTruncated ZeroNegativeBinomial
3. Generate arandom number from distribution U (0,1) method Mixed-congruential method as the random number that will be raised from the distribution ZeroTruncated NegativeBinomial
4. Determine the area of the definition of a random variable X distributed ZeroTruncated NegativeBinomial
5. Identify the location ofthe value of the random numberU(0.1) in thearea ofthe definition ofa random variable X. and Therandom variable X with distribution Zero Truncated Negative Binomial success fully obtained

ISBN : 978-602-99849-2-7

Deterministic and Probabilistic Seismic Hazard Risk Analysis in Bantul Regency

Septianusa[1], Maulina Supriyaningsih[1], and Atina Ahdika[1]

[1]Department of Statistics, Islamic University of Indonesia

septianusa@gmail.com, maulinasupriyaningsih@gmail.com, and atina.a@uii.ac.id

Abstract: There are high impacts of the damage after the earthquake in 2006, one of them is because most of buildings are not resistant to earthquakes. Located in the critical region, Bantul Regency is expected to face a high-intensity earthquake in the next 50 years. Therefore, design rules and specific regulations related to earthquake resistant buildings are required. A seismic hazard approach and the relation to building infrastructure are necessary to solve the problems. This paper gives the results of the evaluation of the seismic hazard in deterministic and probabilistic. Hopefully, this reference can be used as a basis in determining the policy of building.
Keywords: Earthquake; DSHA; PSHA; Bantul Regency

1. Introduction

On May 27, an earthquake struck the very heartland of Indonesia with its epicenter in the Indian Ocean at about 33 kilometers south of Bantul district, it measured 5.9 on the Richter Scale and lasted for 52 seconds. The earthquake was relatively shallow at 33 kilometers under ground, shaking on the surface was more intense than deeper earthquakes of the same magnitude, resulting in major devastation, in particular in the districts of Bantul in Yogyakarta Province and Klaten in Central Java Province [1]. With 4.5 million inhabitants, these six districts are very densely populated.

On an earthquake case, the damage pattern on buildings is determined by the earthquake factor itself and the environment conditions where the building is built [5]. Large-scale damage to buildings is associated with a lack of adherence to safe building standards and basic earthquake resistant construction methods [1]. Because man-made failures to build earthquake resistant structures. Most of the private homes used low-quality building materials and lacked essential structural frames and reinforcing pillars and collapsed easily as a result of lateral shaking movements. Clearly, there was minimal enforcement of building codes.

Geologically, Bantul Regency is adjacent to an active subduction zone of south Java Island—a part of the Indo-Australian tectonic plate that has subducted beneath the Eurasian plate (Figure 1) [4]. Located in such a critical geologic setting regarding seismic hazards, it is expected to experience another earthquake of high intensity within the next 50 years [3]. Certainly the threat of this risk could not be ignored. One solution that can be used to estimate the magnitude and earthquake acceleration for some period in the future reliably. It aims to evaluate the rules that already exist and include safety scope and design of the building in the district of Bantul.

The earthquake hazard assessment and its relationship with the damage of building infrastructure can be performed using the acceleration value in the ground or peak ground acceleration (PGA) [5]. The purpose ofthis studywas to assess theseismic hazardespeciallyfor the districtof Bantul. The estimatedhazardzonesbuilt using Gumbel I method [7] and attenuation function of Youngs et. al. [9]are suitable foractivesubductionarea, appropriatewith the typicalareaof BantulRegency.

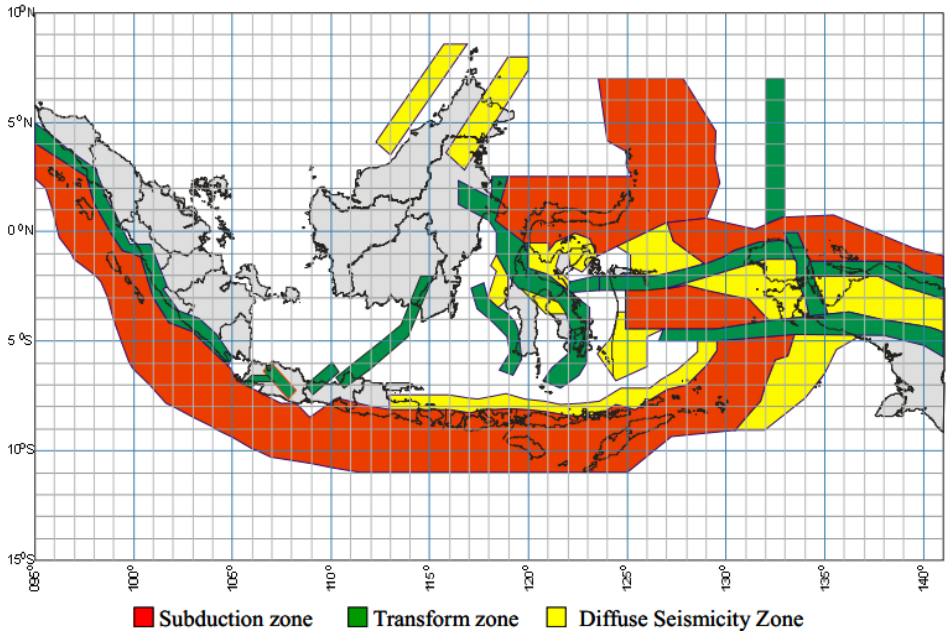


Figure 1. Seismic Source Zone for Indonesia Region [4]

2. Related Works/Literature Review

In the dissertation of [2] explores the use of contingent claims mechanisms, index based insurance specically, in addressing natural disaster risk exposure and nancing within the setting of Indonesian earthquake risk. The approach uses measures of ground motion intensity as the basis for the index. This dissertation's purposes are provide some insights and new approaches for enhancing the ability of financial institutions and individuals to address their risk exposure in a way that contributes to greater institutional and livelihood resiliency and ultimately to development and poverty reduction.

In [3], Hizbaron, et al. (2012) argue that preventive measurements i.e., risk-based spatial plan, building code regulation and other measurement have become critical to reduce future impact of natural disasters. This research aims to assess urban vulnerability due to seismic hazard through a risk based spatial plan. The research area covers six sub-districts in Bantul, Indonesia where experienced 6.2 Mw earthquakes on May, 27th, 2006, suffered a death toll of 5700, economic losses of up to 3.1 billion US$, and damage to nearly 80% of a 508 km^2 area. The research reveals that (1) SMCE-SV (spatial multi criteria evaluations for social vulnerability) and SMCE-PV (spatial multi criteria evaluations for physical vulnerability) are empirically possible to indicate the urban vulnerability indices; and (2) integrating the urban vulnerability assessment into a spatial plan requires strategic, technical, substantial and procedural integration.

In [8] an attempt has been made to estimate seismic hazard of South India at rock level in terms of peak horizontal acceleration (PHA) and spectral acceleration (SA) using probabilistic seismic hazard analysis (PSHA).

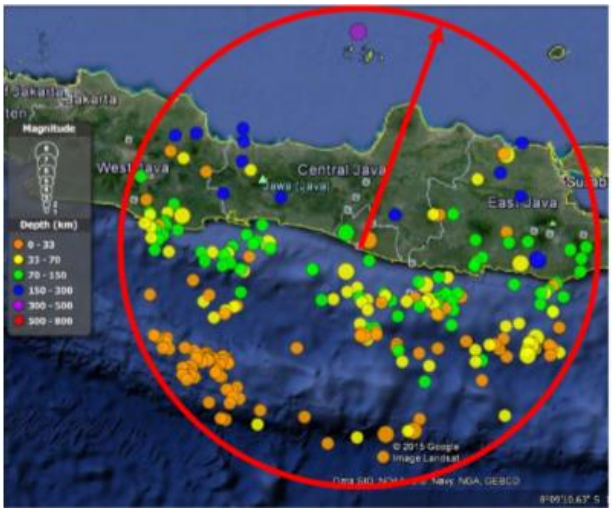
In [6], Lin & Lee (2008) studied about ground-motion attenuation characteristics of subduction zone earthquakes that occur in northeastern Taiwan. This study confirms that subduction zone earthquakes have lower attenuation than crustal earthquakes. The estimated ground-motion level obtained in this study is lower for all magnitudes and all periods when compared with the subduction zone earthquake attenuation results obtained using worldwide data by Youngs et al. (1997).

3. Material & Methodology

3.1 Data

Seismic data collection is done by taking from the earthquake catalog of United States Geological Survey's (USGS) Earthquake Hazards Program. Data can be downloaded via http://earthquake.usgs.gov/earthquakes/search/.

ISBN : 978-602-99849-2-7



Several input when retrieving earthquake data from the USGS include:
Timescales : 31/12/1944 23:59:59 to 25/08/2015 23:59:59
Magnitude : 5.0 – 9.0 SR
Depth : 0 – 500 km
Latitude : -7.8981899
Longitude : 110.3627566
Outer radius : 300 km

Figure 2. Earthquake distribution within radius 300 km from Bantul Regency

3.2 Method

Bantul earthquakes recorded by the USGS has a different magnitude scale reference. Therefore it is necessary to conversion into the same magnitude scale (Mw), to homogenize the output in the seismic risk analysis. For earthquakes that occur in Indonesia, Irsyam et al. (2010) in [10] provide a correlation between the conversion of some magnitude for the region of Indonesia.

Table 1. Conversion correlation

Conversion correlation		
Mw	=	$0.143M_s^2 - 1.051M_s + 7.285$
Mw	=	$0.114m_b^2 - 0.556m_b + 5.560$
Mw	=	$0.787M_E - 1.537$
mb	=	$0.125M_L^2 - 0.389M_L - 3.513$
M_L	=	$0.717M_D + 1.003$

Analysis of seismic risk in Bantul begins by calculating the distance of the earthquake epicenter and the earthquake hypocenter. Earthquake epicenter distance calculations done using Haversine formulation proposed by Sinnott by modeling a simple ball [10]. The equation is given as follows [7]:

$$r = cos^{-1}(\sin(lat\ 1) \times \sin(lat\ 2) + \cos(lat\ 1) \times \cos(lat\ 2) \times \cos(long\ 2 - long\ 1)) \times R \quad (1)$$

Where Latitude and Longitude in radians; Point 1 is the city that reviewed; Point 2 is the location of the earthquake source; R = the diameter of the earth = 6378.1 km.

As for the earthquake hypocenter distance calculation is done by using the Pythagoras theorem as follows [11]:

$$R = \sqrt{D^2 + H^2} \quad (2)$$

Where R is hypocenter distance; D is epicenter distance to point of location that reviewed; H is epicenter distance.

3.3 Ground Motion Model (Attenuation Relationship)

The amount of the earthquake intensity (maximum acceleration, maximum speed) for such a location depends on the magnitude of the earthquake and the epicenter distance from the location of the damage, which is often expressed as the weakening pattern (attenuation). The most important basic in attenuation function selection is the occurrence mechanism of earthquakes [12]. The mechanism of the earthquakes in Bantul district is a mechanism of subduction. Hence, the attenuation function used is the equation proposed by Youngs et al. (1997). Then calculate the maximum earthquake acceleration using Gumbel I method with attenuation function used is the attenuation function of Youngs et al. (1997) for soil, as follows:

ISBN : 978-602-99849-2-7

$$\ln(PGA) = -0.6687 + 1.438M - 2.329\ln(R + 1.097exp^{0.617M}) + 0.00648H + 0.3846Zt \quad (3)$$

Where PGA is Peak Ground Acceleration (gals); M is moment magnitude M ≥ 5; R is epicenter distance; H is depth (10-500 km); Zt is type of earthquake source (0 for interface and 1 for interslab). Afterwards, calculate the constants of Gumbel I distribution such as A, B, α and β using this formula:

$$G(M) = e^{(-\alpha.e^{-\beta M})} \quad (4)$$

Which α is the average number of earthquakes per year, β is a parameter that express the relationship between the distribution of earthquakes with magnitude, and M is magnitude. Then, after getting the constants of A, B, α and β in each district of Bantul Regency, the next process is computing the correlation between the return period T and the acceleration a using the following formula:

$$a = \frac{\ln(T.\alpha)}{\beta} \quad (5)$$

3.4 The Probabilistics Calculation

Computation of exceedance probability of PGA [14]. First, we get the distribution of possible ground-motion levels for a scenario of seismic source from the attenuation relationship:

$$P_i(\ln P_{GA}) = \frac{1}{\sigma_n\sqrt{2\pi}} exp\left\{-\frac{[\ln P_{GA} - g(m_i,d_i)]^2}{2\sigma_n}\right\} \quad (6)$$

Where $g(m_i,d_i)$ and σ_i are the mean and standard deviation of lnPGA. Then, we get the exceedance probability by integration:

$$P_i(>\ln P_{GA}) = \frac{1}{\sigma_n\sqrt{2\pi}}\int_{\ln P_{GA}}^{\infty} exp\left\{-\frac{[\ln P_{GA} - g(m_i,d_i)]^2}{2\sigma_n}\right\} d\ \ln P_{GA} = 1 - \phi\left(\frac{lnPGA - g(m_i,d_i)}{\sigma_n}\right) \quad (7)$$

Then, summing over all scenarios of seismic sources, we get the total annual rate of exceeding each lnPGA:

$$R_{tot}(>\ln P_{GA}) = \sum_{t=1}^{N} R_i(>\ln P_{GA}) = \sum_{t=1}^{N} r_i P_i(>\ln P_{GA}) \quad (8)$$

Then, using the Poissonian distribution, we can compute the exceedance probability of each ground-motion level within the T years:

$$P(>\ln P_{GA}.T) = 1 - \exp(-R_{tot}\ T) \quad (9)$$

4. Results and Discussion

4.1 Result

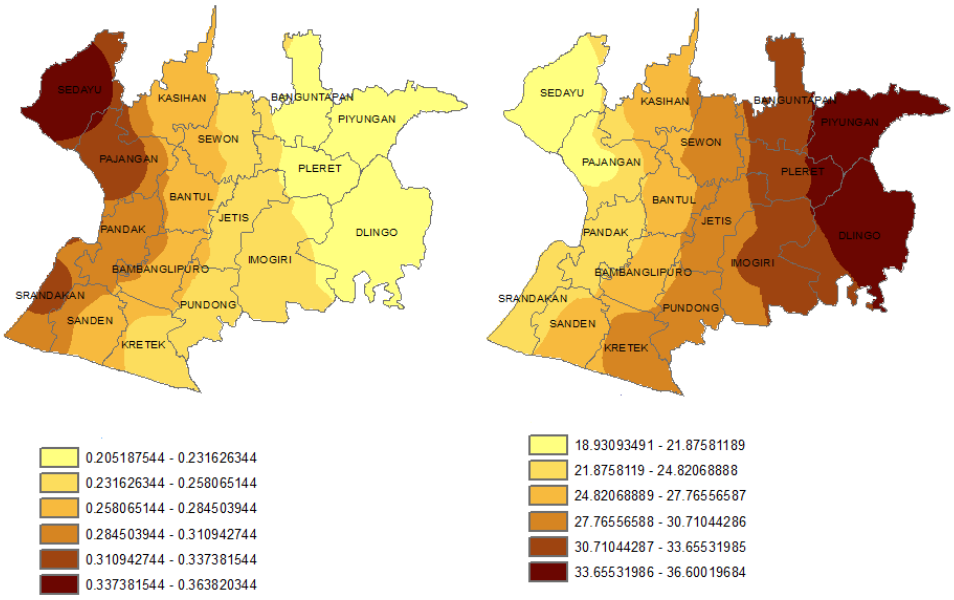


Figure 3.Relation between return period for 50 years(left) and 100 years (right) and its acceleration (a) in gals

ISBN : 978-602-99849-2-7

PGA values in each sub-district of Bantul that have been processed shown in Table 2. This result is the peak ground acceleration events above the soil surface. Based on 'Modified Mercally Intensity 1956' MMI (Table 3), for the return period 50 years most of sub-districts of Bantul including MMI > VII zone where threatened by earthquakes with hazard class is high. MMI is used to measure the damage caused by earthquakes. The highest acceleration values for the return period 50 years occur in Sedayu and the lowest occur in Piyungan. This also happens in the return period of 100 years. The longer a return period, then the acceleration value will be higher.

Table 2. PGA values in each sub-district of Bantul Regency

	A	**B**	**B**	**α**	**Relation between return period (T) and acceleration (a) in gals**		**Probability of 1 gal Exceeded in T Years**	
					50 years	**100 years**	**50 years**	**100 years**
Banguntapan	-0.20771	-16.5767	16.5767	0.812442	0.2234647	0.2652792	0.567394	0.812852
Jetis	-0.21763	-15.0255	15.0255	0.804421	0.2458740	0.2920052	0.619778	0.855432
Pleret	-0.18896	-16.692	16.6919	0.827822	0.2230454	0.2645712	0.607171	0.845685
Bambanglipuro	-0.25866	-12.923	12.923	0.772083	0.2827022	0.3363390	0.621214	0.856521
Sewon	-0.24969	-14.2176	14.2176	0.779039	0.2575913	0.3063440	0.583837	0.826809
Imogiri	-0.1981	-15.7458	15.7457	0.820286	0.2358678	0.2798889	0.630627	0.863564
Kretek	-0.20184	-15.0059	15.0059	0.817226	0.2472479	0.2934394	0.637404	0.868524
Sanden	-0.25085	-13.0934	13.0934	0.778142	0.2796196	0.3325582	0.620042	0.855632
Srandakan	-0.29628	-11.2617	11.2616	0.743577	0.3210659	0.3826151	0.612112	0.849543
Sedayu	-0.34722	-9.79798	9.79797	0.706653	0.3638310	0.4345749	0.574859	0.819255
Pandak	-0.2912	-11.6616	11.6615	0.747366	0.3104918	0.3699304	0.61371	0.85078
Pajangan	-0.3132	-11.0409	11.0409	0.731106	0.3259541	0.3887340	0.596188	0.836936
Kasihan	-0.2679	-13.835	13.8350	0.764988	0.2633984	0.3134993	0.556198	0.80304
Piyungan	-0.16521	-18.2605	18.2604	0.847715	0.2051869	0.2431458	0.588618	0.830765
Bantul	-0.25523	-13.5455	13.5455	0.774737	0.2699627	0.3211344	0.605121	0.844071
Pundong	-0.19961	-15.2337	15.2337	0.819047	0.2436971	0.2891980	0.640732	0.870926
Dlingo	-0.15926	-17.3858	17.3858	0.852773	0.2158517	0.2557203	0.644447	0.873582

Table 3. Class of Earthquake Intensity Indicator [15]

Number	**Hazard Classes**	**Intensity (MMI)**	**Acceleration Value (gals)**
1	Low	< VI	< 0.15
2	Moderate	VI-VII	0.15 – 0.20
3	High	> VII	> 0.20

ISBN : 978-602-99849-2-7

5. Conclusion

In this paper is described how the above methods can be used as a tool for assessing vulnerability due to natural disasters in Bantul Regency. Hopefully, these results can be used as consideration in future as specific rules especially related to the design and building resilience against earthquakes.

References

[1] BAPPENAS, BAPEDA, World Bank, ADB, GTZ, JBIC, . . . UN Habitat, "Preliminary Damage and Loss Assesment Yogyakarta and Central Java Natural Disaster," *In the 15th Meeting of The Consultative Group on Indonesia*, Jakarta, 2006.

[2] Hartell, J., "*Earthquake Risk in Indonesia: Parametric Contingent Claims for Humanitarian Response and Financial Institution Resiliency*," Doctoral Dissertations, Agricultural Economics, University of Kentucky, Lexington, United States, 2014.

[3] Hizbaron, D. R., Baiquni, M., Sartohadi, J., and Rijanta, R., "Urban Vulnearability in Bantul District, Indonesia-Towards Safer and Sustainable Development," *Sustainability* 4, 2022-2037 (2012).

[4] Irsyam, M., Dangkua, D. T., Kusumastuti, D., and Kertapati, E. K., "Methodology of Site-Specific Seismic Hazard Analysis For Important Civil Structure," *Civil Engineering Dimension* 9 (2), 103-112 (2007).

[5] Irwansyah, E., Winarko, E., Rasjid, Z. E., and Bekti, R., "Earthquake Hazard Zonation Using Peak Groud Acceleration (PGA) Apporach," *Journal of Physics*, 1-9 (2013).

[6] Lin, P.-S., dan Lee, C.-T., "Ground-Motion Attenuation Relationships for Subduction-Zone Earthquakes in Northeastern Taiwan," *Bulletin of the Seismological Society of America* 98, 220-240 (2008).

[7] Sa'adah, U., Purwana, Y. M., and Djarwanti, N., "Earthquake Risk Analysis in Surakarta City with Gumbel Method Apporach," *e-Jurnal Matrik Teknik Sipil*, 30-35 (2015).

[8] Vipin, K. S., Anbazhagan, P., and Sitharam, T. G., "Estimastion of Peak Ground Acceleration and Spectral Acceleration for South India with Local Site Effects: Probabilistic Approach," *Natural Hazards and Earth System Sciences* 9, 8865-878 (2009).

[9] Youngs, R. R., Chiou, S. J., Silva, W. J., and Humphrey, J. R., "Strong Ground Motion Attenaution Relationships for Subduction Zone Earthquake," *Seismological Research Letters* 68 (1), 58-73 (1997).

[10] Putra, R. P., "*Peak Ground Acceleration Studies of the Indonesia Earthquake Zone Maps in the Special Region of Yogyakarta*," Undergraduate Theses, Civil Engineering, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia, 2012.

[11] Sari, H.Y., "*Study of Maximum Earthquake Acceleration for Eatrhquake Map Zone Of Indonesia in Banda Aceh City*," Undergraduate Theses, Civil Engineering, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia, 2012.

[12] Ringkasan Hasil Studi Tim Revisi Peta Gempa Indonesia, http://www.preventionweb.net/files/14654_AIFDR.pdf, Retrieved 29 August, 2015.

[13] Search Earthquake Archives, http://earthquake.usgs.gov/earthquakes/search/, Retrieved 28 August, 2015.

[14] Li, Lian-Fa., Wang, Jin-Feng., Leung,Hareton., "Using Spatial Analysis and Bayesian Network to Model the Vulnearability and make Insurance Pricing oc Catastophhic Risk" *International Journal of Geographical Inforamtion Science* 12 (24), 1759-1784 (2010).

[15] Oktariadi, O., "Penentuan Peringkat Bahaya Tsunami dengan Metode Analytical Hierarchy Process (Studi kasus: Wilayah Pesisir Kabupaten Sukabumi)," *Jurnal Geologi Indonesia* 4 (2), 103-116 (2009).

ISBN : 978-602-99849-2-7

Applying Extrapolation Technique to Flexible Binomial Model for Efficiency of American Option Valuation

Arum Handini Primandari[1], Indira Ihnu Brilliant [2]

[1]Statistics Department, Universitas Islam Indonesia
[2]Statistics Department, Universitas Islam Indonesia

primandari.arum@uii.ac.id, indira.i@uii.ac.id

Abstract: Binomial model is one of discrete model used to valuing American option. This model is simpler than Black-Scholes which is a continue model. However discrete model is slow to reach its convergence. We applied repeated Richardson extrapolation on flexible binomial model to accelerate the pace of its convergence. A number of time step used in this scheme are based on the stepsize characterized by sequence of integers. We carried out pricing option using Indonesian stock as underlying asset. As the result, repeated Richardson extrapolation technique works on flexible binomial model can be used to accelerate the sequence of approximation produced by this scheme so that we merely need a less of time step for pricing option.

Keywords: Richardson extrapolation; flexible binomial model; American option.

1. Introduction

Option is derivative financial product which gives the owner right to buy (call) or sell (put) their underlying asset during certain period at specified strike price. Option became interesting because of its benefit as tool for hedging asset. There are two types of common option based on time to exercise, European option that may only be exercised on its expiry and American option that may be exercised anytime on or before expiration. That is to why American option is more appeal than European option.

Pricing American option take more effort than pricing European option especially using Black-Scholes that is a continuous model. We need to know the best time to exercise this option since we allow to do early exercise. In 1979, Cox, Ross, and Rubinstein [2] proposed binomial model as a method for pricing American option. In further discussion, it's called CRR binomial model.

CRR binomial model became an effective way to valuing American option because it provides us possibility of stock movements within its period. We need to define N number of time step parting the period equally. Logically, the more time step we take, the more accurate the result to continue model. But, this case does not work in binomial. That makes CRR binomial model reach its convergence erratically.

In 1999, Tian [10] developed binomial model by adding tilt parameter that changes the form of binomial tree. This model is called flexible binomial model. Compared to CRR binomial model, flexible binomial model has a smoother convergence.

Another problem while using binomial model is the pace to reach its convergence. In other words, we need to take a lot of time steps in order to close to the actual result. The more time steps we use, the more time we need. Thus we applied Repeated Richardson Extrapolation (RRE) as a technique to reduce the elapsed time during computational.

The remainder of the paper is organized as follows. In Section 2, we introduce related work, section 3 and 4 sequentially describes flexible binomial model and RRE. The application of RRE on flexible binomial model is explained in section 5. Numerical result is described in section 5, while section 6 indicate conclusion. Finally, we write up future work in section 7 and reference in the last section.

2. Related Work

In 1984 Geske-Johnson [4] proposed a method applying extrapolation for pricing option. This contribution showed that American option can be valued by using option sequence exercised in some node of Bermudan option.

According to Omberg [1], Geske-Johnson method may be caused un-uniform convergence in some case. Consider $P(n)$ is the price of Bermudan option which can be exercised at one of *n* equal interval. Thus we take $P(1)$, $P(2)$, and $P(3)$ to approximate American option price. Applying Geske-Johnson method, this case $P(1) < P(2) > P(3)$, may happened. In order to solve the problem of un-uniform convergence, Chang, et al [1] modified the work of Geske-Johnson by substituted the stepsize from arithmetic sequence to geometric sequence. While Geske-Johnson used $P(1)$, $P(2)$, and $P(3)$, Chang, et al used $P(1)$, $P(2)$, and $P(4)$. The use of geometric sequence assured this condition $P(4) \geq P(2) \geq P(1)$ was hold.

In 2008, Barzanti, et al [7]applied RRE in several numerical schemes for the valuation of American option. Their research indicated that RRE can be used as an effective technique to improve the precision of the approximation. Moreover, the use of Romberg sequence as stepsize provides high accuracy.

3. Flexible Binomial Model

Binomial model assumes that stock price follows multiplicative binomial process during discrete period. Consider an option which has length of period τ. This time to maturity is partitioned into N-equal length of subintervals $\Delta t = \tau / N$. If we have S_0 as the price at $t = 0$, then this price at $t = 1$ can jump upward to *uS* with probability *p* or jump downward to *dS* with probability *1 – p*, where $0 < d < 1 < u$ and $0 \leq p \leq 1$. Binomial model is specified by parameters *u*, *d*, and *p*. By those parameters, we can build binomial tree.

Label node in the tree as (i, j) where *i* indicates time period (t_i) and *j* is the number of up move from (0,0) to (i, j). Asset price in node (i, j) is given by:

$$S(i, j) = S_0 u^j d^{i-j},\ 0 \leq j \leq i,$$

where S_0 is the asset price at time t_0. Pricing option is carried out using backward recursive procedure. Fisrt, we calculate the payoff at maturity date. Second, we determine option value in prior period. The payoff function at maturity for call option is given by

$$f(S_T) = \max[S(T, j) - K, 0],$$

and for put option is:

$$f(S_T) = \max[K - S(T, j), 0]$$

where *K* is strike price and *S(T,j)* is price of underlying asset at time of maturity T.

American option can be exercised at any time before its expiry or at its expiry itself, so we allow to do early exercise. Because of that, American option value can be obtained by comparing the payoff at exercise node and option value expectation if it is not exercised. The maximum value from the both becomes the option value at that node. American call option value at $0 \leq i < T$ is given by

$$V(i, j) = \max\left[e^{-r\Delta t}\left(pV(i+1, j+1) + (1-p)V(i+1, j)\right), \max\left(S(i, j) - K, 0\right)\right],$$

while American put option is

$$V(i, j) = \max\left[e^{-r\Delta t}\left(pV(i+1, j+1) + (1-p)V(i+1, j)\right), \max\left(K - S(i, j), 0\right)\right].$$

Tian [10] proposed the jump parameters in flexible binomial model as follow:

$$u = e^{\sigma\sqrt{\Delta t} + \lambda\sigma^2\Delta t},$$

$$d = e^{-\sigma\sqrt{\Delta t} + \lambda\sigma^2\Delta t},$$

where λ is arbitrary constant which is called tilt parameter. This parameter can be positive, negative, or zero, as long as it is bounded.

The arbitrary parameter λ can be regarded as degree of tilt in the binomial tree. When $\lambda = 0$, binomial tree will have symmetrical form so that the centre of the tree will draw horizontal line such

ISBN : 978-602-99849-2-7

as CRR binomial tree. In other words, the upward and downward movements bring the asset on the same level as it started. If $\lambda > 0$, binomial tree tilts upward, meaning that the upward movement raise the level of asset price. It is vise versa for $\lambda < 0$.

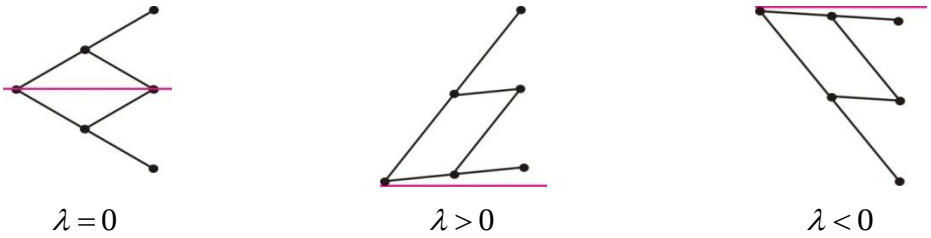


Figure 1. The Effect Of Tilt Parameter In Binomial Tree

CRR binomial model converges to Black-Scholes in erratic fashion, meaning that the convergence of this model is not smooth. Below is figure illustrates the convergence of CRR binomial model and flexible binomial model for pricing European call option.

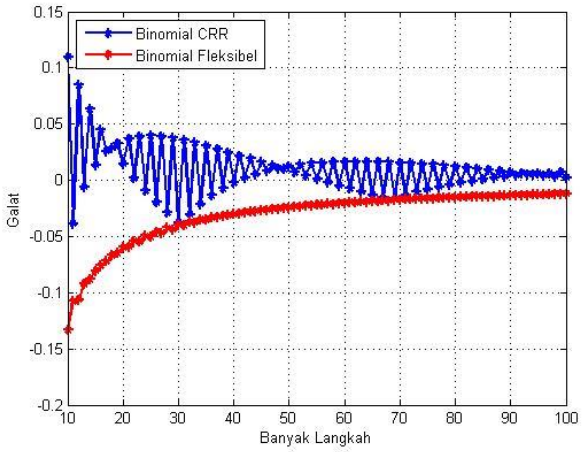


Figure 1. The convergence of CRR binomial and flexible binonmial. Both model used for pricing option that has information: the asset price is $100, the strike price is $95, time to maturity is 6 months, volatility is 20%, and risk free risk is 6%. Pricing error is defined as the different between the binomial price and Black-Scholes price.

We need a model that has smooth convergence so that we can apply Richardson extrapolation on it. Unfortunately, CRR binomial model does not satisfy this condition. Using flexible binomial model, there is a simple way to achieve smooth convergence. This is done by selecting a tilt parameter such that a node in the tree coincides exactly with the strike price at the maturity of option. The formula of λ define as follow:

$$\lambda = \frac{2(\eta - j_0)\sqrt{\Delta t}}{\sigma T}, \tag{1}$$

where

$$\eta = \frac{\log\left(\frac{K}{S_0}\right) - N\log d_0}{\log u_0 - \log d_0} \tag{2}$$

ISBN : 978-602-99849-2-7

$$j_0 = \left[\frac{\log\left(\frac{K}{S_0}\right) - N \log d_0}{\log u_0 - \log d_0} \right] \tag{3}$$

Where [□] in equation (3) denotes the closet integer to its argument. Equation (1) is the formula to determine the value of λ, as $|\eta - j_0| < 0.5$ by definition, $\lambda \to 0$ as $\Delta t \to 0$. The non negative probabilities are ensured for sufficiently small Δt.

4. Repeated Richardson Extrapolation

In the numerical analysis, the unknown quantity (P_0) can be approximated with calculated quantity $(P(h))$ depending on parameter *stepsize* $h > 0$ such that

$$\lim_{h \to 0} P(h) = P(0) = P_0. \tag{4}$$

Under the assumption that $P(h)$ is sufficiently smooth function, we can write:

$$P(h) = a_0 + a_1 h^{\gamma_1} + a_2 h^{\gamma_2} + \ldots + a_k h^{\gamma_k} + O\left(h^{\gamma_{k+1}}\right), \tag{5}$$

Where $0 < \gamma_1 < \gamma_2 < \ldots$ we assume an known parameter, $a_1, a_2, \ldots,$ unknown parameter and $h > 0$. In particular, we have $a_0 = P_0$. According to Schmidt [1], the following algorithm can be built whenever $\gamma_j = \gamma j, j = 1, \ldots, k$.

Repeated Richardson extrapolation can be done by choosing a constant $\omega \in (0,1)$ and $h_0 \in (0, b]$, and let $h_i = h_0 \omega^i; i = 1, 2, \ldots$. Obviously $\{h_i\}$ is a decreasing sequence in $(0, b]$, also $\lim_{i \to \infty} h_i = 0$.

a. Define $P_0^{(j)} = P(h_j), j = 0, 1, 2, \ldots$

b. Define $c_n = \omega^{\gamma_n}$ and calculate $P_n^{(j)}$ for $j = 0, 1, 2, \ldots$ and $n = 1, 2, \ldots$ with recursive process:

$$\begin{aligned} P_n^{(j)} &= \frac{P_{n-1}^{(j+1)} - c_n P_{n-1}^{(j)}}{1 - c_n} \\ &= P_{n-1}^{(j+1)} + \frac{P_{n-1}^{(j+1)} - P_{n-1}^{(j)}}{\frac{1}{c_n} - 1}. \end{aligned} \tag{6}$$

Determining the value of γ, we use Taylor series $P(h)$ around $P(0)$ such that we have $\gamma = 1$ and $\gamma_n = \gamma n = \{1, 2, 3, \ldots, k\}$ for $n = 1, 2, \ldots, k.$ After that, we can build extrapolation table that is:

$$\begin{matrix} T_{1,1} & & \\ T_{2,1} & T_{2,2} & \\ T_{3,1} & T_{3,2} & T_{3,3} \end{matrix}$$

Recursive algorithm can be generalized using the triangle rule in extrapolation table:

a. Define $T_{i,1} = P(h_i); i = 1, 2, \ldots.$

b. For $i \geq 2$ and $j = 2, 3, \ldots, i$ calculate

ISBN : 978-602-99849-2-7

$$T_{i,j} = T_{i,j-1} + \frac{T_{i,j-1} - T_{i-1,j-1}}{\frac{h_{i-j+1}}{h_i} - 1}. \tag{7}$$

Recursion (7) is special case of recursion (6). This algorithm is known as Aitken-Neville algoritm. Let $\omega = \frac{1}{2}$, the recursive equation (7) become:

$$T_{i,j} = T_{i,j-1} + \frac{T_{i,j-1} - T_{i-1,j-1}}{2^{j-1} - 1}; j = 2,3,...,i. \tag{8}$$

We take $h_0 = H$, called H as basic step. Because of $\omega = \frac{1}{2}$, we have sequence $h_i = \frac{H}{n_i}; i = 1,2,...$, where $\{n_i\} = \{2^i\} = \{2,4,8,16,32,...\}$. This sequence is Romberg sequence.

5. Repeated Richardson Ekstrapolation on Flexibel Binomial Model

The procedure to apply repeated Richardson extrapolation on flexible binomial model binomial as follows:

a. Establish the basic step and the number of repetition

Basic step (H) is used to determine the sequence of approximation:

$$\varphi(H;h_1), \varphi(H;h_2), \varphi(H;h_3), ...,$$

for stepsize

$$h_1 > h_2 > ... > 0.$$

Stepsize h_i can be derived using basic step H, that is:

$$h_i = \frac{H}{n_i} \text{ untuk } i = 1,2,...,$$

Thereby $\{h_i\}$ is characterize using integers sequence $\{n_i\}$. In this case $\{n_i\}$ is Romberg sequence,

$$\mathbf{F}_R := \{2,4,8,16,32,...\}.$$

b. Define the sequence of option price

Option values in the first coloumn in extrapolation table are calculated using flexible binomial model.

c. Extrapolate the sequence

We use Aitken-Neville algorithm to do extrapolation.

The accuracy of Richardson extrapolation can be measured using root-mean-squared (hereafter RMS) relative error given as:

$$RMS = \sqrt{\frac{1}{m}\sum_{i=1}^{m} \varepsilon_i^2}, \tag{9}$$

where $\varepsilon_i = \frac{P_i^* - P_i}{P_i}$, P_i^* is estimated option price, while P_i is the true option price. In this state, P_i is obtained from binomial model without extrapolation. This price is assumed as the convergent one.

6. Numerical Result

In 2004, Jakarta Stock Exchange (BEJ) inaugurated the trade of option of five stocks. But, in 2009, Indonesian Stock Exchange (IDX) stopped the trading in order to refine the technical of option. In the beginning of 2014, IDX concerned to restarting option, therefore flexible binomial model can become proposed method for valuing option.

ISBN : 978-602-99849-2-7

We use data from PT Astra Agro Lestari, Tbk., (AALI.JK) for study case. Historical prices are taken from August 17th 2015 till August 17th 2014. We collect some information from market such as stock price at September 17th (S_0) is Rp 19,125,-, risk free rate in Indonesia (r) is 7.5%. We took strike price (K) at Rp 19.000,- and time to maturity (τ) is 0.25.

We calculated option price using binomial model by taking 10000 time steps, while Richardson extrapolation using 5 repetition and 100 as initial time steps (N_0). The sequence of option price which is defined by flexible binomial model used time steps as follow {200, 400, 800, 1600, 3200}. The volatility of AALI.JK estimated using historical volatility method and assumed to be normal. The results are showed in the following table.

Table 1. American Option Price of AALI.JK

Model	Call Option Price	Time Elapsed (second)	Put Option Price	Time Elapsed (second)
CRR Binomial	1,675.80	31.7	1,226.99	29.4
Flexible Binomial	1,675.74	31.2	1,226.94	30.1
Flexible Binomial with Richardson Extrapolation	1,675.77	8.3	1,226.96	7.9

Those three models carried off the same result of option price at the level of 10^{-1}. Among those three, flexible binomial model with RRE needed the least time to carry on the calculation that was a mere 8 seconds. Meanwhile binomial models spent around four times as much time as that model with RRE.

Flexible binomial model with Richardson extrapolation is done by taking 100 initial time steps and 5 repetitions. The RMS relative error of this model used both historical and implied volatility method of estimation is given:

Table 2. RMS Model Flexible Binomial Model with Repeated Richardson Extrapolation

Repetition	n(i)	RMS Call Option	RMS Put Option
1	200	10×10^{-4}	8.28×10^{-4}
2	400	5×10^{-4}	4.14×10^{-4}
3	800	4×10^{-4}	2.76×10^{-4}
4	1600	3×10^{-4}	2.07×10^{-4}
5	3200	2×10^{-4}	1.65×10^{-4}

According to the results in **Table 2.**, the more repetitions we do the smaller RMS relative error on diagonal price we have.

7. Conclusion

Based on the preceding discussion and numerical result, we successfully applied RRE as a technique to fasten the computing process of obtaining American option price using flexible binomial model. REE improve the pace of option value sequence to reach its convergence, hence it need less amount of time to do the numerical schemes. Moreover, the more repetition we take, the more accuracy result we have. Thus the method becomes an alternative way to valuing an option.

8. Future Works

While using binomial model, we need an underlying asset with normal distributed return. Meanwhile, many Indonesian stocks have non normal distributed return. In order to overcome this problem, we should have a model in which has robustness to normality.

9. Acknowledgement

We would like to thank to Research and Community Service Department, Islamic University of Indonesia (DPPM UII) for financial support.

10. Reference

[1] Chang, C.C., Chung, S.L., and Stapleton, R.C., Richardson Extrapolation Technique for Pricing American-Style Options, *Working Paper, Management School, National Central University, Taiwan*. (2002).

ISBN : 978-602-99849-2-7

[2] Cox, J., Ross, S., dan Rubinstein, M., "Option Pricing: A Simplified Approach", *Journal of Financial Economics*, 7, 229-263 (1979).
[3] Deuflhard, P., "Order and Stepsize Control in Extrapolation Method", *Numerische Mathematik*, 41, 399-422 (1983).
[4] Geske, R., and Johnson, H., "The American Put Option Valued Analytically", *Journal of Finance*, 39, 1511-1524 (1984).
[5] Hull, C. J., *Option, Futures, and Other Derivatives 6th edition*, Pearson Education Inc , 2006.
[6] Joyce, D. C., "Survey Of Extrapolation Processes In Numerical Analysis", *SIAM Review*, 13(4) 435-483 (1971).
[7] Barzanti, L., Corradi, C., and Nardon, M., "On The Efficient of The Repeated Richardson Extrapolation Technique to Option Pricing", *Working Paper 147, Department of Applied Mathematics, Università Ca' Foscari Venezia* (2008).
[8] Shapiro, B. E.,"Introduction to Numerical Analysis", *Math 418A Lecture NoteCalifornia State University* (2008).
[9] Avram S., *Practical Extrapolation Methods: Theory and Applications*, Cambridge University Press, 2003.
[10] Tian, Y.S., "A Flexible Binomial Option Pricing Model", *Journal of Futures Markets*, 19, 817-843 (1999).

ISBN : 978-602-99849-2-7

Small Area Estimation Considering Skewness Data and Spatially Correlated Random Area Effects

Dian Handayani[1], Anang Kurnia[2]Asep Saefuddin[2]and Henk Folmer[3]

[1]Department of Mathematics, State University of Jakarta - Indonesia
[2]Department of Statistics, Bogor Agricultural University - Indonesia
[3]Faculty of Spatial Sciences, University of Groningen - The Netherlands

dian99163@yahoo.com,, anangk@apps.ipb.ac.id,
asaefuddin@gmail.com

Abstract: Linear mixed models are frequently used to obtain small area estimates. However, such models are not suitable when the response variable has skewed distribution and there is spatial dependence among small areas. This paper discusses Spatial Empirical Best Prediction (SEBP) for small area estimation of mean taking into both positively skewness data and spatially correlated random area effects. To account for the positively skewness of the response variable, we take logarithm transformation such that the result of transformation has symmetric distribution. Then, we assume that the logarithm transformation of response variable is linear with some auxiliary variables.
Keywords: Small Area Estimation; Spatial Empirical Best Predictor; spatial dependence; skewness data; poverty.

1. Introduction

Survey data is one of the data source to make decision or policy. Survey is commonly designed for providing statistics in large population. In fact, after the survey was conducted, the data obtained from the survey is also frequently utilized to estimate some parameter in sub population. Therefore, it is often met the sample size from the sub population is very small (or might be zero). Certainly, the very small sample size could yield invalid direct estimates because its variance is very large. The sub-population which the sample selected from it is not large enough to produce direct estimates with adequate precision is called small area (Rao, 2003). The field of statistics dealing with adequate estimates in small areas is called small area estimation (SAE).

Standard SAE method (i.e. Empirical Best Linear Unbiased Prediction/EBLUP) assumes that the variable of interest is normally distributed. However, in practice, especially for socio economic research which concerns about income or expenditure, the normality assumption is frequently not satisfied. The welfare variables (i.e.: income or expenditure) often follows positively skewed distribution. To make the data is more symmetric, logarithm transformation is often used. Then, the standard method could be applied for the data which is yielded from the transformation. However, it often could produce the biased estimates. Karlberg (2000a,2000b) derived the bias correction based on the assumption that logarithm transformation of the variable of interest follows normal distribution.Kurnia and Chambers (2011) adopted the bias correction for the SAE context.

The standard SAE model also assumed that small areas are independent. However, in reality, this assumption is frequently violated in that some characteristics of a certain area is similar to those in neighboring small areas. In other words, there is spatial dependence between small areas. Petrucci and Salvati (2004a, 2004b), Salvati (2004), Petrucci and Salvati (2006), Pratesi and Salvati (2008), Molina et al (2009)introducedSpatial Empirical Best Linear Unbiased Prediction (SEBLUP) which relaxes the assumption of independence between small areas.

In this paper, we propose the Spatial Empirical Best Predictor (SEBP) to estimate a small area mean for highly skewed data and when there is spatial dependence among the small areas. In this case, we assume that the variable of interest will be normal after using log transformation and that the relationship between the log transformed variable and the auxiliary variables is linear

The paper is organized as follows. Section 2 introduces EBLUP and SEBLUP which assume that variable of interest follows normal distribution but for the SEBLUP, besides the normality, between small areas are assumed spatial dependence. Section 2also describes empirical best prediction (EBP). Section 3 describes development of spatial empirical best prediction (SEBP) which assume that the

variable of interest has skewed distribution and also consider spatial dependence between small areas. The final remarks are shown in Section 4.

2. A Brief of Empirical Best Linear Unbiased Predictor, Spatial Empirical Best Linear Unbiased Predictor and Empirical Best Prediction

Indirect estimation in small area is frequently model-based. In SAE standard method, estimation parameter is based on linear mixed model which assume that variable of interest follows normal distribution and there is independence among small areas. Consider auxiliary information $x_{ij} = \left(x_{ij1,ij2,\ldots xijp}\right)^T$ is available for each unit population and the population mean $\overline{X}$ is also known. The relationship between response variable and auxiliary variables is given by:

$$y_{ij} = x_{ij}^T \beta + z_{ij} v_i + e_{ij} \quad j = 1,2\ldots N_i \; ; \; i = 1,2\ldots m \tag{1}$$

where β is the vector of regression parameters, z_{ij} is a known positive constant, v_i and e_{ij} are area random effect and sampling error, respectively, with distributions $v_i \sim N\left(0,\sigma_v^2\right)$ and $e_{ij} \sim N\left(0,\sigma_e^2\right)$. Furthermore, v_i and e_{ij} are assumed mutually independent. Hence, the response variable follows the normal distribution $y_{ij} \sim N\left(x_{ij}^T \beta, z_{ij}^2 \sigma_v^2 + \sigma_e^2\right)$.

The mean for i^{th} small area is defined by $\mu_i = N_i^{-1} \sum_{\forall_j} y_{ij} = N_i^{-1} \left[\sum_{s_i} y_{ij} + \sum_{r_i} y_{ij} \right]$ for $i = 1,2\ldots m \; ; j = 1,2\ldots N_i$ where s_i is element of population which was selected as sample whereas r_i is element of population that was not selected as sample. If σ_v^2 and σ_e^2 are known then Best Linier Unbiased Predictor (BLUP) for μ_i is given by

$$\hat{\mu}_i^{BLUP} = \frac{1}{N_i} \left[\sum_{s_i} y_{ij} + \sum_{r_i} \hat{y}_{ij}^{BLUP} \right] \; ; \; i = 1,2\ldots m \; ; \; j = 1,2\ldots n_i \tag{2}$$

where $\hat{y}_{ij}^{BLUP} = x_{ij}^T \hat{\beta} + z_{ij} \hat{v}_i = x_{ij}^T \hat{\beta} + z_{ij} \gamma_i \left(\overline{y}_{is} - \overline{x}_{is}^T \hat{\beta} \right)$ and $\gamma_i = z_{ij}^2 \sigma_v^2 \left(z_{ij}^2 \sigma_v^2 + \sigma_e^2 / n_i \right)^{-1}$.

In practice, the variances σ_v^2 and σ_e^2 are usually unknown. However, they could be estimated from the sample data using restricted maximum likelihood (REML) or maximum likelihood (ML). By replacing $\left(\sigma_v^2, \sigma_e^2\right)$ in (2) with their estimates $\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right)$, it leads to the empirical best linear unbiased predictor (EBLUP) of μ_i:

$$\hat{\mu}_i^{EBLUP} = N_i^{-1} \left[\sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \hat{y}_{ij}^{EBLUP} \right] \; ; \; i = 1,2\ldots m \; , j = 1,2\ldots n_i \tag{3}$$

where $\hat{y}_{ij}^{EBLUP} = x_{ij}^T \hat{\beta} + z_{ij} \hat{v}_i = x_{ij}^T \hat{\beta} + z_{ij} \hat{\gamma}_i \left(\overline{y}_{is} - \overline{x}_{is}^T \hat{\beta} \right) \; ; \; \hat{\gamma}_i = z_{ij}^2 \hat{\sigma}_v^2 \left(z_{ij}^2 \hat{\sigma}_v^2 + \hat{\sigma}_e^2 / n_i \right)^{-1}$.

We now consider the spatial dependence among neighboring areas by specifying a linear mixed model with spatially correlated random area effects (Cressie, 1993; Anselin, 1992) as follows:

$$Y = X\beta + Zu + \varepsilon \text{ and } u = \rho W u + v \tag{4}$$

with ρ the spatial autoregressive coefficient, W a $\left(m \text{ x } m\right)$ proximity or weights matrix, and v is the $\left(m \text{ x } 1\right)$ vector of independent error terms distributed v: $N\left(0,\sigma_v^2 I_m\right)$ with I_m the $\left(m \text{ x } m\right)$ identity matrix. Since $u = \left(I_m - \rho W\right)^{-1} v$, with $E\left(v\right) = 0$ and $Var\left(v\right) = \sigma_v^2 I_m$, we have $E\left(u\right) = 0$ and $Var\left(u\right) = \sigma_v^2 \left[\left(I_m - \rho W\right)\left(I_m - \rho W^T\right) \right]^{-1} = D$. Then model (4) can be rewritten as:

ISBN : 978-602-99849-2-7

$$Y = X\beta + Z\left(I_m - \rho W\right)^{-1} v + \varepsilon \tag{5}$$

with covariance matrix of Y, $V = \sigma_e^2 I_n + ZDZ^T$. Under (5) and if σ_v^2, σ_e^2 and ρ are known, the spatial BLUP of y_{ij} is $\hat{y}_{ij}^{SBLUP} = x_{ij}^T\hat{\beta} + z_{ij}\hat{u}_i$ where : $\hat{u}_i = b_i^T\left(Z^T\left(\sigma_e^2 I_n\right)^{-1} Z + D^{-1}\right)^{-1} Z^T\left(\sigma_e^2 I_n\right)^{-1}\left(Y - X\hat{\beta}\right)$ and b_i^T is $1 \times m$ vector (0,0…0,1,0..0) with 1 in the i-th position. In practice, σ_v^2, σ_e^2 and ρ are unknown. Replacing the parameters by the estimates $\hat{\sigma}_v^2, \hat{\sigma}_e^2, \hat{\rho}$, the spatial empirical best linear unbiased predictor (SEBLUP) of μ_i is given by:

$$\hat{\mu}_i^{SEBLUP} = N_i^{-1}\left[\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} \hat{y}_{ij}^{SEBLUP}\right] \tag{6}$$

where $\hat{y}_{ij}^{SEBLUP} = x_{ij}^T\hat{\beta} + z_{ij}\hat{u}_i^*$; $\hat{u}_i^*$ is $\hat{u}_i$ with parameters σ_v^2, σ_e^2 and ρ replaced by their estimates.

In many socio-economic problems, the response variable has highly skewed distribution. If the log transformed of response variable follows a normal distribution and the relationship between the transformed variable and the auxiliary variables is linear, we have:

$$l_{ij} = \log y_{ij} = x_{ij}^T\beta + z_{ij}v_i + e_{ij} \; ; i = 1,2...m \; ; j = 1,2...n_i \tag{7}$$

where v_i and e_{ij} are mutually independent error terms with $v_i \sim N\left(0, \sigma_v^2\right)$ and $e_{ij} \sim N\left(0, \sigma_e^2\right)$. The l_{ij} follows a normal distribution with mean $x_{ij}^T\beta$ and variance $z_{ij}^2\sigma_v^2 + \sigma_e^2$. Hence: $l_{ij} = \log y_{ij} \sim N\left(x_{ij}^T\beta, z_{ij}^2\sigma_v^2 + \sigma_e^2\right)$. Furthermore, the mean of y_{ij} is $E\left(y_{ij}\right) = e^{x_{ij}^T\beta + \frac{1}{2}\left(z_{ij}^2\sigma_v^2 + \sigma_e^2\right)}$ and its variance is $Var\left(y_{ij}\right) = e^{2\left[x_{ij}^T\beta + \frac{1}{2}\left(z_{ij}^2\sigma_v^2 + \sigma_e^2\right)\right]}\left[e^{\left(z_{ij}^2\sigma_v^2 + \sigma_e^2\right)} - 1\right]$. If $\mu_i = E\left(y_{ij}\right)$ and it could be written as $\mu_i = N_i^{-1}\left[\sum_{s_i} y_{ij} + \sum_{r_i} y_{ij}\right]$ then the best predictor (BP) for μ_i is given by: $\hat{\mu}_i^{BP} = N_i^{-1}\left[\sum_{s_i} y_{ij} + \sum_{r_i} \hat{y}_{ij}^{BP}\right]$ where $\hat{y}_{ij}^{BP} = E\left(y_{ij}\right) = e^{x_{ij}^T\hat{\beta} + \frac{1}{2}\left(z_{ij}^2\sigma_v^2 + \sigma_e^2\right)}$.

The empirical best predictor (EBP) for μ_i is obtained by replacing the unknown parameter σ_v^2 and σ_e^2 with their estimators :

$$\hat{\mu}_i^{EBP} = N_i^{-1}\left[\sum_{s_i} y_{ij} + \sum_{r_i} \hat{y}_{ij}^{EBP}\right] \text{where } \hat{y}_{ij}^{EBP} = e^{x_{ij}^T\hat{\beta} + \frac{1}{2}\left(z_{ij}^2\hat{\sigma}_v^2 + \hat{\sigma}_e^2\right)}. \tag{8}$$

If y_{ij} is not strictly log normal distributed then $\hat{\mu}_i^{EBP}$ will be biased. Following Karlberg's (2000), Kurnia and Chambers (2011) proposed the following bias correction for (9) is $c_{ij}^K = 1 + \frac{1}{2}x_{ij}Var\left(\hat{\beta}\right)x_{ij}^T + \frac{1}{8}\nabla\left(z_{ij}^2\hat{\sigma}_v^2 + \hat{\sigma}_e^2\right)$ where $\nabla(.)$ is an asymptotic variance-covariance matrix. Then, the EBP for μ_i with the bias correction is:

$$\hat{\mu}_i^{KEBP} = N_i^{-1}\left[\sum_{s_i} y_{ij} + \sum_{r_i} \hat{y}_{ij}^{KEBP}\right] \tag{9}$$

where $\hat{y}_{ij}^{KEBP} = \left(c_{ij}^K\right)^{-1} e^{x_{ij}^T\hat{\beta} + \frac{1}{2}\left(z_{ij}^2\hat{\sigma}_v^2 + \hat{\sigma}_e^2\right)}$; $c_{ij}^K = 1 + \frac{1}{2}x_{ij}Var\left(\hat{\beta}\right)x_{ij}^T + \frac{1}{8}\nabla\left(z_{ij}^2\hat{\sigma}_v^2 + \hat{\sigma}_e^2\right)$.

ISBN : 978-602-99849-2-7

3. Spatial Empirical Best Predictor (SEBP)

The EBP (8) as well as the EBP using Karlberg's bias correction (9) are derived from model (7) under the assumption that the area random effects v_i are independent. Model (7) can be extended to allow for spatially dependent area effects as follows. Let

$$l_{ij} = \log y_{ij} = x_{ij}^T \beta + z_{ij} u_i + e_{ij} \; ; i = 1,2...m \; ; j = 1,2...n_i \tag{10}$$

where u_i is the random area effect which is assumed to follow a SAR process with spatial autoregressive coefficient ρ andweights matrix W or in matrix notation with $Y^* = (\log y_{11}, \log y_{12}, ..., \log y_{1n_1}, ..., \log y_{m1}, \log y_{m2}, ... \log y_{mn_m})^T$:

$$Y^* = X\beta + Zu + \varepsilon \tag{11}$$

where $u = \rho W u + v \Rightarrow u = \left(I_m - \rho W\right)^{-1} v$, v: $N\left(0, \sigma_v^2 I_m\right)$; ε: $N\left(0, \sigma_e^2 I_n\right)$ and u: $N\left(0, \sigma_v^2 \left[\left(I_m - \rho W\right)\left(I_m - \rho W^T\right)\right]^{-1}\right)$.

Under model (11), the spatial best predictor (SBP) for μ_i is given by

$$\hat{\mu}_i^{SBP} = N_i^{-1}\left[\sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \hat{y}_{ij}^{SBP}\right]; \; i = 1,2...m \; ; \tag{12}$$

where $\hat{y}_{ij}^{SBP} = e^{x_{ij}^T \hat{\beta} + \frac{1}{2}\left(z_{ij}^2 \tau_i^2 + \sigma_e^2\right)}$; $\tau_i^2 = b_i^T D b_i$; $D = \sigma_v^2 \left[\left(I_m - \rho W\right)\left(I_m - \rho W^T\right)\right]^{-1}$; b_i^T is a m vector (0,0,…0,1,0,0…) with 1 referring to i^{th} area.

The spatial empirical best predictor (SEBP) for μ_i, $\hat{\mu}_i^{SEBP}$, is derived by replacing $\left(\sigma_v^2, \sigma_e^2, \rho\right)$ in (12) with estimators $\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2, \hat{\rho}\right)$ and we could apply Karlberg's bias correction so that we have :

$$\hat{\mu}_i^{SEBP} = N_i^{-1}\left[\sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \hat{y}_{ij}^{SEBP}\right]; i = 1,2...m \tag{13}$$

where : $\hat{y}_{ij}^{SEBP} = \left(c_{ij}^{SEBP}\right)^{-1} e^{x_{ij}^T \hat{\beta} + \frac{1}{2}\left(z_{ij}^2 \hat{\tau}_i^2 + \hat{\sigma}_e^2\right)}$,

$$c_{ij}^{SEBP} = 1 + \frac{1}{2}\left[x_{ij} V\left(\hat{\beta}\right) x_{ij}^T + C_1 V\left(\hat{\rho}\right) + C_2 V\left(\hat{\sigma}_v^2\right) + 2C_3 Cov\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) + \frac{1}{4} V\left(\hat{\sigma}_e^2\right)\right],$$

$$C_1 = b_i^T WD \left\{WD\left[\left(I_m - \rho W\right)\left(I_m - \rho W\right)^T\right]^{-1} + 2W\left[\left(I_m - \rho W\right)\left(I_m - \rho W\right)^T\right]^{-1} + 2\rho\left(I_m - \rho W\right)^{-2}\right\} b_i,$$

$$C_2 = \frac{1}{4} b_i^T \left[\left(I_m - \rho W\right)\left(I_m - \rho W\right)^T\right]^{-2} b_i, \text{ and } C_3 = \frac{1}{4} b_i^T \left[\left(I_m - \rho W\right)\left(I_m - \rho W\right)^T\right]^{-1} b_i.$$

To measure the precision of the SEBP, we consider the mean square error of the SEBP as follows:

$$\begin{aligned} MSE\left(\hat{\mu}_i^{SEBP}\right) &= E\left(\hat{\mu}_i^{SEBP} - \mu_i\right)^2 \\ &= E\left[\left(\hat{\mu}_i^{SEBP} - \hat{\mu}_i^{SBP}\right) + \left(\hat{\mu}_i^{SBP} - \mu_i\right)\right]^2 \\ &= E\left(\hat{\mu}_i^{SEBP} - \hat{\mu}_i^{SBP}\right)^2 + E\left(\hat{\mu}_i^{SBP} - \mu_i\right)^2 + 2E\left(\hat{\mu}_i^{SEBP} - \hat{\mu}_i^{SBP}\right)\left(\hat{\mu}_i^{SBP} - \mu_i\right) \end{aligned} \tag{14}$$

where :

$$\hat{\mu}_i^{SEBP} = N_i^{-1}\left[\sum_{s_i} y_{ij} + \sum_{r_i} \hat{y}_{ij}^{SEBP}\right], \hat{y}_{ij}^{SEBP} = \left(c_{ij}^{SEBP}\right)^{-1} e^{x_{ij}^T \hat{\beta} + \frac{1}{2}\left(z_{ij}^2 \hat{\tau}_i^2 + \hat{\sigma}_e^2\right)}$$

ISBN : 978-602-99849-2-7

$$\hat{\mu}_i^{SBP} = N_i^{-1}\left[\sum_{s_i} y_{ij} + \sum_{r_i} \hat{y}_{ij}^{SBP}\right], \hat{y}_{ij}^{SBP} = \left(c_{ij}^{SBP}\right)^{-1} e^{x_{ij}^T\hat{\beta}+\frac{1}{2}\left(z_{ij}^2\tau_i^2+\sigma_e^2\right)}$$

$$\tau_i^2 = b_i^T D b_i = b_i^T \sigma_v^2\left[\left(I_m-\rho W\right)\left(I_m-\rho W^T\right)\right]^{-1} b_i \; ; D = \sigma_v^2\left[\left(I_m-\rho W\right)\left(I_m-\rho W^T\right)\right]^{-1}.$$

For simplicity, we assume that the cross-product term in (14), $E\left(\hat{\mu}_i^{SBP}-\mu_i\right)\left(\hat{\mu}_i^{SEBP}-\hat{\mu}_i^{SBP}\right)$ is negligible, and the Taylor approximation to the first term of the right hand side of (14) is:

$E\left(\hat{\mu}_i^{SEBP}-\hat{\mu}_i^{SBP}\right)^2 \approx N_i^{-2}\left\{A^2V(\hat{\rho})+\nabla\left(B\hat{\sigma}_v^2+C\hat{\sigma}_e^2\right)\right\}$ where $A=\sum_{r_i}\hat{y}_{ij}^{SBP}b_i^T WD\left(I_m-\rho W\right)^{-1}b_i$,

$B=\sum_{r_i}\frac{1}{4}\hat{y}_{ij}^{SBP}b_i^T\left[\left(I_m-\rho W\right)\left(I_m-\rho W\right)^T\right]^{-2}b_i, C=\sum_{r_i}\frac{1}{2}\hat{y}_{ij}^{SBP}$.

The second term of the right hand side in (14) is :

$$E\left(\hat{\mu}_i^{SBP}-\mu_i\right)^2 = N_i^{-2}\left\{Var\left(\sum_{r_i}\hat{y}_{ij}^{SBP}\right)+Var\left(\sum_{r_i}y_{ij}\right)\right\}$$

$$= N_i^{-2}\left\{e^{\left(z_{ij}^2\tau_i^2+\sigma_e^2\right)}\left[e^{\left(z_{ij}^2\tau_i^2+\sigma_e^2\right)}-1\right]\sum_{r_i}e^{2x_{ij}^T\hat{\beta}}+\sum_{r_{\bar{\partial}}}\left[\left(c_{ij}^{SBP}\right)^{-1}e^{x_{ij}^T\hat{\beta}+\frac{1}{2}\left(z_{ij}^2\tau_i^2+\sigma_e^2\right)}\right]^2\left[x_{ij}Var\left(\hat{\beta}\right)x_{ij}^T\right]\right\}.$$

Finally, the MSE of $\hat{\mu}_i^{SEBP}$:

$$MSE\left(\hat{\mu}_i^{SEBP}\right) \approx G_1\left(\rho,\sigma_v^2,\sigma_e^2\right)+G_2\left(\rho,\sigma_v^2,\sigma_e^2\right)+G_3\left(\rho,\sigma_v^2,\sigma_e^2\right) \quad (15)$$

where

$$G_1\left(\rho,\sigma_v^2,\sigma_e^2\right) = N_i^{-2}\left\{e^{\left(z_{ij}^2\tau_i^2+\sigma_e^2\right)}\left[e^{\left(z_{ij}^2\tau_i^2+\sigma_e^2\right)}-1\right]\sum_{r_i}e^{2x_{ij}^T\hat{\beta}}\right\}$$

$$G_2\left(\rho,\sigma_v^2,\sigma_e^2\right) = N_i^{-2}\left\{\sum_{r_{\bar{\partial}}}\left[\left(c_{ij}^{SBP}\right)^{-1}e^{x_{ij}^T\hat{\beta}+\frac{1}{2}\left(z_{ij}^2\tau_i^2+\sigma_e^2\right)}\right]^2\left[x_{ij}Var\left(\hat{\beta}\right)x_{ij}^T\right]\right\}$$

$$G_3\left(\rho,\sigma_v^2,\sigma_e^2\right) = N_i^{-2}\left\{A^2V(\rho)+\nabla\left(B\sigma_v^2+C\sigma_e^2\right)\right\}.$$

Following Prasad and Rao (1990), an estimator of $MSE\left(\hat{\mu}_i^{SEBP}\right)$ is given by:

$$mse\left(\hat{\mu}_i^{SEBP}\right) \approx G_1\left(\hat{\rho},\hat{\sigma}_v^2,\hat{\sigma}_e^2\right)+G_2\left(\hat{\rho},\hat{\sigma}_v^2,\hat{\sigma}_e^2\right)+2G_3\left(\hat{\rho},\hat{\sigma}_v^2,\hat{\sigma}_e^2\right) \quad (16)$$

where $G_1\left(\hat{\rho},\hat{\sigma}_v^2,\hat{\sigma}_e^2\right)$, $G_2\left(\hat{\rho},\hat{\sigma}_v^2,\hat{\sigma}_e^2\right)$ and $G_3\left(\hat{\rho},\hat{\sigma}_v^2,\hat{\sigma}_e^2\right)$ are defined in (16) but with $\rho,\sigma_v^2,\sigma_e^2$ replaced with their estimates $\hat{\rho},\hat{\sigma}_v^2,\hat{\sigma}_e^2$, respectively.

ISBN : 978-602-99849-2-7

4. Final Remarks

The results of this study suggest that the proposed SEBP allows one to obtain an appreciable improvement of the small area estimates when there is spatial dependence between small areas and the variable of interest is highly skewed.

Acknowledgement. This research is fully supported by Fundamental Research Grant from Kementerian Riset, Teknologi dan Pendidikan Tinggi Dikti The Republic of Indonesia.

References

Anselin, L. (1992). Spatial Econometrics : Method and Models. Boston : Kluwer Academic Publishers.

Chambers, R. Tzavidis, N. and Salvati, N. (2009). Borrowing Strength Over Space in Small Area Estimation : Comparing Parametric, Semi-Parametric and Non-Parametric Random Effects and M-quantile Small Area Models. Working Paper 12-09, Centre for Statistical and Survey Methodology, School of Math and Applied Statistics, University of Wollongong Australia.

Chandra, H., Salvati, N., Chambers, R. and Tzavidis, N. (2010). Small Area Estimation under Spatial Nonstationarity. Working Paper 21-10, Centre for Statistical and Survey Methodology, School of Mathematics and Applied Statistics, University of Wollongong Australia.

Chandra, H. and Chambers, R. (2011). "Small Area Estimation under Transformation to Linearity". *Survey Methodology*, 37 : 39-51.

Cressie, N. (1993). Statistics for Spatial Data. New York : John Wiley and Sons.

Datta, G.S and Lahiri, P. (2000). "A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictors in Small Area Estimation Problems". *Statistica Sinica*, 10 : 613-627

Kackar, R.N. and Harville, D.A. (1984). "Approximation for Standard Errors of Estimation for Fixed and Random Effects in Mixed Models". *JASA*, 79 : 853-862.

Karlberg, F. (2000a). "Population Total Prediction Under a Lognormal Superpopulation Model. *Metron*, LVIII : 53-80.

Karlberg, F. (2000b). "Survey Estimation for Highly Skewed Population in the Presence of Zeroes". *Journal of Official Statistics*, 16: 229-41.

Molina, I. , Salvati, N. and Pratesi, M. (2009). "Bootstrap for estimating the MSE of the Spatial EBLUP". *Comput. Stat* , 24: 441-458.

Patresi, M. and Salvati, N. (2008). "Small Area Estimation : the EBLUP Estimator Based on Spatially Correlated Random Effects". *Stat.Meth. & Appl*. 17: 113-141

Petrucci, A. and Salvati, N. (2004a). Small Area Estimation Using Spatial Information,The Rathbun Lake Watershed Case Study. Working Paper no 2004/02, "G. Parenti" Department of Statistics, University of Florence.

Petrucci, A. and Salvati, N. (2004b). Small Area Estimation Considering Spatially Correlated Errors : The Unit Level Random Effects Model. Working Paper no 2004/10, "G. Parenti" Department of Statistics, University of Florence.

Petrucci, A. and Salvati, N. (2006). "Small Area Estimation for Spatial Correlation in Watershed Erosion Assesment". *JABES*, 11: 169-182.

Prasad, N.G.N. and Rao, J.N.K. (1990). "The Estimation of Mean Squared Errors of Small Area Estimators". *Journal of American Statistical Association*, 85: 163-171.

Rao , J.N.K. (2003). Small Area Estimation. New York : John Wiley and Sons.

Salvati, N. (2004). Small Area Estimation by Spatial Models : the Spatial Empirical Best Linear Unbiased Prediction (Spatial EBLUP). Working Paper no 2004/03, "G. Parenti" Department of Statistics, University of Florence.

Salvati, N., Tzavidis, N., Pratesi, M. and Chambers, R. (2008). Spatial M-quantile Models for Small Area Estimation. Working Paper 15-08, Centre for Statistical and Survey Methodology, School of Math and Applied Statistics, University of Wollongong Australia.

ISBN : 978-602-99849-2-7

Comparison of Binary, Uniform and Kernel Gaussian Weight Matrix in Spatial Autoregressive (SAR) Panel Data Model and The Application.

Tuti Purwaningsih[1], Erfiani[2], and Anik Djuraidah[2]

[1]Department of StatisticsFaculty of Mathematics and Natural Sciences, Universitas Islam Indonesia
[2]Department of Statistics, Faculty of Mathematics and Natural Sciences, Bogor Agricultural University

purwaningsiht@yahoo.com, erfiani_ipb@yahoo.com, and anikdjuraidah@gmail.com

Abstract. Spatial statistics is one of statistical methods to solve problem in statistics when the data come from different locations and has possibility influencing eachothers. Then we have new problem when the data not just come from different locations but also have different period of time. We can say that it is spatial panel data, at this model we have index of location and time. The focus is about correlation between location that we presented with ρ as parameter to measure the strength of correlation between locations. When we applied model of spatial statistics so we can't use ordinary model which assume that between the sample was identic independent stochastically. The new approach will be used, using Spatial Autoregressive Panel Data Model (SAR-PDM). The Model has some parameter to be estimated, one of them is ρ that can influence the goodness of fit model for prediction. ρ build 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 research try to compare Binary, Uniform and Kernel Gaussian W in SAR panel data model to get the best W based on RMSE value. The result is Uniform W has the minimum RMSE value for almost all combination of location number (n) and periods number (t).Then, if we run SAR-Panel Data Model so it will be better if we use the Uniform W to bulid contiguity matrix. Then using Uniform W to estimate poverty model in Indonesia using 3 independent variables.

Keywords: ρ; Binary W;Uniform W;Kernel Gaussian W;SAR 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 involves spatial influence in the analysis panel data that will be mentioned as Spatial Panel Data Model.

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 types 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 tocompare Binary W, Uniform W and Kernel Gaussian W in SAR panel data model using RMSE value which is obtained from simulation.

The rest of this paper is organized as follow. Section 2 describes about related works with this paper, Section 3 describes the rudimentary about this paper ideas, Section 4 describes material and proposed method, Section 5 describes result and discussion about the simulation and finally, the conclusion of this work is described in Section 5.

2. Related Works

Some recent literature of Spatial cross-section data is Spatial Ordinal Logistic Regression by Aidi and Purwaningsih [2], Geographically Weighted Regression [3] and Comparison of Unifrom and Kernel Gaussian Weight Matrix in Generalized Spatial Panel Data Model [8]. Some of the recent literature of Spatial Panel Data is forecasting with spatial panel data [5] and spatial panel models [4].

3. Rudimentary

3.1. Data Panel Analysis

Data used in the panel data model is a combination of cross-section and time-series data. Cross-section 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 cross-section 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:

$$u_{it} = \mu_i + \varepsilon_{it} \quad (2)$$

with μ_i is an individual-specific effect that is not observed, and ε_{it} is a remnant of cross section-i and time series-t [5].

3.2. 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 (nonuniform 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,

$$W_{ij} = \begin{cases} \frac{1}{n_i^{(l)}}, & jisneighborofiinl - thorder \\ 0, & others \end{cases} \quad (3)$$

ISBN : 978-602-99849-2-7

$n_i^{(l)}$ 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 :

$$w\_j\ (i) = \exp[-1/2\ (d\_ij\ /\ b)\text{^}2\ ] \qquad (4)$$

with d is distance between location i and j, then b is *bandwith* which is a parameter for smoothing function.

3.3. Spatial Autoregressive Panel Data Model (SAR-Panel Data)

Autoregressive spatial models expressed in the following equation:

$$y_{it} = \rho \sum_{j=1}^{N} w_{ij} y_{jt} + \mathbf{x}'_{it}\boldsymbol{\beta} + \mu_i + \varepsilon_{it} (5)$$

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]

4. Material and Proposed Method

4.1. Data

Data used in this study got 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. N index indicates the number of locations and T index indicates the number of period in each locations, the structure of data can be look at **Table 1**.

Table 1. Data structure

i	T	Y_{it}	X_{it}
1	1	y_{11}	y_{11}
:	:	:	:
1	T	y_{1T}	x_{1T}
2	1	y_{21}	x_{21}
:	:	:	:
2	T	y_{2T}	x_{2T}
:	:	:	:
N	1	y_{N1}	x_{N1}
:	:	:	:
N	T	y_{NT}	x_{NT}

4.2. Proposed Method

In this section, the proposed methods are described with the following step. The Step are used to build simulation proccess, comparing three of W in SAR Panel Data Model to get the best W for all of combination.

a. Determine the number of locations to be simulated is N = 3, N = 9 and N = 25
b. Makes 3 types of Map Location on step 1

ISBN : 978-602-99849-2-7

c. 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.
d. Creating Spatial Uniform weighted matrix based on the concept of Queen Contiguity of each type of map locations.
e. Making weighted matrix kernel gaussian based on the concept of distance. To make this matrix, previously 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:

$$w_j(i) = exp\left[-\frac{1}{2}\left(d_{ij}/_b\right)^2\right] \text{[3]}$$

f. Specifies the number of time periods to be simulated is T = 3, T = 6, T = 12 and T = 24
g. Generating the data Y and X based on generalized spatial panel data models follows equation (5).
h. Cronecker multiplication between matrix Identtity of time periods and W, then get new matrix named IW.
i. Multiply matrix IW and Y to obtain vector WY.
j. Build a spatial panel data models and get the value of RMSE
k. Repeat steps 7-9 until 1000 replications for each combination on types of W, N, and T.
   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
l. Get the RMSE value for all of 1000 replicationsof each combination between W, N, andT.
m. Determine the best W based on the smallest RMSE for all combinations.

5. Results and Discussion

The simulation generates data for vector Y as dependent variable and X matrix as independent variable matrix. Y and X is generate with parameter initiation. After doing simulation, then get RMSE for each combinations and proccessing it, then calculating RMSE for each W, N and T. The result can be looked at **Table 2** then continued to figure it out into graphs in order to look the RMSE comparison easily.

Table 2.Value of RMSE resulted from Simulation for all the combinations (W, N, T)

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		

ISBN : 978-602-99849-2-7

W types	Location types	Periods types	RMSE	Average RMSE	Average RMSE
	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		

The above table represent the result of simulation combining some parameters. The parameters is number of location, number of periods of time and types of W. The RMSE value from **Table 2** then represented again by bellow graphs on **Figure 1** and **Figure 2.**

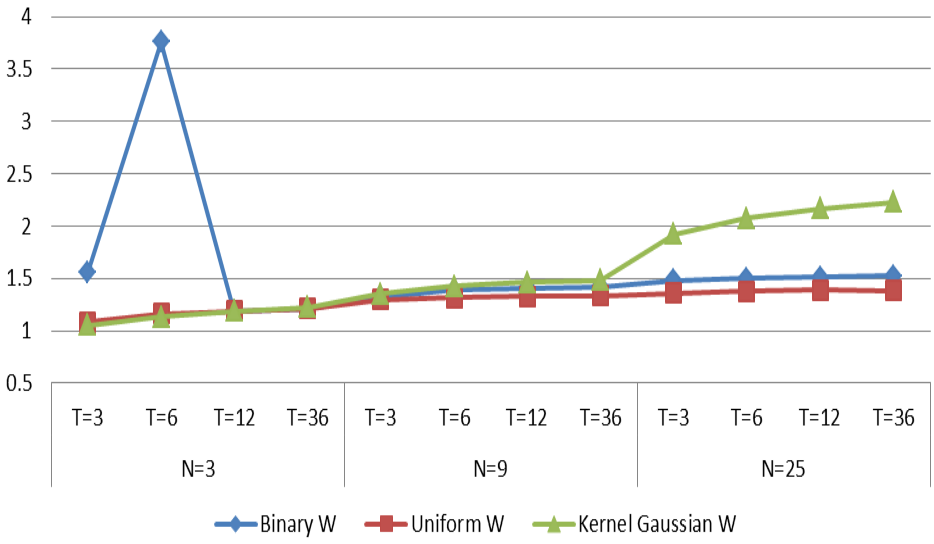


Figure 1.RMSE between Binary, Uniform and Kernel Gaussian Weight for Combinations N and T

Figure 1 show that Uniform W has smaller RMSE value than Binary W and Kernel Gaussian W for almost combination 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 1-4 and Binary from 1-2.5. So can be concluded that Uniform W is better than Binary and Kernel Gaussian W in SAR panel data model.

ISBN : 978-602-99849-2-7

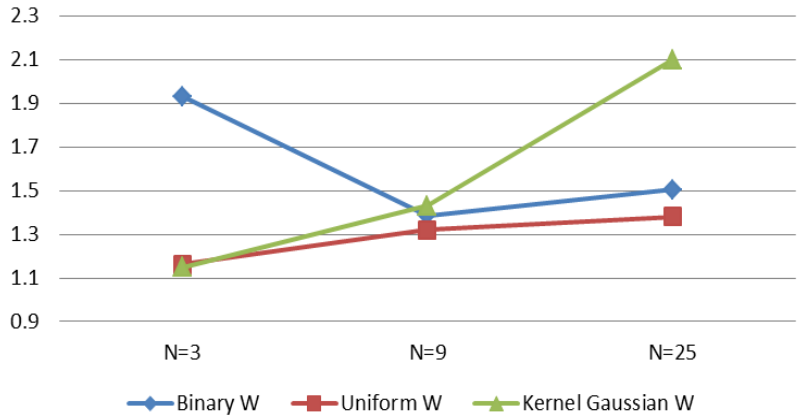


Figure 2. Comparison RMSE of W based on N types

Figure 2 try to analyze diferencies between the W based on N types (3 locations, 9 locations and 25 locations). The graph above show that Uniform W has smallest RMSE value in all N types, and Kernel Gaussian W has trend, if the number of locations increase then followed by the increasing of RMSE value.

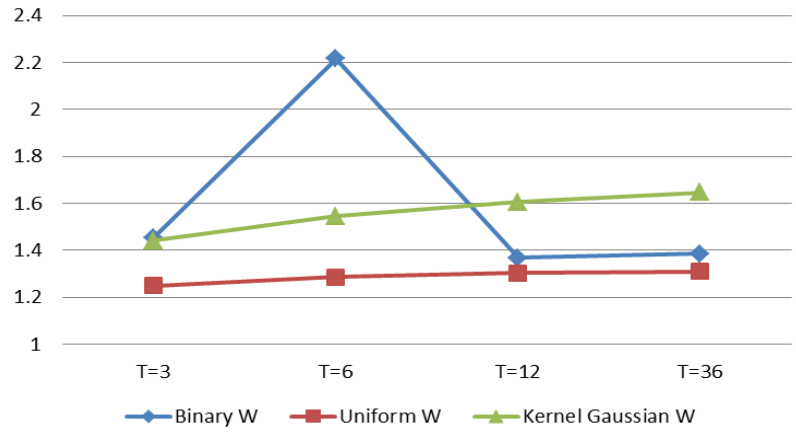


Figure 3. Comparison RMSE of W based on T types

Figure 3 try to analyze diferencies between the W based on T types (3 periods, 6 periods, 12 periods and 36 periods). The graph above show that Uniform W has smallest RMSE in all types of T and Kernel Gaussian W has trend similar when based on locations, if the number of periods increase then followed by the increasing of RMSE value.

Here are the poverty model using 3 independent variables: income, unemployment index and education index.

Variable	Coefficient	Probability
WPoverty	0.516	0.00001
Constant	7246251	0.00001
Income	-0.862	0.17010
Unemployment	193648.1	0.00860
Education	883779.3	0.00038

This model results R-square value at 61.4% and there are only 2 variables influencing significantly at alpha 5%. It is Un employent and education Index. Here is the model :

ISBN : 978-602-99849-2-7

Poverty = 7246251 + 0.516 WPoverty + 193648.1 Unemployment- 883779.3 Education

6. Conclusion

In this study, the comparison of Binary, Uniform and Kernel Gaussian Weight Matrix in Spatial Autoregressive (SAR) Panel Data Model has been presented. Based on simulations result and after explorating the RMSE value, can be concluded that Uniform W is the best W in SAR panel data model and the factors that influence the poverty in Indonesia is unemployment index and education index.

References

[1] Anselin L, Gallo Julie and Jayet Hubbert : "*The Econometrics of Panel Data",* Berlin: Springer (2008).
[2] Aidi, MN, Purwaningsih T : "Modelling Spatial Ordinal Logistic Regression and The Principal Component to Predict Poverty Status of Districts in Java Island", International Journal of Statistics and Aplication 3(1):1-8 (2013).
[3] Fotheringham A.S., Brunsdon C., Chartlon M :"Geographically Weighted Regression, the analysis of spatially varying relationships", John Wiley and Sons, LTD (2012).
[4] Elhorst :"Spatial panel models. Regional Science and Urban Econometric" (2011)
[5] Baltagi BH : "Econometrics Analysis of Panel Data". 3th edition, England: John Wiley and Sons, LTD (2005).
[6] Dubin R. 2009. *Spatial Weights*. Fotheringham AS, PA Rogerson, editor, Handbook of Spatial Analysis. London: Sage Publications.
[7] Elhorst JP. 2010. *Spatial Panel Data Models.* Fischer MM, A Getis, editor, Handbook of Applied Spatial Analysis. New York: Springer.
[8] Purwaningsih T, Erfiani, Djuraidah A : "Comparison of Unifrom and Kernel Gaussian Weight Matrix in Generalized Spatial Panel Data Model", Open Journal of Statistics 5, 90-95 (2015).

ISBN : 978-602-99849-2-7

Persistence Process of Stock Price Movements Based On Markov Chain Analysis
Case Study: Indonesia Stock Exchange(IDX)

Atina Ahdika[1*], Bayun Matsaany[2]

[1]Education Staff at Statistics Program, Faculty of Mathematics and Natural Sciences, Islamic University of Indonesia
[2]Student of Statistics Program, Faculty of Mathematics and Natural Sciences, Islamic University of Indonesia

atina.a@uii.ac.id,bayunmatsaany94@yahoo.com

Abstract. Stocks are securities that show ownership of an investor in a company. The stock price index is very volatile, which make it difficult to predict appropriately. Instead, the movements can be modeled by classifying it into some states using Discrete-Time Markov Chain Analysis (DMCA). DMCA method modeling the transition probability of the stock price movement from one state to another. The transition probability can be used as the consideration to the investor whether to buy the stock or not. In addition, the investor can determine the steps to be taken by looking at the behavior and characteristics of the stock through its persistence. This paper aims to formulate the persistence index of the stock price. This index can accommodate any kind of states in Markov Chain. Moving Average (MA) of order three was used to modeled the stock price to obtain Difference of Price (DoP) at time t and $t-1$ which used to classify the states. The data used in this paper is the historical data of Indonesia Stock Exchange (IDX) from 25 August 2014 to 25 August 2015 and the prediction until one week later. The result shows that the probability of stock price movements to jump down is larger than the probability to jump up or remain stable in each state classification. Persistence index shows that the behavior of stock price unchanged for some initial states and changed for the other, especially when the intial state is stable. It means that when the stock price is in stable condition, the probability to remain stable is smaller than the probability to move to other conditions.
Keywords:discrete-time markov chain analysis; forecasting; persistence index; stock price; transition probability.

1. Introduction

Studies on forecasting stock price were mostly done by researchers. The most common method which has used is a time series model that generates the prediction of future prices. Ayodele A. Adebiyi and Aderemi O. Adewumi [1] predictedthe New York Stock Exchange (NYSE) and Nigeria Stock Exchange (NSE) using ARIMA model, Jun Zhang [9] predicted Shanghai Composite Index using ARIMA and ARCH model, Selene Yue Xu [7]predicted the stock price by combining the conventional time series analysis with information from Google trend website and Yahoo Finance website, and manyother researchers conducted similar work. The results of the forecasting have been used by the investor as the consideration for determining the decision whether to buy the shares or not.

However,forecasting stockprice usingtime series methoddoes notprovidehighaccuracy result because its movement is very dynamics. Based onthis reason, there is another idea todeterminewhether theshareswill be purchasedornotby predicting the probability of the stock movement.For this kind of forecasting, it is not necessaryto get theresultspecifically, butit is enough to predict the stock price movement of the next period,whetheritwill go up, down, or stay stablebased onthe state ofthe stock pricein the previous period.The method used topredict theprobability of the stock price movementis DiscreteTimeMarkov ChainAnalysis(DMCA).Thismethodmodeling theconditions thatmaybe passed bythe stock pricewithin acertain periodandcalculate the transition probability of the stock price movement. The investorspay attention tothe probabilityof thestock price movement based on the result of the DMCA.If theprobability of thestock pricecontinues torisefromanyprevious states, the investorwillmake a decision tobuythe shares.Some researchers who applying the DMCA methods are Milan Svoboda and Ladisva Luke[5], Deju Zhang and Zhang Xiaomin[8],KevinJ.Doubleday and Julius N.Esunge[3], and Qing-xin Zhou[10].

In addition, it is necessary to define another indicator which describing the continuity of the stock behavior. Hence, the idea to formulate this indicator is emerged and named as persistence index of the stock price.

The paper is organized as follows. In Section 2, we review about the DMCA method and its relation with time series model for predicting the probability of stock price movements. In Section 3, we formulate the persistence index of the stock price movements for some kinds of state and the general formulation of the index. Application of DMCA method and the calculation of persistence index will be conducted to the Indonesia Stock Exchange (IDX) in Section 4.

2. Related Works

Basic uses of Markov Chain in forecasting stock price is divided the stock price behavior into some possible states, then build a transition probability matrix. In predicting transition probability for each state, we need the combination of Markov Chain method and time series model. The time series model is necessary to predict the stock price index for the next period, then from the prediction we obtain the difference of price (DoP) of the stock price. The value of DoP is used to classify the states in Markov Chain. There are some time series models which can be used to forecast the stock price index.

Milan Svoboda dan Ladislav Lukas [5] predicted the stock index trend of Prague Stock Exchange (PX) using two time series models and four models of Markov Chain. Model 1 and model 2 used time series model that has been built from the ratio of the stock price at time t and $t+1$ that is $Y_t = P_t / P_{t+1}$ with each model contain of two and eight states. While model 3 and model 4 used time series model which defined as $K_t = K_{t-1} Y_t$, if $(P_{t-2} \le P_{t-1} \le P_t)$ or $(P_{t-2} \ge P_{t-1} \ge P_t)$.

Deju Zhang dan Xiaomin Zhang [8] predicted the stock market in China using Markov Chain with three states; jump up, zero-plus, and jump down and six states that have each growth level. Similar work was done by Kevin J. Double day dan Julius N. Esunge [3] too, they modeled Dow Jones Industrial Average (DJIA).

Meanwhile, Qing-xin Zhou [10] predicted stock price of China Sport Industry using Weighted Markov Chain. Markov Chain was built by classifying the changes of stock price into six states that is plunge, flat plunge, downward flat, upward flat, rise, and soar. Then the results from both models were compared.

3. Material & Methodology

3.1 Data

Data used in this paper is a historical data of Indonesia Stock Exchange (IDX) from August 25, 2014 to August 25, 2015. Stock price index for the next week was predicted using Moving Average smoothing of order three. The data that is used in this work is the daily adjective close and can be downloaded via http://finance.yahoo.com/q/hp?s=IDX+Historical+Prices.

3.2 Method

In forecasting the stock price, the accuracy of the prediction obtained is not good enough because it cannot predict the long-term predictions. It happened because the changes in the stock price are very volatile from time to time so that the forecasting of the stock price is best done to short-term period prediction. The investors then have other alternative method to take the decision whether to buy the shares or not by looking at the probability of the stock price movements and its persistence in the future period. The method used to estimate the probability of the stock price movements is time series model and Discrete Markov Chain Analysis (DMCA). The time series model is used to predict the stock price index while DMCA is used to predict the probability of the its movements. Time series model used in this work is the Moving Average smoothing of order three because the data behaves like those which have stationary property.

Markov Chain is a stochastics process $\{X_t\}$ with $t = 0, 1, 2, \ldots$ which has the following properties [4]:

i. The value of X_t is finite or countable

ISBN:978-602-99849-2-7

ii. Transition probability from state "i" (at time t) to state "j" (at time $t+1$) is P_{ij} that is

$$P\left(X_{t+1}=j \middle| X_t=i, X_{t-1}=i_{t-1},\ldots,X_1=i_1, X_0=i_0\right)=P\left(X_{t+1}=j \middle| X_t=i,\right)=P_{ij}$$

iii. Conditional probability of X_{t+1} given the past states $X_0, X_1,\ldots,X_{t-1}$ and the present state X_t is only depend on the present state (Markov property)

P_{ij} represents the probability that the process, when in state i, next make transition into state j

$$P_{ij}\geq 0,\ i,j\geq 0;\ \sum_{j=0}^{\infty}P_{ij}=1,\ i=0,1,\ldots$$

The transition probability from state i to state j is represented to the form of Markov Chain or the transition probability matrix. The transition chain was shown in Figure 1

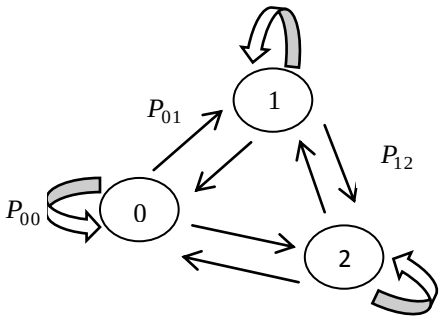


Figure 1.Markov Chain

Figure 1 illustrates the Markov Chain with three states in which the arrows indicate the possibility of a transition from one state to another. The other representation of Markov Chain is the transition probability matrix P which denote one-step transition probabilities P_{ij}

$$P=\begin{pmatrix} P_{00} & P_{01} & P_{02} & \cdots \\ P_{10} & P_{11} & P_{12} & \cdots \\ \vdots & & & \\ P_{i0} & P_{i1} & P_{i2} & \cdots \end{pmatrix}$$

For an irreducible ergodic Markov Chain $\lim_{n\to\infty}P_{ij}{}^{n}$ exist and is independent of i. Suppose

$$\pi_j=\lim_{n\to\infty}P_{ij}{}^{n},\quad j\geq 0,$$

then π_j is the unique nonnegative solution of

$$\pi_j=\sum_{i=0}^{\infty}\pi_i\,P_{ij}{}^{n},\quad j\geq 0,\ \text{with}\ \sum_{j=0}^{\infty}\pi_j=1.$$

Pierre Vallois and Charles Tapiero [6] in their paper construct the persistence index to claim process and insurance. Those index indicate the pattern of claim process to insurance client based on the pattern of change in the provision for claims using Markov Chain. The persistence index formulated in their work was constructed to Markov Chain with two states only. Based on those idea, we will formulated the persistence index of stock price movements to Markov Chain with more than two states. The formulation of the index is given in Proposition 1

Proposisi 1. Suppose $\{X_t, t\geq 0\}$is a sequence of random variables which represent the states of the stock price movements that has been classified from the Difference of Price (DoP) at time t. Suppose that the transition probability of stock price movements is written in matrix P**.** Then, the persistence index of the stock price movements is

$$\rho_i=(n-1)P_{ii}-\sum_{j\neq i}P_{ji},\quad \rho_i\in\left[-(n-1),(n-1)\right]$$

with n is the size of the states.

ISBN:978-602-99849-2-7

- If $\rho_i > 0$, then $P_{ii} > \frac{\sum_{j \neq i} P_{ji}}{n-1}$. This process is a persistence process, where the state probability of the stock price at the next period will be equal with the previous period is greater than all of the other states.
- If $\rho_i \leq 0$, then $P_{ii} \leq \frac{\sum_{j \neq i} P_{ji}}{n-1}$. This process is not a persistence process, where the state probability of the stock price at the next period will be equal with the previous period is smaller than all of the other state.

Proof.

Suppose that there are n possible states in which the initial state is i, than the stock would have a persistence property if $P_{ii} > P_{ji}$ with $j \neq i$ for all j, hence

$$\begin{aligned} \left(P_{ii} - P_{j_1 i}\right) + \left(P_{ii} - P_{j_2 i}\right) + \ldots + \left(P_{ii} - P_{j_n i}\right) > 0, \quad j \neq i \\ (n-1)P_{ii} - \left(P_{j_1 i} + P_{j_2 i} + \ldots + P_{j_n i}\right) > 0 \\ (n-1)P_{ii} - \sum_{j \neq i} P_{ji} > 0 \end{aligned}$$

4. Results and Discussion

The prediction of the stock price using Moving Average smoothing of order three (MA(3)) is shown by Figure 2,

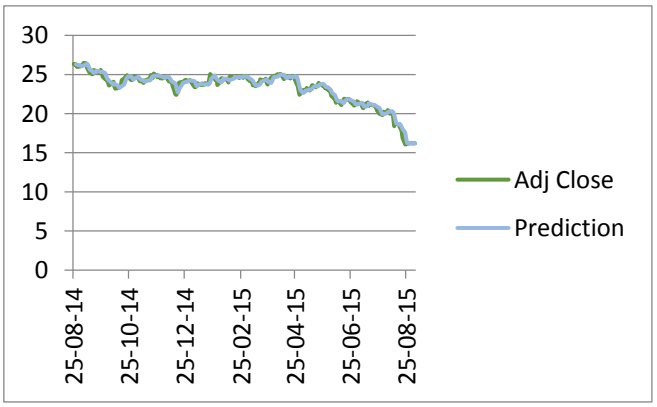


Figure 2.Stock Price Prediction Using MA(3)

Figure 2 shows that the prediction of the stock price using MA(3) is closed enough with the actual price. It is because the data have a stationary properties so that MA(3) method is appropriately applied to this kind of data. Following are the prediction of the transition probability of stock price movements using DMCA.

Model 1: Markov Chain with 3 States

The states defined in Model 1 are '0' : jump down $(X_t < 0)$, '1': stable $\left(X_t = 0\right)$, and '2' : jump up $\left(X_t > 0\right)$. The result of Model 1 is

ISBN:978-602-99849-2-7

$$P = \begin{pmatrix} 0.754 & 0 & 0.246 \\ 0.333 & 0 & 0.667 \\ 0.306 & 0.027 & 0.667 \end{pmatrix}$$

Then we obtain the long-run probability of each state by solving the equations

$$\pi_0 = 0.754\pi_0 + 0.333\pi_1 + 0.306\pi_2 \qquad \pi_2 = 0.246\pi_0 + 0.667\pi_1 + 0.667\pi_2$$

$$\pi_1 = 0.027\pi_2 \qquad \sum_{j=0}^{2} \pi_j = 1$$

The solution of those equations is obtained by using software *Matlab* and given in the vector

$$\pi = [0.5549 \quad 0.0117 \quad 0.4334]$$

Based on the transition probability matrix, the persistence properties for each state are shown in Table 1

Table 1. Persistence Properties of Model 1

Initial State	Persistence Index	Persistence Properties
Jump Down	$\rho_0 = 2(0.754) - (0.333 + 0.306) = 0.869$	Persistence
Stable	$\rho_1 = 2(0) - (0 + 0.027) = -0.027$	Not Persistence
Jump Up	$\rho_2 = 2(0.667) - (0.246 + 0.667) = 0.421$	Persistence

Table 1 shows that the initial state 'jump down' gives the greatest persistence index, it means that if the stock price is in state 'jump down' then the probability to tend to fall is greater than the probability to stay stable or to go up. It is also applied to the initial state 'jump up'. Meanwhile, it is not happened if the initial state is 'stable'. It means that if the initial state is 'stable', then the probability to stay stable is very small than to change to another state.

Model 2: Markov Chain with 5 States

The states defined in Model 2 are '0': drastically down $(X_t \le -0.2)$, '1': jump down $(-0.2 < X_t < 0)$, '2': stable $(X_t = 0)$, '3': jump up $(0 < X_t < 0.2)$, and '4': drastically up $X_t \ge 0.2$. The result of Model 2 is

$$P = \begin{pmatrix} 0.55 & 0.375 & 0 & 0.075 & 0 \\ 0.147 & 0.539 & 0 & 0.294 & 0.02 \\ 0 & 0.333 & 0 & 0.667 & 0 \\ 0.033 & 0.315 & 0.033 & 0.522 & 0.098 \\ 0 & 0.105 & 0 & 0.474 & 0.421 \end{pmatrix}$$

Then we obtain the long-run probability of each state by solving the equations

$$\pi_0 = 0.55\pi_0 + 0.147\pi_1 + 0.033\pi_3$$
$$\pi_1 = 0.375\pi_0 + 0.539\pi_1 + 0.333\pi_2 + 0.315\pi_3 + 0.105\pi_4$$
$$\pi_2 = 0.033\pi_3$$
$$\pi_3 = 0.075\pi_0 + 0.294\pi_1 + 0.667\pi_2 + 0.522\pi_3 + 0.474\pi_4$$
$$\pi_4 = 0.02\pi_1 + 0.098\pi_3 + 0.421\pi_4$$

ISBN:978-602-99849-2-7

$$\sum_{j=0}^{4} \pi_j = 1$$

The solution of those equations is given in the vector π below

$$\pi = [\,0.1564 \quad 0.3981 \quad 0.0119 \quad 0.3591 \quad 0.0745\,]$$

Table 2 shows the persistence properties of Model 2

Table 2. Persistence Properties of Model 2

Initial State	Persistence Index	Persistence Properties
Drastically Down	$\rho_0 = 4(0.55) - (0.147 + 0.033) = 2.02$	Persistence
Jump Down	$\rho_1 = 4(0.539) - (0.375 + 0.333 + 0.315 + 0.105) = 1.028$	Persistence
Stable	$\rho_2 = 4(0) - 0.033 = -0.033$	Not Persistence
Jump Up	$\rho_3 = 4(0.522) - (0.075 + 0.294 + 0.667 + 0.474) = 0.578$	Persistence
Drastically Up	$\rho_4 = 4(0.421) - (0.02 + 0.098) = 1.566$	Persistence

Table 2 shows that the greatest persistence index is when the initial state of the stock price movement is in the state 'drastically down'. Furthermore, the state will also be persistence if the initial state of the stock price is in states 'jump down', 'jump up', and 'drastically up'. Similar to the Model 1, Model 2 also gives the result that the process will not be persistence if the initial state of the stock price movement is in state 'stable'. It means that whenever the price is in 'stable' condition, the probability to remain 'stable' is smaller than to move to the other state.

5. Conclusion

Discrete-Time Markov Chain Analysis (DMCA) is one of some methods which is used by the investors in making a decision whether to buy shares or not. This method is more effective than using another method that predicts the stock price index because the DMCA predict the transition probability of the stock price movements which is less volatile. In addition, we can observe the tendency of the pattern of stock price movements by looking its persistence in the next period. In this paper have been formulated the persistence index of the stock price movements which was proposed in Proposition 1. This index can indicate the behavior of the shares if it is in a certain initial states. The results of the application of DMCA and persistence process show that, in the period time 25 August 2014 until 25 August 2015, the stock price has a tendency to persist like the previous state except when the initial state is stable. It means that if the initial state is either down or up, then it will also be down or up in the next period with greater probability. This analysis gives a general result about persistence process. In addition, we can classify the level of the persistence based on certain criteria. This study can be conducted as a continuation of the discussion which has been done in this paper.

References

[1] Adebiyi, A.A., Adewumi, A.O., "Stock Price Prediction Using the ARIMA Model," *In the Proceeding of AMSS 16th International Conference on Computer Modelling and Simulation*, IEEE Computer Society, 105-111 (2014).

ISBN:978-602-99849-2-7

[2] Dharmawan, Y.V., "*Forecasting Loss Based on Stochastic Processes,*" Bachelor thesis, Department of Mathematics, Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung, 2012.
[3] Doubleday, K.J., Esunge, J.N., "Application of Markov Chains to Stock Trends," *Journal of Mathematics and Statistics* 7(2), 103-106 (2010).
[4] Ross, S.M., "*Introduction to Probability Models, 9th Edition,"* Academic Press, Elsevier, 2007.
[5] Svoboda, M., Lukas, L., "Application of Markov Chain Analysis to Trend Prediction of Stock Index," *In the Proceeding of 30th International Conference Mathematical Methods in Economics,* Silesian University in Opava, School of Business Administration, 848-853 (2012).
[6] Vallois, P., Tapiero, C.S., "A Claim Persistence Process and Insurance," *Insurance: Mathematics and Economics, Elsevier* 44(3), 367-373 (2009).
[7] Xu, S.Y., "Stock Price Forecasting Using Information from Yahoo Finance and Google Trend," https://www.econ.berkeley.edu/sites/default/files/Selene Yue Xu.pdf Retrieved 02 September, 2015.
[8] Zhang, D., Zhang, X., "Study on Forecasting the Stock Market Trend Based on Stochastic Analysis Method," *International Journal of Business and Management* Vol. 4, No. 6, 163-170 (2009).
[9] Zhang, J., Shan, R., Su, W., "Applying Time Series Analysis Builds Stock Price Forecast Model," *Modern Applied Science* Vol. 3, No. 5, 152-157 (2009).
[10] Zhou, Qing-xin, "Application of Weighted Markov Chain in Stock Price Forecasting of China Sport Industry," *International Journal of u- and e- Service, Science and Technology* Vol.8, No. 2, 219-226 (2015).

ISBN:978-602-99849-2-7

Application of Fuzzy Logic to Diagnose Severity of Coronary Heart Disease: Case Study in dr. Zainoel Abidin General Hospital, Banda Aceh Indonesia

Zurnila Marli Kesuma, Hizir, Izazi

Mathematics Department, Faculty of Mathematics and Natural Science Syiah Kuala University Banda Aceh 23111 Indonesia

kesumaku@yahoo.com, hizir@unsyiah.ac.id, izazi@gmail.com

Abstract: Fuzzy logic is a mathematical tool for dealing with uncertainty. The application of fuzzy logic method has covered various fields, including in medical application. This study aims to diagnose the severity of coronary heart disease (CHD) using Mamdani inference method. The designed system based on patients' medical records in Dr. Zainoel Abidin General Hospital, Banda Aceh. Risk factors that consist of LDL cholesterol, age, blood pressure, fasting blood sugar and smoking histories were used to examine the rate of severity. The result indicated that there were 45% of patients who were high to very high severity, and 55% were mild to moderate Thus, the severity of CHD is associated with risk of myocardial infarction.
Keywords: Fuzzy Logic; Mamdani methods; Coronary heart disease; Myocardial infarction; Risk factor

1. Introduction

Fuzzy logic is a branch of mathematics that is growing quite rapidly. It has been applied in various fields, including the medical field. Most of the medical information is vague, imprecise and uncertain. One of the problems in medicals that could be solved using fuzzy logic is disease diagnosis.

In the construction of fuzzy logic, Samples were taken and selected to represent a disease, then the rule of fuzzy were conducted accurately to strengthen the accuracy of disease diagnosis. This condition not only could reduce the time of diagnosis but also could achieve the low mortality rate.

One of the diseases that become the number one leading causes of death in the world is coronary heart disease (CHD) [1]. According to a survey conducted by AIA [2], heart disease (53%) is a medical condition that most feared by most adults in Indonesia. It was ranked eight of the ten major diseases of leading cause of death in Indonesia [3]. Government declared that 15 provinces in Indonesia have high heart disease prevalence above the national rate (1.5%), including Aceh Province that occupy the top five positions (2.3%) [4].

On the other side, heart disease also creates a high economic burden for the country through health financing. It is one of the most prevalent diseases in the outpatient and inpatient in Indonesia using National Health Insurance in 2012. Total costs incurred for outpatient advanced heart disease is Rp.3,264,033,343, while for inpatient are Rp.19,731,040,425 [5].

To reduce losses caused by heart disease, it is necessary to plan early prevention system through early diagnosis based on the severity of the risk factors.

This study aims to perform early diagnosis of heart disease severity on outpatients in January 2014 at the Regional General Hospital Dr. Zainoel Abidin (RSUDZA), Banda Aceh using fuzzy logic with risk factors as input variable.

2. Literature Review

Early diagnosis on heart disease is important aspect to deal with reducing mortality rate. Experiences in different countries showed that fuzzy expert system has been functioning effectively related to the diagnosis of heart disease. several variables such as chest pain type, blood pressure, cholesterol, resting blood sugar, resting maximum heart rate, sex, electrocardiography

ISBN:978-602-99849-2-7

(ECG), exercise, old peak (ST depression induced by exercise relative to rest), thallium scan and age were used for inputs and the status of the patients as healthy or sick were defined as output [6].

The heart disease can be ascertained by using seven attributes of patient as input values from the Clevand database as implementation of fuzzy rule base. The data were partitioned into several intervals based on certain intermediate values of the available data values [7].

Based on the fuzzy value of the output variable, Kumar and Kaur [8] have designed system for CMC and Civil Hospital Amritsar. The system uses 6 attributes for input and 2 attributes for result. Input fields (attributes) are chest pain type, blood pressure, cholesterol, resting blood sugar, resting maximum heart rate, old peak (ST depression induced by exercise relative to rest). The output field refers to the presence of heart disease in the patient and the Precautions according to the risk. Integer values were used from 0 (no presence) to 1. Increasing value shows increasing heart disease risk.

In their study, they use low density lipoprotein (LDL) cholesterol and systolic blood pressure. In a dataset, fields were divided into some sections and each section has a value. For instance, in dataset, chest pain has 4 section (very low, low, normal, high and very high), resting blood sugar has five section (very low, low, normal, high and very high).

All the works explained above used the data to develop membership function. However, in this study, each variable is partitioned based on the gold standard used in medicals, with a few changes. For example, blood pressure variables divided based on JNC7 [9]. In JNC7, they were divided into four sets, but in this research they were divided into three sets, i.e. normal (90-139), hypertension 1 (140-159) and hypertension 2 (> 160). The system to diagnose severity of CHD was implemented by using fuzzy logic (Mamdani). We used five input variables that are LDL cholesterol, age, blood pressure, fasting blood sugar and smoking histories. Severities of CHD which have 4 fuzzy sets were used as output variable. They are mild, moderate, severe and very severe.

3. Material & Methodology

Data

The study used data from outpatients of RSUDZA, Banda Aceh in January 2014. Total data that was found for this study were 200 patients. They must have complete information to include in this study. In the first step, patients' identification was found from administration unit. For the second round, we found that only 55 patients had all of the variables that would be need in this study in their medical record.

Eligibility Criteria for Study Participants

Eligibility criteria included patients that had symptoms of CHD, including information the level about LDL, blood pressure and fasting blood sugar, aged between 20 years old to over. They must have normal to over level of those variables. For the patients who have lower than normal level of the variables were excluded in this study.

Analysis

Determination of severity of heart disease in this study was done by using MATLAB. MATLAB provides the tools Fuzzy Logic Toolbox (FLT) to create fuzzy inference system (FIS). Thus, we can check the diseases and risks in the patient according to the values of the variables. If the values of the attributes or inputs are high, then the patient would have high risk and vice versa. Similarly, if the values are normal then the patient and results showed that the patient was normal.

Membership function

The first step of designing system in fuzzy logic is determination of input and output variables. There were five input and one output variables. From that, the membership functions (MF) of all variables were designed. MF was used to determine the membership of objects to fuzzy sets. For example, the following is the membership function for LDL variable.

ISBN:978-602-99849-2-7

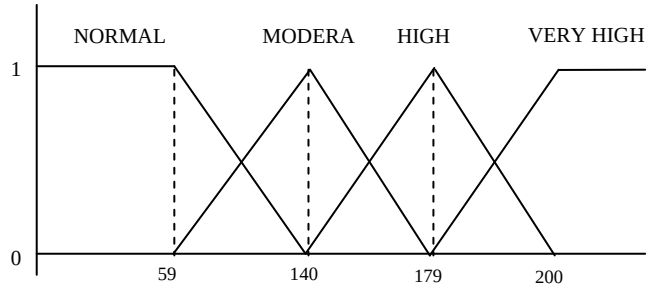


Figure 1 Membership functions of blood pressure

$$\mu_{NORMAL}(x) = \begin{cases} 1 & ; \quad x \leq 59 \\ \dfrac{140 - x}{81} & ; \quad 59 < x < 140 \\ 0 & ; \quad x \geq 140 \end{cases}$$

$$\mu_{SEDANG}(x) = \begin{cases} 0 & ; \quad x \leq 59 \; atau \; x \geq 179 \\ \dfrac{x - 59}{81} & ; \quad 59 < x \leq 140 \\ \dfrac{179 - x}{39} & ; \quad 140 < x \leq 179 \end{cases}$$

$$\mu_{TINGGI}(x) = \begin{cases} 0 & ; \quad x \leq 140 \; atau \; x \geq 200 \\ \dfrac{x - 140}{39} & ; \quad 140 < x \leq 179 \\ \dfrac{200 - x}{21} & ; \quad 179 < x \leq 200 \end{cases}$$

$$\mu_{SANGAT\ TINGGI}(x) = \begin{cases} 0 & ; \quad x \leq 179 \\ \dfrac{x - 179}{21} & ; \quad 179 < x < 200 \\ 1 & ; \quad x \geq 200 \end{cases}$$

Fuzzy rule base

A rule base is the main part in fuzzy inference system and quality of results in a fuzzy system depends on the fuzzy rules. It consists of a set of fuzzy IF-THEN rules. This system included 144 rules.

Fuzzification and defuzzification.

In the fuzzification crisp data values for each variable is converted into fuzzy values (degree of membership) through membership functions. The membership functions of shoulder which is a combination of membership function of linear down, triangle, and linear up were used in this study. Centroid method was applied for defuzzification here, due to its good plausibility and continuity [10].

ISBN:978-602-99849-2-7

4. Results and Discussion

Result

Table 1 presents the different values of different input variables and their results accordingly. If the values of the inputs lie in their low ranges then the risk is also low, as the consequences of minimum value. For the values of the input, would also go to the same way. The severity was quite high (79.4%) for patients who have very high LDL cholesterol, middle age, high blood pressure (hypertension 2), fasting blood sugar exceeds normal limits (diabetes), although they did not smoke.

Table 1. System Testing

LDL Cholesterol	Age	Blood Pressure	Fasting Blood Sugar	Smoking History	Severity
45	55	110	112	no	15,5%
116	61	110	99	no	23,5%
42	77	120	175	no	40%
156	61	90	187	no	69,2%
118	57	140	297	yes	73%
239	46	120	176	yes	74,5%
205	50	190	267	no	79,4%

Table 2.Classification of the severity of CHD patients

Severity	Range	Number of Patients
Mild	15,5% - 31,5%	10
Moderate	31.6% - 47,5%	20
High	47,6% - 63,5%	17
Very High	63,6% - 79,4%	8

Table 2 shows the range of patients' severity. In general, patients have high severity. 45% of them had high to very high severity; with the largest percentage was 79.4%.

Table 3.. Characteristics of patients with very high severity

No	LDL Cholesterol	Age	Blood Pressure	Fasting Blood Sugar	Smoking history
1	Very high	Middle Age	Hypertension 2	Diabetic	No
2	Very high	Middle Age	Normal	Diabetic	Yes
3	Moderate	Middle Age	Hypertension 1	Diabetic	Yes
4	Moderate	Middle Age	Normal	Diabetic	Yes
5	High	Middle Age	Normal	Diabetic	Yes
6	High	Old	Normal	Diabetic	Yes
7	High	Middle Age	Normal	Diabeteic	Yes
8	High	Young	Normal	Diabetic	Yes

Table 3 indicates patients' characteristics with the very high severity. Patients who had very high LDL cholesterol, diabetics, and Hypertension 2 would be included in very high severity even though they did not have smoking history. On the other hand, the one who had high cholesterol, had diabetic and had smoking history would also had very high severity whereas they were still young and had normal blood pressure.

Discussion

ISBN:978-602-99849-2-7

In this study, the lowest of severity score of coronary heart disease was 15. 5% and the highest one was 7.4%. These results were conformed to the study from Kumar [8]. As we mentioned above that the severity definition in this research is the prediction of someone's having a myocardial infarction. Furthermore, all of data in this study were classified with these provisions.

The severity of high and very high of CHD patients can cause the patients to have risk of myocardial infarction greater than mild and moderate severity. As declared by Wong [11], that nonfatal or fatal MI or sudden coronary deaths is typically included as 'hard' CHD end points in clinical trials. However, smoking as a risk factor cannot be ignored. From Table 2, out of eight patients, only one did not have smoking history that has risk of high severity. This is corresponded with the study conducted by Teo [12], that tobacco use is one of the most important causes of acute myocardial infarction globally, especially in men. Also Prescott mentioned that relative risk of myocardial infarction increased with tobacco consumption in both men and women [13]. On the other hand, 79.4% are patients who have very high LDL cholesterol, middle age, high blood pressure (hypertension 2), fasting blood sugar exceeds normal limits (diabetes), although the patient did not smoke.

A study from USA showed that the prevalence of acute myocardial infarction in young patients (<46 years), was more than 10%, and 1 in 10 patients with AMI were young [14]. This result was in line with the results on this study (the eighth row from table 3). It is possible that someone with young age have a very high severity. For blood pressure factor, only two of the eight patients with high severity were get suffer hypertension, perhaps through the fact that 96% of the blood pressure data in this study were within the normal range. It may not have the same line with some references, as blood pressure was included in the most influential risk factor. And that hypertension were significantly affected the risk of myocardial infarction [15].

5. Conclusion

This study has been successfully built a system to diagnose the severity of heart disease using fuzzy logic. The membership functions, input variables, output variables and rule base were conducted by involving a competent expert in Coronary Heart Disease field. There were 45% of the patients had a severity high to very high, and 55% are mild to moderate. The cause of the severity of high and very high is due to the complications of other diseases is hyperlipidemia, hypertension and diabetes.

The first limitation of this study is that the result of severity status of patients was not validated by the hospital standard, due to unavailability of data. Then, the accuracy of the system was not measured. Secondly, there are several medical and non-medical procedures that can be done to prevent the severity of CHD included modifying lifestyles to reduce the risk factors, gender, body mass index, etc. [16] that were not included in this study. Thirdly, due to the small sample size in this study, our result can not be generalizable to all hospital.

Therefore, further research is expected to perform the accuracy of this system. The sample size and the scope of the research by involving more other risk factors should be further improved. The determination is needed since a healthy lifestyle is predicated upon long-term behavior [17].

Acknowledgement.

We sincerely acknowledge the support of dr. Nurkhalis, Sp,JP and all staff in RSUZA Banda Aceh and Mr. Marzuki, M.Si a colleague in Faculty of Mathematics and Natural science Syiah Kuala University, for a very valuable suggestion in this study.

References

[1] WHO, *projections of mortality and causes of death 2015 and 2030,* http://www.who.int/healthinfo/global_burden_disease/projections/en/ Retrieved 05 Maret, 2015.

ISBN:978-602-99849-2-7

[2] AIA. http://www.aia.com/en/resources/30f22200423d273fa2b8ea0f2cbf0f90/AIA_Healthy_Living Index_Survey_2013.pdf Retrieved 07 April, 2015.

[3] Indonesian Health Departement, "*Profil Kesehatan Indonesia 2005,"* Jakarta, (2007).

[4] Department of Research and Development of Health, "*Riskesdas dalam Angka Provinsi Aceh 2013,"* Jakarta, (2013).

[5] The Ministry of Health of the Republic of Indonesia, "*Situasi Kesehatan Jantung,"* Center of Data and Information of the Ministry of Health, South Jakarta, (2014).

[6] Adeli, A., and Neshat, M., "A Fuzzy Expert System for Heart Disease Diagnosis,". *In The Proceedings of the International Multi Conference of Engineers and Computer Scientists*, IMECS, 136-139 (2010).

[7] Barman, M., and Choudgury, J.P., "A Fuzzy Rule Base System fot the Diagnosis of Heart Disease," *International Journal of Computer Applications* 57(7), 46-53 (2012).

[8] Kumar, S., and Kaur, G., "Detection of Heart Diseases using Fuzzy Logic," *International Journal of Engineering Trends and Technology* 4(6), 2694-2699 (2013).

[9] National High Blood Pressure Education Program, "The Seventh Report of the Joint National Committee on Prevention, Detection and Treatment of High Blood Pressure," 2004.

[10] Wang, L., "*A Course in Fuzzy Systems and Control*," Prentice-Hall International, 1997.

[11] Wong, N.D., "Epidemiological Studies of CHD and the Evolution of Preventive Cardiology*," Nature Reviews Cardiology* 11(5), 276-289 (2014).

[12] Teo, K.K., et al., "Tobacco Use and Risk of Myocardial Infarction in 52 countries in the interheart Study: A Case-Control Study," *The Lancet*. 368(9536), 647-658 (2006).

[13] Prescott, E., et al., "Smoking and Risk of Myocardial Infarction in Women and Men: Longitudinal Population Study," *BMJ*. 316(7137), 1043 (1998).

[14] Doughty, M., et al., "Acute Myocardial Infarction in the Young- The University of Michigan experience," *American Heart Journal*. 143(1), 56-62 (2002).

[15] Deshpande J.D. and Dixit J. V. "Risk Factors for Acute Myocardial Infarction: A Hospital-Based Case Control Study," *Health and Population Perspectives* 31(3), 164-169 (2008).

[16] Fulcher, G.R., Conner, G.W and Amerena, J.V., "Prevention of Cardiovascular Disease: An Evidence-Based Aid 2004," *Medical Journal of Australia* 181(6), 1-14 (2004).

[17] P.L.da Luz., M. Nishiyama and A.C.P. Chagas., "Drugs and Lifestyle for the Treatment and Prevention of Coronary Artery Disease – Comparative Analysis of the Scientific Basis," *Brazilian Journal of Medical and Biological Research* 44(10), 973-991 (2011).

ISBN:978-602-99849-2-7

Statistics Application On Terestrial Phenomena Of Metallic Mining's Activity

Eddy Winarno[1], Ira Mughni Pratiwi[2], Abdul Rauf[1]

[1]Lecture of Mining Department of UPN 'Veteran' Yogyakarta
[2]Student of Mining Department of UPN 'Veteran' Yogyakarta
winarnoeddy@gmail.com ; mughniira@gmail.com; abdulrauf_nuke@yahoo.co.id

Abstract: Stages of mining activities, especially mineral and coal start from the determination of activities prospected area (prospecting), the quantity and quality (exploration), feasibility of mining operation (exploitation), processing metallurgically (processing), and marketing (marketing). The linkages of mining activities require a whole series of data accuracy, either at the time of data acquisition and processing, data analysis and interpretation. Terrestrial phenomenon of the existence of a mining commodity (genesis) is something unique and specific, linked between one parameter with others that influence the determination of the location, amount and accuracy of sampling (sampling data), as well as the procedure of sample treatment.
At this research, statistical application will be assessed against terrestrial phenomena that can provide data accurately, hence the value of conservation minerals and coal can be exploited optimally and has added value (optimally added value). Statistics application are applied to terrestrial phenomena then known as geostatistics. Moreover, It requires a good understanding of the data, normal distribution, data probability, and sampling techniques. Geostatistic output then is being based in quality control (statistical quality control) in the mining activity, metallurgical process, and marketing (statistical trend analysis). As a case study, will be conducted on metallic mineral mining activity.
Keywords: terrestrial phenomena, data accuracy; conservation; and value added.

1. Introduction

Mineral and coal mining activities that started from the prospected area determination (prospecting) to the ready-to-be-sold product for the market (marketing) is strongly influenced by the statistical data exploration accuracy. Purposes of statistical science at every stages of mining activity are shown in Table 1 below.

Table 1. Mine Stages and Statistics Correlation

Mine Stages	Results	Statistics Science
Prospecting	Prospected Area	Probability
Exploration	Resources (m^3 or ton)	Statistics and Geostatistics
Feasibility Study	Feasible or not tobe mining	Inferential Statistics
Exploitation	Reserves (m^3 or ton)	Geostatistics
Metalurgical Processing	Concentrates or Processed-Metal	Quality Control Statistics and Trend Analysis
Marketing		

The existence of mineral and coal in nature, known as the genesis of minerals and coal, is unique and specific due to the specific parameter form (e.g levels of a precipitate). In addition, is strongly influenced by the presence of other minerals forming parameters (heat, temperature, and others). Therefore, the accuracy of determining the location of the sample (sampling techniques), the amount of samples, and gathering sample (sampling data) and sample treatment procedure are a requirement that must be met.

Uniqueness and specific properties of a mineral and coal are not random, but there is a relation between genesisly. Therefore, a statistical approach that is widely used in terrestrial phenomena also known as geostatistics.
Geostatistics is statistic-science, which is applied for terrestrial phenomena; with the nearby data has a huge influence. The farther distance, the less data influenced.

2. Scale of Theory

2.1. Resources and Reserves

Classification of resources and the reserve has been published by the Australasian IMM / AMIC base on classification accuracy improvement and the results of geological investigations (Table 2).
Table 2 indicates that the increased resources into reserves to account for economic factors, mining, processing, market, environment, and government regulation.

Table 2. AIMM/AMIC Classification of identified minerals resources

Identified mineral resources (in situ)	Ore reserves (mineable)	Increasing level of geological knowledge and confidenec
Inferred **Indicated** ⟷ **Probable** Consideration of economic, mining, metallurgical, marketing, enviromental, social and govermental factors **Measured** ⟷ **Proved**		↓

Sources :Mineral Deposit Evaluation, A.E. Annel (1991)

Base on Table 2, Diehl and David in A.E. Annels (1991) develop a classification of ore deposits involves a degree of uncertainty (assurance) and degree of accuracy (error tolerance) for each of the different deposits. It was stated comprehensively in Table 3 below.

Table 3. Ore Reserve Classification

Identified				Undiscovered	
Demonstrated					
Measured		Indicated			
Proved	Probable	(Possible)	Inferred	Hypothetical	Speculative
± 10%*)	± 20%	± 40%	± 60%		
> 80%$)	60-80%	40-60%	20-40%	10-20%	< 10%
Economically significant resources				Resources base	

Sources :Mineral Deposit Evaluation, A.E. Annel (1991)
*) : *Error tolerance* $) : *Assurance*

Table 3 state that for proven reserves has an error tolerance of ± 10% and above 80% accuracy rate.

2.2. Error Tolerance and Level of Accuracy

In statistical theory, the probability of a confidence levels of the data distribution can be formulated by

$$P(\overline{X} - Z_{\alpha/2}.\overline{\sigma} < \mu < \overline{X} + Z_{\alpha/2}.\overline{\sigma}) = 1-\alpha$$

with: :

$\overline{X}$ = Average of sample data
$\overline{\sigma}$ = Standard deviation of population data
μ = population mean
n = a mount of samples
$1-\alpha$ = level of confidence
α = fault tolerance (accuracy)
$Z\alpha/2$ = normal table value

ISBN:978-602-99849-2-7

Various values of the confidence level can be described by a normal distribution table and summarized in Table 4.

Table 4. Z value at Various Confidence Levels

Level Confidence (1-α)	Accuracy (α)	Value Z α/2	Deviation Reserve
65%	35%	0.935	$0.935.\overline{\sigma}$
80%	20%	1.282	$1.282.\overline{\sigma}$
90%	10%	1.645	$1.645.\overline{\sigma}$
95%	5%	1.960	$1.960.\overline{\sigma}$
99%	1%	2.575	$2.575.\overline{\sigma}$

Reserve estimates, especially for the ore, it is generally formulated as follows:

$$\text{Re}\,serves = Volume.x.Density..x.Grade. \pm Deviation$$

Deviation is the backup has been taken of the level of confidence (accuracy) and the estimation error tolerance.

2.3. Geostatistics

Geostatistics is the statistics of spacially or temporally correlated data. Geostatistics is concerned with the study of phenomena that fluctuate in space. Geostatistics offers a collection of deterministic and statistical tools aimed at understanding and modelling spacial variability.

2.3.1. Variography

The semivariogram is the basic geostatistical tool for visualizing, modelling, and exploiting the spacial autocorrelation of a regionalized variable. As the name implies, a semivariogram is measure of variance, illustrated in figure 2 and the formula.

$$\gamma^*(h) = \sum [Z(x) - Z(x+h)]^2/2n$$

with :

$\gamma^*(h)$ = semivariogram at site h units distance h = distance, m

$Z(x+h)$ = Z variable between site x and a site h units distance 2n = number of variable pairs

For modelling the semivariogram through iterative or least-square methods, practitioners recommend actual inspection of the observed semivariogram and the fitted model. If the surface represented by two points is continuous and h is small distance, one expects that the difference is small as well. If h were very large relative to the spacial degree of change in the variable, then the difference might be expected to increase.

The difference both the expected variance (model, γ^*) and the real inspection (γ) will be small, that can be expressed with :

$Deviation = \sum_{i=1}^{n} [\gamma^* - \gamma]$ is small

2.3.2. Kriging

In the geostatistics theory, the deviation with small value, can be expressed by Kriging standard deviation (illustrated in figure 2).

ISBN:978-602-99849-2-7

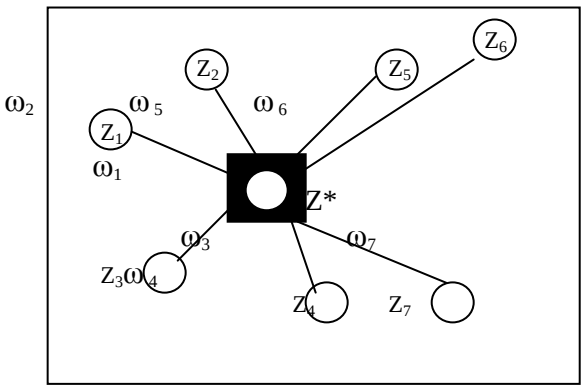


Figure 2. Point Kriging Orientation

Grade of A block can be estimated with $Z^* = \sum_1^n \omega_i . Z_i$;

with : ω : weigted factor ; Z_i : real block with grade-i ; Z* : estimate block

Figure 3 show that :

a. Var[Z,Z*] = $\sigma^2 + \sum_{i=1}^{n}\sum_{i=1}^{n} \omega_i . \omega_j . Cov(z_i, z_j) - 2. \sum_{i=1}^{n} \omega_i . Cov(z_i, z^*)$

b. To reach the maximum or minimum value of Var[Z,Z*] = 0, need to support with the Lagrange Function : $2.\mu.(\sum_{i=1}^{n} \omega_i - 1)$ with $\sum_{i=1}^{n} \omega_i = 1$.

c. Var[Z,Z*] = $\sigma^2 + \sum_{i=1}^{n}\sum_{i=1}^{n} \omega_i . \omega_j . Cov(z_i, z_j) - 2. \sum_{i=1}^{n} \omega_i . Cov(z_i, z^*) + 2.\mu.(\sum_{i=1}^{n} \omega_i - 1)$

d. The differencial of Var [Z,Z*] : $\frac{d}{d\mu}\sigma_k^2 = 0$ and $\frac{d}{d\omega}\sigma_k^2 = 0$

reachminimum value, and the result is : $\sigma_k^2 = \sigma^2 - \sum_{i=1}^{n} \omega_i . Cov(z_i, z^*) + \mu$

e. This function can be writen in matrix formula :

$$\begin{vmatrix} \gamma(Z_1,Z_1) & \gamma(Z_1,Z_2) & \ldots\ldots & \gamma(Z_1,Z_n) & \omega_1 \\ \gamma(Z_2,Z_1) & \gamma(Z_2,Z_2) & \ldots\ldots & \gamma(Z_2,Z_1) & \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \\ \gamma(Z_n,Z_1) & \gamma(Z_n,Z_1) & \ldots\ldots & \gamma(Z_n,Z_n) & \omega_3 \\ 1 & 1 & & 1 & \mu \end{vmatrix} \begin{vmatrix} \\ \omega_2 \\ \\ \\ \end{vmatrix} = \begin{vmatrix} \gamma(Z_1,Z^*) \\ \gamma(Z_2,Z^*) \\ \ldots\ldots \\ \gamma(Z_3,Z_n) \\ 1 \end{vmatrix}$$

Note : $\gamma(Z_1,Z_n)$ = variogram point Z_1 to point Z_n

f. Solution of matrix get the value of ω (weigted factor) and μ (Langrange factor).

3. Reserves Estimate

3.1. Characteristics of Ore

Reserve estimation carried out simulations on the " X nickel deposits". Exploration activity that has been done is taking samples with a regular distance of 25 m by using a rotary drilling tool. Topography of the hills with a slope of 30^0-50^0 and 50-230 meters above sea level. Based on how the formation, geology of ore deposits is a nickel laterite ore, mineral deposit is the result of the weathering of ultra basic rock peridotite, in general, contain elements of iron, cobalt and klorium. This ultramafic rock outcrops generally have undergone weathering, yellow-brown mottled gray, black or white with a greenish tint on the outer edge or rim. In this area there are also small cracks, fractures are commonly filled by secondary minerals (silica and magnesite).

In general profiles ore deposits in the study area are as follows:

a. Top Soil, ground cover is reddish brown, there are the rest of the herbs.

ISBN:978-602-99849-2-7

b. Limonite, is the result of weathering of the soil soft yellowish brown color containing nickel and iron in the ratio is not necessarily.
c. Saprolite, is highly weathered soils have yellowish brown to greenish with many veins garnierit and onyx, has a relatively high nickel content.
d. Bed Rock, a peridotite host rock that has not weathered serpentinite.

3.2. Estimate Simulation

Estimated reserves of nickel ore deposits base on data illustrated in the figure 2 and Table 5.

Table 5. Drillhole Data

Sample		Coordinate		Grade,	Distance
No.	Code	E	N	ppm	From Z (63E, 137N) to no sample
1	225	61	139	477	4,5
2	437	63	140	696	3,6
3	367	64	129	227	8,1
4	52	68	128	646	9,5
5	259	71	140	606	6,7
6	436	73	141	791	8,9
7	366	75	128	783	13,5

a. Determined semivariogram value from one point to another, $\gamma(Z_1,Z_1)$; $\gamma(Z_1,Z_2)$; $\gamma(Z_1,Z_3)$; $\gamma(Z_1,Z_4)$; $\gamma(Z_1,Z_5)$; $\gamma(Z_1,Z_6)$; $\gamma(Z_1,Z_7)$; $\gamma(Z_2,Z_2)$; $\gamma(Z_2,Z_3)$; $\gamma(Z_2,Z_4)$; $\gamma(Z_2,Z_5)$; $\gamma(Z_2,Z_6)$; $\gamma(Z_2,Z_7)$; $\gamma(Z_3,Z_3)$; $\gamma(Z_3,Z_4)$; $\gamma(Z_3,Z_5)$; $\gamma(Z_3,Z_6)$; $\gamma(Z_3,Z_7)$; $\gamma(Z_4,Z_5)$; $\gamma(Z_4,Z_6)$; $\gamma(Z_4,Z_7)$; $\gamma(Z_5,Z_6)$; $\gamma(Z_5,Z_7)$; $\gamma(Z_6,Z_7)$; $\gamma(Z_7,Z_7)$; $\gamma(Z_0,Z_1)$; $\gamma(Z_0,Z_2)$; $\gamma(Z_0,Z_3)$; $\gamma(Z_0,Z_4)$; $\gamma(Z_0,Z_5)$; $\gamma(Z_0,Z_6)$; and $\gamma(Z_0,Z_7)$

b. From output of Geostatistical program (GS$^+$7), Solution the kriging matrices :

$$\mathbf{C} = \begin{bmatrix} \bar{C}_{11} & \bar{C}_{12} & \bar{C}_{13} & \bar{C}_{14} & \bar{C}_{15} & \bar{C}_{16} & \bar{C}_{17} & 1 \\ \bar{C}_{21} & \bar{C}_{22} & \bar{C}_{23} & \bar{C}_{24} & \bar{C}_{25} & \bar{C}_{26} & \bar{C}_{27} & 1 \\ \bar{C}_{31} & \bar{C}_{32} & \bar{C}_{33} & \bar{C}_{34} & \bar{C}_{35} & \bar{C}_{36} & \bar{C}_{37} & 1 \\ \bar{C}_{41} & \bar{C}_{42} & \bar{C}_{43} & \bar{C}_{44} & \bar{C}_{45} & \bar{C}_{46} & \bar{C}_{47} & 1 \\ \bar{C}_{51} & \bar{C}_{52} & \bar{C}_{53} & \bar{C}_{54} & \bar{C}_{55} & \bar{C}_{56} & \bar{C}_{57} & 1 \\ \bar{C}_{61} & \bar{C}_{62} & \bar{C}_{63} & \bar{C}_{64} & \bar{C}_{65} & \bar{C}_{66} & \bar{C}_{67} & 1 \\ \bar{C}_{71} & \bar{C}_{72} & \bar{C}_{73} & \bar{C}_{74} & \bar{C}_{75} & \bar{C}_{76} & \bar{C}_{77} & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 10.00 & 5.11 & 0.44 & 0.20 & 0.49 & 0.26 & 0.05 & 1.00 \\ 5.11 & 10.00 & 0.36 & 0.20 & 0.91 & 0.49 & 0.06 & 1.00 \\ 0.44 & 0.36 & 10.00 & 2.90 & 0.20 & 0.11 & 0.36 & 1.00 \\ 0.20 & 0.20 & 2.90 & 10.00 & 0.24 & 0.15 & 1.22 & 1.00 \\ 0.49 & 0.91 & 0.20 & 0.24 & 10.00 & 5.11 & 0.22 & 1.00 \\ 0.26 & 0.49 & 0.11 & 0.15 & 5.11 & 10.00 & 0.19 & 1.00 \\ 0.05 & 0.06 & 0.36 & 1.22 & 0.22 & 0.19 & 10.00 & 1.00 \\ 1.00 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00 & 0.00 \end{bmatrix}$$

The D matrix is

$$\mathbf{D} = \begin{bmatrix} \bar{C}_{10} \\ \bar{C}_{20} \\ \bar{C}_{30} \\ \bar{C}_{40} \\ \bar{C}_{50} \\ \bar{C}_{60} \\ \bar{C}_{70} \\ 1 \end{bmatrix} = \begin{bmatrix} 2.61 \\ 3.39 \\ 0.89 \\ 0.58 \\ 1.34 \\ 0.68 \\ 0.18 \\ 1.00 \end{bmatrix} \qquad \mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \\ w_7 \\ \mu \end{bmatrix} = \mathbf{C}^{-1} \cdot \mathbf{D} = \begin{bmatrix} 0.173 \\ 0.318 \\ 0.129 \\ 0.086 \\ 0.151 \\ 0.057 \\ 0.086 \\ 0.907 \end{bmatrix}$$

c. Estimated value $Z^* = \sum_1^n \omega_i . Z_i$

Z* = (0,173)(477) + (0,318)(696) + (0,129)(227) +(0,086)(646) + (0,151)(606) + (0,57)(791) +(0,086)(783) =592 ppm

d. Kriging variance $\sigma_k^2 = \sigma^2 - \sum_{i=1}^n \omega_i . Cov(z_i, z^*) + \mu$

Z_k^2 = 10 – (0,173)(2,61) – (0,318)(3,39) – (0,129)(0,89) – (0,086)(0,58) – (0,151)(1,34) – (0,057)(0,68) – (0,086)(0,18) + 0,907 Z_k^2 = 8,96 ppm^2 and Z_k = 2,99 ppm

e. The reserves of nickel ore on Z* block for one ton resources is : (592 $\pm$ 2,99) ppm

4. Discussion

Reserves estimate of nickel ore on Z* block (see figure 2), can be discribed with 3 methods are :

a. Nearest Neightbor Point (Poligon), for :
   - Grade of nickel ore, equal to nearest point with minimum distance. Here, grade of nickel ore on 0 block equal with Z_2 = 696 ppm (distance point 0 to point 2 is nearest distance, Table 5)
   - Deviation standard are computed to all point, point 1 until point 7, σ = 198,11ppm
b. Inverse Distance Weight (IDS), for :
   - Grade of nickel ore are estimated base on distance weighted factor. The estimated value is

$$Z^* = \sum_1^n \frac{1}{d_i} . Z_i \Big/ \sum_{i=1}^n \frac{1}{d_i}$$

Here, grade of nickel ore on 0 block equal with Z_2 = 488,87 ppm
   - Deviation standard are computed to all point, point 1 until point 7, σ = 198,11ppm

ISBN:978-602-99849-2-7

c. Ordinary Kriging
   - Grade of nickel ore on 0 block , Z_2 = 592 ppm (sub chapter 3.2).
   - Deviation standard, σ = 2,99 ppm

From the result estimate of grade nickel ore above, Ordinary Kriging has moderat value, not to big and not to small. And deviation standard value has smallest value, that can be concluded has best accurate value.
That concluded can be reached if :

1. The base data must be valid and reliable, both in prospecting stage (sampling tecnics), exploration (procedure of data processing), exploitation (prediction and validation on target area), metalurgical processing (feed, process and product quality control), and marketing (product quality control).
2. In the future advantage, resources or reserves can be improved to increase added value with optimally exploitation.

5. Conclusion

To reached the higher accuracy prediction of resources or reserves metallic deposits can be expreseed with standard deviation value. Too smallest standard deviation value, too accurate prediction result.

References

[1] Winarno, E., Aplikasi dan Pengaruh Cut-Off Grade dalam estimasi cadangan bahan galian , *Jurnal Ilmu Kebumian Teknologi Mineral,* Vol.23 Nomer 2 Mei-Agustus 2010, UPN "Veteran" Yogyakarta, 73-80 (, 2010).

[2] Winarno, E., K. Gunawan, T. Wahyuningsih, R.Z. Mirahati, *Accuracy Statement Of Ore Deposits Reserves Estimation,* International Symposium on Earth Science and Technology, Bandung, (2012).

[3] Annel, A.E, *Mineral Deposit Evaluation : A practical approach*, Chapman & Hall, London, p.99-212 (1991).

[4] Clark., I., *Practical Geostatistics,* Chapman & Hall, London, (1979)

[5] Carras, S., *Sampling Evaluation And Basic Principles of Ore Reserve Estimation,* Carras Mining & Associates, Australia, p.40-83 (1986).

ISBN:978-602-99849-2-7

Empirical Study of Student's Stage Thinking According To Bloom And Van Hiele Learning Theories In Mathematics Instruction

Noening Andrijati, Budi Harjo, Zaenal Arifin

Semarang State University, Semarang
SMA Islam Al Azhar 7 Solo Baru, Sukoharjo
SMK NU Dukuhturi, Tegal

andrijt06@gmail.com, budisaroh2001@yahoo.co.id., arifin.zaenal36@yahoo.com

Abstract: Problems to be solved by this study are the teachers' understanding of the application of learning theory according to Bloom and Van Hiele in mathematics, how the results of empirical testing and simulation of the application of the theory of learning, as well as how the consistency of both the theory learned in classifying stages of student's thinking. The population in this study is the whole second semester of eighth grade students of SMP Negeri 1 Dukuhturi Tegal 2013/2014 school year, with a total of 315 students. Instruments in this study using multiple choice test on the subject matter of the circle with the number of questions for each device test as many as 20 items. Empirical data collected through documentation, tests and interviews, as well as a simulation of the 1000 samples in ten replication for each form of the test. Analysis of the level of difficulty of a test item on empirical data using Rascal. Simulation difficulty level of data generated with real data using WinGen3 program. From empirical research on real data and simulation results are as follows: (1) a good understanding of teachers on the application of learning theory Bloom and Van Hiele; (2) classification of the thinking stage of students to learn Bloom's theory is more consistent than the Van Hiele; and (3) consistency in the application of learning theories in Bloom and Van Hiele can be proved empirically.
Keywords: simulation; empirical testing; Bloom and Van Hiele learning theories

1. Introduction

Learning is a complex thing. Learning complexity can be seen from two point of view, student's and teacher's. From student's point of view, learning is a series of processes while on the teacher side, learning is performed in learning attitude. The aim of learning and teaching process is to improve the student's cognitive, affective and psychomotoric ability.

According to Bloom's taxonomy, the purpose mapping should always refers to three domains which are attached in students, they are 1. Cognitive domain 2. Affective domain and 3. Psychomotor domain. In the context of learning output, those three domains should always be the purpose of the learning output. Cognitive domain covers the mental activity (brain). According to Bloom, the whole effort dealing with brain deals with cognitive domain. In the cognitive domain, there are six step of thinking process. They are knowledge, understanding, application, analysis, syntesis, and evaluation.

Circle is a material in geometry unit. In geomatry learning, there lies a theory specifically classifies stuident's thinking steps. It is the theory of Van Hiele. Van Hiele talks about student's yhinking steps in learning about geometry. Student may not come to the next level without passing the previous lower level. There are three aspects in this theory, they are existence in every level, characteristic in every level and the movement of one level to another. This theory contains students main thinking ability in geometry accordingly, they are : thinking level 0: introduction to thunking level 1: analysis of thinking level 2: sorting level thinking 3: deduction of thinking level 4: accuration [7].

Basically, Bloom and Van Jiele are different learning theory. Bloom is a humanistic leraning theory which emphasises on content or what is being learnt, while Hiele is a cognitive learning theory which emphasises on learning process. From the explanation above, it acceptable that there will be different result in mathematics learning from the two theories.

Therefore, this research will answer the following questions 1) How to apply Bloom and hiele theory in mathematics teaching process 2) How to empirically test the learning theory using item response theory and by using the simulation way 3) which leraning theory is more sonsistent in classifyng the student's thinking ability. The answer found in thie research will inform the teachers on how to formulate learning and scoring in mathematics learning process and how to effectively apply Bloom and Hiele theory

The dicussion in the research covers preface, bibliography, methods, result and discussion and conclusion.

2. Related works / literature review

Classification on student's thinking level in cognitive domain according to Bloom is devided into six levels, knowledge, comprehension, application, analysis, syntesis and evaluation. Knowledge is the basic aspect which ia also called recall. In this level, students are required to recognize concepts, facts, terms etc without having to understand nor use it. There is ususally a pressure in comprehension aspect. Students are required to comprehend what is being taught, to know what is being communicated, and use the content without having to connect with any other things. Comprehension ability can be devided into three groups. Translating, interpreting and extrapolating. Application, this levels requires the acceptance of new ideas, procedures, methods, principals, and theories in the new concrete situation. This situation needs to use new principals as well ideas unles the scoring will be based on the recalling instead of application. Analysis elaborates a certain situation into substances or the building components. The situation gets clearer in this part. Analysis ability is classified into three sections, substance analysis, relation analysis, organized principal analysis. Synthesis, in this steps students are required to make something new by combining the existing aspects. The result of this step can be in the form of written materials, planning or mecanism. Evaluation, in this step students are asked to evaluate the condition, situation, statement and concept based on certain criteria. In evaluation step, the most important matter is how to make student able to develop criteria, standard, or measurement to evaluate something. Evaluation competence is the highes competence according to Bloom.

According to Hiele, there are three main aspect in teaching geometry, time, learning materials, and the applied teaching method. If organized well, this aspect will increase student ability to the next level. According to Hiele, there are five stages in learning geometry. Introduction, analysis, sorting, deduction and accuration. Introductio starts when student see the shape of geometry but they are not familiar with the characteristic of the geometry shape. The result of introduction stage is the classes or group of similar looking shape. Sample of introduction activities are geometry identification, geometry description, and manipulating physical model. In analysis stage, students are familiar with the caharcteristic of the geometry. The students are able to mention the pattern in geometric shape they observe. The result of analysis stage is the geometrics characteristics. Sample of analysis activities are classifying thre geometric shape and defining the character of flat geometry. Sorting stage, student comprehend and recognize the characteristics of geometric shape and are able to sort one geometrics shape to another. The results are the characters of geomatrics objects. Sample of sorting stage is defining the enough condition and require condition of geomtric shape and testing a hypothesis. Deduction stage, students are able to conclude deductively, from the general matter to the more specific ones. The are also able to comprehend the role of the undefined substance and the defined ones as well. The result of this stage is the basic deductive systems of geometry. Sample activity of the stage is proving by using axioms and theorems. Accurationn stage, students are aware of the importance of basic principles which bases a proof. For example, they understand how important of axioms and postulats of euclid geometry. Accuratuion step is high, complicated and complex. The result in this stage is the comparation of and difference in early geometry systems [10]

ISBN:978-602-99849-2-7

The exixtence of Bloom learning theory has been empirically performed in Anshori's research (2012) about students thinking competence according to Bloom's taxonomy and resulted that students thinking level are on five Bloom's taxonomy level eventhough they are not in the same percentage. Meanwhile, based on Sudarsono's research (2012) using Hiele theory results in the achievement of 100% in introduction, 78,4% for analysis, 67,6% for sorting, 86,5% for deduction, 56,8% for accuration. From the research we can see that students are on the introduction level. Those theory show that students competence are on every level of the theory.

Learning is a process to create the study condition to develop student's competence maximally so that the competence and purpose of the learning are achieved. In the learning procee, there is an interaction which somehow involves the extrinsic matter of the student and the teacher, including environment. According to the law number 20 year 2003 about national education systemwhich elaborates that learning is an interaction among student, teacher and learning source in a learning environment [6]. Meanwhile, mathematics does not deal with numbers and operation only but also quantity. Pointing mathematics in quantity (number and quality) is not enough, has not come to mathematics dealing with relation, pattern, shape and structure. According to [8] mathematics play a very important role in the development of logics so that mathematics and logics are inseparable. Numbers in operation and geometry concept and the relation of both of them can be stated as tghe sample of logics using in mathematics. Mathematics is a logical study of shape, sort, magnitude, and the related concept. Mathematics is usually classified in three field, aljabar, analysis and geometry but the classification itself is not relly clear since the branches have mixed up. Mathematics can be defined as the lesson that learns about number, shapes and concepts that related to logics using common symbols and applictionin other fields. By doing so, mathematics is an interaction process between teacher and student so that the change will take place which will take to the comprehension on the sitematically organized abstract ideas.

The classic item difficulty rate depends on the respondent ability. For the high baility respondent, the test item is not too difficult. For the low level respondent, the test item is too difficult. The difficulty level of the item in modern measurement is directly connected with item characteristic. Modern measurement can be used to check the student's ability and the item characteristic. Item characteristic is defined by the response of the respondent as it is called item response theory. Basically, item characteristic is regression between competence and probability in answering correctly.Formulalogistic model1P(one parameter)in IRTused in this studyis

$$P_i(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}}$$

The difficulty level of the item is a parameter in item response theory. Difficulty level is the ability to answer correctly in the certain level of ability stage [3] which is usually in the form of index. Difficulty index is commonly stated in the form of proportion which rates from 0,00 – 1,00. The bigger the index, the easier the item is. According to the most followed agreement, difficulty index is often classified as follow: 1) $0,00 < P < 0,30$, difficult item 2) $0,31 < P < 0,70$, medium item and 3) $0,71 < P < 1,00$ easy item. The function of the test difficulty usually depends on the purpose of the test. Medium test is used in semester test, high difficulty test for selection and easy test for diagnostic test. Item difficulty level has the function for the teacher and the function of teaching process. For the teacher (a) as the concept introduction and feedback for the students for their study (b) get the information on curriculum pressure or suspicious about bias question. And for learning are (a) concept introduction and re-teaching (b) sign of strength and weakness of curriculum (c) give feedback to the student (d) the chance of bias item. (e) assembling test with data item match.

ISBN:978-602-99849-2-7

3. Material and method

This research uses quantitaive approach with survey descriptive research. Survay descriptive research is a research which describe an object or subject to give an informationof the observed matter. Thje population of this research is all student of VIII even semester SMP Negeri 1 Dukuhturi study year of 2013/2014 with totl 315 students. The premier data of the research is the result of mathematics test about circle with catagory of Bloom and Van Hieleand interview data with the teacher about the application of Bloom and Van Hiele. The interview was to see how far the teacher understand about Bloom and Van Hiele. Data collecting method is done by using difficulty index and rised using software WinGen3. Simulation is done on variable 2 (0,1) multiple choice with the catagory of Bloom an Van Hiele.this simulation is conducted sample as many as 1000 test participants and the participants' competence is considered similar and there are 20 items in each test. Data analysis data is done to interview datato get information baout teacher's understanding to Bloom and Van Hiele. Item difficulty analysis is done empirically using Item Response Theory (IRT) with logistic approach one paramater with the help of RASCAL program.

4. Results and Discussion

Teachers' understanding about Bloom's learning theory is illustrated through the activities in the classroom. Mathematics learning activities guided by the teacher in the class have applied Bloom's learning theory gradually based on the learning continuum and tried to be adjusted with cognitive, affective and psychomotor domain. Moreover, related to developing students' thinking level about geometry,the teacher has applied Van Hiele learning theory. The development of mathematical thinking on students is done gradually and adjusted with the chronology of their thinking ability. Bloom and Van Hiele learning theory have its advantages and weakness. The advantage of Bloom's learning theory is that it facilitates students' understanding in learning Mathematics. It also facilitates teachers in arranging and designing learning activities to improve students' thinking level. Meanwhile, the weakness of Bloom's learning theory is that it assumes that students have the same ability, so for the students who have weak memory capacity and do not actively participate will be left behind. The advantage of Van Hiele's learning theory is that the material mastery is faster because with the use of model, the teacher can understand in more detail how far the achievement of students' thinking level in learning geometry and the students' knowledge quality is determined by the thinking process that they use. The weakness of Van Hiele's learning theory is that this theory is only used on the material about geometry and if the direct learning is started on a certain level without considering students' thinking level then the students will undergo difficulties to go to the next thinking level.

The calculation of difficulty level index is done to every item. Principally, the average score that are obtained by the students on the item is called item difficulty. The function of the item difficulty level is in general connected with the purpose of the test. For example for the end semester test, items which have the medium difficulty level are used, for the need of selection, items which have high difficulty level are used, and for the need of diagnostic test, items which have low difficulty level are used. The level of items difficulty for the 20 items of Mathematics questions related to the material about circle is done based on the classification of Bloom and Van Hiele level of thinking on the students of grade VIII of SMP Negeri 1 Dukuhturi Tegal in the academic year of 2013/2014. The real data of items are mostly classified into good difficulty level both for items with the classification of Bloom and Van Hiele. Meanwhile, the data of items for simulation are mostly in the medium difficulty level both for items with the classification of Bloom and Van Hiele.

Empirically, it is discovered that most of the students for Bloom's classification are 86% in the level of knowledge and 26% in the level of analysis. For Van Hiele's classification, 84% of students are in the level of introduction and 29% of them are in the level of deduction. With the simulation wat for the level of thinking based on Bloom's learning theory, 84% of students are in the level of knowledge and 31% of them are in the level of analysis. Based on Van Hiele's classification, 84% of students are in the level of introduction and 32% of them are in the level of deduction.

Based on the analysis result with the help of *WinGen3* software, if it is examined from the level of

ISBN:978-602-99849-2-7

difficulty and students' thinking level based on Bloom and Van Hiele, there is a consistency difference in the classification based on the two learning theories. The item number 17 and 19 in the level of application based on the category they have lower difficulty level than in the level of understanding and there is also number 16, an item in the level of application, which has higher difficulty level than in the level of analysis, and also number 14, an item in the level of analysis which has lower difficulty level than in the level of application. The item number 20 in the Van Hiele's classification have lower difficulty level than in the level of introduction and item number 11, an item in the level of deduction, has lower difficulty level than in the level of order.

In general from the result of calculation about the average of difficulty level in each students' thinking level, it is found out that those theories of learning have similar consistency. If it is viewed from the thinking level of Bloom and Van Hiele on each item, Bloom's learning theory is more consistent than Van Hiele's learning theory. Therefore, it can be concluded that Bloom's learning theory is more consistent than Van Hiele's learning theory in classifying the students' thinking level.

After the difficulty level is ordered from the lower to higher category of difficulty level, the significant value which indicates the increase of difficulty level is obtained based on Bloom's theory, Van Hiele's theory, *Item Response Theory* (IRT) and simulation. So, the classification of learning theories based on Bloom and Van Hiele can be proven empirically.

The data analysis with simulation by using difficulty level of real data has a weakness, that is the result of the simulation is in the form of difficulty level of classical test theory so that the data of difficulty level that are obtained are based on the respondents' ability. In fact, there is the result of simulation in the form of students' worksheets that can be analyzed using IRT approach. However, because of the limitation of the researcher's ability in using the other programs for calculating the difficulty level of the result of worksheets from the simulation then the researcher only uses the data that exist in the form of classical difficulty level.

5. Conclusions

The conclusions of this research are: 1) the teacher has mastered the application of Bloom's learning theory in Mathematics learning; 2) the teacher has mastered the application of Van Hiele's learning theory in Mathematics learning; 3) the classification of Bloom's and Van Hiele's thinking level can be proven empirically (real data and simulation) in Mathematics learning; 4) Bloom's learning theory is more consistent than Van Hiele's learning theory in classifying the students' thinking level in learning Mathematics on the material about circle empirically (real data and simulation).

References

[1] Bloom, Benjamin S. , *Evaluation in Education*, Allyn and Bacon, 1981
[2] Daryanto, *Evaluasi Pendidikan*, Rineka Cipta, 2001.
[3] Hambleton, Swaminathan, dan Rogers, *Fundamentals of Item Response Theory*, Sage Publication, 1991.
[4] Karso, dkk., *Pendidikan Matematika 1*, Universitas Terbuka., 2009.
[5] Masyhuri dan Zaenudin. *Metodologi Penelitian Pendekatan Praktis dan Aplikati*, Refika Aditama. 2008
[6] Prastowo, Andi. *Pengembangan Bahan Ajar Tematik*. Diva Press, 2013.
[7] Purwoko, *Teori Belajar Van Hiele* dalam Aisyiah dkk., "Pengembangan Pembelajaran Matematika SD", Dirjen Dikti, 2007.
[8] Russel, Bertrand, *The Principles of Mathematics*, Cambridge University Press, 1903
[9] Sudijono, Anas, *Pengantar Evaluasi Pendidikan*, Raja GrafindoPersada, 2012
[10] Van De Walle, John A, *Matematika Sekolah Dasar dan Menengah*, Erlangga, 2008.

ISBN:978-602-99849-2-7

Service Quality in Religious and Common Tourism

Fety Ilma Rahmillah[1] and Andi Rahadiyan Wijaya[2]

[1]Department of Industrial Engineering, Islamic University of Indonesia,
[2]Department of Industrial Engineering, University of Gadjah Mada

[1]fety_rahmillah@yahoo.com

Abstract: Syariah tourism is growing worldwide in line with the increasing of Muslim amount in the world. Religious tourism as one of syariah tourism in Indonesia has a potency to be the main tourism destination since 89% of Indonesian people are Muslim. This study aims to explore service quality in religious and common tourism objects. 172 data from eight sacred sites and 31 data from one common site were collected by using retrospective method. SERVQUAL is used to identify the differences between experiences of religious tourists and common tourists. Findings showed that there were differences between the results of religious and common tourism for the most important service item, the most unimportant service item, and the minimum service gap. Moreover, common tourists have higher importance score in almost all services that can indicate the higher expectation than religious tourists. This gives a new insight on understanding needs of visitors so that the strategic business for religious tourism should be differentiated from common tourism.
Keywords: Religious tourism; Common tourism; SERVQUAL.

1. Introduction

Syariah tourism is growing worldwide in line with the increasing of Muslim amount in the world. The first conference of syariah tourism is held on June 2-3, 2014 in Jakarta and yield 13 recommendations. One of them is the necessity to focus on research to develop syariah tourism [1]. Indonesian domestic tourism industries reached 245 million visitors in 2013 [2] and the number of transactions reached in about 171.70 trillion rupiahs [3][4]. Religious tourism as one of syariah tourism in Indonesia has a potency to be the main tourism destination since 89% of Indonesian people are Muslim [5]. Moreover, so many figures had a significant role in distributing Islam to all over Indonesia, so that keeping the proofs and historical things is an important thing to do. In addition, for areas that do not have a common tourist attraction can be seeded, but has religious tourism object, it could be an alternative of new source income to be optimized.

Religious tourism is tourism which is motivated by specific purposes related with people's belief [4]. It can be a way of enriching knowledge and transferring religious values into global humanism [5]. Moreover, religious tourism is a constructive tool to strengthen local economy and generating employment opportunities by promoting variegated spiritual [6]. Generally, religious tourism objects are also heritage tourism objects [7] for example, mosque, cathedral, temple, shrine, and so on. Hajj is excluded since it is an obligation for Muslims who are able. Religious tourism objects has been assessed here are cemetery of Islamic figures of Sunan Kudus, Sunan Gresik, Sunan Giri, Sunan Ampel, Sunan Kalijaga, Raden Patah, and so on.

This paper is part of research "The development of business strategy for religious tourism" and the objective of this paper is to explore the service quality of religious and common tourism sites. It will be useful for academicians and researchers on how to improve the domestic economic contribution. It also could be as a recommendation for Indonesian Tourism Government on improving service quality of sacred tourism.

2. Literature Review

Hengky [8] had been envisaged the potency of sustainable sacred tourism in Java. The 304 data of respondents were tabulated by content analysis and after passed six stages then it was resulted Kappa's values. The result stated that 57.89 % categorized as sustainable sacred tourism. The rest 42.11 % should be improved by implementing sustainable sacred tourism concept, where 36.84 percent showed lack of sustainable sacred tourisms and 5.26 % is categorized as unsustainable sacred

tourisms. The benefit of sustainable sacred tourism is not only to increase tourist visitation, but also environmental performance.

One way to increase domestic tourist visitation as well as income is by increasing tourist satisfaction. Based on field observation, there was uniformity parking rates, inadequate facilities, and irregular administration. It indicates that the management of existing religious tourism does not optimal yet [4] [8].The key to satisfy visitors is by identifying what kind of factors that can fulfill their needs. SERVQUAL [9] look at five service dimensions which are tangibles, reliability, responsiveness, assurance, and empathy. Tangibles include all physical facilities and personnel appearances such as exterior and interior facilities and social dimensions (e.g., employee characteristics) whereas reliability is an organizations ability to perform the promised services accurately. Responsiveness refers to willingness to help customers; assurance is about knowledge and courtesy of employees; while empathy is about care and attention of employees.

3. Methodology

Table 1 presents the number of participants involved in this study. For religious tourism, there were 172 people consisted of 75 males and 97 females, while 31 participants of common tourism consisted of 9 males and 22 females. Most of respondents of religious tourism (73.23%) were a first-time visitor and 81.85% participants were students. The frequency of visiting sacred site was categorized as follow: 15.38% more than once per year, 32.92% at least once per year, and 19.08% at least once per two years, and 32.62% stated that it is not necessarily. Then, the companion was categorized as group (not family) 87.69%, 6.46% with a spouse or child or family, and 5.85% with friends and nobody alone. While for common tourism, all is college students.

Table 1. Number of Participants

Number of respondents								
Religious tourism								**Common tourism**
Cemetery of Sunan Giri	Cemetery of Sunan Ampel	Cemetery of Sunan Gresik	Cemetery of Sunan Kudus	Cemetery of Sunan Kalijaga	Cemetery of Gus Dur	Cemetery of Mbah Kholil	Cemetery of Raden Patah	Taman Pelangi
31	24	39	12	14	30	18	4	31

This study used SERVQUAL method and the questionnaire consists of thirty four service items (Table 2) with three columns for each service (column importance, expectation, and perception). Each service item statement is measured by a 5-point Likert-type scale (for level of importance, 1 = not important at all and 5 = absolutely important; for level of expectation, 1 = not essential at all and 5 = absolutely essential; for level of perception, 1 = strongly disagree and 5 = strongly agree). Retrospective method is used to collect the data during January until June 2014. Retrospective is a method which uses information from past experiences.

There are two types of questionnaires, for religious visitors and common visitors. Both of questionnaires are similar and only have little adaptation. Table 2 provides list of statements for common questionnaire. Reliability test is used to evaluate the internal consistency of constructs. It is shown in Table 3. The instrument is reliable since all of constructs have cronbach's alpha value more than 0,6 [10], while standard value more than 0.25 is used for validity [11]. All questions of religious questionnaire are valid while some items in common questionnaire are not valid.

Table 2. Description of Variables

Service Quality	**Code**	**Description**
Tangible	SQA1	Parking area is clean and safe
	SQA2	Market is clean and comfortable
	SQA3	Transportation cost is affordable
	SQA4	Food and products sold are affordable
	SQA5	There is comfortable and adequate rest area
	SQA6	There is a clear direction
	SQA7	Toilet and ablution are clean and comfortable
	SQA8	Praying equipment are clean and adequate
	SQA9	Shoe rack is clean and safe

ISBN:978-602-99849-2-7

	SQA10	Trash cans are clean and adequate
	SQA11	There is special way for disable
	SQA12	There is a safe children playground
	SQA13	There is an interesting historical display and figure
	SQA14	The lighting is appropriate
	SQA15	There is pleasant atmosphere
	SQA16	The originality of historical building is well maintained
	SQA17	Indoor and outdoor are well maintained
	SQA18	There is separation between entrance and exit
	SQA19	There is a site map so that visitors will not lost
Reliability	SQB1	Information of visitors flow can be observed clearly
	SQB2	Employees provide clear guidance
	SQB3	Management always strive to give best services
	SQB4	Water is always available and running smoothly
	SQB5	Overall, you get what you want
Responsiveness	SQC1	Employees provide services quickly
	SQC2	Employees are always willing to help visitors
	SQC3	Employees are able to arrange visitors to avoid long queues
	SQC4	Information of services and facilities are provided
Assurance	SQD1	Employees have enough knowledge to answer visitors' questions
	SQD2	Employees have well understanding about condition and interesting thing about tourism object
	SQD3	Tourism object provides a safe environment
Empathy	SQE1	Employees are helpful, friendly, and respectful
	SQE2	There is convenient visiting hours
	SQE3	The visit to tourism object gives benefit

Table 3. Reliability and Validity Test

Constructs	**Item**	**Cronbach's alpha value**	
		Religious Tourism	**Common tourism**
Service Quality	34;33	0,907	0.961
Tangible_SQ	19;18	0,849	0.926
Reliability_SQ	5;5	0,704	0.885
Responsibility_SQ	4;4	0,738	0.866
Responsiveness_SQ	3;3	0,684	0.837
Empathy_SQ	3;3	0,614	0.819

4. Results and Discussion

There are some points can be figured out. First, 'The tourism object gives benefits' (SQE3) becomes the most important item and has the minimum service gap for religious tourism visitors. It is understandable since religious benefit is something most searched. Based on Focus Group Discussion with some visitors, the religious benefits are the tendency to acquire the tranquility near the site, searching for the God's blessing with the religious leader as mediator, believing that the site is the more prominent to pray, and remembering the death. SQE3 can also has the minimum service gap due to the initial formation of the religious tourism, already has a spiritual appeal, so it does not need to be improved anymore. In other side, for common tourism visitors, benefit is not really important since the important thing is enjoy, happy, and entertained. Figure 1 provides importance score for all services in religious and common tourism whereas figure 2 shows service gap of services.

ISBN:978-602-99849-2-7

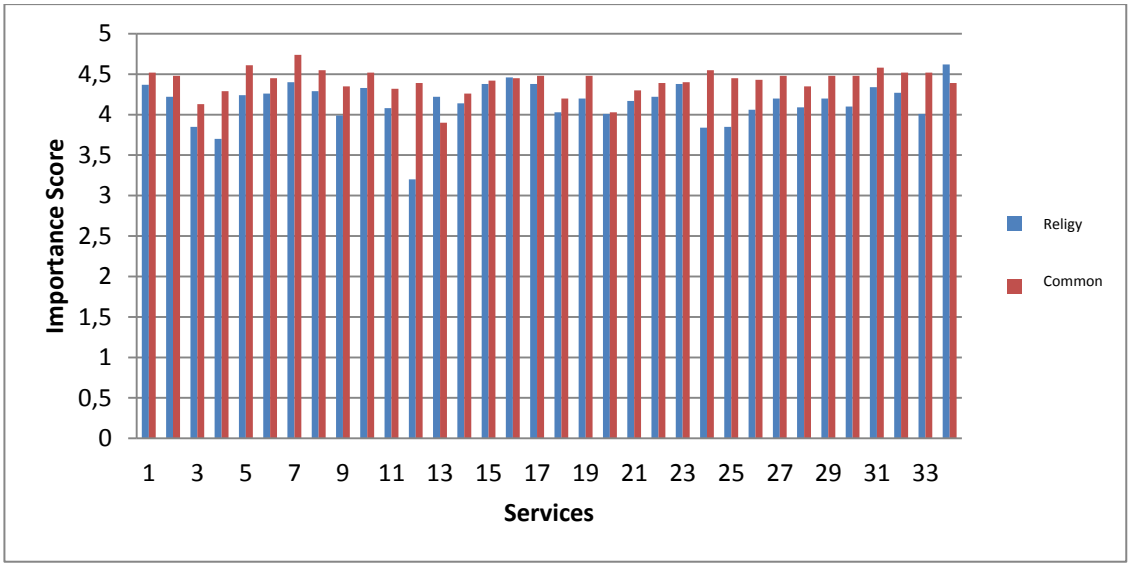


Figure 1. Importance Services in Religious and Common Tourism

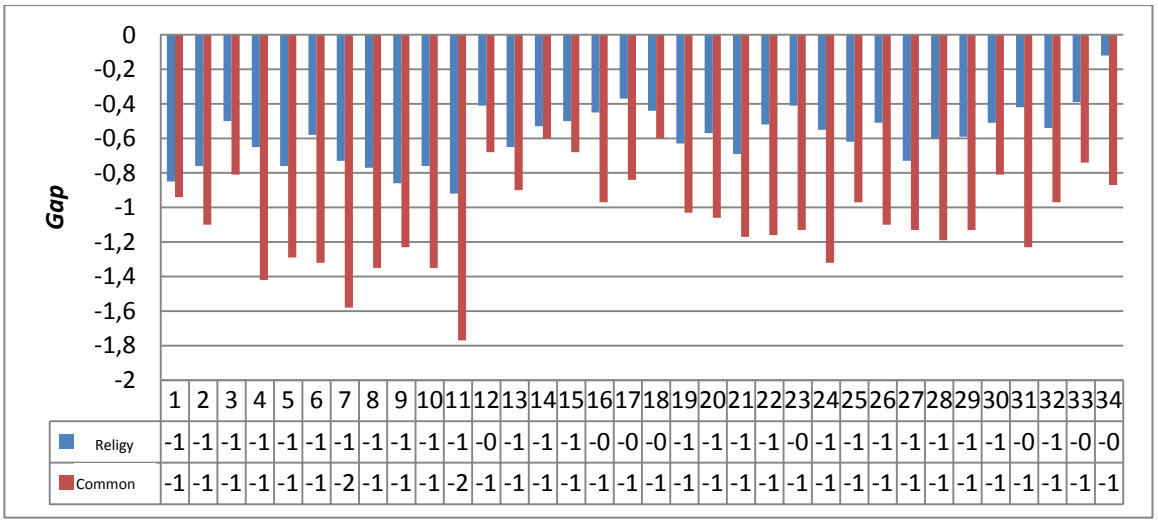


Figure 2. *Service Gap* in Religious and Common Tourism

Second, SQA12 'There is a safe play area for children' becomes the most unimportant service. It has a strong relationship with religious motivation of visitors. Even though most of visitors are parents who bring their children, but children playground area is not important since the goal is for spirituality. It is different with common tourism which really a place where children can play and recreation, the important score is 4.39. Third, the cleanliness of the toilets and the place in general is the main thing in common tourism. It is proven since SQA7 "toilet and ablution are clean and comfortable" becomes the most important thing (4.74) different with religious tourism (4.4). Religious tourists do not really care but they will very welcome if the cleanliness is fulfilled.

Fourth, 'the special way for visitors with physical limitations (disable)' SQA11 has maximum service gap for religious tourism. This is in line with the fact that there was not yet special way for disable in religious tourism object which can be a friendly handgrip along stairs to the location. Even, this service item also has the highest gap score in common tourism even though the question is not considered as valid due to score of corrected item-total correlation which less than 0.25 [11].Corrected item-total correlation is used to test validity as an early analysis to select feasible items used in the test as a whole. Thus, all items that have a correlation less than 0.25 can be set aside and the items to be included in the test are items that have a correlation more than 0.25. The closer the value to one, the better the consistency is.

ISBN:978-602-99849-2-7

Table 4. Highlight of Importance and Service Gap between Religious and Common Tourism

Code	Service item	Religious tourism		Common tourism	
		Importance	Gap	Importance	Gap
SQA7	Toilet and ablution are clean and comfortable	4.4	-0.73	**4.74**	-1.58
SQA11	There is special way for disable	4.08	**-0.92**	4.32	**-1.77**
SQA12	There is a safe play area for children	**3.2**	-0.41	4.39	-0.68
SQA13	There is interesting historical display and figure	4.22	-0.65	**3.9**	-0.9
SQA14	The lighting is appropriate	4.14	-0.53	4.26	**-0.6**
SQA18	There is separation between entrance and exit	4.03	-0.44	4.2	**-0.6**
SQE3	Tourism object gives benefits	**4.62**	**-0.12**	4.39	-0.87

*Bold type indicates maximum or minimum score

Fifth, common tourists have higher importance score in all services except point SQA13 (There is interesting historical display and figure) and SQE3 (The visit to tourism object gives benefit). Even, SQA13 has the lowest important score. This is something unusual since theory said that an attractive display will make people interesting and force to come, but why it is not exist in this study. Common tourism object in this study is located in one area with Monumen Jogja Kembali in which full of historical things and open during morning till afternoon. It is very different with Taman Pelangi which opens only in the night and has modern concept with so many kinds of entertainment things. There is actually an interesting display, but maybe visitors tend to differentiate between Taman Pelangi as an entertain tourism object and Monumen Jogja Kembali as a historical tourism, even though both are located in the same area.

Last but not least, the importance of tangible service items such as parking area, market, directions, rest area, praying equipment, map, entrance and exit are similar between religious and common tourisms. All services have value above four out of five in range scale which means the availability of those services is important.

Finally, it can be said that common visitors have higher expectation on services rather than religious visitors (Figure 1) since the importance score is higher in almost all aspects. It indicates that more difficult to satisfy tourists in general than religious tourists. There is possible reason such as general tourists usually have to buy ticket to visit tourism object, so that the demand for services is higher than religious tourists whereas religious tourists tend to be receptive with the condition of tourism object since *infaq* is only for people who wanted. Other, the spiritual motivation of religious tourists also has big influence toward receptive behavior.

5. Conclusion

Based on the discussion above, some services have different importance for religious and common tourism, but some others also have similarity. There were differences between the results of religious and common tourism for the most important service item, the most unimportant service item, and the minimum service gap. Religious tourists do not expect service quality as high as common tourists so that it will be easier to develop strategic business for religious tourism. This study only uses SERVQUAL method which has linear assumption, it will be better to complete the shortcoming by using Kano model.

Acknowledgement. This research is part of thesis conducted in Gadjah Mada University and it is fully supported by Scholarship from DIKTI.

References

[1] Respati, Y., http://mysharing.co Retrieved 10 June, 2014.
[2] Tirani, E., http://www.metrotvnews.com Retrieved 7 December, 2013.
[3] Pusdatin Kemenparekraf & BPS, http://www.budpar.go.id Retrieved 7 December, 2013.
[4] UNWTO, 2011, Religious Tourism in Asia and the Pacific, World Tourism Organization, Madrid, Spain.
[5] Singh, R.P.B., Pilgrimage-Tourism: Perspective and Vision. Chapter 9 from the book of *Hindu Tradition of Pilgrimage: Sacred Space and System.* Dev Publishers, New Delhi. ISBN (13): 978-93-81406-25-0; pp. 305-332 (2013).

ISBN:978-602-99849-2-7

[6] Eugene, J., Holidays in a Holy Land: Spiritual Tourism in Placid Puducherry, *International Journal of Humanities and Social Science Invention*, Vol. 2(5), pp. 17-22 (2013).
[7] Irimias, A., Michalko, G., Religious Tourism in Hungary-an Integrative Framework, *Hungarian Geographical Bulletin*, Vol. 62(2), pp. 175-196 (2013).
[8] Hengky, S.H., Envisaged the Potential of Sustainable Sacred Tourism in Java Indonesia, *International Journal of Business and Social Science*, Vol. 4(12) (2013).
[9] Parasuraman, A., Berry, L.L., Zeithaml, V.A., SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality, *Journal of Retailing*, Vol. 64(1), pp. 12-40 (1988).
[10] Churchill, G., A Paradigm for Developing Better Measures for Marketing Constructs, *Journal of Marketing Research*, Vol. 16(1), pp. 64-73 (1979).
[11] Azwar, S., *Reliabilitas dan Validitas*, Pustaka Pelajar, Yogyakarta, 2009.

ISBN:978-602-99849-2-7

Maximum Likelihood Estimation In Intervention Analysis Model Multy Input Step Function: the impact of sea highway policies on stock price movements in field of shipping company (TMAS.JK)

Wigid Hariadi[1], Abdurakhman[2]

[1]Post Graduate mathematics of the Mathematics and Science Faculty, Gadjah Mada University
[2]Mathematics Lecturer at Gadjah Mada University

wigid_hariadi@yahoo.co.id, rachmanstat@ugm.ac.id

Abstract: ARIMA model is a popular model and is often used in time series analysis. This model is able to represent well on data that has a trend, a trend either up or down trend. However, the ARIMA model is no longer suitable for time series data if there is a change in the data that extreme (shock data), so that resulted in a change in the pattern of mean out of the ordinary. This data shock often occurs because of the intervention (the influence of external factors). Intervention analysis was used to evaluate the effects of external events on a time series data. In this paper, the authors wanted to model the impact of sea highway policies on stock price movements company in the field of shipping, in this paper were sampled stock of TMAS.JK. After analyzing the data, it is evident that the case of intervention on daily stock price TMAS.JK. Where intervention namely: first interventions, on August 11, 2014, allegedly as a result of the election of Jokowi-JK pair as President and vice President Republic of Indonesia on July 22, 2014. second interventions, Jokowi speech at the APEC forum on marine highway program, and offers investment in the construction of ports to foreign nations on November 10, 2014. In this analysis, the authors use the method of maximum likelihood estimation ,where the models obtained are as follows:

$$X_t = \frac{0{,}2185 + 0{,}1582\,B - 0{,}1734\,B^2}{1 - 0{,}6915\,B} S_{1,t} + \frac{0{,}0296}{1 - 0{,}8731\,B} S_{2,t-3} + \frac{\varepsilon_t}{(1-B)\,(1 + 0{,}2734\,B + 0{,}2327\,B^2)}.$$

Keywords: intervention analysis; step function; sea highway polcies; maximum likelihood estimation;

1. Introduction

Model autoregressive integrated moving average (ARIMA) is a popular method for use in forecasting univariate time series data. ARIMA model can represent well on data that has a trend, either up or down trend trend. However, what if there is a change in the data that extreme (shock data), so that resulted in a change in the pattern of mean out of the ordinary, whether the ARIMA model is still suitable for use ?. In this case ARIMA method is no longer suitable for use. Due to changes in the pattern of extreme (shock data) can lead to errors in the identification of models which resulted in obtaining the wrong model (less precise) for a time series data. This data shocks occur due to an intervention. Where, according to Box, et al (1994) time series is often affected by specific events or circumstances such as policy changes, labor strikes, ad campaigns, environmental regulations, and similar events, where events like this are often called the event of intervention.

To overcome the above problems, one method that can be used is the analysis model of intervention. According Makridakis, et al, (1995) This analysis has been widely disseminated through the article titled "intervention analysis with applications to economic and environmental problems" writings Box and Tiao (1975). They propose an approach to recognize the intervention of the independent variable on the dependent variable.

According to Wei (2006), Analysis of interventions used to evaluate the effects of external events on a time series data. Analysis of these interventions have been successfully used to examine the impact of air pollution control and economic policy (Box and Tiao, 1975), the impact of an oil embargo arab (Montgomery and Weatherby, 1980), the impact of the New York balckout (Izenman and Zabell, 1981), and many again other events.

The paper was written with the purpose to provide the results of theoretical studies and application of a step function intervention model using maximum likelihood estimation. Applied study carried out on a time series data is data daily closing price TMAS (Pelayaran Tempuran Emas

Tbk), which of companies engaged in shipping. Data observed starting on January 1, 2014 until December 19, 2014. The election of Jokowi-JK pair as President and Vice President Republic of Indonesia, as well as some policies are made, allegedly has upgraded the share price interventions in the field of shipping. To the authors wanted to apply the step function intervention model using maximum likelihood estimation for modeling the daily stock price movement TMAS.

2. Related Works/Literature Review

Application of the intervention analysis model is quite extensive. Where there is a time series data, it is possible intervention models can be applied. Here are some of the research analysis model of intervention in several areas of science. According to Box and Tiao (1975), a dynamic model for this intervention can be classified into two kinds of functions, namely the function step and pulse function. For a step function can be shown by the indicator $S_t^{(T)}$, while for the indicator pulse function can be demonstrated by $P_t^{(T)}$.

According to Poirier, et al (2007), the analysis model can be applied in the field of health interventions. In France, Salmonellosis is one of the major causes of bacterial infection with serotypes default on food enteritis (SE) and Typhimurium (ST) accounted for some 70% of all cases. French government implement a control program SE and ST on poultry meat and eggs since October 1998. There was a decrease Salmonellosis by 33% since the policy is applied. Want researched and evaluated the impact of the program since the control is applied. It was alleged that there were two interventions namely SE and ST. By using a model of intervention for the ARMA model, the result that there is a relationship between Salmonella control program and a decrease in the target observation for two serotypes.

According to Cole, et al (2013), the analysis model can be applied in the field of government intervention, namely the impact of government policy on the collection of household waste. In the regional administration, frequent events interventions and policy changes regarding service of household waste. These changes include the policy of separation between organic and inorganic waste. In his writings, in this case, there have been interventions as much as 2 times. And to model the noise, using the model AR (1), AR (2), SAR (1), and SAR (2). Where the intervention analysis model that formed was able to predict the impact of the amount of recyclables at the time of seasonal and regular day of work.

3. Material & Methodology

Data

Data used in this case study is a secondary data. the data in the form of stock data that has been provided by Independent Sekutitas. The data will be analyzed in the form of daily stock data TMAS (Pelayaran Tempuran Emas Tbk) from January 1, 2014 until December 19, 2014. Where the data is the data's closing stock price (close).

Method

methodology in this study are:

- The data is divided into two parts, before the intervention and after intervention.
- Stock data before the intervention were analyzed using ARIMA method with estimation method is maximum likelihood estimation.
- Then calculated residual, to make residuals chart. and define the order of *b*, *s* and *r* for the first intervention analysis model.
- Stock data after the first intervention in the analysis using analytical models of intervention, using the order of *b, s, r* above. with estimation method is maximum likelihood estimation.
- Then calculated residual, to make residuals chart. and define the order of *b, s* and *r* for the second intervention analysis model.
- Stock data after the second intervention in the analysis using analytical models of intervention, using the order of *b, s, r* above. with estimation method is maximum likelihood estimation.
- Then do a diagnostic test.

ISBN:978-602-99849-2-7

Intervention Analysis Model

According to Box, et al (1994), the analysis of the intervention, it is assumed that the event had occurred at the time of intervention to-T from a set of time series. It becomes important to determine for sure if there is a change or the effect of expectation on the time series data Yt where the incident occurred intervention. Based on research conducted by Box and Tiao, models of intervention analysis is as follows:

$$Y_t = \frac{\omega(B)B^b}{\delta(B)}\xi_t + N_t$$

Where $\mathcal{Y}_t = \delta^{-1}(B)\,\omega(B)B^b\xi_t$ represents the effect of the intervention on the input deterministic events ξ_t series, and $0 < \delta < 1$. While N_t is a represents noise series, that represents background Yt observation without intervention effect. In this case it is assumed that N_t follow ARIMA (p, d, q). For exogenous variables ξ more than one, can be represented by a dynamic model, where the model (Box, 1975):

$$f(\delta, \omega, \xi, t) = \sum_{j=1}^{k} \mathcal{Y}_{tj} = \sum_{j=1}^{k} \left\{\frac{\omega_j(B)}{\delta_j(B)}\right\}\xi_{tj}$$

According to Wei (2006), for the analysis of multi input intervention, the general model is as follows:

$$Z_t = \theta_0 + \sum_{j=1}^{k} \frac{\omega_j(B)B^{b_j}}{\delta_j(B)} I_{jt} + \frac{\theta(B)}{\psi(B)}\varepsilon_t$$

where:

$\omega_j(B) = \omega_{sj}(B) = \omega_{0j} - \omega_{1j} B - \omega_{2j} B^2 - ... - \omega_{Sj} B^S$

$\delta_j(B) = \delta_{rj}(B) = 1 - \delta_{1j} B - \delta_{2j} B^2 - ... - \delta_{rj} B^r$

I_{jt} , j = 1, 2,..., k is a intervention is variable, which may be a step function or pulse function

$\psi(B) = (1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p)$

$\theta(B) = (1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q)$

θ_0 is a constant, which is usually not always written into the model. $\frac{\theta(B)}{\psi(B)}\varepsilon_t$ is a model of noise. *B* represents the backwshiff operator, namely $B^k\, Y_t = Y_{t-k}$. $\omega_S(B)$ is an operator of order *s*, where *s* order stating the length of time it takes for the effect of the intervention becomes stable. $\delta_r(B)$ is an operator of order *r*, where *r* order stating the time required to effect the intervention showed a clear pattern. Order *b* states the effect of an intervention, where the order of *b* is a time delay (delay) from influential intervention of *X* to *Y*.

According to Box, et al (1994), there are two common types of input deterministic ξ_t variables that can be used to represent the impact of intervention on the events of the time series data. The first is a step function, the indicator can be shown by $S_t^{(T)}$, with: T: time of the start of the intervention, in which:

$$S_t^{(T)} = \begin{cases} 0, & t < T \\ 1, & t \geq T \end{cases}$$

Which is usually used to represent the effects of interventions that have alleged (hope) that the effect will occur a long time after time to-T. Other types of the pulse function, it can be shown by the indicator $P_t^{(T)}$, which:

$$P_t^{(T)} = \begin{cases} 0, & t \neq T \\ 1, & t = T \end{cases}$$

ISBN:978-602-99849-2-7

Which is usually used to represent the effects of interventions that have alleged y (hope) that the effect will occur while and will be completed after a time to-T.

Maximum Likelihood Estimation Method (MLE)

Definition (Bain, 1992) : *Joint density function of n random variables X_1,X_2, ... , X_n is estimated by x_1, x_2, ... , x_n is denoted by $f(x_1, x_2, \dots, x_n; \theta)$ with θ is an unknown parameter , then the likelihood function of θ is:*

$$L(\theta) = f(x_1;\theta)f(x_2;\theta) \dots f(x_n;\theta)$$

Definition (Bain, 1992) : *Let $L(\theta) = f(x_1;\theta)f(x_2;\theta) \dots f(x_n;\theta)$ is the probability density function together of random variables X_1,X_2, ... , X_n for the set of observations x_1, x_2, ... , x_n. The value of θ that maximizes $L(\theta)$ is called the maximum likehood estimator (MLE) of θ. $\hat{\theta}$ is the value of θ that satisfy:*

$f(x_1, x_2, \dots, x_n;\hat{\theta}) = \max_{\theta \in \Omega} f(x_1, x_2, \dots, x_n; \theta)$

According Subanar (2013), maximum likelihood method is the most popular method of producing estimator. Suppose X_1, ... , X_n is i.i.d. a sample of the population density $f(x \mid \theta_1 \dots \theta_k)$. Function of probability (likelihood) is defined as:

$$L(\underline{\theta}|\underline{x}) = L(\theta_1 \dots \theta_k|x_1 \dots x_n) = \prod_{i=1}^{n} f(x_i|\theta_1 \dots \theta_k)$$

Definition: *For each sample point $\underline{x}$, let $\hat{\theta}(\underline{x})$ is the price parameter which $L(\underline{\theta} \mid \underline{x})$ as a function $\underline{\theta}$, assuming constant $\underline{x}$ reaches it is maximum. Maximum likelihood estimator (MLE) of the parameter θ based on a sample X is $\hat{\theta}(\underline{X})$.*

When the likelihood function differentiable (in θ_i), then the candidate MLE which is probably the prices ($\theta_1 \dots \theta_k$) such that:

$$\frac{\partial L(\underline{\theta} \mid \underline{x})}{\partial \theta_i} = 0 \ , i = 1,2, \dots, k$$

Estimation Parameter

Intervention Model AR(2) Step function (b=0,s=2,r=1)

the intervention single input model can be write:

$Y_t = \frac{\omega_2(B)B^0}{\delta_1(B)} S_t + \frac{\varepsilon_t}{\phi(B)}$, with $\varepsilon_t \sim N(0,\sigma^2)$

$Y_t = \frac{\omega_0 - \omega_1 B - \omega_2 B^2}{1 - \delta_1 B} S_{1,t} + \frac{\varepsilon_t}{(1 - \phi_1 B - \phi_2 B^2)}$

$\delta_1(B).\phi(B).Y_t = \phi(B).\omega_2(B).S_t + \delta_1(B)\varepsilon_t$

$\varepsilon_t = \phi(B).Y_t - \phi(B).\omega_2(B).(\delta_1(B))^{-1}.S_t$

$\varepsilon_t = (1 - \phi_1 B - \phi_2 B^2)Y_t - (1 - \phi_1 B - \phi_2 B^2).(\omega_0 - \omega_1 B - \omega_2 B^2).(\delta_1(B))^{-1}.S_t$

And then we can write:

$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2)$

ISBN:978-602-99849-2-7

$f(\varepsilon_t) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2} (\frac{\varepsilon_t - 0}{\sigma})^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2} (\frac{\varepsilon_t}{\sigma})^2)$

$L = \prod_{t=p+1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2} (\frac{\varepsilon_t}{\sigma})^2) = \frac{1}{2\pi^{(n-p)/2}\sigma^{n-p}} \exp(-\frac{1}{2\sigma^2} \sum_{t=p+1}^{n} {\varepsilon_t}^2)$

$$\text{Log L} = -\left(\frac{n-p}{2}\right) \log(2\pi) - (n-p) \log \sigma - \frac{1}{2\sigma^2} \sum_{t=p+1}^{n} (\phi(B).\mathrm{Y_t} - \phi(B).\omega_2(\mathrm{B}).(\delta_1(B))^{-1}.\mathrm{S_t})^2$$

- Calculate forestimator σ^2

$\frac{\partial \log L}{\partial \sigma} = 0$

$$\frac{-(n-p)}{\sigma} - (2(-2\sigma^{-3}) \sum_{t=p+1}^{n} (\phi(B).\mathrm{Y_t} - \phi(B).\omega_2(\mathrm{B}).(\delta_1(B))^{-1}.\mathrm{S_t})^2) = 0$$

$$\frac{-(n-p)}{\sigma} + \frac{1}{\sigma^3} \sum_{t=p+1}^{n} (\phi(B).\mathrm{Y_t} - \phi(B).\omega_2(\mathrm{B}).(\delta_1(B))^{-1}.\mathrm{S_t})^2) = 0$$

$$\sigma^2 = \frac{\sum_{t=p+1}^{n} (\phi(B).\mathrm{Y_t} - \phi(B).\omega_2(\mathrm{B}).(\delta_1(B))^{-1}.\mathrm{S_t})^2}{(n-p)}$$

- Calculate forestimator ω_0

$\frac{\partial \log L}{\partial \omega_0} = 0$

$$-\frac{1}{2\sigma^2} .2. \sum_{t=p+1}^{n} (\phi(B).\mathrm{Y_t} - \phi(B).(\omega_0 - \omega_1\mathrm{B} - \omega_2\mathrm{B}^2).(\delta_1(B))^{-1}.\mathrm{S_t}).\ \phi(B)(\delta_1(B))^{-1}.\mathrm{S_t} = 0$$

$$\frac{1}{\sigma^2} \sum_{t=p+1}^{n} (\phi(B).\mathrm{Y_t} - \phi(B).(\omega_0 - \omega_1\mathrm{B} - \omega_2\mathrm{B}^2)(\delta_1(B))^{-1}\mathrm{S_t})\ \phi(B)(\delta_1(B))^{-1}\mathrm{S_t} = 0$$

$$\frac{1}{\sigma^2.\ \delta_1(B)} \phi(B).\ \mathrm{S_t} \sum_{t=p+1}^{n} (\phi(B).\mathrm{Y_t} - \phi(B).(\omega_0 - \omega_1\mathrm{B} - \omega_2\mathrm{B}^2).(\delta_1(B))^{-1}.\mathrm{S_t}) = 0$$

$$\omega_0 = (\sum_{t=p+1}^{n} (\phi(B).\mathrm{Y_t}\ \phi(B).\ \mathrm{S_t} - \omega_1\mathrm{B} \sum_{t=p+1}^{n} \phi(B)^2.(\delta_1(B))^{-1}\ (\mathrm{S_t})^2 - \omega_2\mathrm{B}^2 \sum_{t=p+1}^{n} \phi(B)^2.(\delta_1(B))^{-1}\ (\mathrm{S_t})^2\ ) / \sum_{t=p+1}^{n} \phi(B)^2.(\delta_1(B))^{-1}\ (\mathrm{S_t})^2$$

- Calculate forestimator ω_1

$\frac{\partial \log L}{\partial \omega_1} = 0$

$$-\frac{1}{2\sigma^2} .2. \sum_{t=p+1}^{n} (\phi(B).\mathrm{Y_t} - \phi(B).(\omega_0 - \omega_1\mathrm{B} - \omega_2\mathrm{B}^2).(\delta_1(B))^{-1}.\mathrm{S_t}).\ \text{-B}\ \phi(B)(\delta_1(B))^{-1}.\mathrm{S_t} = 0$$

$$\frac{1}{\sigma^2} \sum_{t=p+1}^{n} (\phi(B).\mathrm{Y_t} - \phi(B)(\omega_0 - \omega_1\mathrm{B} - \omega_2\mathrm{B}^2)(\delta_1(B))^{-1}\mathrm{S_t})\ \phi(B)(\delta_1(B))^{-1}\mathrm{S_{t-1}} = 0$$

$$\frac{1}{\sigma^2.\ \delta_1(B)} \phi(B).\ \mathrm{S_{t-1}} \sum_{t=p+1}^{n} (\phi(B).\mathrm{Y_t} - \phi(B).(\omega_0 - \omega_1\mathrm{B} - \omega_2\mathrm{B}^2).(\delta_1(B))^{-1}.\mathrm{S_t}) = 0$$

$$\omega_1 = (\sum_{t=p+1}^{n} (\phi(B).\mathrm{Y_t}\ \phi(B).\ \mathrm{S_{t-1}} - \omega_0 \sum_{t=p+1}^{n} \phi(B)^2.(\delta_1(B))^{-1}\mathrm{S_t}\ \mathrm{S_{t-1}} - \omega_2\mathrm{B}^2 \sum_{t=p+1}^{n} \phi(B)^2.(\delta_1(B))^{-1}\mathrm{S_t}\ \mathrm{S_{t-1}}\ ) / \sum_{t=p+1}^{n} \mathrm{B}\ \phi(B)^2.(\delta_1(B))^{-1}\mathrm{S_t}\ \mathrm{S_{t-1}}$$

- Calculate forestimator ω_2

$\frac{\partial \log L}{\partial \omega_2} = 0$

$$-\frac{1}{2\sigma^2} .2. \sum_{t=p+1}^{n} (\phi(B).\mathrm{Y_t} - \phi(B).(\omega_0 - \omega_1\mathrm{B} - \omega_2\mathrm{B}^2).(\delta_1(B))^{-1}.\mathrm{S_t}).\ \text{-}B^2\phi(B)(\delta_1(B))^{-1}.\mathrm{S_t} = 0$$

ISBN:978-602-99849-2-7

$\frac{1}{\sigma^2}\sum_{t=p+1}^{n}(\phi(B)\mathrm{Y_t} - \phi(B)(\omega_0 - \omega_1\mathrm{B} - \omega_2\mathrm{B}^2)(\delta_1(B))^{-1}.\mathrm{S_t})\ \phi(B)(\delta_1(B))^{-1}.\mathrm{S_{t-2}} = 0$

$\frac{1}{\sigma^2.\ \delta_1(B)}\phi(B).\ \mathrm{S_{t-2}}\sum_{t=p+1}^{n}(\phi(B).\mathrm{Y_t} - \phi(B).(\omega_0 - \omega_1\mathrm{B} - \omega_2\mathrm{B}^2).(\delta_1(B))^{-1}.\mathrm{S_t}) = 0$

$\omega_2 = (\sum_{t=p+1}^{n}(\phi(B).\mathrm{Y_t}\ \phi(B).\ \mathrm{S_{t-2}} - \omega_0\sum_{t=p+1}^{n}\ \phi(B)^2.(\delta_1(B))^{-1}\mathrm{S_t}\ \mathrm{S_{t-2}} -$

$\omega_1\sum_{t=p+1}^{n}\ \phi(B)^2.(\delta_1(B))^{-1}\mathrm{S_t}\ \mathrm{S_{t-2}}\ ) / \sum_{t=p+1}^{n}\mathrm{B}^2\ \phi(B)^2.(\delta_1(B))^{-1}\mathrm{S_t}\ \mathrm{S_{t-2}}$

- Calculate forestimatorδ_1

$\frac{\partial \log L}{\partial \delta_1} = 0$

$-\frac{1}{2\sigma^2}.2.\sum_{t=p+1}^{n}(\phi(B).\mathrm{Y_t} - \phi(B).(\omega_0 - \omega_1\mathrm{B} - \omega_2\mathrm{B}^2).(1-\delta_1 B)^{-1}.\mathrm{S_t}).\ -(1-\delta_1 B)^{-2}\phi(B)(\omega_0 - \omega_1\mathrm{B} - \omega_2\mathrm{B}^2)\ \mathrm{S_t} = 0$

$\frac{1}{\sigma^2(1-\delta_1 B)^2}\sum_{t=p+1}^{n}(\phi(B).\mathrm{Y_t} - \phi(B).(\omega_0 - \omega_1\mathrm{B} - \omega_2\mathrm{B}^2).(1-\delta_1 B)^{-1}.\mathrm{S_t}).\ \ \phi(B)(\omega_0 - \omega_1\mathrm{B} - \omega_2\mathrm{B}^2)\ \mathrm{S_t} = 0$

$(\phi(B)(\omega_0 - \omega_1\mathrm{B} - \omega_2\mathrm{B}^2)\ \mathrm{S_{t-1}}).\sum_{t=p+1}^{n}\big(\phi(B).\mathrm{Y_t} - \phi(B).(\omega_0 - \omega_1\mathrm{B} - \omega_2\mathrm{B}^2)\big)(1-\delta_1 B)^{-1}.\mathrm{S_t} = 0$

$\sum_{t=p+1}^{n}(\phi(B).\mathrm{Y_t}\ \phi(B)(\omega_0 - \omega_1\mathrm{B} - \omega_2\mathrm{B}^2)\ \mathrm{S_t} -$

$\sum_{t=p+1}^{n}\ (\phi(B)(\omega_0 - \omega_1\mathrm{B} - \omega_2\mathrm{B}^2)\ \mathrm{S_t}\ )^2(1-\delta_1 B)^{-1} = 0$

$(1-\delta_1 B)^{-1} = \frac{-\sum_{t=p+1}^{n}(\phi(B).\mathrm{Y_t}\phi(B)(\omega_0 - \omega_1\mathrm{B} - \omega_2\mathrm{B}^2)\ \mathrm{S_t}}{-\sum_{t=p+1}^{n}\ (\phi(B)(\omega_0 - \omega_1\mathrm{B} - \omega_2\mathrm{B}^2)\ \mathrm{S_t}\ )^2}$

$1-\delta_1 B = \frac{\sum_{t=p+1}^{n}\ (\phi(B)(\omega_0 - \omega_1\mathrm{B} - \omega_2\mathrm{B}^2)\ \mathrm{S_t}\ )^2}{\sum_{t=p+1}^{n}(\phi(B).\mathrm{Y_t}\phi(B)(\omega_0 - \omega_1\mathrm{B} - \omega_2\mathrm{B}^2)\ \mathrm{S_t}}$

Intervention Model Multy Input AR(p) Step function (b>0,s>0,r>0)

The intervention model multy input can be write:

$\mathrm{Y_t} = \frac{\omega_{s,1}(\mathrm{B})\mathrm{B}^b}{\delta_{r,1}(B)}\mathrm{S_{1,t}} + \ldots + \frac{\omega_{s,k}(\mathrm{B})\mathrm{B}^b}{\delta_{r,k}(B)}\mathrm{S_{k,t}} + \frac{\varepsilon_t}{\phi(B)}$, with $\varepsilon_t \sim \mathrm{WN}\ (0,\sigma^2)$

$\mathrm{Y_t} = \frac{\omega_{s,1}(\mathrm{B})\mathrm{B}^b}{\delta_{r,1}(B)}\mathrm{S_{1,t}} + \frac{\omega_{s,2}(\mathrm{B})\mathrm{B}^b}{\delta_{r,2}(B)}\mathrm{S_{2,t}} + \frac{\varepsilon_t}{\phi(B)}$

$\delta_{r,1}(B)\delta_{r,2}(B)\phi(B).\mathrm{Y_t} = \phi(B)\omega_{s,1}(\mathrm{B})\mathrm{B}^b\delta_{r,2}(B)\mathrm{S_{1,t}} +$

$\phi(B)\omega_{s,2}(\mathrm{B})\mathrm{B}^b\delta_{r,1}(B)\mathrm{S_{2,t}} + \delta_{r,1}(B)\delta_{r,2}(B)\varepsilon_t$

$\delta_{r,1}(B)\delta_{r,2}(B)\varepsilon_t = \delta_{r,1}(B)\delta_{r,2}(B)\phi(B).\mathrm{Y_t} - \phi(B)\omega_{s,1}(\mathrm{B})\mathrm{B}^b\delta_{r,2}(B)\mathrm{S_{1,t}} -$

$\phi(B)\omega_{s,2}(\mathrm{B})\mathrm{B}^b\delta_{r,1}(B)\mathrm{S_{2,t}}$

$\varepsilon_t = (\ \delta_{r,1}(B)\delta_{r,2}(B)\phi(B).\mathrm{Y_t} - \phi(B)\omega_{s,1}(\mathrm{B})\delta_{r,2}(B)\mathrm{S_{1,t-b}} -$

$\phi(B)\omega_{s,2}(\mathrm{B})\delta_{r,1}(B)\mathrm{S_{2,t-b}}\ ) / (\delta_{r,1}(B)\delta_{r,2}(B))$

with :

$\mathrm{c'(B)} = \delta_{r,1}(B)\delta_{r,2}(B)\phi(B)$

ISBN:978-602-99849-2-7

$d'(B) = \phi(B)\omega_{s,1}(B)\delta_{r,2}(B)$

$f'(B) = \phi(B)\omega_{s,2}(B)\delta_{r,1}(B)$

and then we can write:

$$\varepsilon_t = \frac{c'(B)\, Yt - d'(B)\, S_{1,t-b} - f'(B)\, S_{2,t-b}}{\delta_{r,1}(B)\delta_{r,2}(B)}$$

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

$$f(\varepsilon_t) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{1}{2}\left(\frac{\varepsilon_t - 0}{\sigma}\right)^2\right) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{1}{2}\left(\frac{\varepsilon_t}{\sigma}\right)^2\right)$$

$$L = \prod_{t=p+1}^{n}\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{1}{2}\left(\frac{\varepsilon_t}{\sigma}\right)^2\right) = \frac{1}{2\pi^{(n-p)/2}\sigma^{n-p}}\exp\left(-\frac{1}{2\sigma^2}\sum_{t=p+1}^{n}\varepsilon_t{}^2\right)$$

$$\text{Log } L = -\left(\frac{n-p}{2}\right)\log(2\pi) - (n-p)\log\sigma - \frac{1}{2\sigma^2}\sum_{t=p+1}^{n}\varepsilon_t{}^2$$

$$= -\left(\frac{n-p}{2}\right)\log(2\pi) - (n-p)\log\sigma - \frac{1}{2\sigma^2}\sum_{t=p+1}^{n}\left(\frac{c'(B)\, Yt - d'(B)\, S_{1,t-b} - f'(B)\, S_{2,t-b}}{\delta_{r,1}(B)\delta_{r,2}(B)}\right)^2$$

- Calculate forestimatorσ^2

$$\frac{\partial \log L}{\partial \sigma} = 0$$

$$\frac{-(n-p)}{\sigma} - \left(2(-2\sigma^{-3})\sum_{t=p+1}^{n}\left(\frac{c'(B)\, Yt - d'(B)\, S_{1,t-b} - f'(B)\, S_{2,t-b}}{\delta_{r,1}(B)\delta_{r,2}(B)}\right)^2\right) = 0$$

$$\frac{-(n-p)}{\sigma} + \frac{1}{\sigma^3}\sum_{t=p+1}^{n}\left(\frac{c'(B)\, Yt - d'(B)\, S_{1,t-b} - f'(B)\, S_{2,t-b}}{\delta_{r,1}(B)\delta_{r,2}(B)}\right)^2 = 0$$

$$\sigma^2 = \frac{\sum_{t=p+1}^{n}(c'(B)Y_t - d'(B)S_{1,t-b} - f'(B)\, S_{2,t-b})^2}{(n-p).\sum_{t=p+1}^{n}(\delta_{r,1}(B)\delta_{r,2}(B))^2}$$

to find the value estimator $\phi_1, \phi_2, \ldots, \phi_p$, $\omega_{s,1}$, $\omega_{s,2}$and $\delta_{r,1}$, $\delta_{r,2}$ can be done by performing the differentiation function *log L*to the above parameters in models.

4. Results and Discussion

Result

Stock price movements TMAS.JK has alleged there is the influence of the intervention. it can be seen from the movement that moves is very high. The following interventions were thought to influence the rate of movement of the stock price TMAS.JK. Where the intervention is as follows:

1. On August 11, 2014, allegedly as a result of the election of Jokowi-JK pair as President and Vice President Republic of Indonesia on July 22, 2014
2. Jokowi speech at the APEC forum on marine toll program, and offers investment in the construction of ports to a foreign nation: 10 November 2014

Statement of results

From Figure 4.1 above, can be explained that the first intervention symbolized by S1, S2 is symbolized by a second intervention. it appears that after the first intervention increase the stock price, it is declared that the decision of the election of the President Jokowi as a positive impact. After second intervention was seen movement of stock prices rise high enough. This is why the authors to apply the methods of intervention on the analysis of the time series data.

ISBN:978-602-99849-2-7

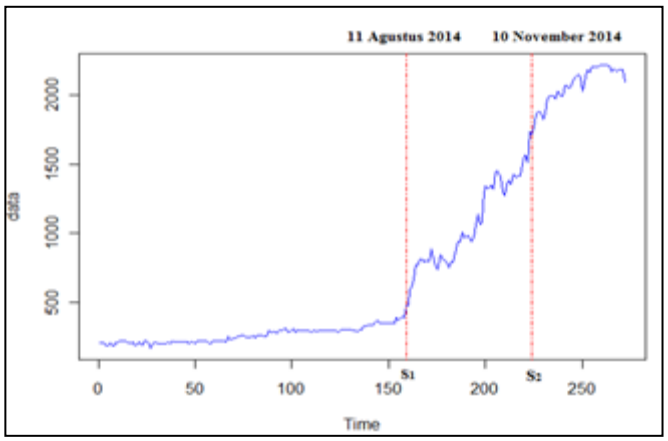


Figure 4.1: Graphic of Stock Price TMAS.JK

Explanatory text

After processing the data through the stages of identification, parameter estimation and diagnostic check, then to the data obtained before any intervention ARIMA (2,1,0) as the best model. Mathematically, this model can be written as follows:

$$(1 - B)(1 - \phi_1 B - \phi_2 B^2) X_t = \varepsilon_t$$

$$X_t = \phi_1 X_{t-1} + X_{t-1} - \phi_1 X_{t-2} + \phi_2 X_{t-2} - \phi_2 X_{t-3} + \varepsilon_t$$

$$X_t = -0.4222\, X_{t-1} + X_{t-1} + 0.1394\, X_{t-2} + 0.2828\, X_{t-3} + \varepsilon_t$$

After we find model before intervention event, now we can do the analysis of the first intervention.

First Intervention Analysis

Enactment events Jokowi-Jk as the winner of the election of 2014. The commencement is dated August 11, 2014. In this case, the event is a step function form. The first modeling is to determine the order alleged b, s, and r of the first intervention model. To determine the order of the first intervention can be seen through the residual diagram in Figure 4.2 below.

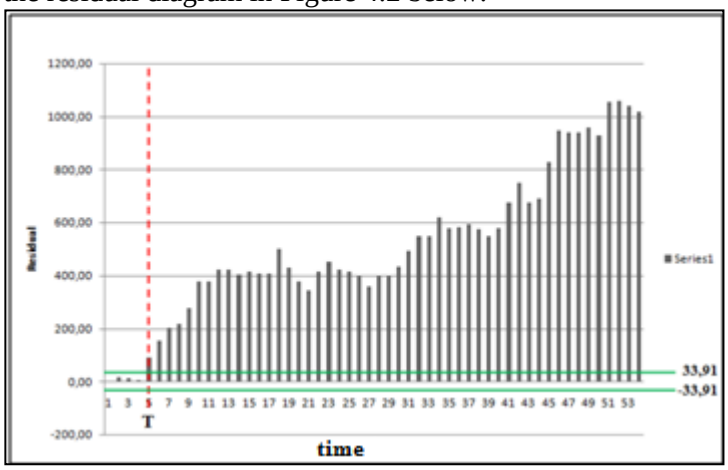


Figure 4.2: Diagram Residual for First Intervention

From Figure 4.2, it can be presumed that the intervention model order is a step function b = 0, s = 5, r = 1. This is evident from the current residual T lines had crossed the line 3 RMSE = 33.91. So that the initial allegations intervention analysis model is a step function (b = 0, s = 5, r = 1) with a model before the intervention is ARIMA (2,1,0). But After processing the data through the stages of

ISBN:978-602-99849-2-7

identification, parameter estimation and diagnostic checks, then the best model is the model of intervention analysis step function (order b = 0, s = 2, r = 1) with a model before the intervention is ARIMA (2,1 ,0).

```
> pvalue(model.2)

Coefficients:
                      s.e.        t    sign.
ar1       -0.2793   0.0665  -4.2000   0.0000
ar2       -0.2340   0.0676  -3.4615   0.0006
step-AR1   0.6919   0.0789   8.7693   0.0000
step-MA0   0.2183   0.0485   4.5010   0.0000
step-MA1  -0.1584   0.0690  -2.2957   0.0226
step-MA2   0.1732   0.0562   3.0819   0.0023
```

Figure 4.3: Output of First Intervention Model

Mathematically, this model can be written as follows:

$$X_t = \frac{\omega_2(B)B^0}{\delta_1(B)} S_{1,t} + \frac{\varepsilon_t}{(1-B)(1-\phi_1 B-\phi_2 B^2)}$$

$$X_t = \frac{0{,}2183-(-0{,}1584)\,B-0{,}1732\,B^2}{1-0{,}6919\,B} S_{1,t} + \frac{\varepsilon_t}{(1-B)(1-(-0{,}2793)\,B-(-0{,}2340)\,B^2)}$$

Second Intervention Analysis

Having obtained the intervention analysis model first, then the next can be obtained residual, to make residual diagram to determine the order of b, s, r analysis of second interventio. The residual diagram is as follows:

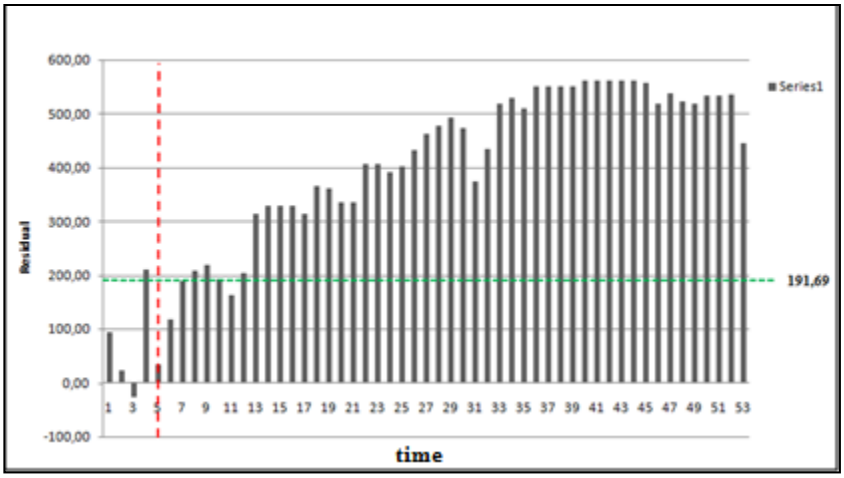


Figure 4.4: Diagram Residual for Second Intervention

From Figure 4.4, it can be presumed that the intervention model order is a step function b = 3, s = 2, r = 0. Thus obtained the alleged model: a model before the intervention is ARIMA (2,1,0), the first intervention (function step (b = 0, s = 2, r = 1)), a second intervention (function step (b = 3, s = 2, r = 0).After various tests, ahirnya elected one good model to use. Namely the intervention model a step function of the order of (b = 0, s = 2, r = 1).

ISBN:978-602-99849-2-7

```
> pvalue(model.3)

Coefficients:
                      s.e.        t   sign.
ar1         -0.2734  0.0617  -4.4311  0.0000
ar2         -0.2327  0.0626  -3.7173  0.0002
step-AR1     0.6915  0.0754   9.1711  0.0000
step-MA0     0.2182  0.0461   4.7332  0.0000
step-MA1    -0.1582  0.0653  -2.4227  0.0161
step-MA2     0.1734  0.0533   3.2533  0.0013
step.1-AR1   0.8731  0.1964   4.4455  0.0000
step.1-MA0   0.0296  0.0283   1.0459  0.2966
```

Figure 4.5: Output of Second Intervention Model

Where the model can be written as follows:

$$X_t = \frac{\omega_2(B)B^0}{\delta_1(B)} S_{1,t} + \frac{\omega_0 B^3}{\delta_1(B)} S_{2,t} + \frac{\varepsilon_t}{(1-B)(1-\phi_1 B - \phi_2 B^2)}$$

$$X_t = \frac{\omega_0 - \omega_1 B - \omega_2 B^2}{1-\delta_1 B} S_{1,t} + \frac{\omega_0 B^3}{1-\delta_1 B} S_{2,t} + \frac{\varepsilon_t}{(1-B)(1-\phi_1 B - \phi_2 B^2)}$$

$$X_t = \frac{0{,}2185 + 0{,}1582\,B - 0{,}1734\,B^2}{1 - 0{,}6915\,B} S_{1,t} + \frac{0{,}0296}{1 - 0{,}8731\,B} S_{2,t-3} + \frac{\varepsilon_t}{(1-B)(1+0{,}2734\,B + 0{,}2327\,B^2)}$$

Discussion

From the analysis above, it is proof that there has been intervention in the movement of stock prices TMAS.JK. Where his intervention I was: On August 11, 2014, allegedly as a result of the election of Jokowi-JK pair as President and Vice President Republic of Indonesia on July 22, 2014. As a result of these events, increase the stock price significantly. This can be seen from just 3 days after the intervention, TMAS share price increased by (55.61%), dated 10 August 2014 (Rp. 392), dated August 14th, 2014 (Rp. 610). The price increase continued higher.

Event the second intervention also contributed to the increase in stock prices TMAS.JK. Jokowi speech at the APEC forum on marine toll program, and offers investment in the construction of ports to a foreign nation: 10 November 2014 turned out to affect the movement of stock prices TMAS.JK. As a result of these events, share prices rebounded TMAS. Within 6 trading days, there has been a rise by (24.50%), dated 6 November 2014 (Rp. 1510), dated 14 November 2014 (Rp. 1880). The price increase continued higher.

From the description above, seen that government policy can intervene in the stock market in Indonesia. As for the TMAS.JK stock price movements can be modeled with ARIMA (2,1,0), intervention I (function step (b = 0, s = 2, r = 1)), intervention II (step function (b = 3, s = 0, r = 1)).

5. Conclusion

A conclusion should give a summary of:

a. The function of the log likelihood of intervention analysis multy input is:

$$\text{Log L} = -\left(\frac{n-p}{2}\right)\log(2\pi) - (n-p)\log\sigma - \frac{1}{2\sigma^2}\sum_{t=p+1}^{n}\left(\frac{c'(B)\,Yt - d'(B)\,S_{1,t-b} - f'(B)\,S_{2,t-b}}{\delta_{r,1}(B)\delta_{r,2}(B)}\right)^2$$

with:

$c'(B) = \delta_{r,1}(B)\delta_{r,2}(B)\phi(B)$

$d'(B) = \phi(B)\omega_{s,1}(B)\delta_{r,2}(B)$

$f'(B) = \phi(B)\omega_{s,2}(B)\delta_{r,1}(B)$

ISBN:978-602-99849-2-7

b. There are two events that lead to intervention TMAS.jk stock price movements, namely:
   1. On August 11, 2014, allegedly as a result of the election of Jokowi-JK pair as President and Vice President Republic of Indonesia on July 22, 2014
   2. Speech at the APEC forum Jokowi toll on marine program, and offers investment in the construction of ports to a foreign nation: 10 November 2014

c. The analysis model of intervention that is formed is a multi intervention analysis model multy input step function, with the model before the intervention is ARIMA (2,1,0), the first intervention (step function (b = 0, s = 2, r = 1)) , the second intervention (step function (b = 3, s=0, r = 1). Where the model can be written as follows:

$$X_t = \frac{0{,}2185 + 0{,}1582\,B - 0{,}1734\,B^2}{1 - 0{,}6915\,B} S_{1,t} + \frac{0{,}0296}{1 - 0{,}8731\,B} S_{2,t-3} + \frac{\varepsilon_t}{(1 - B)\,(1 + 0{,}2734\,B + 0{,}2327\,B^2)}$$

References

[1] Box, G.E.P. and Tiao, G.C. 1975. *Intervention Analysis With Applications to Economic and Environmental Problems*. Journal of American Statistical Association. Marc 1975, Volume 70, Number 349 Invited Paper, Theory and Methods Section.

[2] Cole,C. Quddus, M. Wheatley, A. Osmani, M. And Kay, K. 2013. *The Impact of Local Authorities' Interventions on Household Waste Collection: A case Study Approach Using Time Series Modelling*. Journal Waste Management. Elsevier Ltd.

[3] Poirer, E. Watier, L. Espie, E. Weill, F-X. Devalk, H. And Desenclos, J-C. 2007. *Evaluation of the Impact on Human Salmonellosis of Control Measures Targeted to Salmonella Enteritidis and Typhimurim in Poultry Breeding Using Time-Series Analysis and Intervention Models in France.* Cambridge University Press. United Kingdom

[4] Bain, L.J. dan Engelhardt, M. 1992. *Introduction to Probability and Mathematical Statistics*. Duxbury Press. Boston

[5] Box, G.E.P. Jenskins, G.M. and Reinsel, G.C. 1994. *Time Series Analysis Forecasting and Control: Third Edition*. Prentice-Hall International, Inc. United States of America.

[6] Makridakis, S. et al. 1995. *Metode dan Aplikasi Peramalan, Edisi Kedua, Jilid 1*. Penerbit Erlangga. Jakarta.

[7] Subanar. 2013. *Statistika Matematika*. Graha Ilmu.Yogyakarta

[8] Wei, W.W.S. 2006. *Time Series Analysis: Univariate and Multivariate Methods: Second Edition*. Addison-Wesley Publishing Company, Inc. California

ISBN:978-602-99849-2-7

Indonesia's Province Segmentation Based On Flood Disaster Impact With *Self Organizing Maps* (*Som*) Algorithm

Muhammad Muhajir, Berky Rian Efanna, Reza Aditya Pratama

Department of Statistics Universitas Islam Indonesia

muhammad.muhajir.stat89@gmail.com, berkyeki@gmail.com, and pratamareza001@gmail.com

Abstract: This Research aim to clustering Indonesia's Province based on flood disaster impact since 2000-2015. Clustering, mapping and grouping of areas could be done to decrease the number of casualties in flood, so the Government was able to prioritize certain areas. To clustering special case in Indonesia, we use Self Organizing Maps because it canreduce high dimensional data into 2 dimensional data and ignoring outlier data. As a result of SOM algorithm, 7 cluster formed differently. Majority of province is member of cluster 5th that the floods occur rarely, but south Sulawesi Selatan is the most special province, big number of harm and a lot of settlement is in hard damage if flood happen. It means, Indonesia as a country that have a big problem with flood should arrange regulation to against flood, reducing lost, and mitigation with decentralised policy, because every province have a different problem and characteristics.

Keywords : Flood, Impact, SOM.

Introduction

Most dominating disaster in Indonesia since 1815-2015 is flood. Until February 2015, the flood's occurrence about 31% compared to other disasters [3]. According to Sutopo, Indonesia become the most vulnerable country in the world particularly for natural disaster based on data issued by United Nations International Strategy for Disaster Reduction (UN-ISDR) [2]. Position of Indonesia as most vulnerable country is calculated from the number of endangered human lives, the risk of losing when natural disasters occur. Indonesia, accordingHead of Information Data and Public Relation of Indonesian National Board for Disaster Management, Sutopo Purwo Nugroho told reporters BBC Indonesia, Yusuf Arifin, was ranked sixth for the threat of catastrophic flooding.

Floodingis an overflowing of a large amount of water beyond its normal confines, especially over what is normally dry land, normally or because of no longer accommodated by the rivers and dams as well as cannot be absorbed by the tree, and the land of which result in land being flooded. It can impact human life, damages, destruction facilities and infrastructure(buildings, houses, and bridges), destroying transportation, vulnerabledeveloping, eliminate property supplies as well as equipment, damaging the economy, disrupt daily activity, causing erosion, landslide even bother even disturb the future live.

Across Indonesia, recorded 5.590 stem rivers and 600 of them potentially cause flooding. Flood-prone areas covered main rivers reach 1.4 million hectares. Of the various studies that have been done, the floods that hit cartilage area, basically caused by three things. First, human activities that cause of occurrence of spatial change and have an impact on the change of nature. Second, natural events such as very high rainfall, rising sea levels, Hurricane, etc. Third, 1 environmental degradation such as loss of vegetation cover the soil at catchment area, the superficiality of the river due to sedimentation, flow constriction the river and so on [1]. One of the factors supporting Indonesia disaster-prone countries is a human error that occur because people incomprehension to understand the environment. Cultural development in Indonesia are just business

oriented without regarding environmental aspects so that the lack of green area especially in the capital often resulting of flooding. In other case flood happened because illegal logging, a reduced number of water catchment like sedimentation, size decreasing of rivers, lakes, or other natural problem.

From economic side, according to Mauleny by that published DPR told that flood have a risk to increasing the price of necessities from 10 til 20 percent [8]. First, the impact of flood is a direct impact, like the damage of economic assets (residences, businesses, factories, infrastructure, land agricultural estates and so on). In economic terms, it categorized as stock value, flood causes a decrease of. Second, indirect impacts include a cessation of the production process, the loss of output. In economic terms, this case categorized as flow value. Third, the secondary impact it can make a slow economic growth, disruption of development plans and increasing of public debt and poverty.

Nowadays floods occur in a relatively short time and repeated each year, demanding great efforts to anticipate, so that losses can be minimized. Several government efforts that are structural (structural approach), it has not been fully able to fix flood problem in Indonesia. Flood mitigation, as long as it is more focused on providing flood control for the physical building to reduce the impact of disasters.

Sectoral, centralised or top down policy can not implemented in Indonesia, because it is not appropriate for decentralised condition. Beside government act to prevent the disaster, it is important to includes community participation to response flood. Some mistakes can make some individuals/ groups interest more dominant, then the benefit exploited to become negative. As a result of bad policies set out are not effective. Thus, flood mitigation solely physical development (structural approach), should synergy with the construction of a non physical (non-structural approach), which provide a wider space for the community participation, so the result more optimal [1].

Knowing the bad impact of flooding is not small, then we need for mitigation efforts to reduce flood losses. Mapping and grouping of areas could be done to decrease the number of casualties in flood disaster. so the Government was able to prioritize certain areas. With the help of technology, especially in the field of GIS, were able to help us in making a maps. Mapping made must not arbitrarily and in accordance with existing conditions. However, there is currently no special mapping is done by the Government with regard to the handling of the floods.

Combining GIS and statistical method can make a better maps that will describe the special characteristics from every region/province. Statistical method especially clustering can divide some province with different characteristics and merge same province in same cluster. Cluster result can be implemented in maps with GIS method, so people can understand easily. Maps result can become one of element for stakeholder to make a better policy to against flood.

There are several clustering method like K Means, K Medoid, etc but Indonesia's condition is unique, some island have different characteristics (outlier). Unfortunately, those cluster method will not work good for the data with outlier value. Clustering method with Data Mining approach like SOM can handle it. SOM also reduce high dimensional data and make a more simple mapping result without deleting real data characteristics. It is very suitable to handle Indonesia's flood case and the result can be combined with GIS mapping method.

ISBN:978-602-99849-2-7

Related Works

Bappenas said that several lack to ward off flood disaster, like participation of people as a part of stakeholder, local government regulation to confront flood is still limited, and budget for mitigation still dependent with APBN and APBD [1]. In Jember, Maliasari and friends are analyzing causal factor fo flood. And solution to prevent this disaster is with river capacity normalization, and reforestation [7]. Lestari told that to reduce flood damage, specially in Jakarta is to make weather modification. It can decreasing water surface in Pesanggrahan, Ciliwung ans Sunter [6].

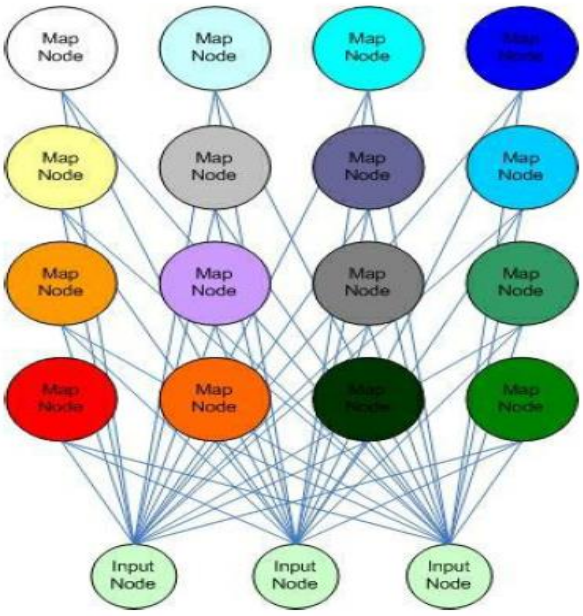
GIS is presented by Pratomo, he use GIS to knowing characteristics of flood and vulnerability of Sengkarang River Flow Territory. In his research it divide 4 class with 25 area unit [9].

Material and Methodology

Population of this research is all flood damage in province of Indonesia, and the sample is the data from 2000 until 2015. The data taken from Indonesian National Board for Disaster Management . Variables of this research is Number of Flood occur, dead victim, Lost damage, harm victim, residential Light and hard damage, Eduction Facility damage, Education facility damage. This research using *Self Organizing Maps* (*SOM*) which is number of cluster is 7, it based on the segmentation of sustainable Indonesia development with noticing island condition of Indonesia.

Self Organizing Maps (*SOM*) is a spatial form of Cluster analyze K Means. Every unit fit with a cluster and number of cluster is specified by grid size that set in rectangular or hexagonal form. SOM use grid in process of mapping, when 2 unit have a same characteristics, it will adjacent. This algorithm concentration for a biggest similarity [10].

The structure of a SOM is fairly simple, and is best understood with the use of an illustration such as picture bellow [4].



That picture is a 4x4 SOM network (4 nodes down, 4 nodes across). It is easy to overlook this structure as being trivial, but there are a few key things to notice. First, each map node is connected to each input node. For this small 4x4 node network, that is 4x4x3=48 connections. Secondly, notice that *map nodes are notconnected to each other*. The nodes are organized in this manner, as a 2-D grid makes it easy to visualize the results. This representation is also useful when the SOM algorithm is used. In this configuration, each map node has a unique (i,j) coordinate. This makes it easy to reference a node in the network, and to calculate the distances between nodes. Because of the connections only to the input nodes, the map nodes are oblivious as to what values their neighbors have. A map node will only update its' weights (explained next) based on what the input vector tells it. The following relationships describe what a node essentially is:

ISBN:978-602-99849-2-7

1. *network⊂mapNode⊂float weights* [*numWeights*]
2. *inputVectors⊂inputVector⊂float weights*[*numWeights*]

1 says that the network (the 4x4 grid above) contains map nodes. A single map node contains an array of floats, or its' weights. numWeights will become more apparent during application discussion. The only other common item that a map node should contain is it's (*i,j*) position in the network. 2 says that the collection of input vectors (or input nodes) contains individual input vectors. Each input vector contains an array of floats, or its' weights. Note that numWeights is the same for both weight vectors. The weight vectors must be the same for map nodes and input vectors or the algorithm will not work.

The Self-Organizing Map algorithm can be broken up into 6 steps [5]:

Step one, each node's weights are initialized.

Step two, a vector is chosen at random from the set of training data and presented to the network.

Step three, every node in the network is examined to calculate which ones' weights are most like the input vector. The winning node is commonly known as the *Best Matching Unit* (BMU).

$$DistFromInput^2 = \sum_{i=0}^{i=n} (I_i - W_i)^2$$

with
I = current input vector
W = node's weight vector
n = number of weights

Step four, the radius of the neighborhood of the BMU is calculated. This value starts large. Typically it is set to be the radius of the network, diminishing each time-step. The formula is :

Radius of the neighbourhood

$$\sigma(t) = \sigma_0 e^{(-\frac{t}{\lambda})}$$

Time constant

$$\lambda = numIterations/mapRadius$$

With
t= current iteration
λ= time constant
σ_0=radius of the map

Step five, any nodes found within the radius of the BMU, calculated in step 4, are adjusted to make them more like the input vector with this formula

New Weight of a node

$$W(t+1) = W(t) + \Theta(t)L(t)(I(t) - W(t))$$

Learning rate is

$$L(t) = L_0 e^{(-\frac{t}{\lambda})}$$

The closer a node is to the BMU, the more its' weights are altered with this formula

$$\Theta(t) = e^{(\frac{-distFromBMU^2}{2\sigma^2(t)})}$$

Step 6, repeat step 2 for N iterations.

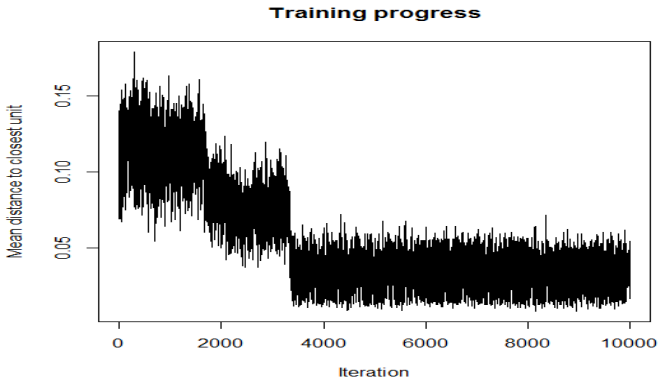
Result

Clustering with *Self Organizing Maps*

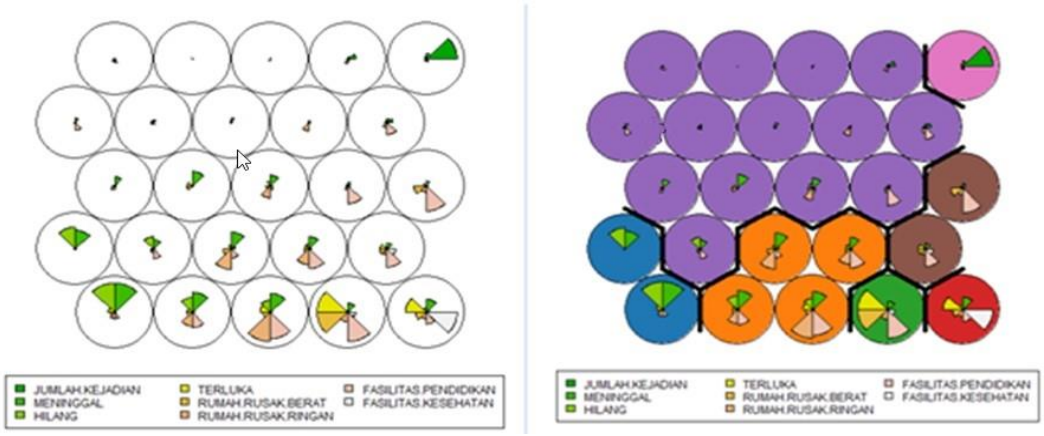
In Self Algorithm Maps, meed iterations to get best cluster result. Picture below is training progress, shows number of training progress that describe how much of iteration and will affect with lowest mean distance to closest unit. More iteration can make a better cluster result, because each unit have lower closest distance. More often SOM executing the iteration, the result become more appropriate and have a little different with previous value. It mean more iteration, the result of

ISBN:978-602-99849-2-7

training progress is more convergence. If we look this picture of Training progress, iteration begin to convergence after 4000th iteration.



More iteration caused lower mean of distance cluster unit and better cluster result. After 4000 time of iteration show that training progress become more stable, if researcher use 10000 time of iteration, mean distance to closest unit value is stopped with value under 0.1 and stable around 0.05. SOM algorithm can give some model. If SOM algorithm executed with R Program, it will result a diagram contain several circle and topologically will contiguous each other with other circle that have same characteristics.



To understand every diagram in SOM algorithm easily, Luckily if we works with R, it automatically give special colour and border for every vectors in visualization of maps plot.

According to Picture, big damage of flood in particular for health facility for example, associated with sample projection in right bottom corner of SOM maps represented with red circle. Green circle in left side of red one, associated with cluster that have a big percentage of harm and severely damage for education facility and settlement if flood happen. Orange circle associate with cluster that have a low percentage of victim but big percentage of education facility and light damage settlement. Blue circle associate with a cluster that have a big lost, more victims dead or lost when flood happen. Brown circle associate with big damage of education facility, purple circle associate with cluster that rarely flood disaster. Most often flood disaster happen in province that become cluster member of pink circle, even no serious damage recorded here.

Cluster	Number of Cluster	Cluster Member
1	2	Sumatra Utara and Papua Barat
2	3	Jawa Timur, Banten and Jawa Tengah
3	1	Sulawesi Selatan
4	1	Jawa Barat
5	25	Aceh, Jambi, Sumatra Selatan, Bengkulu, Lampung, Bangka Belitung, Kepulauan Riau, Jakarta, DIY, Bali, NTB, NTTT,

ISBN:978-602-99849-2-7

		Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tenggara, Sulawesi Tengah, Gorontalo, Sulawesi Barat, Maluku, Maluku Utara and Papua
6	1	Riau
7	1	Sumatra Barat

Map result from cluster analyze with Self Organizing Maps is

After find the cluster result that contain province as cluster members, researcher will divide every cluster member. Using descriptive statistics, researcher can understand every cluster characteristics.
Table. Profile of SOM cluster based on mean.

Cluster	Num. of Disaster	Dead	Lost	Harm	Hard Damage	Light Damage	Education Fac.	Health Fac.
1	156.83	217.00	137.00	6229.50	1854.00	2674.00	79.00	11.00
2	574.67	168.33	49.00	13883.67	5054.00	22122.33	493.33	34.00
3	260.00	86.00	10.00	60406.00	20349.00	8127.00	691.00	249.00
4	658.00	124.00	23.00	39787.00	6899.00	5506.00	279.00	1033.00
5	99.08	34.64	7.76	1216.64	1059.36	2202.84	100.52	21.52
6	98.95	48.00	1.00	5789.00	7998.00	5551.00	604.00	69.00
7	5720.58	55.00	13.00	56.00	745.00	1950.00	74.00	36.00

Cluster 1, contain Sumatra Utara and Papua Barat have a medium number of flood disaster. The characteristics is unique because this cluster members has a lowest level of damage for healthy facility if flood happen but biggest number of victims because of dead and lost comparing with other cluster members.

Cluster 2, 3 province in Jawa Island that is Jawa Timur, Banten and Jawa Tengah. This cluster member have a bigest number of light damage for settlement if flood happen. This cluster is 3th most often flood case, and 2th with big quantities of dead victims.

Cluster 3, Sulawesi Selatan is the only member. The characteristics of this cluster is a big number of harm and a lot of settlement is in hard damage if flood happen. Number of damage for healthy facility is biggest compare with other cluster.

Cluster 4, Jawa Barat , in this cluster have not specific condition , although a big case of flood occur in Jawa Barat. Number of damage is still lower than cluster 3. But this cluster have a big damage with health facility.

Cluster 5, 25 province in Indonesia is cluster member of 5^{th} cluster. Flood rarely happen in this cluster, so the number of dead people as flood victims is lowest in Indonesia. There are no prominent flood effect in this cluster.

Cluster 6 have one cluster member, that is Riau. It recorded as province that have a lowest case of flood. That is only one victim whose lost and only 48 people dead because of flood disaster.

ISBN:978-602-99849-2-7

Cluster 7 is Sumatra Barat, as a cluster that have a biggest number of flood case, Sumatra Barat have lowest number of harm, settlement and education facility damage caused by flood.

In General, most of the provinces in Indonesia have the characteristics as described in cluster 5 which is rarely flood happened. Based on a map, Sumatra region the most islands region with different characteristics, Sumatra Barat that has largest number of cases however, the amount of damage for very little. The conditions on the islands of Java has a big number of cases but Sulawesi Selatan denoted province with rare floods case but the number of injured, homes destroyed and damaged of education facility is big.

Conclusion

Number of cluster from this research is 7 cluster, every province have a different characteristics based on flood damage. SOM is different with other cluster method because it can calculate without any classical assumption test like no outlier, collinearity, etc. Sulawesi Selatan as only member of cluster 3 have most special characteristics because this province occur flood rarely but the number of damage is big specially in harm and harm damage. Suggestion for the next research, can add more variables because SOM can reduce a big dimensional data. More variables, the cluster result is become more specific. Government should optimizing the action of mitigation to preventing more damage. This cluster research can help government to arrange special regulation and make a special policy for some province to get a particular step decreasing a damage level and increasing local people and government awareness for flood mitigation.

References

[1] Bappenas, "Kebijakan Penanggulangan banjir di Indonesia", Deputi Bidang Sarana dan Prasarana, Direktorat Pengairan dan Irigrasi, 2014.

[2] BBC, 2011, Indonesia Negara Rawan Bencana. .http://www.bbc.com/indonesia/berita_indonesia/2011/08/110810_indonesia_tsunami.shtml.Retrived 20 august, 2015.

[3] BNPB. 2015. *Data Banjir*.http://dibi.bnpb.go.id/DesInventar/dashboard.jsp. Retrived 29 august, 2015.

[4] Chesnut , C., "Self Organizing Map AI for Pictures", *generation 5*, 2004., http://www.generation5.org/content/2004/aisompic.asp. Retrived 25 august, 2015.

[5] Guthikonda, S.M., "*Kohonen Self-Organizing Maps*", Wittenberg University, 2005.

[6] Lestari. S., "Analisis Kerugian Banjir dan Biaya Penerapan Teknologi Modifikasi Cuaca dalam Mengatasi Banjir di DKI Jakarta". *Journal Sains & Teknologi Modifikasi Cuaca*, Vol.3 (2): 155-159, 2002.

[7] Maliasari, R.D., Pudyono and Devia, Y.P., "Analisis Penyebab dan Penanggulangan Banjir di Kecamatan Panti Kabupaten Brawijaya", 2012.

[8] Mauleny, A.T. 2014. "Perspektif Ekonomi Kebijakan Penanggulangan Banjir". *Info Singkat Ekonomi dan Kebijakan Publik*. Volume VI, No.03/I/P3DI/Februari/2014.

[9] Pratomo, J.K., "*Analisis Kerentanan Banjir Di Daerah Aliran Sungai Sengkarang Kabupaten Pekalongan Provinsi Jawa Tengah Dengan Bantuan Sistem Informasi Geografis*", Thesis, Department Geography, Universitas Muhammadiyah Surakarta, 2008.

[10] Wehrens, Ron dan Buydenss, Lutgarde M.C.. Self and Super-organizing Maps in R : The Kohonen Package. *Journal of Statistical Software*. Volume 21. Issue 5 (2007).

ISBN:978-602-99849-2-7

The Utilization Density Functional Theory in Structure Determination And Hydrogen Storage Properties of $Ca(BH_4)_2 \cdot 2NH_3$ Compounds

Muhammad Arsyik Kurniawan S.

Chemistry Department
Faculty of Mathematics and Natural Science, Universitas Islam Indonesia

E-mail: m.arsyik@gmail.com

Abstract: This paper addresses the use of density functional theory framework for calculating energy, structure and hydrogen storage properties of $Ca(BH_4)_2 \cdot 2NH_3$ compounds.The method allows good-quality first-principle calculations to be performed. Using this method to generate basis-set, we show that convergence energy and geometry is a necessary condition to obtain the correct model of crystal structure and lattice parameters.

Keywords:density functional theory; first-principle; convergence energy.

1. Introduction

The efficient and safe storage of hydrogen now is known asone of the key technological challenges in the transition towardsa hydrogen-based energy economy [1]. Whereas hydrogen fortransportation applications is currently stored using cryogenicsor high pressure, there is important research and developmentactivity in the use of condensed-phase materials.However, the multiple-target criteria accepted as necessary for the successful implementation of such stores have not yet been met by any single material.

From the research effort conducted in solid-state materials capable of storing hydrogen, the NH_3BH_3 compound called ammonia borane, with an ideal storage capacity of 19.5 wt.% H_2 and a reported release [2]. Ammonia borane upon dehydrogenation, accompanied volatile compounds, e.g., ammonia, diborane, and borazine, are also evolves [3,4], which lead to a reduction of hydrogen capacity and are fatal for fuel cell application. Recent efforts have been made on the ammonia complexesof metal borohydrides and ammonia boron derivatives that uses for hydrogen storage.

In order to better provide some insights into the understanding of the bonding characters and dehydrogenation properties, our objective is to utilize of projector augmented-wave method in density functional theory framework to determine structure and hydrogen storage properties of calcium borohidride diammonia $Ca(BH_4)_2 \cdot 2NH_3$. We first investigate the basis set as descriptor of electron wave function of atoms in $Ca(BH_4)_2 \cdot 2NH_3$compound and then explore the crystal structure until define hydrogen storage properties.

Details of this article is start by literature review in Sec. II, followed by material and methodology in Sec. III. The explain of discussion of our results in Sec. IV. A summary of our conclusions is given in Sec. V.

2. Literature Review

2.1. DensityFunctionalTheory

In the liquid and gas phase, the atoms or molecules are able to move freely and collisions with each other, butthe solid material has a characteristic that the atoms and molecules can't move freely. Solid material has the form of a structured and rigid, so the distance between atoms is very close. This causes the movement of electrons around atoms resembling a cloud of electrons.

The movement of the electron does not only occur in the metal material, but also occurs in any condensed solid material. The movement of electrons is determined the structure and properties of materials.

Each electron in any material can be represented by

A wave function ψ

And each of the wave function(ψ) has its own electron density ρ

$$|\psi|^2 = \rho,$$

orit can be written as

$$\rho = n(r).$$

This method isknown as the Density Functional Theory [5].

The basic idea behind DFT is that the energy of an electron system can be written in terms of the electron probability density $n(r)$. Density functional theory is used to handle the many electron interaction,for a system of *N* electron, $n(r)$ donates the total electron density at a particular point in space r. The energy of *N* electrons in an external potential $V_{ext}(r)$ is written as a functional of the electron density $n(r)$.The total electronic energy E is said to be a functional of the electron density, donated $Etot[n(r)]$, in the sense that for a given function $n(r)$, as written as below,

$$E_{tot}[n(r)] = T[n(r)] + E_{xc}[n(r)] + E_H[n(r)] + \int Vext(r)n(r)\,dr \quad (1)$$

where $T[n(r)]$ is the non-interacting kinetic energy, $E_{xc}[n(r)]$ the exchange and correlation energy, $E_H[n(r)]$ the electrostatic or Hartree energy from the electron-electron repulsion. The ground state corresponds to the minimum of $E_{tot}[n(r)]$, is found from the self-consistent solution of Kohn-Sham equation [6].

Density functional theory is widely used electronic structure method which can be divided into few class: (i) the linear methods [7,8], ranging from linear augmented-plane-wave (LAPW) method to the linear muffin-tin orbital (LMTO) method, (ii) the pseudopotentials method based on norm-conserving ab initio pseudopotentials [9]. The next class, uses Gaussian basis set to expand the full wave functions.

In the linear method, like projector augmented-wave (PAW) [10], it uses plane-wave wave function as representation of electron wave function (basis set). Basis set is used as mathematical representation of electron wave function, which plane-wave basis set has form

$$\psi_{j,k}(r) = \sum_{\mathrm{G}} c_{j,k+\mathrm{G}} \cdot \mathrm{e}^{i(k+\mathrm{G})\cdot \mathrm{r}} \quad (2)$$

with expansions coefficients$c_{j,k+\mathrm{G}}$ for periodic boundary condition in real space G and reciprocal space *k*. The expansions coefficientswill decreased exponentially if kinetic energy of plane-wave increase.

$$\frac{(\mathrm{k+G})^2}{2} \leq E_{cut} \quad (3)$$

This plane-wave basis set only depended by lattice dimension and cutoff kinetics energy.

This method is widely used to study, properties of photocatalytic material TiO2 and metal doping effect [11], chemical reaction mechanism [12], and much more. Studied the ability of the hydrogen molecule dehydrogenation of ammonia-borane with DFT method [13] and also studied the structure, electronic properties of complex metal dehydrogenation of ammonia-borane derivative[14].

Densityfunctionaltheorymethodis the best usedfor studyingthe determination ofpropertiesof the moleculesthatare staticor lesschangein their structuresassolids[15]. This method have proved to be a

ISBN:978-602-99849-2-7

reliable and computationally tractable tool in materials science, condensed matter physics and chemistry, and now impacted virtually every side of modern science and technology.

3. Material & Methodology

3.1 Material

Materialinthis study were drawnfromthe experimental results conducted by Chu et.al [16], and use ABINIT Code [17] to implemented the modeling and GnuPlot for analysis and plotting data.

3.2 Method

3.2.1. Determination of structure

Study of convergeenergybeginsby testing of basis setto obtain thebeststructure. Samplingbasis set which is testedusing a standardof kineticenergycutoffrangingfrom10 to46Ha, toobtainthe energy differenceof 0.001Ha. If thedifference inenergy of the systemhas been reached,then the structureobtained hasconvergedand obtained the best structural model of $Ca(BH_4)_2{\cdot}2NH_3$ crystal.

3.2.2. Hydrogen storage properties

This stageisrelated tothe previousresults, where it isfollowed bythe analysis ofthe crystal structureandproperties ofdehydrogenation$Ca(BH_4)_2{\cdot}2NH_3$.

4. Results and Discussion

Energyconvergencestudyby testing of electron wave function aims to get good full basis set in modelingcrystalstructure and get energyof systemswith increasingof kinetic energyin thesystem of$Ca(BH_4)_2{\cdot}2NH_3$. The energy of $Ca(BH_4)_2{\cdot}2NH_3$compound show in Fig. 1 is about -285,00 Ha or 178,838.355 kcal/mol (theoretically).

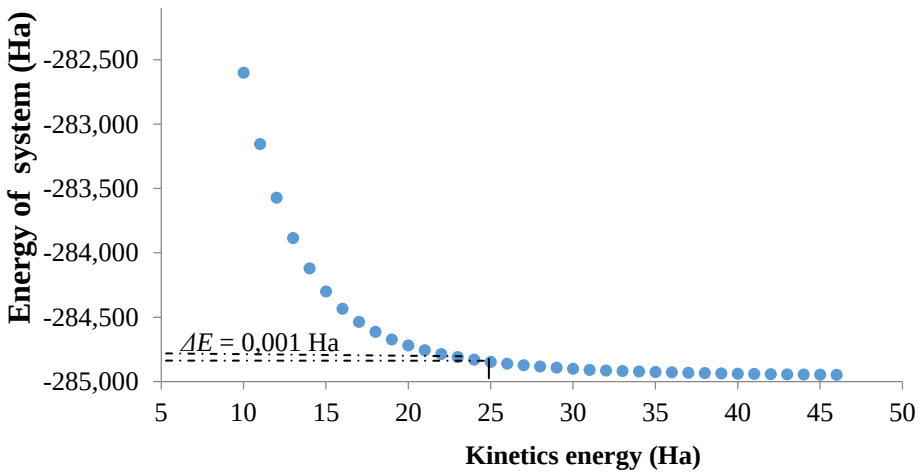


Fig. 1. Convergence energy study of $Ca(BH_4)_2{\cdot}2NH_3$ crystal structure

Convergence criteria are taken in the form of the difference intwo energy of the system inside iteration processand reaches0.001Ha. The result of this study is get the energy system has not changed much to the energy difference with 25.7 Ha of cutoff energy or700eVis quite acceptable as a convergence of basis set parameters.

Optimization geometry using density functional theory method inside ABINIT code with generalized gradient approximation (GGA) as exchange and correlation energy [18] produce the geometry structure of $Ca(BH_4)_2{\cdot}2NH_3$compound as lattice crystal which contains 76 atoms show in Fig. 2. This crystal lattice has orthorhombic space group (Pbcn:60), lattice parameter compare between experimental and calculation is show in Table 1.

ISBN:978-602-99849-2-7

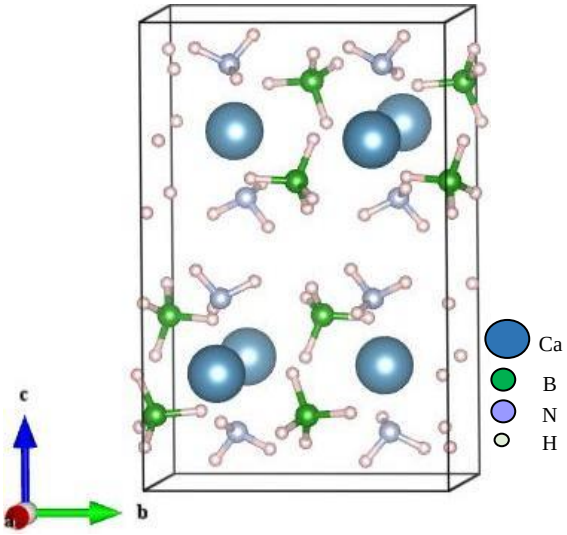


Fig. 2. Structure of $Ca(BH_4)_2 \cdot 2NH_3$ crystal lattice

Table 1. Lattice Parameter between calculation and experimental

Lattice parameter	Calculation	Experimental*
a (Å)	6,492	6,416
b (Å)	8,317	8,390
c (Å)	12,683	12,702
Volume ($Å^3$)	684,912679	683,751708
α (deg)	90	90
β (deg)	90	90
γ (deg)	90	90

*Source [16]

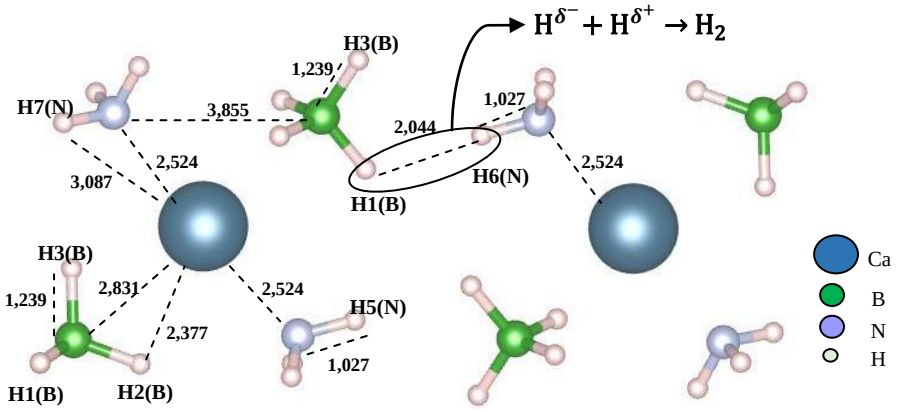


Fig. 3 Bond and interaction length inside of $Ca(BH_4)_2 \cdot 2NH_3$ compounds

ISBN:978-602-99849-2-7

Lattice structure in Table 1. shown that both calculation and experimental result does not havemuch of a difference, it indicatesthe results ofoptimization hassucceeded in obtainingthe beststructural model of $Ca(BH_4)_2{\cdot}2NH_3$ crystal.

In order to better provide some insights into the understanding of hydrogen storage properties of $Ca(BH_4)_2{\cdot}2NH_3$ crystal, the further study is about the complex structure of this compound as octahedral shaped with Ca^{2+} ionic as center and surrounds by four of BH_4^- molecules in side position and two NH_3 molecule in vertical position. The detail visualization is show in Fig. 3.

Bond length of N-H has longer than N-H inside NH_3 molecule in gas phase (1,027:1,0170), because electron density of N has increased for interaction with Ca, this indicates interaction between nitrogen (N) atom and calcium (Ca) as acid-base complex interaction, bond density of N-H has decreased and bond length between N-H increased. The structure of BH_4^- molecule is tetrahedral with B-H bond length is 1,239 Å, if compared with borane as BH_3 with bond length of B-H 1,190 Å, it indicates that B-H bond length in BH_4^- has increased. The atoms H(B) are nearby calcium (Ca) in octahedral structure, it showed atom H(B) has negative charge that used for interaction with positive charge in Ca.

Inside Fig. 3 showed the interaction between two adjacent hydrogen atom, that is H6(N) and H1(B), closest interaction distance is about 2,044 Å, this interaction is called dihydrogen interaction between $H_N^{\delta+}$ from NH_3 molecule with $H_B^{\delta-}$ from BH_4^- molecule, which potentially as source of hydrogen molecule (H_2). Interaction between two adjacent hydrogen atom as dihydrogen interaction is similar with hydrogen bond but differ in assumption, dihydrogen interaction in which thehydrogen atomsnormally would not becontiguouswiththe otherhydrogenatomwith a distance oflessthan 2.4Å, the characteristic of dihydrogenbondingis the distance betweenthe hydrogen atomsapproaching1.8Å[19].

5. Conclusion

Density functional theoryis one of theoretical chemistry that can be usedto explain the phenomenathat occur inchemical systemsin micro scale, which can not beexplained inthe experimental section. In this case, modelinga good geometry structureof $Ca(BH_4)_2{\cdot}2NH_3$compound is a necessary condition to obtain the correct understanding of $Ca(BH_4)_2{\cdot}2NH_3$crystal structure, lattice parameters and its properties as hydrogen storage material. This method very useful for research in most chemical system and phase, or simulating the development of new research in physics, chemist, biologist, and much more.

References

[1] Schlapbach, L. & Zuttel, A. Hydrogen-storage materials for mobile applications. *Nature* **414**, 353–358 (2001).

[2] Baitalow, F. Baumann, J., Wong, G.,Thermal decomposition of B–N–H compounds investigated by using combined thermoanalytical methods, Thermochimica Acta 391, 159–168, (2002)

[3] Hu, M. G., Geanangel, R. A. &Wendlandt,W.W. The thermal decomposition of ammonia borane. *Thermochim. Acta* **23**, 249–255 (1978).

[4] Baumann, J., Baitalow, F., Wong, G. Thermal decomposition of polymeric aminoborane $(H_2BNH_2)_x$ under hydrogen release. Thermochimica Acta, 430, 9-14 (2005)

[5] Hohenberg, P., Kohn, W., Inhomogeneous Electron Gas. Phys. Rev. 136, B864–B871. doi:10.1103/PhysRev.136.B864 (1964)

[6] Kohn, W., Sham, L.J.,. Self-Consistent Equations Including Exchange and Correlation Effects. Phys. Rev. 140, A1133–A1138. doi:10.1103/PhysRev.140.A1133 (1965)

[7] Andersen, O.K.,. Linear methods in band theory. Phys. Rev. B 12, 3060–3083. doi:10.1103/PhysRevB.12.3060, (1975)

[8] Slater, J.C.,. Wave Functions in a Periodic Potential. Phys. Rev. 51, 846–851. doi:10.1103/PhysRev.51.846, (1937)

ISBN:978-602-99849-2-7

[9] Hamann, D.R., Schlüter, M., Chiang, C.,. Norm-Conserving Pseudopotentials. Phys. Rev. Lett. 43, 1494–1497. doi:10.1103/PhysRevLett.43.1494, (1979)

[10] Blöchl, P.E.,. Projector augmented-wave method. Phys. Rev. B 50, 17953–17979. doi:10.1103/PhysRevB.50.17953, (1994)

[11] *Xu et al*, Band structures of TiO2 doped with N, C and B,. *J Zhejiang Univ SCIENCE B 7(4):299-303 (2006)*

[12] Borowski T, Georgiev V, Siegbahn PE. Catalytic reaction mechanism of homogentisate dioxygenase: a hybrid DFT study. J Am Chem Soc. Dec 14;127(49):17303-14 (2005)

[13] Miranda, C.R., Ceder, G.,. Ab initio investigation of ammonia-borane complexes for hydrogen storage. J. Chem. Phys. 126, 184703. doi:10.1063/1.2730785 (2007b)

[14] Wu, H., Zhou, W., Yildirim, T.,. Alkali and Alkaline-Earth Metal Amidoboranes: Structure, Crystal Chemistry, and Hydrogen Storage Properties 14834–14839 (2008)

[15] Leach, A.R.,. Molecular modelling: principles and applications, 2nd ed. ed. Prentice Hall, Harlow, England ; New York (2001)

[16] Chu, H., Wu, G., Xiong, Z., Guo, J., He, T., Chen, P.,. Structure and Hydrogen Storage Properties of Calcium Borohydride Diammoniate. *Chem. Mater*. 22, 6021–6028. (2010)

[17] Gonze, X., Amadon, B., Anglade, P.-M., Beuken, J.-M., Bottin, F., Boulanger, P., Bruneval, F., Caliste, D., Caracas, R., Côté, M., Deutsch, T., Genovese, L., Ghosez, P., Giantomassi, M., Goedecker, S., Hamann, D.R., Hermet, P., Jollet, F., Jomard, G., Leroux, S., Mancini, M., Mazevet, S., Oliveira, M.J.T., Onida, G., Pouillon, Y., Rangel, T., Rignanese, G.-M., Sangalli, D., Shaltaf, R., Torrent, M., Verstraete, M.J., Zerah, G., Zwanziger, J.W.,. ABINIT: First-principles approach to material and nanosystem properties. *Comput. Phys. Commun*. 180, 2582–2615.(2009)

[18] Perdew, J.P., Burke, K., Ernzerhof, M.,. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett*. 77, 3865–3868. (1996)

[19] Custelcean, R., Jackson, J.E.,. Dihydrogen Bonding: Structures, Energetics, and Dynamics. *Chem. Rev*. 101, 1963–1980. (2001)

ISBN:978-602-99849-2-7

Statistical Analysis of the difference Absolut Neutrophil Count (ANC) in the level sepsis patients

Suharyanto[1], Rizka Asdie[2]

[1]Department of Internal Medicine, Medical Faculty, Gadjah Mada University
[2]Department of Internal Medicine, Medical Faculty, Gadjah Mada University

prharyjo2003@yahoo.com,rizka_asdie@gmail.com

Abstract: Sepsis is a common problem in critically ill patients and cause of death in intensive care units. Sepsis cases occur worldwide is estimated to 1.8 million cases per year. The mortality rateof sepsis is still high ranging from 30-70%.The mortality rate is higher in the elderly, immunocompromised and critically ill.Early diagnosis of infection and sepsis before it develops into organ dysfunction has significant implications of the sepsis managementand outcomes. The gold standard diagnosis of sepsis is positive blood culture results, its need many days. There is no gold standard for detection of sepsis which the process is fast, cheap and widely available. The research to know the *Absolut Neutrophil Count* (ANC)value at the Sepsis level, moderate, severe sepsis and syock septic patients.
The research method is a prospective cohort. The patients with the inclusion and exclusion criteria of sepsis were taken informed concent and examination of laboratory. The *Absolut Neutrophil Count* (ANC)will be analysis for moderate, severesepsisand syock septic.

Result

The analysis of mean and standart deviation of ANC of sepsis level are Sepsis (14.40±7.40), Severe Sepsis (19.01±4.03), (11.47±6.79), The result shown that the the differences of ANC of sepsis level are (sepsis vs severe sepsis p=0.076) and not significant. The result compare of sepsis vs syock p=0.245 and not significant. The compare of severe sepsis vs syock p=0.005 and significant.

Conclusion

The analysis of mean and standart deviation of ANC of sepsis level are not significant. Need the research with more samples.

Keywords:ANC; Sepsis; moderateSepsis; severeSepsis.

1. Introduction

Sepsis is a common event in USA with an estimated 751,000 cases [1]. The large observational Sepsis Occurrence in Acutely Ill Patients (SOAP) about 30% of an intensive care unit (ICU) [2]. Padkin et al. reported that 27% of adult ICU patients met severe sepsis criteria in the United Kingdom country[3]. Diagnosis of sepsis, increase of the early resuscitation and therapy of severe sepsis is difficult [4–6], to identify a marker of sepsisto rapidly diagnosis. The good marker of infection should be sensitive to detect of infection in patients with minimal host response and should be rapidly and conveniently measured of prognostic significance [7]. The criteria of SIRS are temperature of body >38°C or <36°C; respiratory rate (RR) >20 breaths per minute/PCO2 <32 mmHg; heart rate (HR) >90 beats per minute; white blood cell count <12×109/l or <4.0×109/l) [8] 2,3,4,5-7,8,21

General signs and symptoms
Rigours, fever (sometimes hypothermia)
Tachypnoea/respiratory alkalosis
Positive fluid balance, oedema
Generalised haematological/inflammatory reaction
Increased (sometimes decreased) white blood cell count
Increased inflammatory markers (C-reactive protein, procalcitonin, interleukin-6)
Haemodynamic alterations
Arterial hypotension
Unexplained tachycardia
Increased cardiac output/low systemic vascular resistance/high SvO_2
Altered skin perfusion
Decreased urine output
Unexplained hyperlactataemia/increased base deficit
Signs of organ dysfunction
Hypoxaemia (acute lung injury)
Altered mental status
Unexplained alteration in renal function
Hyperglycaemia
Thrombocytopaenia/disseminated intravascular coagulation
Unexplained alteration in liver function tests (hyperbilirubinaemia)
Intolerance to feeding (altered gastrointestinal motility)

The current definitions of sepsis are, therefore:

- Infection is a pathologic process caused by the invasion of pathogenic microorganisms to the tissue or fluid or body cavity.
- Sepsis is clinical syndrome defined by infection symptom and a systemic inflammatory response
- Severe sepsis is sepsis that complicated by organ dysfunction
- Septic shock is severe sepsis plus a state of acute circulatory failure (persistent arterial hypotension (systolic pressure < 90 mmHg, a mean arterial pressure <60 mmHg or a reduction in systolic pressure of >40 mmHg from baseline).

Diagnosis

Diagnosis of infection and sepsis patientis often difficult because of the frequently multiple and complex underlying disease. The list of signs of sepsis in the 2001 Sepsis Definitions Conference are a useful guide to diagnosis of sepsis. Fever of ICU patient is causes, by infectious and non-infectious. a raised white blood cell count can be found in many inflammatory processes[9].

Management

Basic standard of care and individual organ support, the management of the patient with severe sepsis are 4 factors: infection control, haemodynamic support, immunomodulatory interventions and metabolic/endocrine support.

Infection control

Two component of infection Control are: removal of an infected focus and appropriate antimicrobial therapy. The earlier implementation of management may be associated with the survival [10].

2. Material & Methodology

Data

The data was collected from subjects. The subjects are septic patients at Sardjito Hospital. The data are demography and laboratory results. The data was analyse with SPSS version 17.

ISBN:978-602-99849-2-7

Method

The research method is a prospective cohort. The subjects with the inclusion and exclusion criteria of sepsis were taken informed concent and examination of laboratory. The ANC value will be analysis for, moderate severe sepsis and syock septic.

3. Results and Discussion

Result

There are 38 subjects consist of 19 male and 19 female ,the outcome deeath are 18 and life are 20 subjects (Table. 1)

Tabel 1. Characteristic subjects.

Variables	N	%
Sex		
Male	19	50.0
Female	19	50.0
Death		
Yes	18	47.4
No	20	52.6
Intensive Care Unit		
Yes	21	55.3
No	17	44.7

The mean and standart deviation of te variable are: age (51,82±15.82), SOFA (7.30±4.04), PCT (19.06±29.51), WBC (17.46±7.31), ANC (14.85±6.87) (table 2).

Table 2. The mean and standard deviation of variable.

Variables	N	Min	Max	Mean	Std. Deviation
Age	38	20.00	87.00	51.82	15.82
SOFA	38	1.00	15.00	7.03	4.04
PCT	38	.51	100.00	19.06	29.51
WBC	38	.7	32.9	17.46	7.31
ANC	38	.42	26.74	14.85	6.87

The distibusion of subject was anayised with QQ Plot and the result distribution is normal. The Figure 2 shown the Distribution of Subject data.

ISBN:978-602-99849-2-7

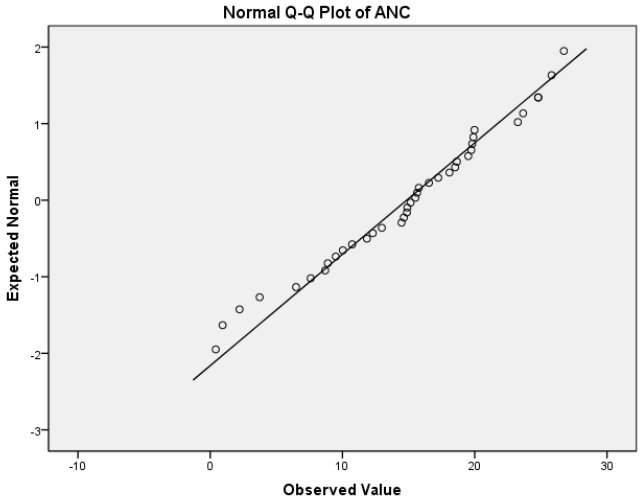


Figure 2. Distribution of Subject data.

The analysis of mean and standart deviation of ANC of sepsis level are Sepsis (14.40±7.40), Severe Sepsis (19.01±4.03), (11.47±6.79),

Table 4. The mean and standart deviation of ANC of sepsis level

Severity	Mean	Std. Deviation
Sepsis	14.40	7.40
Severe Sepsis	19.01	4.03
Syock	11.47	6.79
Total	14.85	6.87

To compare the differences of ANC osef sepsis level used ANOVA methods. The result of analysis at table 5.

Table 5. The compare of mean and standart deviation of ANC of sepsis level.

(I) Severity	(J) Severity	Mean Difference (I-J)	Sig.	95% Confidence Interval Lower Bound	Upper Bound
Sepsis	Severe Sepsis	-4.60808	.076	-9.729	0.512
	Syok	2.92462	.245	-2.092	7.942
Severe Sepsis	Sepsis	4.60808	.076	-0.512	9.729
	Syok	7.53269	.005	2.412	12.653
Syok	Sepsis	-2.92462	.245	-7.942	2.092
	Severe Sepsis	-7.53269	.005	-12.653	-2.412

ISBN:978-602-99849-2-7

The result shown that the the differences of ANC of sepsis level are (sepsis vs severe sepsis p=0.076; sepsis vs syock p=0.245; severe sepsis vs syock p=0.005).

Discussion

The analysis of mean and standart deviation of ANC of sepsis level are Sepsis (14.40±7.40), Severe Sepsis (19.01±4.03), (11.47±6.79), The result shown that the the differences of ANC of sepsis level are (sepsis vs severe sepsis p=0.076) and not significant. The result compare of sepsis vs syock p=0.245 and not significant. The compare of severe sepsis vs syock p=0.005 and significant. The reprodusibility of granulosit /PMN used as parameter to indication of infection or sepsis (Kim *et al.*, 2011).

4. Conclusion

The analysis of mean and standart deviation of ANC of sepsis level are not significant. Need the research with more samples.

References

[1]. Angus DC, Linde-Zwirble WT, Lidicker J, Clermont G, Carcillo J, Pinsky MR (2001) Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. Crit Care Med 29:1303–1310

[2]. Vincent JL, Sakr Y, Sprung CL, Ranieri VM, Reinhart K, Gerlach H, Moreno R, Carlet J, Le Gall JR, Payen D (2006) Sepsis in European intensive care units: results of the SOAP study. Crit Care Med 34:344–353

[3]. Padkin A, Goldfrad C, Brady AR, Young D, Black N, Rowan K (2003) Epidemiology of severe sepsis occurring in the first 24hrs in intensive care units in England, Wales, and Northern Ireland. Crit Care Med 31:2332–2338

[4]. Rivers E, Nguyen B, Havstad S, Ressler J, Muzzin A, Knoblich B, Peterson E, Tomlanovich M (2001) Early goal-directed therapy in the treatment of severe sepsis and septic shock. N Engl J Med 345:1368–1377

[5]. Vincent JL, Bernard GR, Beale R, Doig C, Putensen C, Dhainaut JF, Artigas A, Fumagalli R, Macias W, Wright T, Wong K, Sundin DP, Turlo MA, Janes JM (2005) Drotrecogin alfa (activated) treatment in severe sepsis from the global open-label trial ENHANCE. Crit Care Med 33:2266–2277

[6]. Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, Suppes R, Feinstein D, Zanotti S, Taiberg L, Gurka D,Kumar A, Cheang M (2006) Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Crit Care Med 34:1589–1596

[7]. Chan YL, Tseng CP, Tsay PK, Chang SS, Chiu TF, Chen JC (2004) Procalcitonin as a marker of bacterial infection in the emergency department: an observational study. Crit Care 8:R12–R20

[8]. Levy MM, Fink MP, Marshall JC, Abraham E, Angus D, Cook D, Cohen J, Opal SM, Vincent JL, Ramsay G (2003) 2001 SCCM/ ESICM/ACCP/ATS/SIS international sepsis definitions conference. Intensive Care Med 29:530–538

[9]. Circiumaru B, Baldock G, Cohen J (1999) A prospective study of fever in the intensive care unit. Intensive Care Med 25:668–673

[10]. Zambon M, Ceola M, Castro R, Gullo A, Vincent JL (2008) Implementation of the Surviving Sepsis Campaign guidelines forsevere sepsis and septic shock: We could go faster. J Crit Care (inpress)

ISBN:978-602-99849-2-7