Mixed Geographically and Temporally Weighted Regression with Cluster in West Java's Poverty Cases

I Made Sumertajaya, Muhammad Nur Aidi, Winda Nurpadilah

Department of Statistics, IPB University

DOI: 10.29322/IJSRP.10.12.2020.p10865
http://dx.doi.org/10.29322/IJSRP.10.12.2020.p10865

Abstract- Geographically and temporally weighted regression (GTWR) is a method used when there is spatial and temporal diversity in an observation. GTWR model just consider the local influences of spatial-temporal independent variables on dependent variable. In some cases, the model is not only about local influences but there are the global influences of spatial-temporal variables too, so that mixed geographically and temporally weighted regression (MGTWR) model is more suitable to use. The smaller the regional approach used, the better the resulting solution. This study examines the effects of clustering on MGTWR modeling. The cluster approach is used to reduce the modeling area and make the objects in the same clusters are more similar to each other than objects in other clusters. This study aimed to determine the model to be used in West Java's poverty cases in 2012 to 2018. The result showed that the GRDP, the percentage of literacy, the percentage of expenditure per capita on food, and health index are global variables. Whereas the variable expected years of schooling and households buying rice for poor (*raskin*) are local variables. Furthermore, based on Root Mean Square Error (RMSE), Akaike Information Criterion (AIC), and R^2 was showed that MGTWR with cluster (MGTWRC) better than MGTWR when it used in West Java's poverty cases.

Index Terms- Cluster, Global and Local, MGTWR, Poverty, West Java

I. INTRODUCTION

Spatial regression is a regression model that considers the spatial effects of the data. Some spatial regression methods have been developed, such as the general spatial model (GSM), the spatial autoregressive model (SAR), the spatial error model (SEM) and the geographically weighted regression model (GWR). According to Anselin and Getis (1992), spatial effect among locations can be caused by spatial dependence and spatial heterogeneity. Spatial dependence is defined as the dependence between a location and its surrounding. Spatial heterogeneity, meanwhile, is the variation caused by the differences in the effect of explanatory variables on the response among different locations. One of the spatial regression approaches used to solve the issue of spatial heterogeneity is geographically weighted regression (GWR) (Fotheringham et al. 2002). For each observation, the GWR model produces local parameter estimates (Purhadi and Yasin 2012).

In its analysis, GWR model only considers spatial effect of the data without taking into account the effect of time. In fact, some variables are not only observed based on its geographical location, but also based on its time. Temporal data can be used to provide information about the dynamics of spatial processes and to allow more relevant parameter estimates. Geographically and temporally weighted regression (GTWR) is the advancement of the GWR model that takes into account the effect of location and time. For each location and time, the GTWR model generates a local model (Huang et al. 2010). GTWR modeling utilizes a weight matrix W whose magnitude depends on the proximity of the location and time. The closer the location and time, the greater the weight is.

Not all explanatory variables used in GWR and GTWR have a spatial effect. The effects of certain explanatory variables are global and those of others are local. Therefore, the GWR model was developed into the mixed geographically weighted regression (MGWR) (Fotheringham et al. 2002). The MGWR model produces parameter estimates that are partly global and partly local according to the location. It also applies to the GTWR model that can be developed into the mixed geographically and temporally weighted regression model (MGTWR).

Poverty is one of the fundamental problems that has become the subject of government attention, especially in Indonesia. The government has made numerous attempts to minimize poverty rates. One of those is to recognize the factors that influence Indonesia's poverty rates. Moreover, the government also plays a role in classifying the needs of each regions according to their characteristics, especially in areas with high poverty rates. From year to year, the rate of poverty is changing. Therefore, it is important to study the poverty pattern in each region and to study the clusters of poverty year by year (Dewi 2011). Cluster analysis and regression analysis involving spatial and temporal aspects are quite relevant to be used in this case.

According to Dhiyaa'ulhaq (2017), the smaller the regional approach used, the better the resulting solution will be. Clustering, in this study, was used to decrease the modeling area and make the objects in the data have a high degree of similarity.

A research conducted by Nuramaliyah (2019) noted that the MGTWR model was a better model than the GTWR model for estimating the percentage of poverty in the province of North Sumatra. The analysis by Andrytiarandy (2017) examined the methods of geographically weighted regression fuzzy cluster (GWRFC) and GWR. In overcoming problems in malaria prevalence cases, the

This publication is licensed under Creative Commons Attribution CC BY.

GWRFC method was better than GWR method. Therefore, the effect of clustering on MGTWR model was discussed in this study. Modeling in a smaller area, such as modeling on the clustered data, was expected to produce a better model. The best model would be selected based on the comparison of the RMSE, AIC, and R^2 values.

II. MATERIALS AND METHOD

2. 1 DATA

The data used in this analysis was secondary data from the Central Bureau of Statistics 2012-2018 publication (BPS, 2018a, 2018b, 2018c). The data consisted of percentage of poverty as the response variable and six explanatory variables that were thought to affect the percentage of poverty. The selection of explanatory variables was based on previous studies. The variables used in this study are presented in Table 1. The districts and cities in the West Java Province, which consists of 17 districts and 9 cities, were the scope of this study.

Table 1: The variables used this the study

Variables	Note
Y	Poverty rate (%)
X_1	Gross regional domestic product (Billion Rupiah)
X_2	Literacy rate
X_3	Expenditure per capita on food (%)
X_4	Households buying rice for poor (%)
X_5	Expected years of schooling (Years)
X_6	Health index

2. 2 ANALYSIS PROCEDURE

Steps of data analysis using R Studio are carried out as follows:

1. Explore the percentage of poverty in West Java in 2012–2018 using descriptive statistics.
2. Calculate pearson correlation coefficient to evaluate the relationship pattern of the explanatory variables that are thought to affect the response variable.
3. Conducting global regression modeling which includes:
   a. Parameter estimation
   b. Simultaneous and partial testing of regression parameters
   c. Assumption testing
4. Performing multicollinearity test between explanatory variables using VIF value. If VIF > 10, then multicollinearity was present between variables.
5. Identifying the spatial and time heterogeneity in the data using Breusch-Pagan (BP) test. The hypothesis used was as follows:
   H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_p^2 = \sigma = 0$ (homoscedastic disturbances)
   H_1: At least there is one i where $\sigma_i^2 \neq 0$ (heteroscedastic disturbances)
   Breusch and Pagan (1979) formulated the heterogeneity test statistics for cross-section data by deriving the lagrange multiplier (LM) formula. The Breusch-Pagan test statistics are:
   $$BP = \frac{1}{2}\boldsymbol{f}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{f} \sim \chi_p^2 \text{ with } f_i = \left(\frac{\varepsilon_i^2}{\sigma^2} - 1\right)$$
   Test criteria: $BP = \begin{cases} \leq \chi^2_{(p-1)}, \text{ accept } H_0 \\ > \chi^2_{(p-1)}, \text{ reject } H_0 \end{cases}$
6. Performing cluster analysis.
   a. Conducting cluster analysis with k-means method based on location coordinates for each year.
   b. Using elbow method to determine the best number of clusters.
7. Performing MGTWR modeling on the overall data and the clustered data which includes the following stages:
   a. Determine global and local variables using BP test
   b. Calculate spatial-temporal distance (d_{ij}^{ST}) on coordinates (u_i, v_i, t_i).
   c. Determine parameter estimates of spatial-temporal ratio (τ), spatial (λ), temporal (μ), and bandwidth of spatial-temporal distance (h_{ST}) using the minimum CV.
   d. Calculate weight matrix

This publication is licensed under Creative Commons Attribution CC BY.

e. Calculate the esimated value of global and local parameter.
f. Test the global and local parameters model.

8. Comparing the goodness of the model based on the values of RMSE, AIC, and R^2.
9. Interprete the results by making a map based on explanatory variables that significantly affected the percentage of poverty in West Java's Province.

III. RESULTS

3.1 DATA EXPLORATION

Poverty is characterized as an economic inability to meet basic food and non-food needs, as measured in terms of expenditure, according to the Central Bureau of Statistics (2018). The percentage of poverty is one of the indicators that can be used to define the state of poverty in a region. Figure 1 shows the distribution map of poverty in West Java during 2012-2018.

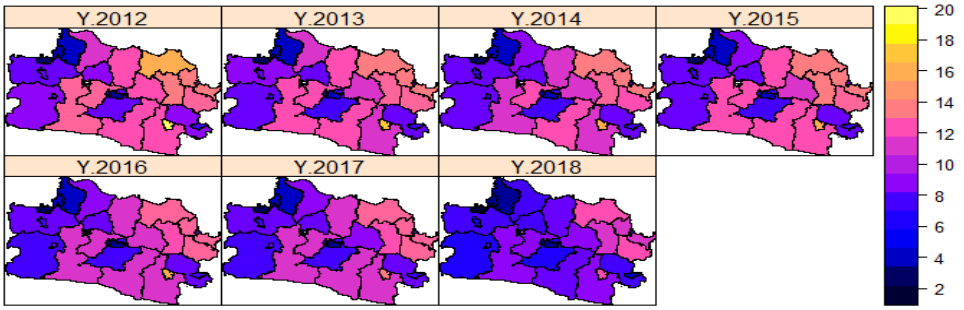

Figure 1: Distribution map of the West Java's poverty rate in 2012-2018

There are 26 regencies/cities consisting of 17 districts and 9 cities in West Java. In terms of location, each region in the same year had different percentages of poverty. Figure 1 indicates that there was a decline in the percentage of poverty that existed in many regions in the period from 2012 to 2018. It can be seen from the increasing number of blue and purple areas on the map. Figure 1 indicates that regions with about the same percentage of poverty tend to be in groups or have close proximity. This shows that the proximity between regions could affect poverty. Poverty in a region was affected not only by the factors in that particular region, but also by factors in other regions.

The linear relationship between each explanatory variable and the response variable was measured using pearson correlation coefficient. The correlation coefficient of each explanatory variable and the response variable are shown in Table 2. X_3 and X_4 had a positive linear relationship with the Y variable. This meant that the higher the percentage of expenditure per capita on food (X_3) and the percentage of households that have bought *raskin* (X_4) in a region in West Java, the higher the percentage of poverty in that region. X_1, X_2, X_5 and X_6 had a negative linear relationship with the Y variable. This meant that the higher the gross regional domestic product (X_1), the literacy rate (X_2), the expected number of years of schooling (X_5), and the health index (X_6) in a region in West Java, the lower the percentage of poverty in the region.

Table 2: The correlation coefficient

Varibles	Pearson correlation
Y and X_1	-0.3921
Y and X_2	-0.2689
Y and X_3	0.6752
Y and X_4	0.5911
Y and X_5	-0.5183
Y and X_6	-0.6030

One of the statistical tests used to see an indication of a risky linear correlation between explanatory variables is multicollinearity test. Multicollinearity test between variables was carried out based on the VIF (Variance Inflation Factor) value. Larger VIF value indicates a more risky linear correlation between explanatory variables. If the corresponding VIF value is greater than ten (VIF> 10), the explanatory variables are multicollinear. Table 3 shows that all the explanatory variables' VIF values were below 10, so it was presumed that there was no multicollinearity between the explanatory variables. This indicated that the linear association between explanatory variables had a low risk.

Table 3: The explanatory variables VIF values

Variables	X_1	X_2	X_3	X_4	X_5	X_6

This publication is licensed under Creative Commons Attribution CC BY.

VIF	1.2603	1.2556	2.5252	1.6217	1.8605	2.3404

3.2 SPATIAL HETEROGENEITY ANALYSIS

Testing for spatial heteroscedastic using the Breusch-Pagan (BP) test. This test was carried out on 26 regions in West Java in the period of 2012 to 2018 with an error rate of 10%. The results shown in Table 4 show that simultaneous test of the data in the period of 2012 to 2018 had p-values that was less than the error rate, so it could be inferred that the percentage of poverty in each area in West Java was spatially heterogeneous. Thus, the spatial regression method using geographically weighted regression is more appropriate for the poverty rate in West Java.

Table 4: Breusch-Pagan test

Year	Breusch-Pagan statistic value	P-value
2012-2018	12.1152	0.0594*

Note: *) significant at $\alpha = 10\%$

3.3 CLUSTER ANALYSIS

Elbow method was used to find the optimal k value in cluster analysis. Figure 2 shows the plot of the elbow method. The x-axis is the number of clusters and the y-axis is the sum of squares within the cluster. The optimum number of clusters was four since the plot shows that in the 4th cluster there was no longer a dramatic decline in the sum of squares within the cluster.

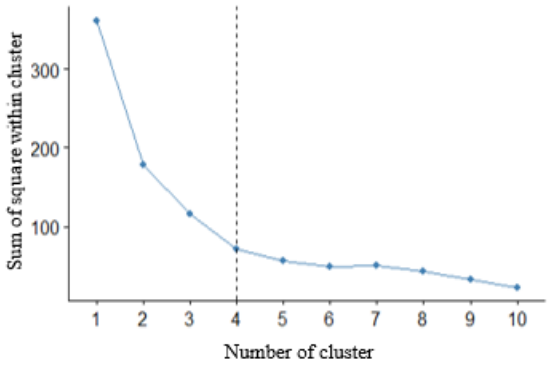

Figure 1: Elbow method based on the location coordinates

The results of clustering based on the location coordinates can be seen in Table 4. Based on the table, it can be seen that cluster 1 and cluster 2 each consisted of 5 regions, while cluster 3 and cluster 4 each consisted of 8 regions. The size of clustered data used in modeling was equal to the number of cluster members times the number of time.

Table 5: The results of clustering based on the location coordinates

Cluster	Cluster members	Number of members
1	Purwakarta, Karawang, Bekasi, Bekasi City, Depok City	5
2	Bogor, Sukabumi, Cianjur, Bogor City, Sukabumi City	5
3	Tasikmalaya, Ciamis, Kuningan, Cirebon, Majalengka, Cirebon City, Tasikmalaya City, Banjar City	8
4	Bandung, Garut, Sumedang, Indramayu, Subang, Bandung Barat, Bandung City, Cimahi City	8

3.4 MIXED GEOGRAPHICALLY AND TEMPORALLY WEIGHTED REGRESSION (MGTWR)

The model produced a total of 26 locations × 7 times = 182 models. This implied that each location had 7 models that were differentiated by year, starting from 2012 to 2018. The MGTWR model consisted of two kinds of variables, namely global and local variables. The determination of the variables was based on Breusch-Pagan test between the response variable (Y) and all the explanatory variables (X). Table 6 reveals that the variables X_1, X_2, X_3, and X_6 had p-values that were more than the error rate, so it could be inferred that these variables were global variables. The variables X_4 and X_5 respectively, had p-values that were smaller than the error rate, so that these variables were local variables.

Table 6: The Breusch-Pagan test for the determination of variables

Peubah	BP test value	P-value	Note
Y and X_1	1.5246	0.2169	Global
Y and X_2	1.8023	0.1794	Global

This publication is licensed under Creative Commons Attribution CC BY.

Y and X_3	0.9599	0.3272	Global
Y and X_4	2.8246	0.0928*	Lokal
Y and X_5	3.1404	0.0764*	Lokal
Y and X_6	0.2751	0.5999	Global

Note: *) significant at α = 10%

A weight matrix constructed from a euclidean distance matrix using the best kernel function was needed for MGTWR modeling. Exponential kernel function had the smallest Cross-Validation (CV) value, so it could be concluded that the exponential kernel function was the best kernel function to be used in the weight matrix. The distance matrix in the modeling used the interaction between spatial distance and temporal distance. In order to resolve the unit difference between spatial distance, temporal distance, and spatial-temporal distance, a balancing parameter between spatial distance and temporal distance was required. The balancing parameters used were spatial distance (λ), temporal distance (μ), and spatial-temporal distance ratio (τ) parameters. These parameters were obtained based on the smallest CV value. Based on the exponential kernel function, the parameter values for each data can be seen in Table 7.

Table 7: The parameters value of each data

	h_s	τ	λ	μ	h_{ST}
Overall	0.057330	0.000541	1.763584	0.000954	0.013410
Cluster 1	0.039304	0.000191	0.716689	0.000137	0.031625
Cluster 2	0.201451	0.003086	0.985350	0.002937	0.064279
Cluster 3	0.055421	0.000253	1.051998	0.000267	0.013865
Cluster 4	0.134703	0.000345	1.243425	0.000429	0.122321

Table 8 shows the results of MGTWR model's global parameters estimation. The parameter estimates of the gross regional domestic product (GRDP) (X_1) of the overall data and each clustered data, were negatives. This indicated that if the GRDP in a region increased, the percentage of poverty would decrease. The estimated literacy rate (X_2) and health index (X_6) parameters in the cluster 1 data were positive, while those from the other clusters were negative. This illustrated that if the literacy rate and the health index in the Cluster 1 regions increased, the percentage of poverty would also increase. Meanwhile, in the other clusters, if the literacy rate and the health index in a region increased, the percentage of poverty in that region would decrease. The estimated parameter of the percentage of expenditure per capita on food (X_3) was positive in the overall data. This meant that if the percentage of expenditure per capita in region increased, the percentage of poverty would also increase.

Table 8: Global estimated parameter of the MGTWR model of each data

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_6$
Overall	-0.0746	-0.0648	0.0338	-1.7548
Cluster 1	-0.0858	0.1359	-0.0042	0.0981
Cluster 2	-0.1657	-0.0235	-0.0616	-0.2786
Cluster 3	-0.1019	-0.0502	-0.0116	-0.9233
Cluster 4	-0.0217	-0.0141	-0.0807	-2.1259

Table 9 shows the summary of the estimated local parameters of the MGTWR model in the overall data. $\hat{\beta}_4$ and $\hat{\beta}_5$ were the estimated parameters for the percentage of households buying *raskin* (X_4) and the expected number of years of schooling (X_5), each of which were local variables. Based on the results, X_4 and X_5 had a significant effect on the percentage of poverty among districts and cities in West Java

Table 9: Summary of Local estimated parameter in MGTWR model in overall data

Parameter	Minimum	Maximum	Average	Standard Deviation
$\hat{\beta}_0$	31.2100	682.3433	160.8183	67.5522
$\hat{\beta}_4$	-0.1502	0.0742	-0.0249	0.0352
$\hat{\beta}_5$	-45.2483	9.6022	-0.3153	5.5053

Based on Figure 3 and Table 9, the estimated parameter for the percentage of households buying *raskin* (X_4) and the expected number of years of schooling (X_5) had different values. In the overall data and in each cluster data, the estimated parameters for the percentage of households buying *raskin* (X_4) had a negative mean. The estimated parameters for the expected number of years of schooling (X_5) had a positive mean in cluster 2, while the other clusters had a negative mean.

This publication is licensed under Creative Commons Attribution CC BY.

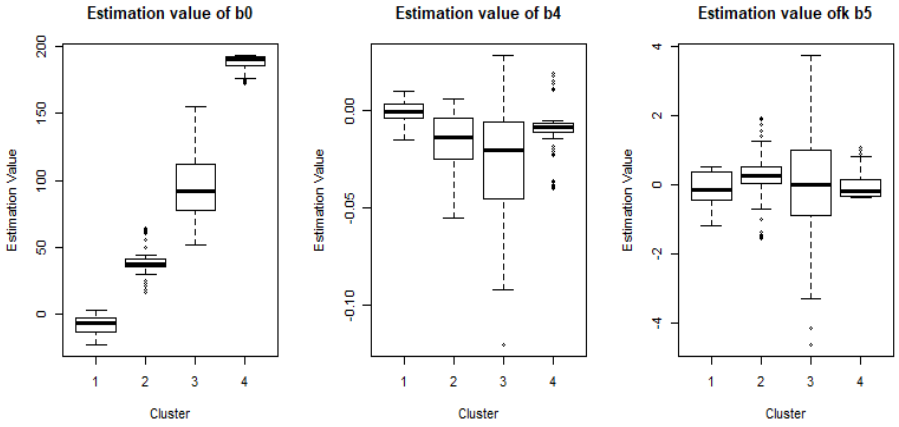

Figure 2: Boxplot of estimated local parameters of the MGTWR model in cluster data

Partial test for the global and local variables was carried out using t-test. The partial test on the global variables with an error rate (α) of 5% indicated that the gross regional domestic product (GRDP) (X_1) had a significant effect on the percentage of poverty in the 1st, 2nd, and 4th clusters. In each clusters, the literacy rate (X_2) had no significant effect on the percentage of poverty. The percentage of expenditure per capita on food (X_3) and health index (X_6) only had a significant effect on the percentage of the poverty in the cluster 4 data. The results of partial test of the local variables in the clustered data are presented in the form of distribution maps which can be seen in Figure 4.

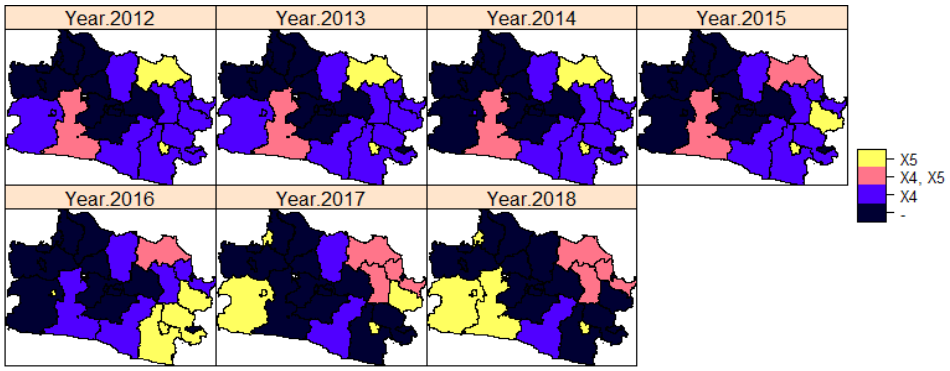

Figure 3: The distribution map of significant local varibles in MGTWRC

3.5 COMPARISON GOODNESS OF MODEL

Based on Table 10, it can be seen that the MGTWRC model had the smallest RMSE value. The small RMSE value in the model indicated that the variance of the MGTWRC model was close to the variance in the observed data. The MGTWRC model also had a lower AIC value as compared to the MGTWR model. In other words, the MGTWRC model was more precise than the former MGTWR model.

Table 10: The comparison of goodness of model

Model	RMSE	AIC	R^2
MGTWR	0.3187	1156.9910	0.9952
MGTWRC	0.2464	810.2995	0.9923

The MGTWR with cluster model had an R^2 value of 0.9923. Based on the value of R^2, the model's explanatory variables could explain the response variable by 99.23%, while other factors that were not included in the model explained the rest. Based on the resulting RMSE, AIC, and R^2 values, it could be concluded that the MGTWRC model was better than the other models.

This publication is licensed under Creative Commons Attribution CC BY.

IV. CONCLUSION

The percentage of poverty varied spatially and temporally so that geographically and temporally weighted regression (GTWR) modeling could be carried out. The gross regional domestic product (GRDP) (X_1), literacy rate (X_2), percentage of per capita expenditure on food (X_3), and health index (X_6) had a global effect. Meanwhile, the percentage of households that bought *raskin* (X_4) and the expected number of years of schooling (X_5) had a local effect.

The MGTWR with cluster (MGTWRC) model was a better model than the former MGTWR model for estimating the percentage of poverty in West Java. The smaller the regional approach used, the better the resulting solution. In this study, clustering was used to decrease the modeling area and make the objects in the data have a high degree of similarity. Based on the resulting RMSE, AIC, and R^2 values, it could be concluded that the MGTWR with cluster model was better than the other models. Therefore, the MGTWR with cluster model was the best model in modeling the percentage of poverty in West Java in 2012-2018.

The partial test on the global variables showed that the gross regional domestic product (GRDP) (X_1) had a significant effect on the percentage of poverty in the 1st, 2nd, and 4th clusters. Overall, the literacy rate (X_2) had no significant effect on the percentage of poverty. The percentage of expenditure per capita on food (X_3) and health index (X_6) only had a significant effect on the percentage of the poverty in the cluster 4 data. In addition, the partial test on the local variables showed that the effect of the percentage of households buying *raskin* (X_4) and the expected number of years of schooling (X_5) on the percentage of poverty varied in each cluster in West Java.

ACKNOWLEDGMENT

This work is fully supported by Kemenristek DIKTI (Kementerian Riset Teknologi dan Pendidikan Tinggi) of Indonesia.

REFERENCES

Andrytiarandy, W. 2017. Analisis regresi terboboti *fuzzy cluster* geografis pada prevalensi malaria di Indonesia [tesis]. Bogor: Institut Pertanian Bogor.

Anselin, L., Getis, A. 1992. Spatial statistical analysis and geographic information systems. *The Annals of Regional Science* **26**(1):19-33.

[BPS] Badan Pusat Statistika. 2018. *Statistik Indonesia 2018*. Jakarta: BPS.

[BPS] Badan Pusat Statistika Jawa Barat. 2019. *Provinsi Jawa Barat Dalam Angka*. Bandung: BPS.

[BPS] Badan Pusat Statistika Jawa Barat. 2019. *Kemiskinan Kabupaten/Kota Di Jawa Barat*. Bandung: BPS.

Dhiyaa'ulhaq, M. 2017. Analisis penyebaran kemiskinan dan pengaruh industri mikro dan kecil terhadap kemiskinan di Daerah Istimewa Yogyakarta [tesis]. Bogor: Institut Pertanian Bogor.

Dewi, AN. 2011. Penggerombolan dan identifikasi trend kabupaten di Indonesia berdasarkan data dan informasi kemiskinan tahun 2002-2009 [skripsi]. Bogor: Institut Pertanian Bogor.

Fotheringham, AS., Brunsdon, C., Charlton, M. 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester: Wiley.

Huang, B., Wu, B., Barry, M. 2010. Geographically and temporally weighted regression for modeling spasio-temporal variation in house prices. *International Journal of Geographical Information Science* **24**(3):383–401.

Nuramaliyah. 2019. Pengaruh peubah gabungan global dan lokal pada model regresi terboboti geografis dan temporal [tesis]. Bogor: Institut Pertanian Bogor.

Purhadi., Yasin, H. 2012. Mixed geographically weighted regression model (case study: the percentage of poor households in Mojokerto 2008). *European Journal of Scientific Research*. **69**(2): 188-196.

AUTHORS

First Author – Dr. Ir. I Made Sumertajaya, M.Si, Lecturer, Departement of Statistics, Faculty of Mathematics and Natural Sciences (FMIPA), IPB University, Bogor,16680, Indonesia, Email: imsjaya.stk@gmail.com

Second Author – Prof. Dr. Ir. Muhammad Nur Aidi, MS, Lecturer, Departement of Statistics, Faculty of Mathematics and Natural Sciences (FMIPA), IPB University, Bogor,16680, Indonesia, Email: nuraidi18081960@gmail.com

Third Author – Winda Nurpadilah, S.Stat, college student, Department of Statistics, Faculty of Mathematics and Natural Sciences (FMIPA), IPB University, Bogor,16680, Indonesia, Email: windanurfadilah17@gmail.com

Correspondence Author – Winda Nurpadilah, S.Stat, college student, Department of Statistics, Faculty of Mathematics and Natural Sciences (FMIPA), IPB University, Bogor,16680, Indonesia, Email: windanurfadilah17@gmail.com

This publication is licensed under Creative Commons Attribution CC BY.