Statistical Downscaling Modeling with Kernel Ridge Regression for Rainfall Prediction in West Java

Muhammad Ihwal*, Aji H. Wigena*, Anik Djuraidah*

* Department of Statistics, IPB University

DOI: 10.29322/IJSRP.10.10.2020.p10638
http://dx.doi.org/10.29322/IJSRP.10.10.2020.p10638

Abstract- The Statistical downscaling (SD) is a technique of analysis using statistical model which can be used to predict local scale as response based on global scale data as predictors. The problem in SD is that the predictors are usually multicollinearity which can be overcome by ridge regression. Rainfall data are generally nonlinear, non-stationary and non-normal. A method that can be used to model the rainfall data is the kernel method. This study aims to apply SD using kernel ridge regression (RKR) for rainfall prediction and to compare the RKR and ridge regression (RR). The data used in this study are the monthly GCM (General Circulation Model) precipitation data with a 5×8 grid (2.5°×2.5° each grid) domain in 1981-2009 as predictor variables and the rainfall data at West Java in 1981-2009 as the response variable. The RKR model was better than RR model for rainfall prediction in West Java at one, two, and three years ahead. This fact was based on the RMSEP value of RKR was smaller than that of RR and the correlation of RKR was higher than that of RR.

Index Terms- statistical downscaling, kernel ridge regression, GCM, rainfall

I. INTRODUCTION

Rainfall greatly affects the activities of human life. The diversity is quite large and characterizes the climate in Indonesia. Global climate change can increase the incidence of extreme rainfall. Analysis is needed to obtain rainfall prediction information that is very useful to reduce the impact of possible extreme rain events.

Statistical downscaling (SD) is an analytical technique with a statistical model that can be used to predict local scale rainfall based on global scale data. Local scale data is used as response variable (rainfall data) with global scale data as predictor variable (Global Circulation Model (GCM) output precipitation data). In general, the basic idea of SD is to estimate the parameters of the relationship between global scale climate variables and local scale climate variables, which are then used for local scale climate predictions.

The problem that commonly occurs in SD is the GCM domain grid, which is a predictor variable containing multicollinearity. This problem can be overcome, among others, by ridge regression (RR). Ridge regression is used to overcome multicollinearity problems through modification of the least squares method (Neter, Waserman and Kutner 1990 in Herwindiati 1997).

Rainfall data as a response variable is generally nonlinear, non-stationary and non-normal, so a method that does not require these conditions is needed. In overcoming this problem, the kernel method is often used because the kernel density estimator is flexible, mathematically easy to work with and has a relatively fast convergence rate (Wahba 1975)

Several studies related to SD include Hadijanti (2016) which examined daily rainfall in Sembalun Station by modeling the kernel nonparametric regression with the reduction method using the classification and Regression Tree (CART) algorithm. The result of his research shows that the statistical downscaling model of the monthly rainfall of Sembalun Station with the resulting Kernel nonparametric regression has consistency in prediction.

Predictions of extreme rainfall have been carried out by Santri (2016) and Mulyati (2018). The result of the research by Santri is SD modeling with quantile regression using the Least Absolute Shrinkage and Selection Operator (LASSO) for rainfall data in Indramayu Regency which result in a predictive value that is more consistent with changes in time when compared to the main component regression model based on the RMSEP criteria. The result of Mulyati's research is that SD modeling using kernel quantile regression model with GCM-lag predictor and kernel quantile regression model using the main components of GCM-lag produce relatively the same and consistent predictions in predicting extreme rainfall.

Based on the consideration that ridge regression can overcome multicollinearity and the kernel method is a method that does not require linear, stationary and normal data properties, in this study, SD modeling with kernel ridge regression (RKR) is carried out in this study. RKR modeling in this study uses the Gaussian Radial Basis Function (RBF) kernel. This study aims to apply SD modeling with kernel ridge regression to predict rainfall in West Java province and to compare the kernel ridge regression model with the ridge regression model.

II. METHOD

A. Data

Data used in this study is the GCM monthly precipitation data released by the Climate Forecast System Reanalysis (CFSR) with GCM domains measuring 5×8 grids (2.5°×2.5° for each grid) in 1981-2009. Local rainfall data in West Java province is used as a response variable that was issued by the Meteorology, Climatology and Geophysics Agency (BMKG) in 1981-2009. Local rainfall data in West Java province is divided into 2 parts based on plain elevation, namely lowlands (0-200 masl) and highlands (> 200 masl).

B. Procedure of Analysis

Data analysis **is** performed using python 3 software. The stages in this study are as follows.

1. Data preparation stage
   a. Identifying predictor variables obtained from the GCM output data and response variables obtained from the BMKG website.
   b. Combining the predictor variables and response variables into new data.
   c. Grouping the data into 2 parts, namely the lowlands (0-200 masl) and the highlands (more than 200 masl).
   d. Dividing lowlands data and highlands data into 2 parts, namely modeling data (rainfall data in 1981-2008, 1981-2007, and 1981-2006) and validating data (rainfall data in 2009, 2008-2009, 2007-2009).
2. Ridge regression modeling stage
   The RR model is as follows:
   $$\hat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$
   Ridge regression modeling is carried out on lowlands data and upland data with the following stages:
   a. Choose the optimum lambda value with the smallest cross validation value.
   b. Ridge regression modeling on modeling data based on selecting the optimum lambda and sigma values..
3. The kernel ridge regression modeling stage (RKR)
   The RKR model is as follows:
   $$\hat{\boldsymbol{y}} = \boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}} + \lambda\boldsymbol{I})^{-\mathbf{1}}\boldsymbol{y}$$
   Ridge regression modeling is carried out on lowlands data and highlands data with the following stages:
   a. Transforming GCM output data by mapping $\phi : \mathbb{R}^p \rightarrow \mathbb{R}^q$
   b. Using the kernel method (Gaussian kernel radius basis function) to solve vector multiplication
   c. Choosing the optimum lambda value and sigma value with the smallest cross validation value.
   d. Kernel ridge regression modeling on modeling data based on selecting the optimum lambda and sigma values.
4. The model comparison stage
   Comparing kernel ridge regression and ridge regression based on the root mean squared error prediction (RMSEP) value with the formula
   $$RMSEP = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$
   and the correlation value between actual rainfall and predicted rainfall with formula
   $$r_{y_i,\hat{y}_i} = \frac{n\sum_{i=1}^{n} y_i\hat{y}_i - (\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} \hat{y}_i)}{\sqrt{\left[n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2\right]\left[n\sum_{i=1}^{n} \hat{y}_i^2 - (\sum_{i=1}^{n} \hat{y}_i)^2\right]}}$$
5. Model consistency test stage
   Test consistency of the model using RMSEP and Correlation

III. RESULT

A. Data Exploration

The highest average rainfall in the lowlands occurs in January, which is 274.56 mm/month with a minimum value of 118.58 and a maximum value of 415.67 and a standard eviation of 88.87. In the highlands, the highest average rainfall also occurs in January at 401.11 with a minimum value of 166.03 and a maximum value of 591.75 and a standard deviation of 108.62. This shows that January is the peak of the rainy season in the lowlands and highlands. This is inversely proportional to August, which has the lowest average rainfall on both plains.

Figure 2 shows that the box-line diagram forms the letter U. This shows that West Java province has a monsoon rain pattern. The monsoon rain pattern is a rain pattern that occurs in areas that have a clear difference between the rainy season and the dry season. Pribadi (2012) states that the rainy season has an average monthly rainfall intensity of greater than 150 mm/month. Therefore, November to April is the rainy season in the lowlands, and October to May is the rainy season in the highlands. The two plains have clear differences between the seasons.

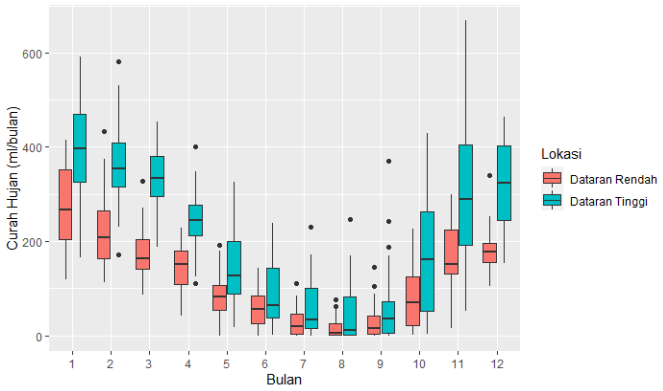


Figure 2 Box diagram of West Java rainfall

B. Kernel Ridge Regression

Kernel ridge regression is a combination of kernel method and ridge regression. The kernel function used in this study is the Gaussian Radius Basis Function (RBF) kernel function. The formation of the RKR model in the low and highlands uses the optimum λ and σ values. Determination of the optimum λ and σ is by selecting the minimum value from cross validation.

The optimum λ and σ values in the lowlands are 0.0001 and 150 for modeling data in 1981-2008, 0.0001 and 200 for modeling data in 1981-2007, 0.0001 and 150 for modeling in 1981-2006. The optimum λ and σ values are used to predict rainfall. Based on the prediction results, the obtained RMSEP value is quite small which is 25.08 for validating data in 2009, 40.86 for validating data in 2008-2009, and 45.71 for validating data in 2007-2009.. Moreover, correlation value obtained is very high which is 0.99 for validating data in 2009, 0.96 for validating data in 2008-2009, 0.95 for validating data in 2007-2009.

The optimum λ and σ values in the lowlands are 0.0001 and 100 for modeling data in 1981-2008, 0.0001 and 150 for modeling data in 1981-2007, 0.0001 and 200 for modeling data in 1981-2006. The optimum λ and σ values are used to predict rainfall. Based on the prediction results, the obtained RMSEP value is quite small which is 38.36 for validating data in 2009, 54.64 for validating data in 2008-2009, and 66.81 for validating data in 2007-2009.. Moreover, correlation value obtained is very high which is 0.98 for validating data in 2009, 0.96 for validating in 2008-2009, 0.93 for validating in 2007-2009. The small RMSEP value indicates that the predicted rainfall value of the RKR model has a small difference with the actual rainfall value, while the high correlation value indicates that the predicted rainfall pattern of the RKR model follows the actual rainfall pattern.

C. Comparison of RKR and RR

Rainfall prediction obtained by RKR will be compared with rainfall prediction obtained by RR. This comparison is done to see the prediction of a better model in terms of the smallest RMSEP value and the largest correlation. Table 1 shows the comparison of RKR and RR on the prediction of rainfall in the lowlands of West Java with validating data in 2009, 2008-2009, and 2007-2009.

Based on Table 1, it can be seen that the RMSEP value of the RKR model is smaller than the RR model. The RMSEP value of the RKR model is 25.08 for validating data in 2009, 40.86 for validating data in 2008-2009, and 45.71 for validating data in 2007-2009. The correlation value of the RKR model is greater than the RR model. This shows that for different validation data, the RKR model is better than the RR model for predicting rainfall in the lowlands of West Java.

Table 1 Comparison of RKR and RR models in lowlands

Modeling Data	Validating Data	RKR		RR	
		RMSEP	Correlation	RMSEP	Correlation
1981-2008	2009	84.29	0.97	25.08	0.99
1981-2007	2008-2009	84.71	0.88	40.86	0.96
1981-2006	2007-2009	93.39	0.93	45.71	0.95

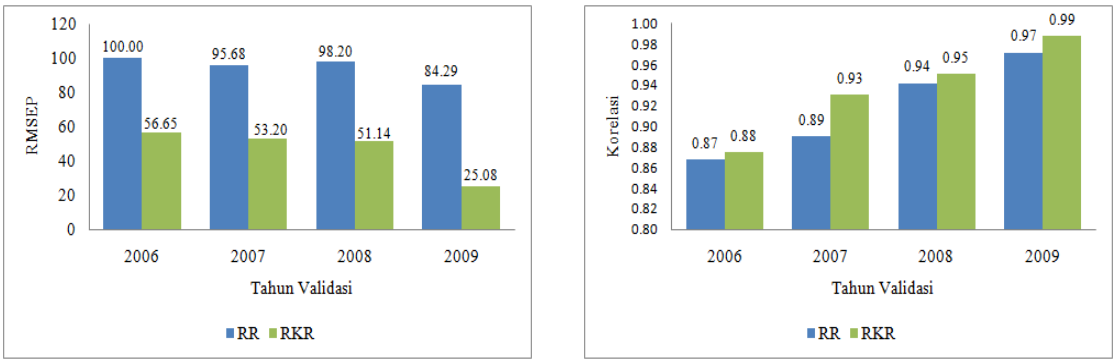
Table 2 shows the comparison of RKR and RR in the prediction of rainfall in the West Java highlands with the validating data in 2009, 2008-2009, and 2007-2009. The RMSEP value of the RKR model is smaller than the RR model. The RMSEP value of the RKR model is 38.36 for valudating data in 2009, 54.64 for validating in 2008-2009, and 66.81 for validating data in 2007-2009. The correlation value of the average RKR model is greater than the RR model. This shows that in different validation data, the RKR model is better than the RR model for predicting rainfall in the highlands of West Java.

Table 2 Comparison of RKR and RR models in highlands

Modeling Data	Validating Data	RKR		RR	
		RMSEP	Correlation	RMSEP	Correlation
1981-2008	2009	121.65	0.98	38.36	0.98
1981-2007	2008-2009	133.62	0.96	54.64	0.96
1981-2006	2007-2009	137.29	0.85	66.81	0.93

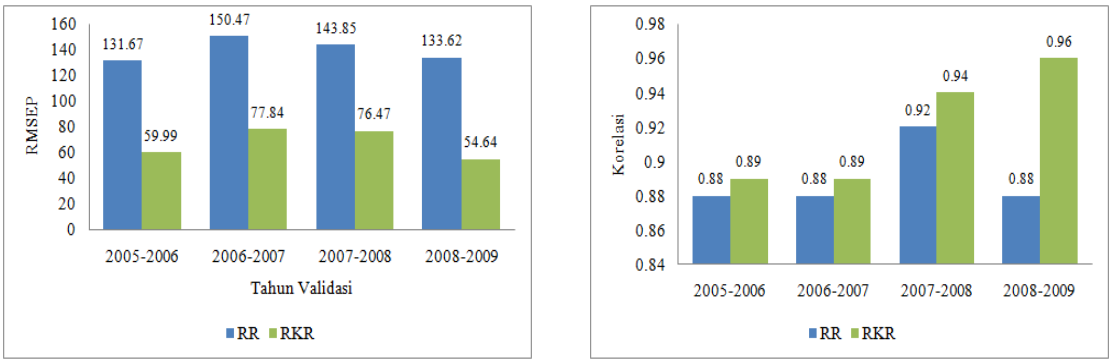
D. Model Validation and Consistency

The RKR model and RR model will determine which model is better in predicting rainfall. Figure 3 and Figure 4 show the RMSEP value and correlation for the RKR and RR models for prediction within one-year in the lowlands and highlands of West Java. Based on Figure 3 and Figure 4, the RMSEP value obtained by RKR is consistently smaller than the RR model in different validating data. Based on Figure 3, it shows that the correlation obtained by the RKR model is consistently greater than the RR model. Figure 4 shows that the correlation values obtained by the RKR and RR models are not significantly different. This shows that the RKR model is consistently better than the RR model for predicting rainfall within 1 year in West Java.



(a) (b)

Figure 3 Bar charts of (a) RMSEP and (b) Correlations for the RKR and RR models in the lowlands of West Java for one-year prediction



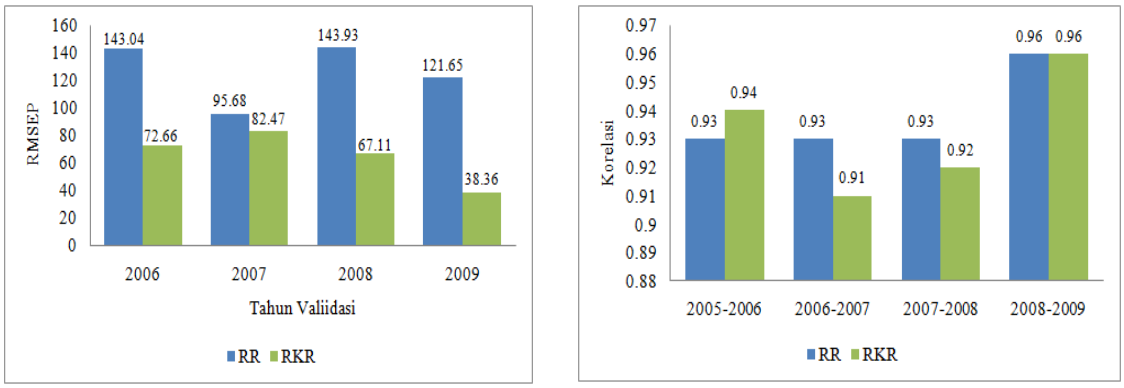
(a) (b)

Figure 4 Bar charts of (a) RMSEP and (b) Correlations for the RKR and RR models in the highlands of West Java for one-year prediction

Figure 5 and Figure 6 show the RMSEP values and correlations for the RKR and RR Models for predictions over a two-year period in the lowlands and highlands of West Java. Based on Figure 5 and Figure 6, the RMSEP value obtained by RKR is consistently smaller than the RR model on different validation data. Figure 5 shows that the correlation obtained by the average RKR model is greater than that of the RR model. Figure 6 shows that the correlation values obtained by the RKR model and RR model are not significantly different. This shows that the RKR model is consistently better than the RR model for predicting rainfall within 2 years in West Java.

(a) (b)

Figure 5 Bar charts of (a) RMSEP and (b) Correlations for the RKR and RR models in the lowlands of West Java for two-year prediction



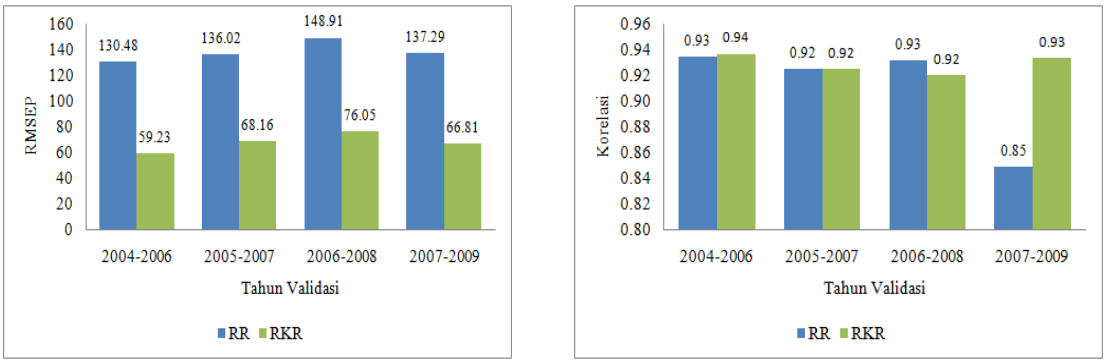
(a) (b)

Figure 5 Bar charts of (a) RMSEP and (b) Correlations for the RKR and RR models in the highlands of West Java for two-year prediction

Figure 7 and Figure 8 show the RMSEP values and correlations for the RKR and RR Models for predictions over a period of 3 years in the lowlands and highlands of West Java. Based on Figure 7 and Figure 8, the RMSEP value obtained by RKR is consistently smaller than the RR model on different validation data. Based on Figure 7, it shows that the correlation obtained by the RKR model is consistently greater than the RR model. Figure 8 shows that the correlation value obtained by the average RKR model is greater than the RR model. This shows that the RKR model is consistently better than the RR model for predicting rainfall within 3 years in West Java.

(a) (b)

Figure 7 Bar chart of (a) RMSEP and (b) Correlation for the RKR and RR models in the lowlands of West Java for three-year prediction



(a) (b)

Figure 8 Bar chart of (a) RMSEP and (b) Correlation for the RKR and RR models in the highlands of West Java for three-year prediction

The consistency of the RKR model can also be measured from the standard deviation of the correlation value in the modeling data and validation data at different times. The smaller the standard deviation, the more consistent the model is (Wigena 2006). The standard deviation obtained by the RKR model for one-year predictions is 0.05 in the lowlands and 0.03 in the highlands. The standard deviation obtained by the RKR model for the two-year prediction is 0.02 in the lowlands and 0.02 in the highlands. The standard deviation obtained by the RKR model for the three-year prediction is 0.02 in the lowlands and 0.04 in the highlands. The standard deviation value obtained is very small. This indicates that the RKR model is quite consistent in predicting rainfall in West Java for the next one year, two years and three years.

IV. CONCLUSION

The RKR model is better than the RR model for predicting rainfall in West Java for the next one, two, and three years. This is based on the RMSEP value of the RKR model which is smaller than the RMSEP model of the RR and the correlation value of the RKR model which is greater than the correlation of the RR model.

REFERENCES

[1] M. Hadijati, "Statistical downscaling Regresi Nonparametrik kernel untuk prediksi Curah Hujan Bulanan Stasiun Sembalun, *Seminar Nasional Matematika II – Bali*, ISSN: 2406-9868, 2016.

[2] D. E. Herwindiati, "Pengkajian Regresi Komponen Utama, Regresi Ridge, dan Regresi Kuadrat Terkecil Parsial untuk Mengatasi Kolinearitas," *Thesis*, 1997.

[3] A. E. Mulyati, "Pemodelan Statistical Downscaling dengan Regresi Kuantil Kernel untuk Menduga Curah Hujan Ekstrim Di Indramayu," *Thesis*, 2018.

[4] D. Santri, "Pemodelan Statistical Downscaling dengan Regresi Kuantil Menggunakan Lasso untuk Pendugaan Curah Hujan Ekstrim," *Thesis*, 2016.

[5] G. Wahba, "Optimal convergence properties of variable knot, kernel, and orthogonal series methods for density estimation," *Annals of Statistics*, 3(1): 15–29, 1975.

[6] A. H. Wigena, "Pemodelan Statistical Downscaling dengan Regresi Projection Pursuit untuk Prediksi Curah Hujan Bulanan," *Disertasi*, 2006.

[7] F.B. Cruz, O. Bousquet, "Kernel Methods and Their Potential Use in Signal Processing," *IEEE Signal Processing Magazine*, *21*(3); 57-65, 2004.

[8] P. Exterkate, P. J. F. Groenen, Heij, D. J. C. Van Dijk, "Nonlinear Foreceasting with Many Predictors using Kernel Ridge Regression," *Tinbergen Instituut Discussion*, pp. 11-007/4, 2011.

[9] T. Hastie, R Tibshirani, J. Friendman. *The Elements od Statistical Learning Data Mining, Inference, and Prediction*. New York (US): Springer, 2009.

[10] A. E. Hoerl, R. W. Kennard, "Ridge Regression: Biased Estimation for Nonortogonal Problems," *Technometrics*, pp. 12: 55-67, 1970.

[11] D. C. Montgomery, E. A. Peck. Introduction to Linier Regression Analysis. New York (US): John Wiley & Sons, 1992.

[12] H. Y. Pribadi, "Variabilitas Curah Hujan dan Pergeseran Musim di Wilayah Banten Sehubungan dengan Variasi Suhu Muka Laut Perairan Indonesia, Samudra Pasifik dan Samudra Hindia" *Thesis*, 2012.

[13] Sutikno, "Statistical downscaling luaran GCM dan pemanfaatannya untuk prediksi produksi padi," *Disertasi*, 2008.

[14] I. Takeuchi, Q. V. Le, T. Sears, A. J. Smola. 2006. Nonparametric Quantile Estimation. *Journal of Machine Learning Research* 7: 1231-1264.

[15] J. Vert, K. Tsuda, B. Scholkopf. *Kernel Method in Computational Biology*. MIT Press: Massachusetts, 2004

AUTHORS

First Author – Muhammad Ihwal, S.Pd, college student, Departement of Statistics, Faculty of Mathematics and Natural Sciences (FMIPA), IPB University, Bogor,16680, Indonesia, and ihwal_muhammad@apps.ipb.ac.id
Second Author – Prof. Dr. Ir. Aji Hamim Wigena, M.Sc, Lecturer, Departement of Statistics, Faculty of Mathematics and Natural Sciences (FMIPA), IPB University, Bogor,16680, Indonesia, and aji_hw@apps.ipb.ac.id

Third Author – Dr. Ir. Anik Djuraidah, MS, Lecturer, Departement of Statistics, Faculty of Mathematics and Natural Sciences (FMIPA), IPB University, Bogor,16680, Indonesia, and anikdjuraidah@apps.ipb.ac.id

Correspondence Author – Prof. Dr. Ir. Aji Hamim Wigena, M.Sc, Lecturer, Departement of Statistics, Faculty of Mathematics and Natural Sciences (FMIPA), IPB University, Bogor,16680, Indonesia, and aji_hw@apps.ipb.ac.id