© 2019 IJSRSET | Volume 6 | Issue 1 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Science and Technology
DOI : https://doi.org/10.32628/IJSRSET196123

Robust Geographically Weighted Regression Modeling using Least Absolute Deviation and M-Estimator

Puteri Pekerti Wulandari, Anik Djuraidah*, Aji Hamim Wigena
Department of Statistics, Bogor Agricultural University, Bogor, West Java, Indonesia

ABSTRACT

Geographically weighted regression (GWR) is development of multiple regression that has spatial varying, so that the estimator of GWR is different for each location. Parameter estimation in GWR uses weighted least square method which is vulnerable to outlier and can cause biased parameter estimation. The robust GWR (RGWR) with LAD and M-estimator is resistance to outliers. This research estimated parameters on RGWR using LAD and M-estimator method and uses data of Java gross domestic product (GRDP) in 2015 containing several outliers. The result showed that RGWR model was better than GWR with M-estimator, and the predictions were closer to the actual values.

Keywords : Geographically Weighted Regression, Least Absolute Deviation, M-Estimator

I. PENDAHULUAN

Geograpically weighted regression (GWR) is a model that can be used for data with spatial varying (Fotheringham *et al.* 2002). The estimator of GWR parameters using weighted least square method is known susceptible to outliers (Zhang dan Wei 2011). In implementation of GWR, outliers are often found. Outliers can be overcome using robust (RGWR) method which is resistance to outliers, such as least absolute deviation (LAD) method and M-estimator.

LAD method introduced by Roger Jseph Boscovich in 1757 used WLS to obtain parameter estimator by minimizing absolute number of residual (Mutan 2009). Application of simplex method uses linier programming as optimum solution and it is considered to be efficient for computing LAD that cannot be solved analytically (Chen 2002).

M-estimator method was introduced by Huber 1973. This method minimizes objective function of residual. Parameter estimation uses weighted least square method iteratively (Chen 2002).

Implementation of LAD method in GWR was studied by Afifah (2015) in the case of Java poverty in 2015, while implementation of M-estimator method was studied by Sari *et al.* (2014) on mapping potential of agriculture of East Java in 2012, and Azizah (2015) compared to least square (LS), median absolute deviation (MAD) and M-estimator methods. Based on the results of this studies, LAD method which minimized the residuals influence of outliers was more robust than GWR and M-estimator.

Gross regional domestic product (GRDP) is one improtant indicator to determine economic conditions in an area and certain periode (BI 2015). GWR and geogrpahically weighted panel regression (GWPR) to estimate GRDP parameters was studied by Fatulloh

(2013) and Handayani (2017). A similar research was studied by Mastuti (2017) in criminal cases. Based on these research, it is necessary to handle outliers having influence on regression coefficient.

The aim of this research is to apply RGWR model using LAD and M-estimator methods on GRDP data in 2015. Explanatory variables were the number of labors, regional income, regional minimum wage, and human development index. Benefits of this research is to be a reference to handle outliers in similar data.

II. METHOD AND MATERIAL

A. Data

Data in this study were Java GRDP in 118 districs/cities in 2015 obtained from website Central Bureau of Statistics (www.bps.go.id). The variables (Table 1) refer to the research by Handayani (2017).

TABLE I
RESPONSE AND EXPLANATION VARIABLES

Variables	Description	Unit
Y	GRDP districs/cities of Java at Contants Price 2010 in 2015	Thousand Rupiah
X_1	Number of Labor	People
X_2	Regional Income	Million Rupiah
X_3	Regional Minimum Wage	Rupiah
X_4	Human Development Index	Percent

B. Procedure Data Analysis

The steps of data analysis are as follows :

1. Data description and exploration.
2. Spatial effects test :
   a. Heterogeneity spatial test
      i. Hypothesis
         $H_0 : {\alpha_2}^2 = \cdots = {\alpha_k}^2 = 0$ (no heterogeneity between location)
         H_1 : minimal ada satu ${\alpha_k}^2 \neq 0$ (there is a heterogeneity between location)
      ii. Test Statistics

$$BP = \frac{1}{2}\boldsymbol{f}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{f} \sim {\chi^2}_{(k-1)}$$

with ; $\boldsymbol{f} = (f_1, \ldots, f_n)^T$

$$f_i = (\frac{{e_i}^2}{e_i} - 1)$$

$$e_i = y_i - \hat{y_i}$$

      iii. Critical region, if BP $> {\chi^2}_{(k-1)}$ than reject H_0
      iv. Conclusion
   b. Spatial autocorrelation
      i. Hypothesis
         H_0 : I = 0 (no autocorrelation between location)
         H_1 : I ≠ 0 (there is a autocorrelation between location)
      ii. Test Statistics

$$\mathbf{Z}_I = \frac{I - E(I)}{\boldsymbol{s}^2_{(I)}}$$

with ;

$$I = \frac{n}{\sum_{i=1}^n \sum_{j=1}^n \boldsymbol{w}_{ij}} \frac{\sum_{i=1}^n \sum_{j=1}^n \boldsymbol{w}_{ij}\,(X_i - \overline{X})(X_j - \overline{X})}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

$$\boldsymbol{S}^2_{(I)} = \frac{n^2 \sum_{ij}^n \boldsymbol{W}^2_{ij} + 3(\sum_{ij}^n \boldsymbol{W}^2_{ij})^2 - n\sum_i^n (\sum_j^n \boldsymbol{W}_{ij})^2}{(n^2-1)(\sum_{ij}^n \boldsymbol{W}_{ij})^2}$$

      iii. Critial region, if $Z > Z_{\alpha/2}$ than reject H_0
      iv. Conclusion
3. Outlier detection using boxplot
4. Analysis GWR model
   a. Determine longitude-latitude coordinates observation area
   b. Calculates euclidean distance (d_{ij}) between observation areas

$$d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2}$$

   c. Determining bandwidth using ACV and CV optimum criteria

$$ACV(\mathrm{h}) = \sum_{i=1}^{n} |y_i - \hat{y}_{i\neq 1}(h)|$$

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

d. Calculates weighting matrixs (w_{ij}) with kernel function

e. Determine $\hat{\beta}^{(0)}$ value as first estimate selected spatial regression parameters.

5. Analysis GWR model using LAD method

a. Using step 4 as first step to determines ACV value

b. Determine regression coefficient by minimizing residual using optimum bandwidth

$$min \sum_{i=1}^{n} |\varepsilon_i| = \sum_{i=1}^{n} \left| y_i - \beta_j(u_0, v_0) x_{ij} \right| w_j(d_{0i})$$

c. Estimating robust GWR LAD parameters

$$y_i = \sum_{j=1}^{p} \beta_j(u_0, v_0) x_{ij} + \varepsilon_i{}^{+} - \varepsilon_i{}^{-}$$

6. Analysis GWR model using M-estimator method

a. Using step 4 as first step to determine CV value, estimate $\hat{\beta}^{(0)}$ and get $\varepsilon_i{}^{(0)}$

b. Calculate robust to get value of influence function

c. Determine objective function and calculate weighting value

$$\boldsymbol{w}_i^*(u_i)^{(0)} = \frac{\psi(u_i)^{(0)}}{u_i{}^{(0)}}$$

d. Determine $\hat{\beta}$ value with WLS method

$$\widehat{\boldsymbol{\beta}}^m = (\boldsymbol{X^T W}^m \boldsymbol{X})^{-1} \boldsymbol{X^T W}^m \boldsymbol{y}$$

e. Set residual step (d) as residual step (b)

f. Iterating IRLS on new weighting until $\widehat{\boldsymbol{\beta}}^m$ convergent

7. Determine the best model based on the MAPE value

$MAPE = \frac{\sum \frac{|e_i|}{y_i} x\ 100\%}{n}$; $|e_i| = y_i - \hat{y}$ and n is amount of data

III. RESULT AND DISCUSSION

A. Data Description

Table 2 shows the highest and lowest GRDP percentage in Java. The highest percentage was in DKI Jakarta and the lowest in West Java. Rate of economic growth between districts/cities in Java shown a varying level.

TABLE 2

THE HIGHEST AND THE LOWEST GRDP PERCENTAGE

Percentage of GRDP	Province	Kabupaten/Kota
The highest	Jakarta	Central Jakarta 6.8%, South Jakarta 6.3%, North Jakarta 5.2%, East Jakarta and West Jakarta 4.7%.
	East Java	Surabaya 6.2%.
	West Java	Bekasi district 3.9%.
The lowest	West Java	Pangandaran district 0.01%, Banjar 0.05%, Indramayu district 0.1%.
	East Java	Pasuruan 0.9%.
	Central Java	Blitar and Mojokerto 0.07%, Magelang 0.1%.

Multicollinearity test uses variance inflation factor (VIF) on each explanatory variables. If VIF value > 5 then there is an indication of multicollinearity. Table 3 shows that VIF value obtained from all explanatory variables are less than 5, so it can be concluded that there is no multicollinearity.

TABLE 3
VIF VALUE EXPLANATORY VARIABLES

Variable	Coefficient
X_1	1.56
X_2	1.34
X_3	1.81
X_4	1.26

B. Spatial Effect Test

Spatial effect to test heterocedasticity uses Breuch-Pagan (BP) test. Breuch-Pagan value is 31.32 with a p-value of 0.00 that is less than 5%, so it can be concluded that there was variance spatial heterocedasticity at 5% significant level.

Spatial effect to test spatial autocorrelation using Moran Index test. The Moran Index value is 0.05 with a p-value of 0.79 which was more than 5%, so it can be concluded spatial autocorrelation at 5 significant level.

C. Outliers Detection

Figure 1 shows standardized residual of GWR model. In the boxplot, there was incicated the outliers. The outliers means that outliers in data cannot be handled by GWR.

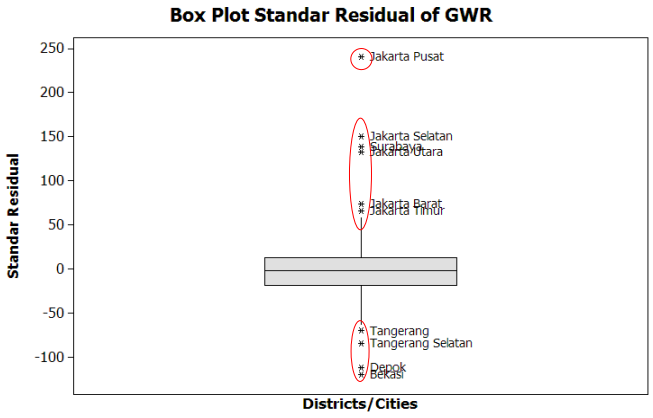
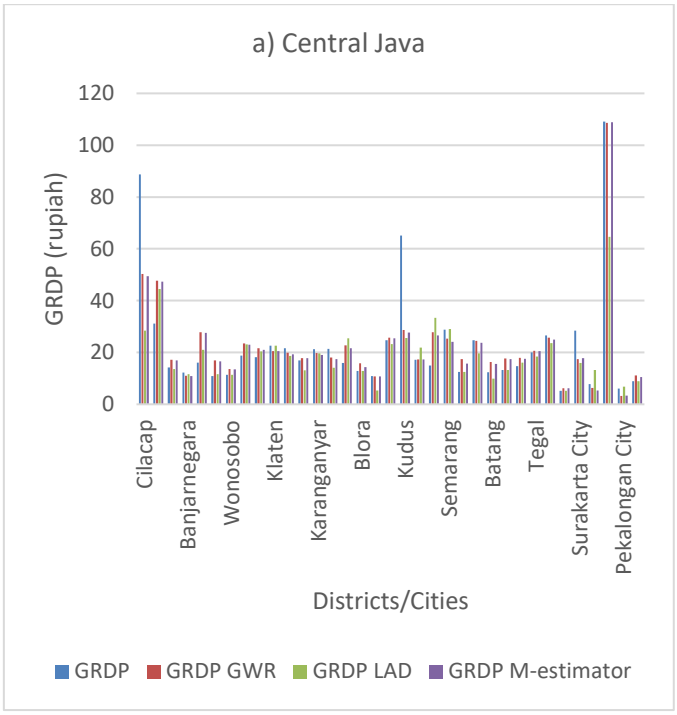
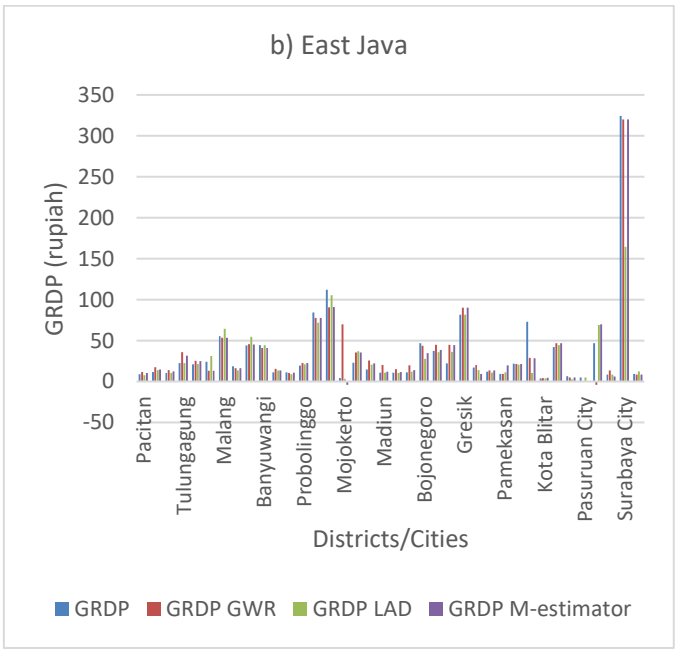


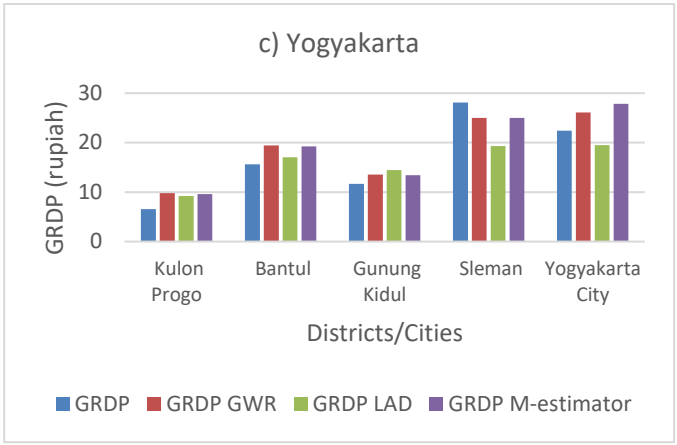
Figure 1 : Residual Boxplot GWR

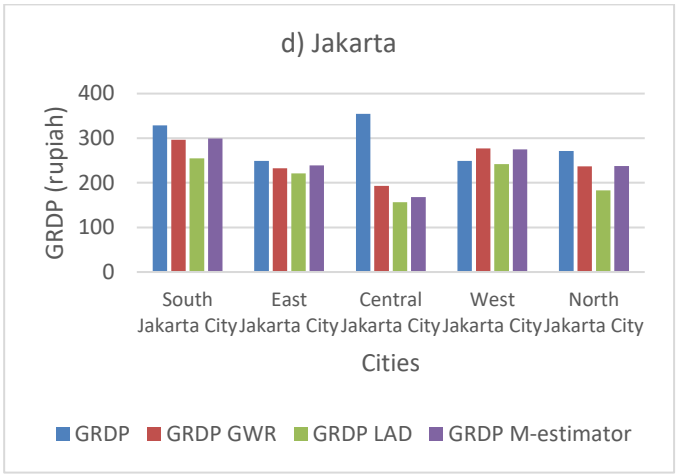
D. Comparison GWR LAD and M-estimator

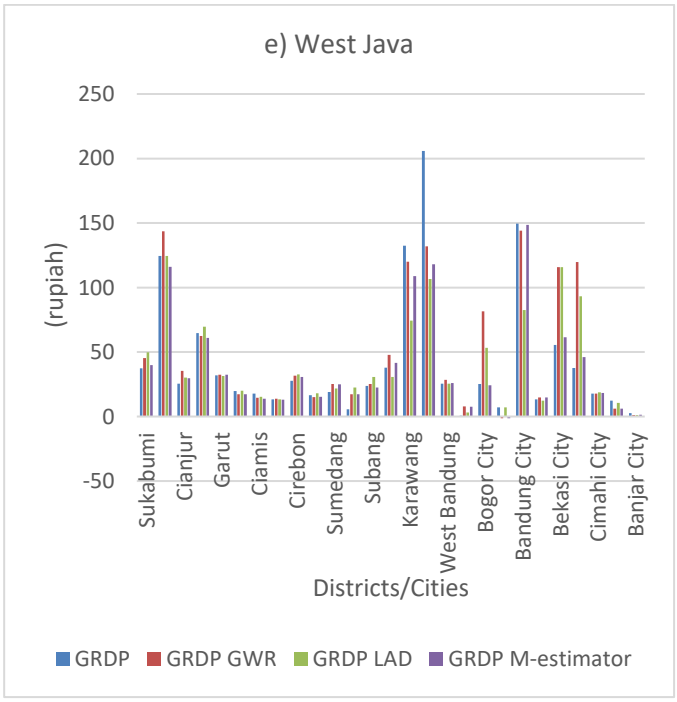
Bar chart in Figure 2 shows the estimation value of GWR, GWR LAD, and GWR M-estimator models for each districts/cities in Java province. M-estimator model had estimate value that was almost same as GWR model. Overall robust model was closer to actual value than GWR.











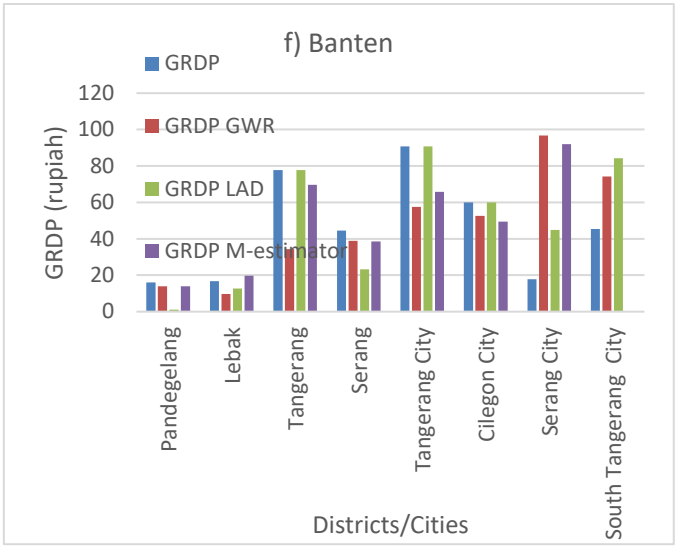
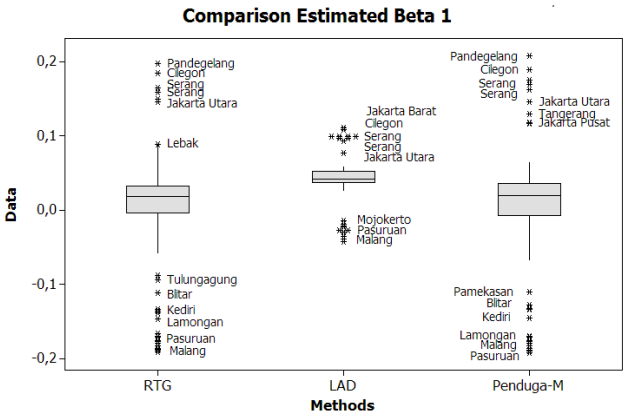


Figure 2 : Comparison of estimation value of Java districts/cities

Figure 3 shows distribution parameter estimation of GWR, LAD and M estimator methods. Measure of dispersion used was range. Changes coefficient estimated using robust method were seen from median, range, and outliers. LAD model had smaller range and higher median position.



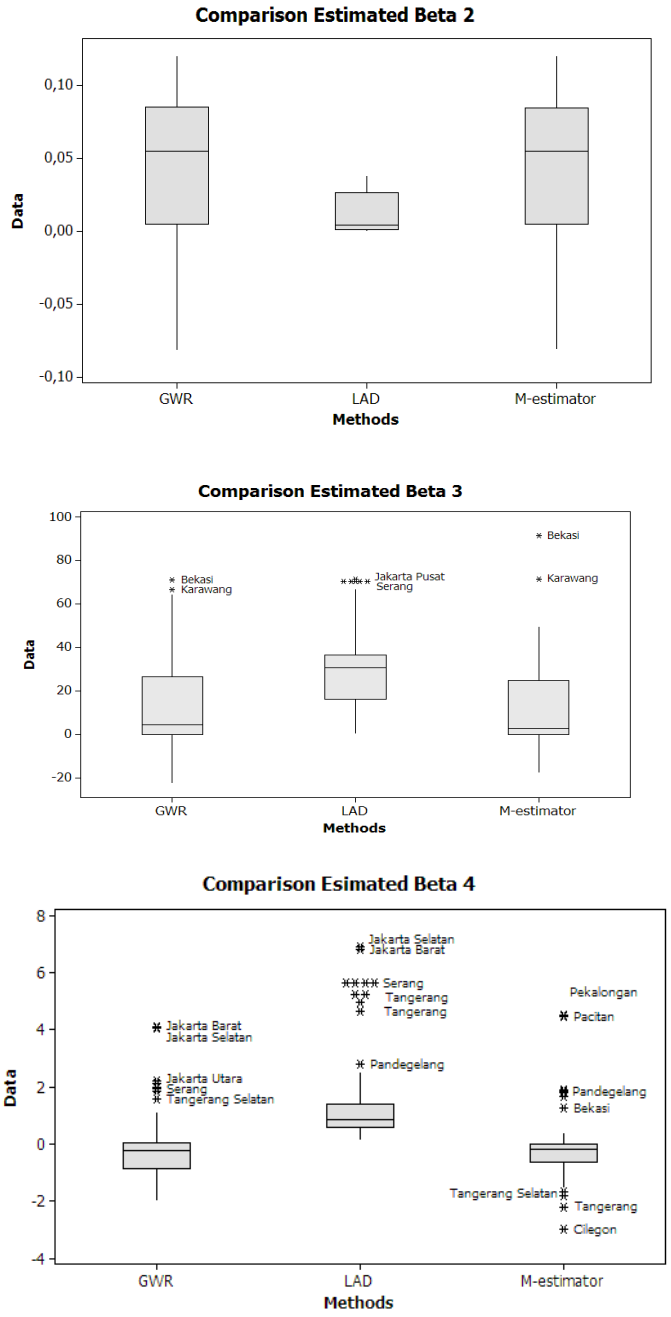


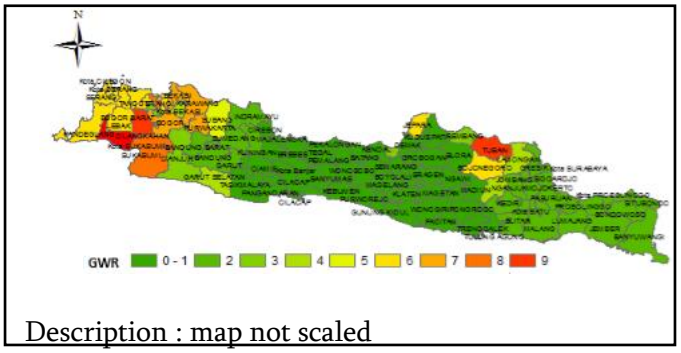
Figure 3 : Boxplot Parameter Estimates of GWR, GWR LAD, and M-estimator

Mean absolute percentage error (MAPE) value also used to compared. Criteria to measure the accuracy of model is smallest MAPE. Smaller value is best model. Table 4 shows the comparison estimation of robust methods. Robust methods had the MAPE value was smaller than GWR, so this model was better applied in outliers data. LAD model had smallest MAPE value and better applied for Java GRDP data in 2015. Coefficient determination of M-estimator model has highest value, it means M-estimator model was be able to explain Java GDP variation Java in 2015 better than LAD. However the M-estimator method in this study had not been as good as LAD method.

TABLE 4
MAPE AND CORRELATION VALUE

	MAPE	R^2
GWR	47.4%	84.53%
RGWR LAD	31.08%	87.21%
RGWR M	37.61%	88.55%

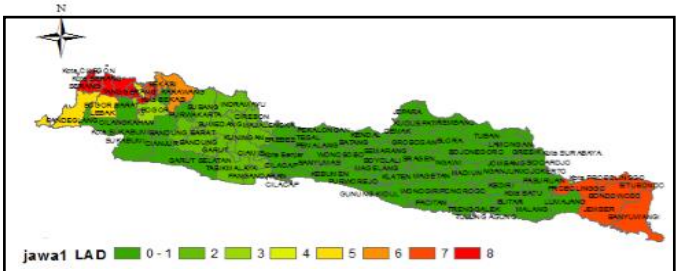
Figure 4, Figure 5, and Figure 6 shows group of factors having significant effect in each location on GWR, LAD and M-estimator methods. Figure 4 shows that group 1 GWR including all factors. Group 2 had X_1, X_3, dan X_4. Group 3 had X_1, X_3 and X_4. Group 4 had X_1 and X_3. Group 5 had X_3. Group 6 had X_2. Group 7 had X_2 and X_4. Group 8 had X_1 dan X_4. However, in group 9 there were no significant factor.



Description : map not scaled

Figure 4 : Variable Group having effect on GWR method

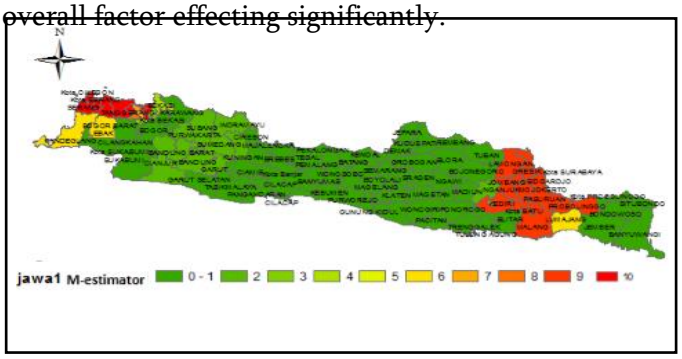
Figure 5 shows that group 1 LAD had overall factors effecting significantly. Group 2 had X_1, X_2, and X_3. Group 3 had X_1, X_2, and X_4. Group 4 had X_1. Group 5 had X_3, X_4. Group 6 had X_2. Group 7 had X_3. Group 8 had overall factor effecting significantly.



Description : map not scaled

Figure 5 : Variable Group having effect on LAD method

Figure 6 shows that group 1 M-estimator had overall factors effecting significantly. Group 2 had X_1, X_2, and X_3. Group 3 had X_2, X_3, dan X_4. Group 4 had X_1 and X_2. Group 5 had X_1 and X_3. Group 6 had X_3 and X_4. Group 7 had X_1. Group 8 had X_2. Group 9 had X_3. Group 9 had overall factor effecting significantly.



Description : map not scaled

Figure 6 : Variable Group having effect on M-estimator method

IV. CONCLUSION

RGWR LAD and M-estimator methods can be used to handle outliers and result estimators closer to actual value. The best estimator of GRDP data was LAD with smallest MAPE value. Factors in explanatory variables, number of labor (X_1), regional income (X_2), regional minimum wages (X_3), and human development index (X_4) had significant influence on each district/cities. Robust methods is an alternative when there is outlier in data, so there is no need to throw out outlier data.

V. REFERENCES

[1] A. B. Author, "Title of chapter in the book," in Title of His Published Book, xth ed. City of Publisher, Country if not

[2] Afifah R. 2017. *Robust Geographically Weighted Regression with Least Absolute Deviation Method in Case of Poverty in Java Island*. 1827(1):020023. doi: 10.1063/14979439.

[3] Anselin L. 1988. *Spatial Econometrics: Method and Models.* Kluwer Academic Publisher. The Netherlands.

[4] Azizah RI. 2015. Pendugaan Parameter Model Dinamik Dengan Metode *Median Absolute Deviation* dan *Bisquare M-Estimation* [Skripsi]. Bogor : Institut Pertanian Bogor.

[5] [BI] Bank Indonesia. 2015. Metadata : Produk Domestik Regional Bruto (PDRB). Jakarta : Bank Indonesia.

[6] Chen C. 2002. *Robust Regression and Outlier Detection with ROBUSTREG Procedurer*. SAS Institute Inc.

[7] Fatulloh. 2013. Penerapan Regresi Terboboti Geografis Untuk Data Produk Domestik Regional Bruto [Skripsi]. Bogor : Institut Pertanian Bogor.

[8] Fotheringham AS, Charlton M, Brundon C. 2002. *Geographically Weighted Regression.* Chicester, UK: John Wiley and Sons.

[9] Handayani LMW. 2017. Penerapan Regresi Panel Terboboti Geografis Pada Produk Domestik Regional Bruto Di Jawa Tengah Tahun 2011-2015 [Tesis]. Bogor : Institut Pertanian Bogor.

[10] Mutan OC. 2009. *A Monte Carlo Comparison of Regression Estimators When the Error Distribution is Long Tailed Symmetric*. Journal of Modern Applied Statistical Methods. 8(1): 161-172. doi: 10.22237/jmasm/1241136780.

[11] Sari RA., Wijayanto H., Indahwati. 2016. Perbandingan Beberapa Metode Kekar Pada Pendugaan Parameter Regresi Linier Sederhana Untuk Data yang mengandung Pecilan [Skripsi]. Bogor : Institut Pertanian Bogor.

[12] Zhang H, Mei C. 2011. *Local Least Absolute Deviation Estimation of Spatially Varying Coefficient Models: Robust Geographically Weighted Regression Approaches*. International Journal of Geographical Information Science. 25(9):1467-1489. doi: 10.1080/13658816.2010.528420.

Cite this article as :

Puteri Pekerti Wulandari, Anik Djuraidah, Aji Hamim Wigena, "Robust Geographically Weighted Regression Modeling using Least Absolute Deviation and M-Estimator", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), ISSN : 2456-3307, Volume 6 Issue 1, pp. 238-245, January-February 2019. Available at doi : https://doi.org/10.32628/IJSRSET196123
Journal URL : http://ijsrset.com/IJSRSET196123