Robust Geographically Weighted Regression Model on Poverty Data in South Sulawesi in 2019

Aqilah Salsabila Rahman*, Georgina Maria Tinungki**, Erna Tri Herdiani***

* Coresponding Author: aqilahsr123@gmail.com
**Corresponding Author: georgina@unhas.ac.id
***herdiani.erna@unhas.ac.id

Department of Statistics, Faculty of Mathematics and Natural Sciences, Hasanuddin University, Makassar, 90245 Indonesia Makassar, Indonesia

Abstract

Geographically Weighted Regression (GWR) is a method of spatial analysis that can be used to perform analysis by assigning weights based on the geographical distance of each observation location and the assumption of having spatial heterogeneity. The result of this analysis is an equation model whose parameter values apply only to each observation location and are different from other observation locations. However, when there are outliers at the observation location, a more robust estimation method is needed. One of the robust methods that can be applied to the GWR model is the Least Absolute Deviation method. In this study, model estimation was carried out on the factors that affect poverty in South Sulawesi in 2019 using Robust Geographically Weighted Regression (RGWR) with the Least Absolute Deviation (LAD) method. Determination of weighting is done by using the adaptive kernel bisquare weighting function. The results obtained are Robust Geographically Weighted Regression (RGWR) models which are different and apply only to each district/city in South Sulawesi. In addition, it was also found that the Robust Geographically Weighted Regression (RGWR) model with the Least Absolute Deviation (LAD) method was the best model for data that experienced spatial heterogeneity and contained outliers.

Keywords: adaptive kernel bisquare, geographically weighted regression, least absolute deviation, poverty, robust;

1. Introduction

Poverty is one of the indicators of people's well-being. According to the Central Bureau of Statistics, poverty is a problem experienced by almost all countries in the world, including even developed countries. According to the World Bank, in 2015, ten percent of the world's population, or 731 million people, lived below the international poverty line. This indicates that poverty remains a global issue that needs to be solved. According to the data of Central Bureau of Statistics, the percentage of the poor population in Indonesia in March 2020 was 9.78 percent, increase 0.56 percent from September 2019 and up 0.37 percent from March

2019. More specifically, the percentage of the poor population in South Sulawesi in March 2020 was 8.72 percent, increase 0.16 percent from September 2019 and increase 0.03 percent from March 2019. The rise in the poverty of the population indicates that the determination of policy still needs to be examined as to the causes.

One of methods that can be used to analyze the factors that cause poverty is multiple linear regression. Multiple linear regression is a method that models the relationship between the response variable and the predictor variable (Montgomery & Peck 1992). The estimation method used to analyze regression is the Ordinary Least Squares (OLS) (Montgomery & Peck 1992). In using a method, there are assumptions that must be fulfilled in order for the method to be used. The OLS method has certain assumptions (Gujarati, 2007). One of the assumptions that must be fulfilled is the homogeneity of the variance. Homogeneity of the variance is difficult to obtain because of characteristic differences in a region that result in the occurrence of spatial heterogeneity (Erda, 2018). Spatial heterogeneity is a condition in which the measurement of the relationship between variables differs between one location of observation and the other. So, we need an analytical approach that takes into account the geographical conditions that in this case are called spatial analysis.

The Geographically Weighted Regression (GWR) model is one of the methods used to estimate data that has spatial heterogeneity with the estimation using the weighted least squares (WLS) method (Fotheringham, 2002). The result of this analysis is an equation model whose parameter values could only be applied on each observation location and differ from other observation location.

All of the research that has just been mentioned uses the estimation method of WLS. However, the method is not resistant to the existence of outlier (Wulandari, 2019). Thus, when there is an outlier on the data, it will result in the creation of biased estimation of parameters and result in confusion in concluding regression relations (Wulandari, 2019). An alternative that can be done to regression analysis to carry the presence of outliers on data, is to use robust regression method. Therefore, the Robust Geographically Weighted Regression (RGWR) analysis method is expected to be suitable for analyzing data that covers several observation locations with the assumption of having spatial heterogeneity and containing outliers.

One of parameter estimation methods was introduced by Roger Joseph Boscovich in 1957, the Least Absolute Deviation method. This method is used by minimizing the absolute number of errors to obtain estimated regression parameters. One study compared the LAD method and the M estimation, then it was concluded that LAD is the best method (Wulandari, 2019).

Based on the above description, researchers are interested in modeling the factors that influence poverty in South Sulawesi in 2019 using Robust Geographically Weighted Regression (RGWR) with the Least Absolute Deviation method. (LAD).

2. Materials and Methods

The data used in this study are secondary data obtained from the Central Bureau of Statistics through the National Social Economic Survey, and the National Labour Force Survey. The data consists 24 districts/cities in South Sulawesi Province in 2019. The variables used in this study are shown in Table 1.

Table 1. Operational definition of variables

Variables	Description
Y	Percentage of poor population
X_1	Human development index
X_2	Unemployment rate

X_3	Literacy rate of poor population
X_4	Percentage of poor people who have not completed elementary school
X_5	Percentage of poor people working in the informal sector
X_6	Percentage of per capita expenditure of the poor population on food
X_7	Percentage of poor households using decent water
X_8	Percentage of poor households using private/shared latrines
X_9	Percentage of poor households receiving government assistance

The data was analyzed by using the Robust Geographically Weighted Regression (RGWR) model using the Least Absolute Deviation method (LAD). Given the latitude and longitude coordinates (u_i, v_i) at the location of i-th observation, then the model of Geographically Weighted Regression at such location is expressed as follows (Huang, 2010).

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \qquad (1)$$

where $i = 1, 2, \ldots, n$.

Since d_{ij} is the distance (u_i, v_i) between the i-th location of observations, then the regression coefficient on the equation (1) is estimated using the Least Absolute Deviation method based on the location (u_i, v_i), which is expressed as

$$\min \sum_{i=1}^{n} W_{ij} |\varepsilon_i| \qquad (2)$$

with the adaptive weighting function of the bisquare kernel W_{ij} as follows.

$$W_{ij} = \begin{cases} \left(1 - \left(\frac{d_{ij}}{h_i}\right)^2\right)^2 & ; d_{ij} \le h_i \\ 0 & ; d_{ij} > h_i \end{cases}$$

When a decomposition is performed on the error ε_i, and it is assumed $\varepsilon_i = y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}_i$ (Wheeler, 2014), then the estimation of the RGWR model parameters with the Least Absolute Deviation method at the i-th location of observation can be expressed as in the equation (3).

$$\hat{\boldsymbol{\beta}}_{i,LAD} = \arg\min \left\{ \sum_{i=1;\varepsilon_i \ge 0}^{n} W_{ij}\left(y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}_i\right) + \sum_{i=1;\varepsilon_i < 0}^{n} W_{ij}\left(y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}_i\right) \right\} \qquad (3)$$

The equation (3) is then solved using the simplex method with the linear optimization problem as follows.
Minimize $\sum_{i=1}^{n} W_{ij} |\varepsilon_i| = \sum_{i=1,\varepsilon_i \ge 0}^{n} W_{ij}(\varepsilon_i^+) + \sum_{i=1,\varepsilon_i < 0}^{n} W_{ij}(\varepsilon_i^-)$

Subject to $\beta_{i0} + \sum_{k=1}^{p} \beta_{ik} x_{ik} + \varepsilon_i^+ - \varepsilon_i^- = y_i$, $\quad \varepsilon_i^+, \varepsilon_i^- \ge 0, i = 1, 2, \ldots, n$.

From the estimation of the parameters of the Robust Geographically Weighted Regression (RGWR) model using the Least Absolute Deviation (LAD) method obtained at the equation (3) then applied to the poverty data in South Sulawesi in 2019 with the estimation results of the RGWR model as follows.

$$\hat{y}_i = \hat{\beta}_{i,0,LAD} + \hat{\beta}_{i,1,LAD} x_{i1} + \hat{\beta}_{i,2,LAD} x_{i2} + \cdots + \hat{\beta}_{n,p,LAD} x_{np}$$
$$\hat{y}_i = \mathbf{x}_i \hat{\boldsymbol{\beta}}_{i,LAD}$$

The initial step in this study was to determine a multiple linear regression model using the OLS method. Next, perform the test of spatial effects which in this case is the effects of spatial heterogeneity using the

Breusch-Pagan test. Next, determine the Geographically Weighted Regression (GWR) model by using the WLS method. Then, do an outlier detection. After that, determine the Robust Geographically Weighted Regression (RGWR) model using the LAD method. Then, determine the best model and interpret the model.

3. Results and Discussion

3.1. Multiple linear regression modeling by using ordinary least squares method

Estimates of the multiple linear regression model parameters were obtained using the ordinary least squares method (OLS), thus obtaining $\widehat{\boldsymbol{\beta}}_{OLS}$ that applies to all districts/cities in South Sulawesi. Furthermore, from $\widehat{\boldsymbol{\beta}}_{OLS}$ an estimate of the multiple linear regression model is obtained as follows.

$$y_i = 26.236 - 0.377x_{i1} + 0.259x_{i2} - 0.008x_{i3} - 0.158x_{i4} + 0.048x_{i5} + 0.209x_{i6} - 0.016x_{i7} - 0.042x_{i8} + 0.099x_{i9}$$

One of the assumptions that must be fulfilled when using the OLS method is a homogenization assumption. So, heterogeneity testing needs to be done to find out if there's any heterogeneity in the data.

3.2. Spatial heterogeneity

Spatial homogeneity is indicated by characteristic differences between one observation location and another observation site. This spatial heterogeneity test can be done using the statistics of the Breusch-Pagan test (BP) with the following hypothesis test procedures.

i. Hypothesis

$H_0 : {\sigma_1}^2 = {\sigma_2}^2 = \cdots = {\sigma_{24}}^2 = \sigma^2$ (Tidak terdapat keragaman spasial)
H_1 : Terdapat ${\sigma_i}^2 \neq \sigma^2$ (Terdapat keragaman spasial)
$i = 1,2,\ldots,24.$

ii. Statistical testing

$$BP = \left(\frac{1}{2}\right)\mathbf{f}^T\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{f} = 17.9327$$

iii. Decision

Since $BP = 17.9327 > \chi^2_{(0.05;9)} = 16.9190$, therefore reject H_0 and accept H_1.

It can be concluded that there is spatial heterogeneity or diversity between the locations of each variable in poverty data. Therefore, regression modeling can be continued by taking into account the location, in this case using the Geographically Weighted Regression model (GWR).

3.3. Geographically weighted regression modeling by using the weighted least squares method

The estimated coefficients of the GWR model parameters were obtained using the weighted least square method (WLS) with the Bisquare adaptive kernel weigher function, thus obtaining $\widehat{\boldsymbol{\beta}}_{i,WLS}$ that applies to every 24 districts/cities in South Sulawesi as follows.

Table 2. The estimated parameter coefficients of the RGWR model

Variables	Max	Min
$\hat{\beta}_{i,0,WLS}$	39.60167871	28.58984779

$\hat{\beta}_{i,1,WLS}$	-0.358079417	-0.511432619
$\hat{\beta}_{i,2,WLS}$	0.540733443	-0.033881214
$\hat{\beta}_{i,3,WLS}$	-0.05076815	-0.11388752
$\hat{\beta}_{i,4,WLS}$	-0.171393947	-0.275857622
$\hat{\beta}_{i,5,WLS}$	0.055597586	0.041668852
$\hat{\beta}_{i,6,WLS}$	0.321181178	0.142539452
$\hat{\beta}_{i,7,WLS}$	0.037580739	-0.017584531
$\hat{\beta}_{i,8,WLS}$	-0.045886364	-0.066528292
$\hat{\beta}_{i,9,WLS}$	0.112528391	0.082146459

Furthermore, from the estimate of the parameter $\hat{\boldsymbol{\beta}}_{i,WLS}$ obtained the estimation of the GWR model for the percentage of the poor population in South Sulawesi in 2019 that applies to each district/city in South Sulawesi.

3.4. Outlier detection

The outlier detection of the GWR model error was performed using the Boxplot method and it was obtained that there was a search on the model error of GWR, i.e. in Wajo district. The existence of this search indicates that the obtained GWR model is not robust to the existence. So, the GWR modeling needs to continue using a more robust method which in this case is later called the Robust Geographically Weighted Regression model.

3.5. Robust geographically weighted regression modeling by using LAD method

The estimated coefficients of the RGWR model parameters were obtained using the Least Absolute Deviation (LAD) method with the bisquare adaptive kernel weighting function, thus obtaining the $\hat{\boldsymbol{\beta}}_{i,LAD}$, which applies to every 24 districts/cities in South Sulawesi as follows.

Table 3. The estimated parameter coefficients of GWR model

Variables	Max	Min
$\hat{\beta}_{i,0,LAD}$	79.82158209	31.26710187
$\hat{\beta}_{i,1,LAD}$	-0.265432205	-0.601283711
$\hat{\beta}_{i,2,LAD}$	0.734236154	-0.474746239
$\hat{\beta}_{i,3,LAD}$	0.013574449	-0.214162747
$\hat{\beta}_{i,4,LAD}$	-0.143106993	-0.323611519
$\hat{\beta}_{i,5,LAD}$	0.106275046	-0.038117891
$\hat{\beta}_{i,6,LAD}$	0.331467393	-0.107133715
$\hat{\beta}_{i,7,LAD}$	0.074456188	-0.012645455
$\hat{\beta}_{i,8,LAD}$	-0.035886859	-0.110285072
$\hat{\beta}_{i,9,LAD}$	0.12213158	-0.003640847

Furthermore, from the estimate of the parameter $\hat{\boldsymbol{\beta}}_{i,LAD}$ is obtained the estimation of the RGWR model for the percentage of the poor population in South Sulawesi in 2019 that applies to each district/city in South.

3.6. Goodness of fit

A comparison of the estimated percentage of the poor population in South Sulawesi in 2019 for each model of Multiple Linear Regression, GWR, and RWGR is presented in Table 4.

Table 4. Comparison of the estimated percentage of poor population in South Sulawesi in 2019

District/City	Y	MLR	GWR	RGWR
Kepulauan Selayar	12.83	12.45	12.71	12.83
Bulukumba	7.26	8.25	7.94	7.93
Bantaeng	9.03	9.49	9.35	9.03
Jeneponto	14.88	15.46	15.29	14.88
Takalar	8.7	8.38	8.72	8.7
Gowa	7.53	7.23	7.85	7.53
Sinjai	9.14	8.9	8.79	9.14
Maros	9.89	9.5	9.61	9.89
Pangkep	14.06	11.79	13.47	14.06
Barru	8.57	7.94	8.01	8.57
Bone	10.06	10.79	10.50	10.06
Soppeng	7.25	7.98	7.47	7.25
Wajo	6.91	8.42	8.28	8.79
Sidrap	4.79	4.88	4.77	4.79
Pinrang	8.46	7.66	7.55	7.39
Enrekang	12.33	12.03	12.25	12.33
Luwu	12.78	12.57	12.57	12.78
Tana Toraja	12.35	12.62	12.71	12.35
Luwu Utara	13.6	13.15	13.55	13.6
Luwu Timur	6.98	6.75	6.80	6.98
Toraja Utara	12.41	11.83	11.87	12.41
Kota Makassar	4.28	2.85	4.04	4.28
Kota Parepare	5.26	6.67	6.15	5.26
Kota Palopo	7.82	9.57	8.15	7.82

As for the comparison chart of the estimated percentage of the poor population in South Sulawesi in 2019 for each model of Multiple Linear Regression, GWR, and RWGR are presented in Fig.1.

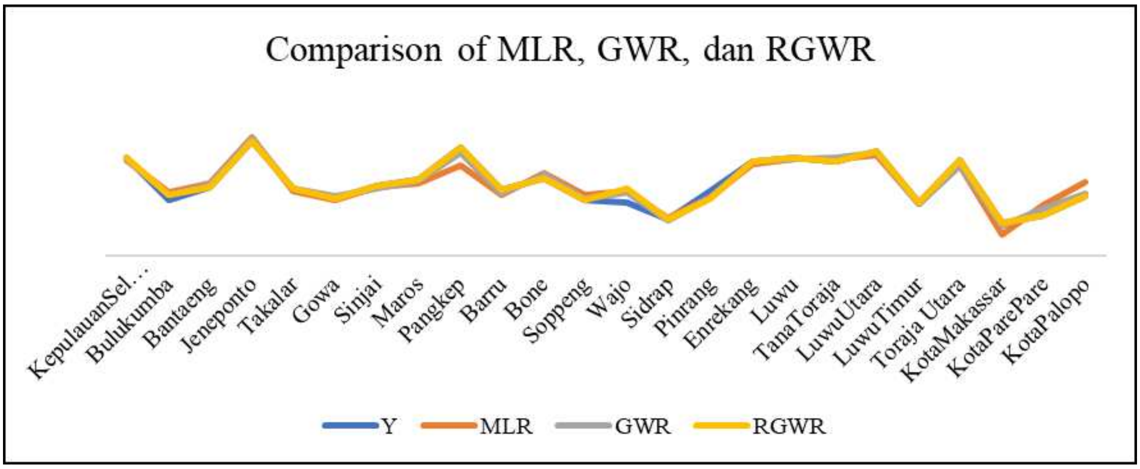

Fig. 1. Comparison of MLR, GWR, dan RGWR

Based on Table 4 and Fig. 1 above, it can be seen that the estimated percentage of poor population in South Sulawesi in 2019 ($\hat{Y}$) on the Robust Geographically Weighted Regression model is closer to the actual percentages of poor people. (Y). Next, the best model measurement needs to be done, namely the Mean Squares of Error (MSE) and Determination Coefficient (R^2) criteria. The MSE value with a degree of freedom of error of 14 and the R^2 value obtained are presented in Table 5.

Table 5. Comparison of the MSE and R^2 values

Mean Squares of Error Values		
MLR	GWR	RGWR
1.396537566	0.436862426	0.367206272
R^2 Values		
MLR	GWR	RGWR
90.84755496	93.86278363	97.51678603

From Table 5 it can be seen that the regression model that produces the smallest MSE value and the largest R^2 value is the Robust Geographically Weighted Regression model. Therefore, it can be concluded that the Robust Geographically Weighted Regression model is the best model to be used in estimating the percentage of the poor population in South Sulawesi in 2019 that is spatially diversified and contains outliers.

4. Conclusion

Based on the results obtained, it can be concluded that the Robust Geographically Weighted Regression model produces a different model estimate for each district/city in South Sulawesi. In addition, the Robust Geographically Weighted Regression model with the Least Absolute Deviation method gives the best estimates on poverty data in South Sulawesi in 2019 which is spatially diversified and contains outliers.

References

Montgomery, D. C., Peck, E. A. 1992. Introduction to Linear Regression Analysis, 2nd edition. New York: John Wiley & Sons, Inc.

Gujarati, D. 2007. Dasar-dasar Ekonometrika, Edisi ketiga, jilid 1. Julius A. Mulyadi, S.E, translator. Jakarta: Erlangga. Translation from: Essentials of Econometrics.

Erda, G. 2018. Pendugaan Model Regresi Terboboti Geografis dan Temporal Kekar Menggunakan Penduga-M. Thesis. Institut Pertanian Bogor.

Fotheringham, et al. 2002. Geographically Weighted Regression : The Analysis of Spatially Varying Relationships. John Wiley and Sons, Ltd. UK.

Wulandari, P., Djuraidah, A., Wigena, A. 2019. Robust Geographically Weighted Regression Modelingusing Least Absolute Deviation and M-Estimator. International Journal of Scientific Research in Science, Engineering and Technology. 6(1): 238-245.

Huang, et al. 2010. Geographically and Temporally Weighted Regression For Modeling Spatio-Temporal Variation In House Prices. International Journal of Geographical Information Science. 24(3): 383-401.

Wheeler, D.C. .2014. Geographically Weighted Regression. In Fischer M., Nijkamp P. (eds) Handbook of Regional Science. Springer, Berlin, Heidelberg.