International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Estimating the Parameters of a Robust Geographically Weighted Regression Model in Gross Regional Domestics Product in East Java

Bayutama Isnaini[a]*, Utami Dyah Syafitri[b], Muhammad Nur Aidi[c]

[a,b,c]*Department of Statistics, IPB University, Bogor, 16680, Indonesia*

[a]*Email: baytma2103@gmail.com*

[b]*Email: utamids@gmail.com*

[c]*Email: nuraidi18081960@gmail.com*

Abstract

Geographically weighted regression (GWR) is a regression parameter estimation method that accommodates location elements. Estimates of regression parameters have problems when there are outliers in the modelled data, including data based on location. This problem can be handled by a robust method of outliers, the robust GWR method (RGWR). M-estimator and S-estimator have high efficiency and high breakdown points. This study aimed to determine the best regression parameter estimation model on gross regional domestic product (GRDP) data in East Java Province in 2015, which is indicated to have various value based on the characteristic of regency/city. The city of Surabaya has very different characteristics from other regions and is detected as outliers based on a GWR model error plot, so RGWR with M-estimator and S-estimator are used. The mean absolute deviation (MAD) show that the best model for data in this study is the RGWR with M-estimator.

Keywords: Outliers; M-estimator RGWR; S-estimator RGWR.

1. Introduction

GWR is a method used to handle the spatial diversity of a regression relationship [1]. GWR illustrates that each location has a different model. This difference is caused by spatial weighting in estimating the parameters.

---

* Corresponding author.

The observations are spatially close to the predicted observation, given more weight than the ones located far away [2]. In some cases, including data that have spatial diversity, there are observations that are far from observations in general, called outliers. Outlier can affect and does not affect the estimation of parameters. Detection of outliers in spatial data is very difficult because in the GWR model, estimating parameters is local for each location, so outliers contained in spatial data are called local outliers [1]. The use of linear regression (ordinary least squares, OLS) on observations containing outliers can cause inaccurate parameter estimation. The authors in [3] noted that handling outliers data can be done by eliminating outliers, then estimating, and other methods, namely repeated weighting. Repeated weighting is used in estimating the robust regression parameters. Robust regression has several estimating methods including the M-estimator method, the least trimmed squares estimator method, the S-estimator method, and the MM-estimator method. The M-estimator was introduced by Huber that is a method known to have high efficiency by minimizing the error functions [4]. While the S-estimator method was first developed by authors in [5] that minimizes the standard deviation of the residuals. The S-estimator can handle a high percentage of outliers at 50% [6]. Research related to spatial regression has been carried out by authors in [7] who examined the robust GWR (RGWR) to measure the spatial relationship between freshwater acidification critical loads and catchme nt attributes. The authors in [1] simulate various data conditions, then apply various developments from the GWR method including the RGWR method with the least absolute deviation (LAD). The simulation results show that in data that are contaminated with outliers, the RGWR with LAD models are better to use. Related research in the field of economics namely in [8] uses the RGWR with the method LAD to model poverty in Java Island, which shows that the RGWR model produces a more accurate fitted value than the GWR model in the data that indicated contain outliers. Besides, the authors in [9] compared the RTGK model with DMT-based RTGK in the case of the percentage of diarrhea incidents in Semarang City in 2015, which concluded that the DMT-based RTGK model produced a smaller mean absolute percentage error (MAPE) than the RTG model. Based on the previous study, it can be said that the RTGK with M-estimators and S estimators is still not widely used, so this study want to apply both of these estimators. Based on previous study, it can be said that the RGWR with M-estimators and S estimators is still not widely used, so this study aimed to apply both of these estimators. Besides, the M estimator has a high efficiency, while the S estimator has a high breakdown point. Gross regional domestic product (GRDP) is the gross of value added that produced by all good and services in a region produced by all unit of economy in a specified period of time regardless of the production factors owned by residents or non-residents [10]. Economic growth has a positive linear relationship with GRDP [11], so the higher the GRDP value, the higher the economic growth in the region. There are two approaches used to compile figures of GRDP, namely GRDP at the current price and the constant price. GDRP at current prices is calculated using prices at that year, whereas GRDP at constant prices using a particular year as a basis. By using GRDP at constant prices, it can be seen whether or not economic growth. East Java Province is a province that consists of regencies/cities known as their industrial areas. East Java Province consists of regencies/cities with industrial areas. The industrial estate plays an important role in the economic progress of the province and the country. The authors in [11] estimate the parameters with the improved model of geographical and temporal weighted regression (IGTWR) on GRDP data in Central Java with variables of the total labor force (TLF), as well as regional/city minimum wages (RMW), and it is found that TLF and RMW have a significant effect on the GRDP value. Besides, the authors in [12] The author uses several regression models to model the GRDP value in regencies/cities in Java with one of

the explanatory variables being the human development index (HDI) that has a real influence. According to [10] the GRDP of East Java province in 2015 amounted to 14.5% of the national economy, this is inseparable from the role of 38 regencies/cities that have geographical and demographic diversity between regions. This study aimed applies and compares several methods for estimating regression parameters including global regression, M-estimator, S-estimator, GWR, RGWR with M-estimator, and RGWR with S-estimator based on GRDP of regency/city in East Java at constant prices with the base year used is 2010 and the factors that influence it, that are TLF, RMW, and HDI.

2. Material and Method

2.1. Data

The data used in this study are GRDP of regency/city in East Java at constant price with base year used is 2010 and the factors that influence it, which are secondary data obtained from the publication of the Badan Pusat Statistik East Java [10]. This study consists 38 observation of regencies/cities in East Java Province. The variables used in the study can be seen in Table 1.

Table 1: Variables used in the study

Variable	Description of variables
Y	GRDP (Billion Rupiah)
X_1	Total labor force (Soul)
X_2	Human development index (Percentage)
X_3	Regional minimum wages (Rupiah)

2.2. Method

The data analysis are as follows:

1. Data exploration of GRDP of regency/city in East Java.
2. Testing the spatial effect using the Breusch-Pagan test.
3. Estimating the regression parameters of the GWR model with the following steps:

   i. Determine the optimum bandwidth (h). The optimum bandwidth is obtained with the smallest cross validation (CV) value. The bandwidth obtained is then used to calculate the spatial weighting matrix.

   ii. Calculates the spatial weighting matrix $\boldsymbol{W}(u_i,v_i)= diag(w_{i1},w_{i2},\ldots,w_{in})$ with the following exponential kernel functions:

$$w_{ij} = exp\left(-\frac{d_{ij}}{h}\right) \quad (1)$$

iii. Calculates the following parameter estimators of the GWR model [3]:

$$\hat{\boldsymbol{\beta}}(u_i,v_i) = [\boldsymbol{X}^T\boldsymbol{W}(u_i,v_i)\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{W}(u_i,v_i)\boldsymbol{y} \quad (2)$$

4. Detecting outliers, if outliers are detected then continue to step 5.
5. Estimating the regression parameters of the M-estimator RGWR model with stages:

i. Calculate $\hat{y}_i = x_i^T\hat{\boldsymbol{\beta}}(u_i,v_i)$, initial values $\hat{\boldsymbol{\beta}}(u_i,v_i)$ are obtained from GWR modeling.

ii. Calculating the value of robust weighting w_i using the Tukey Bisquare weighting function (Mahmood dan Salahuddin 2015):

$$w_i = \begin{cases} \left(1-\left(\frac{\varepsilon_i^*}{c}\right)^2\right)^2, & |\varepsilon_i^*| \le c \\ 0, & |\varepsilon_i^*| > c \end{cases} \quad (3)$$

with value c=4.685, $\varepsilon_i^* = \varepsilon_i / s$, $s = \frac{median\ \{|\varepsilon_i - median(\varepsilon_i)|\}}{0.6745}$, and $\varepsilon_i = y_i - \hat{y}_i$.

iii. Calculate $\hat{\boldsymbol{\beta}}^{\boldsymbol{M}}(u_i,v_i)^l = (\boldsymbol{X}^T\boldsymbol{W}^{l-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{l-1}\boldsymbol{y}$ with $\boldsymbol{W}$ is a weighting matrix multiplied by a robust weighting with an exponential weighting in equation (1).

iv. Repeat step 5(i) to 5(iii) until a convergent $\hat{\boldsymbol{\beta}}^{\boldsymbol{M}}$ is obtained.

6. Estimating the regression parameters of the S-estimator RGWR model with stages:

i. Calculate $\hat{y}_i = x_i^T\hat{\boldsymbol{\beta}}(u_i,v_i)$, initial values $\hat{\boldsymbol{\beta}}(u_i,v_i)$ are obtained from GWR modeling.

ii. Calculating the value of $\varepsilon_i = y_i - \hat{y}_i$, and estimating the standard deviation of error [13]:

$$s = \begin{cases} \frac{median\{|\varepsilon_i - median(\varepsilon_i)|\}}{0.6745}, & \text{Iteration} = 1 \\ \left[\frac{1}{n(0.199)}\sum_{i=1}^{n} w_i\varepsilon_i^2\right]^{1/2}, & \text{Iteration} = 2, 3, 4, \ldots \end{cases}$$

iii. Calculate $\varepsilon_i^* = \varepsilon_i / s$, then calculate the following robust weighting value:

w_i

$$= \begin{cases} \left(1-\left(\frac{\varepsilon_i^*}{c}\right)^2\right)^2, & |\varepsilon_i^*| \le c \\ 0, & |\varepsilon_i^*| > c \\ \frac{\rho(\varepsilon_i^*)}{(\varepsilon_i^*)^2} & \end{cases}$$

Iteration = 1

Iteration = 2, 3, 4, …

with value c = 1.547.

iv. Calculate $\hat{\boldsymbol{\beta}}^S(u_i,v_i)^l = (\boldsymbol{X}^T\boldsymbol{W}^{l-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{l-1}\boldsymbol{y}$ with $\boldsymbol{W}$ is a weighting matrix multiplied by a robust weighting with an exponential weighting in equation (1).

v. Repeat step 6(i) to 6(iv) until a convergent $\hat{\boldsymbol{\beta}}^{\boldsymbol{S}}$ is obtained.

7. Compare models that have been formed based on values MAD $= \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n}$

8. Perform partial testing of the parameters of the best model based on step 7

3. Result and Discussion

3.1. Data Exploration

East Java Province is a fairly large province with the second highest GRDP on Java Island and provides 14.5% on Indonesia's gross domestic product. The province's GRDP is high because it has many sectors, including manufacturing, wholesale and retail trade and motor vehicle, agriculture, forestry and fisheries respirations. GRDP in regencies/cities in East Java can be seen in Figure 1.

Figure 1 shows that the GRDP in East Java Province tends to be diverse and form groups. These groups can be seen from the similarity of colours in the regions that indicate a similarity in the value of GRDP, or in other words, nearby regencies/cities tend to have GRDP values that were not much different. This was possible because of the similarity of characteristics between the neighbouring regions. Figure 1 also shows that regency/city that was located close to the provincial capital, namely Surabaya have a higher GRDP than regency/city located far from the provincial capital. This certainly cannot be separated because the City of Surabaya has a good economic infrastructure. The selection of explanatory variables must be based on whether or not the relationship between each explanatory variable and the response variable. The closeness of the linear relationship of each explanatory variable with the GRDP value can be seen from the Pearson correlation value. Problems arise when using more than one explanatory variable that is causing a correlation between these explanatory variables (multicollinearity). High multicollinearity that is the value of the variance inflation factor (VIF) is more than or equal to five, resulting in inaccurate estimation of parameters. The Pearson and VIF correlation values indicate that the three explanatory variables can be used to explain the GRDP value.

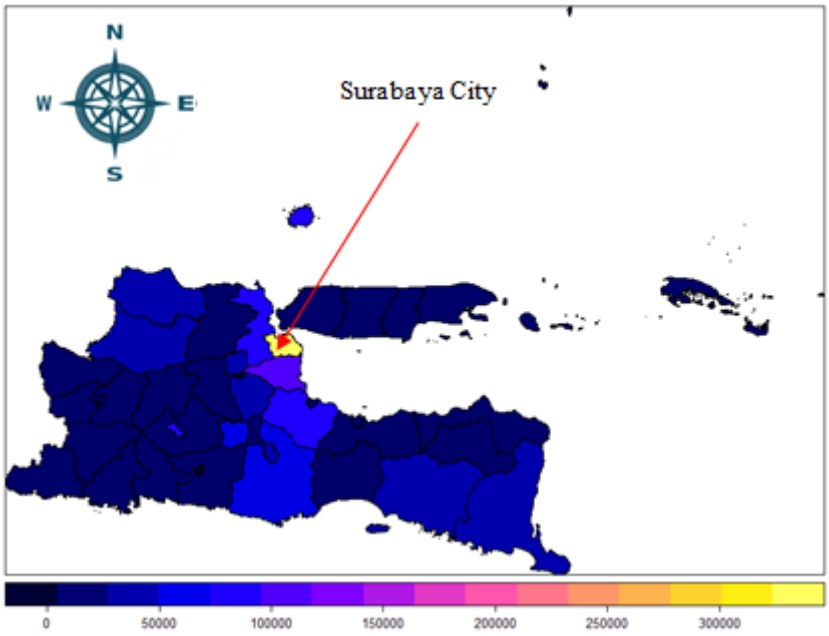

Figure 1: Map of the distribution of GRDP value of East Java Province in 2015

3.2. Geographically Weighted Regression Model

Estimating the parameters of the GWR model begins with testing the existence of spatial diversity to see there are differences in the characteristics of each regency/city or not. The test shows that the test statistic value of 17,893 and p-value of 0.00046. Based on the significance level of 0.05, it can be concluded that there is spatial diversity in the GDRP data and its explanatory variables. Estimation of GWR parameters allows different parameter estimators for each location. This difference is very dependent on the proximity of the i-th location to the surrounding location. The diversity of GWR model parameter estimators is shown in Figure 2. Figure 2 shows that the diversity of HDI variable parameter estimators is the smallest compared to the others. Figure 2 also shows the parameter estimator has a negative value, which indicates that there are regency/city that have a negative parameter estimator.

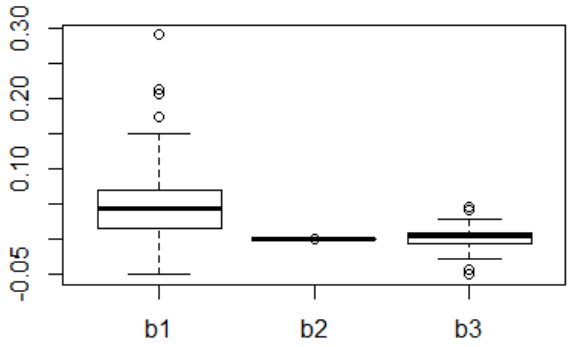

Figure 2: Boxplot of GWR model parameter estimators

3.3. Detecting Outliers

The city of Surabaya has very different characteristics from other regions (it can be seen from the GRDP value). This can indicate the city of Surabaya as an outlier. Outliers in the linear regression model are different from outliers in the GWR model. Outliers in the GWR model are also affected by distance, so outliers detection is quite complicated. The approach is made by creating a plot between explanatory and response variables. Both variables are multiplied by spatial weights.

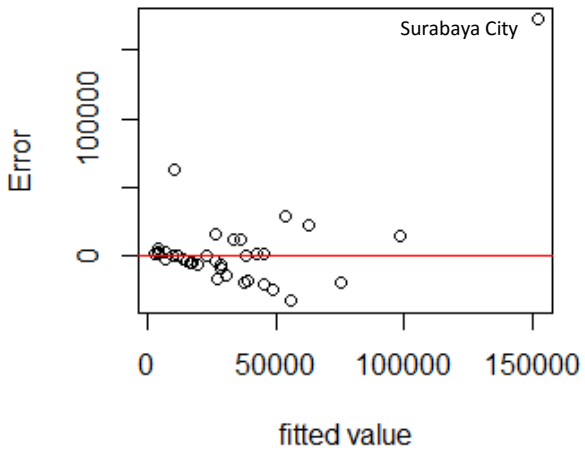

Figure 3: Detection of outliers using plot error against the estimated value of the GWR model

However, if the study have more than one explanatory variable, an error of GWR model plot against the expected value is used, shown in Figure 3. Figure 3 shows that the point identified as outlier is Surabaya City. The detected outlier indicate that the GWR model is not good enough so that robust GWR model is used to improve the goodness of the model.

3.4. Model Comparison

In this study, six models were formed to compare. The model is compared based on the errors variance and the accuracy of the model. Error variance from the six models, namely global regression, M-estimator RR, S-estimator RR, GWR, RGWR with M-estimator, and RGWR with S-estimator can be seen in Figure 4. Figure 4 shows that the six models produce different error variance. OLS (global regression model) has a fairly high error variance compared to other models. This can be caused by the presence of outliers and spatial diversity in the GRDP data and its explanatory variables which cannot be handled by global regression. M-estimator RR and S-estimator RR produce a smaller error variance than the global regression model, even tend to approach zero. This is because RR can handle the effect of outliers on the data in estimating its parameters.

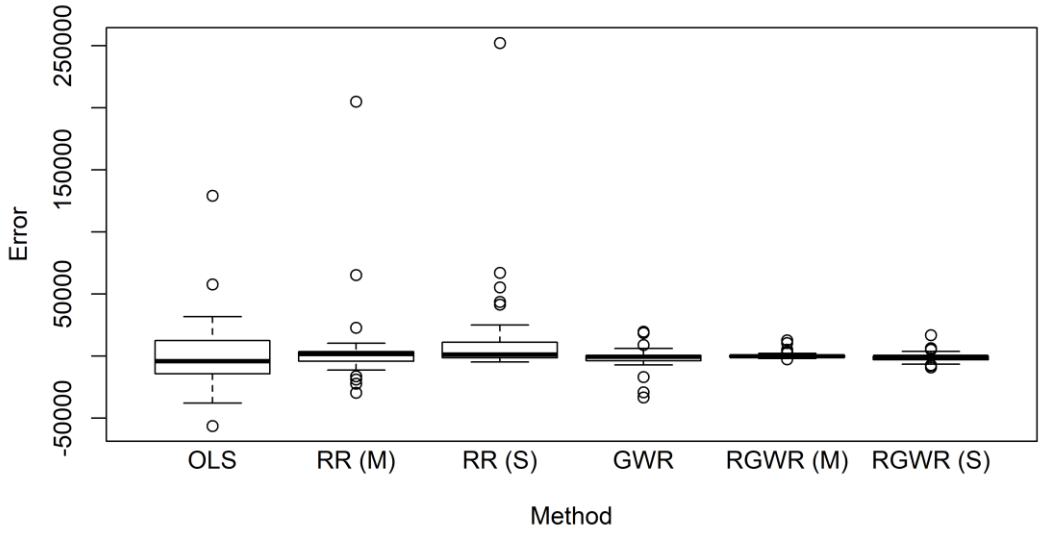

Figure 4: Boxplot of error global regression model, M-estimator RR, S-estimator RR, GWR, RGWR (with M-estimator), and RGWR (with S-estimator)

Figure 4 also shows the errors variance produced by the GWR model is smaller and tends to be close to zero than the global regression model. Besides, errors from the global regression model, M-estimator RR, and S-estimator RR located far from zero are closer to general error collections (close to zero) due to the use of GWR. The diversity that tends to be smaller than global regression can be caused by spatial diversity in data that can be handled by the GWR model, which produces local parameter estimators for each location. The RGWR model (with the M-estimator and the S-estimator) appears to have a small errors variance and tends to approach zero values compared to other models. This is consistent with the ability of the RGWR method that can accommodate spatial diversity and minimize the effect of outliers on data. From the accuracy seen from the mean absolute deviation (MAD), a comparison is made to determine which estimated that is close to the actual data. MAD of the six models are shown in Table 2. Table 2 shows that M-estimator RGWR has the smallest MAD, that is 1546.28, so the best model to be used in the case GRDP of regency/city in East Java Province in 2015 along with TLF, HDI, and RMW variables is the M-estimator RGWR model. Table 2 also shows that the goodness of the GWR model is no better than the RGWR model (with the M-estimator or the S-estimator). It might be caused by the process of estimating RGWR parameters that minimize the effect of outliers.

Table 2: Comparison of model accuracy

Model	MAD
Global	19635.4
M-estimator RR	12951.3
S-estimator RR	15619.8
GWR	5413.3
M-estimator RGWR	1546.3
S-estimator RGWR	3453.1

3.5. Influence Explanatory Variables to M-Estimator RGWR Model

The parameter testing on the model was carried out to find out the variables that significantly affected the GRDP for each location by using a partial test. Each location may have different significant variables. Based on these conditions, the increase in GRDP is realized by taking policies in accordance with the variables that have a significant effect on each location. A map of the distribution of the effect of explanatory variables on GRDP can be seen in Figure 10. Figure 10 shows that neighboring regency/city have similarities in terms of explanatory variables that have a significant effect based on a partial test on the M-estimator RGWR model.

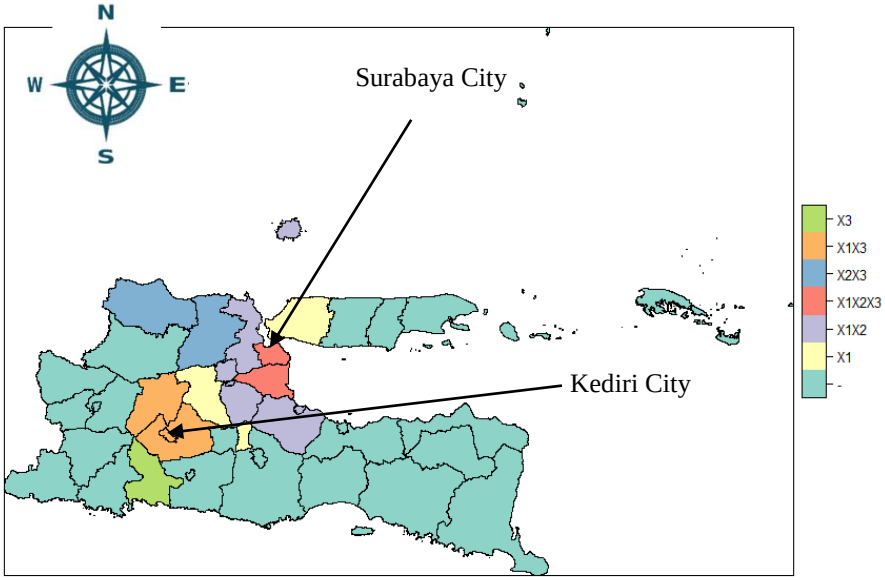

Figure 10: Map of the distribution of the effect of explanatory variables on the response variables of the M-estimator RGWR model

Figure 10 shows that the GRDP of areas located in the eastern part of Java Island and the eastern part of Madura Island does not have a significant explanatory variable. It also happened in some areas in the western part of East Java Province, the three explanatory variables did not have a real influence. No explanatory variable that has a significant effect is very likely to occur due to the absence of diversity in the GRDP, although there are variations in observed values of the explanatory variables, or in other words, there is no significant increase in GRDP despite the observed increase in values of the three explanatory variables. On the other hand, regencies/cities that are geographically close to Surabaya City generally have variables that significantly affect the GRDP. Surabaya City and Sidoarjo Regency have three explanatory variables, namely TLF, HDI, and RMW which have significant influence, while the areas adjacent to Surabaya City and Sidoarjo Regency are Gresik Regency, Mojokerto Regency, Pasuruan Regency, Mojokerto City, and Pasuruan City have TLF variables and HDI that has a real effect. RMW parameter estimator values for Surabaya City, Sidoarjo Regency, Gresik Regency, Mojokerto Regency, and Pasuruan Regency are negative. This means that every increase in RMW value will reduce the GRDP value. This can be caused in these areas there are many small industries which are heavily burdened by the increase in RMW each year so that it results in the number of small industries that are closed and results in the decline in producers of goods and services. The estimator value of HDI parameters in

Surabaya City and surrounding areas is negative, which means an increase in HDI will reduce the value of GRDP. Different things happened in Kediri City and Kediri Regency, which have positive HDI parameter estimator values.

4. Conclusion

The GRDP of regency/city in East Java 2015 along with the explanatory variables is indicated to have spatial diversity and there are outliers in the errors of GWR model. MAD from the six models formed shows that the RGWR model (M-estimator and S-estimator) is better than the global regression model and RTG. This shows that the RGWR model can handle spatial diversity and is robust to outliers. Based on the MAD, the M-estimator RGWR model is better than the other models to see the variables that affect The GRDP of regency/city in East Java 2015. The test results of the M-estimator RTGTK model produce several regions that do not have explanatory variables that have a significant effect on the model. This can be due to the limitations of the explanatory variables used.

5. Suggestion

The test results of the M-estimator RTGTK model produce several areas that do not have explanatory variables that are influential in the model, so there is a need for further studies of these areas.

References

[1]. Zhang and C. Mei. “Local least absolute deviation estimation of spatially varying coefficient models: robust geographically weighted regression approaches.” International Journal of Geographical Information Science. vol. 25(9), pp. 1467-1489, 2011.

[2]. .S. Windle, et al. “Exploring spatial non-stationarity of fisheries survey data using geographically weighted regression (GWR): an example from the Northwest Atlantic.” ICES Journal of Marine Science. vol. 67, pp.145–154, 2010.

[3]. . Fotheringham. et al. Geographically Weighted Regression: The Analysis of Spatially Varying Relationship. Chichester, UK: John Wiley & Sons Ltd, 2002, pp. 27-64.

[4]. Mahmood. and Salahuddin. “Resampling Method for the Data Adaptive Choice of Tuning Constant in Robust Regression.” Pakistan Journal of Statistics. vol. 31(3), pp. 281-294, 2015.

[5]. Rousseeuw and V.J. Yohai. “Robust Regression by Means of S-Estimators, Robust and Nonlinear Time Series Analysis.” Lecture Notes in Statistics. vol. 26, pp. 256–272, 1984.

[6]. Verard and C. Croux. “Robust regression in stata.” The Stata Journal. vol. 9(3), pp. 439-453, 2009.

[7]. Harris. et al. “Robust geographically weighted regression: a technique for quantifying spatial relationships between freshwater acidification critical loads and catchment attributes.” Annals of the Association of American Geographers. vol. 100(2), pp. 286 – 306, 2010.

[8]. Afifah. “Robust geographically weighted regression with least absolute deviation method in case of poverty in Java Island.” American Institute of Physics Conference Proceedings. vol. 1827(1), pp. 020023, 2017.

[9]. . Nurhayati et al. “Robust geographically weighted regression with least absolute deviation (case study: the percentage of diarrhea occurrence in semarang 2015).” Journal of Physics: Cofference Series. vol. 1217, pp. 012099, 2019.

[10]. an Pusat Statistik (BPS). Provinsi Jawa Timur dalam Angka 2017: Jawa Timur, ID, 2017.

[11]. Solihin. “Pengembangan regresi terboboti geografis dan temporal menggunakan interaksi jarak spasial temporal studi kasus pertumbuhan ekonomi di jawa tengah tahun 2011-2015.” M.Sc. thesis, IPB University, Indonesia, 2017.

[12]. ulita. “Pemodelan regresi komponen utama dan LASSO terboboti geografis (global dan lokal) (studi kasus data produk domestik regional bruto (PDRB)) pada 113 kabupaten/kota di pulau jawa.” M.Sc. thesis, Indonesia, 2016.

[13] usanti et al. “M estimation, S estimation, and MM estimation in robust regression.” International Journal of Pure and Applied Mathematics. vol. 91(3), pp. 349-360, 2014.