Application Geographically Weighted Ridge Regression and Geographically Weighted Lasso in Case Correlation of Covariate

Reny Wulandari, Asep Saefuddin, Farit Mochamad Afendi

Abstract— Geographically Weighted Regression (GWR) is a method for nonstationary data due to geographic dependency. However, GWR does not handle a problem local multicolinearity. Geographically Weighted Ridge Regression (GWRR) and Geographically Weighted Lasso (GWL) are aimed to overcome the problem of local multicolinearity. The objectives was to compare among three methods using Locally Generated Revenue (LGR). Detection of local multicolinearity was measured by Variance Inflation Factor (VIF) of more 5 and local correlation. LGR in West Java Province methods of GWR, GWRR and GWL obtained different models in each region. The comparison used R^2 and RMSE. The result showed that GWL was better than the GWR and GWRR. Hence, to analyze LGR in West Java there a GWL was suggested.

Index Terms— Local Multicollinearity, GWL, GWR, GWRR, Spasial Heterogeinity.

1 Introduction

LINEAR regression models in statistics are often used to describe a relationship between the response variable with the explanatory variables (covariates). In a linear regression model there are some classical assumptions regarding multicollinearity. The problem produces and predictions unstable makes difficult interpretation. The regression model also assumes that the same coefficient of linear regression can be applied to all geographic locations. In the area-based models, regression models that apply globally well applied if there is no spatial variation between regions. In other words, the regression model can be applied if the relationship between the response variable and the explanatory variables do not depend on the region are called spatially stationary (Fotheringham et al 2002). The condition is known as spatial effects, which can be divided into two parts the spatial autocorrelation and spatial heterogeneity (Anselin 1988). The spatial effects can not be ignored in estimating the model. Ignoring the effect of spatial information on the data, the result produce different conclusions (Lesage 1997). Therefore, the effects of spatial autocorrelation and spatial heterogeneity must be considered in a model.

One modeling to handle the spatial effect is Geographically Weighted Regression (GWR). GWR development is one method of least squares regression models (OLS) are used to handle the problem of spatial heterogeneity is caused by a condition the location of the other locations are not the same. The Locally Generated Revenue (LGR) in the districts/cities in West Java is one example of data that is influenced by spatial effects (Anggraini 2016). In covariates amount of LGR data causes local multicolinearity on the model. So that multicollinearity on the covariates that affect the LGR causes interpretation becomes difficult in other words the model obtained infeasible. GWR can not resolve the problem local multicollinearity so the addition of Ridge Regression known as Geographically Weighted Ridge Regression (GWRR) and Least Absolute Shrinkage and Selection Operator (Lasso) known as Geographically Weighted Lasso (GWL) on the model is expected to overcome the multicollinearity and more effective in modeling.

On the basis of the background necessary to do an analysis of the various factors of the value of LGR cities and districts in West Java province. The purpose of this study was to compare the model GWRR and GWL to handle local multicollinearity on GWR models and indicated the factors that influenced the LGR value in the districts/cities in West Java Province.

2 Research Method

2.1 Data

The data used in this research is the data collecting of SUSENAS (a trimester national socio-economic survey in Indonesia) 2015, PODES (a census of village/ region potency) 2014 and the publications issued by BPS (Statistic Indonesia) West Java Province. The unit of observation in this study is 18 districts and 9 cities in West Java Province, while the coordinates of points in each district/city that is used to get the distance as forming the weighting matrix derived from www.google.com/maps. Response variable and the explanatory variables used are listed in Table 1.

2.2 Methods of Data Analysis

The stages of data analysis in this research are as follow:

1. Descriptive Analysis
   Descriptive analysis was performed to explore the general description of data pattern that aimed to get the appropriated next analysis.

- *Reny Wulandari is currently pursuing masters degree program in applied statistics in Bogor Agricultural University, Indonesia, PH +6285271678010. E-mail: renywulandari58@gmail.com*
- *Asep Saefuddin is Lecturer, Departement of Statistics, Bogor Agricultural University, Bogor, Indonesia*
- *Farit Mochamad Afendi is Lecturer, Departement of Statistics, Bogor Agricultural University, Bogor, Indonesia*

TABLE 1
LIST FOR OF VARIABLES

Variables	Explanation
Y	Locally Generated Revenue (Billion)
X_1	Population (Million)
X_2	Number of medium and large Manufacturing (Unit)
X_3	Gross Regional Domestic Product (GRDP) At Current Market Prices (Million)
X_4	Number of Restaurant (Unit)
X_5	Number of foreign and domestic tourist visitors (Thousand)
X_6	Number of Hotel (Unit)
X_7	Number of Market (Unit)
X_8	Number of Hospital (Unit)

2. Checking on spatial effect

a. Check Moran's I to detect spatial autocorrelation :

$$I = \frac{n\sum_s \sum_{j\neq s} W_{sj}(x_s - \bar{x})(x_j - \bar{x})}{(\sum_s \sum_{j\neq s} W_{sj})\sum_s (x_s - \bar{x})^2}$$

b. Check Breusch Pagan (BP) to detect spatial heterogeneity :

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

3. Perform GWR modeling stages as follows :

a. Choose bandwidth optimum (b) which minimize Cross Validation (CV) value :

$$CV = \sum_{s=1}^{n} \left(y_s - \hat{y}_{\neq s}(N)\right)^2$$

b. Set the weight matrix **W** (s) = diag [$w_1(s),\dots, w_n(s)$] for each location by using exponential kernel function.

$$w_j(s) = \exp\left(-\frac{d_{sj}}{b}\right)$$

With

$$d_{sj} = \sqrt{(v_s - v_j)^2 + (v_s - v_j)^2}$$

c. Estimate the GWR parameter for each location

$$\hat{\boldsymbol{\beta}}(s) = [\boldsymbol{X}^T\boldsymbol{W}(s)\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{W}(s)\boldsymbol{y}$$

so that is obtained a model of each location

$$\boldsymbol{y}(\boldsymbol{s}) = \boldsymbol{X}(\boldsymbol{s})\boldsymbol{\beta}(\boldsymbol{s}) + \varepsilon(\boldsymbol{s})$$

4. Detection of local multicollinearity in each location with local coefficient correlation and local variance inflation factor (VIF)

a. Local coefficient correlation for two variables in each location

$$r_{k,l}(s) = \frac{\sum_{j=1}^{n} w_{sj}(x_{kj} - \bar{x}_{ks})(x_{lj} - \bar{x}_{ls})}{\sqrt{\sum_{j=1}^{n} w_{sj}(x_{kj} - \bar{x}_{ks})^2 \sum_{j=1}^{n} w_{sj}(x_{lj} - \bar{x}_{ls})^2}}$$

b. Local VIF in each location : $VIF_k(s) = \frac{1}{1 - R_k^2(s)}$

5. Perform GWRR modeling to handle of local multicollinearity :

$$\hat{\boldsymbol{\beta}}(\boldsymbol{s}) = (\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{s})\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{W}(\boldsymbol{s})\boldsymbol{y}$$

6. Perform GWL modeling stages as follows :
   a. estimate the local shrinkage and bandwidth kernel exponential optimum by using CV stages as follows:
      i. calculate **W** used Euclidean distance matrix (d_{sj}) and bandwidth kernel exponential
      ii. for each location i, i= 1, 2,..., n
         - set $\mathbf{W}^{1/2}_{(i)}$= sqrt (diag($\mathbf{W}_{(i)}$)) and $\mathbf{W}^{1/2}_{(i)ii}$= **0** , that is set the (i,i) element of the diagonal weights matrix to 0 to effectively remove observation i
         - set $\mathbf{X}_w$=$\mathbf{W}^{1/2}_{(i)}\mathbf{X}$ and $\mathbf{Y}_W$= $\mathbf{W}^{1/2}_{(i)}\mathbf{Y}$ using the square root of the kernel weights $\mathbf{W}_{(i)}$ at location i.
         - Call Lars algorithm to get solution of lasso that minimizes the error Y_i and save this solution.
   a. estimate the final of parameter fit CV and base on of boundary shrinkage value.
7. Compare the RMSE and R^2 values of GWR, GWRR and GWL estimation results to get the best model.

$$RMSE = \sqrt{\frac{1}{n}\sum_{1=1}^{n}(y_i - \hat{y}_i)^2}$$

$$R^2 = {(JKT - JKG)}/{JKT}$$

3 RESULT AND DISCUSSION

3.1 Descriptive Analysis

The relationship between the response variable with covariates can be known from the resulting correlation coefficient. Correlation is used Pearson Correlation with alpha 0.01. In Table 2 show all covariates (explanatory variables) significantly the response variable at 5% significance level. So it can be used in research.

TABLE 2
PEARSON COEEFICEINT CORRELATION VALUE THE RESPONSE VARIABLE WITH EXPLANATORY VARIABLE

Variable	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Y	0.70	0.61	0.64	0.75	0.56	0.51	0.60	0.87
P- value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Figure 1 is a map of the distribution of the value of LGR districts/cities in West Java Province in 2015. The map shows the districts/cities that have a value of LGR the group lying side by side. This indication spatial effect on the value of LGR in the province of West Java that can incorporate aspects of the spatial model.

Table 3 shows the detection of the covariates correlated (multicollinearity) is shown with a value of Variance Inflation Factor (VIF). Testing criteria ie no covariates correlated if VIF <5. From the table, it contained VIF> 5 that is at the variable X1,X2,X5 and X7. It can be concluded the covariates are correlated.

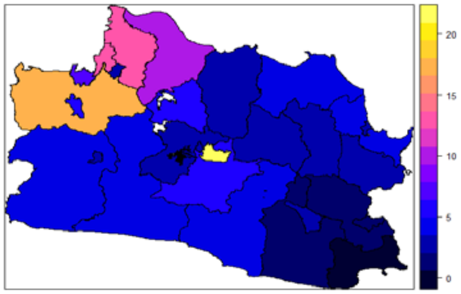

FIGURE 1
LOCALLY GENERATED REVENUE

TABLE 3
VIF VALUE OF DATA

Variables	VIF	Variables	VIF
X_1	5.93	X_5	5.19
X_2	9.01	X_6	3.03
X_3	3.28	X_7	5.91
X_4	4.07	X_8	4.25

3.2 Spasial Effect

Test results spatial autocorrelation Moran's I is obtained the value of the Moran's I by 0.336 with p-value of 0.00046 which is less than of 0.05 in order to obtain a decision reject H_0 and testing of spatial heterogeneity using test Breusch-Pagan (BP) with the value of BP at 18.079 with p-value of 0.02064 which is less than 0.05 real level in order to obtain a decision reject H_0 which means that there are autocorrelation and spatial heterogeneity in the data LGR in the districts/cities in West Java province in 2015.

3.3 Geographically Weighted Regression Model

Results from modeling GWR is a parameter estimator for each location of the observations in other words estimate of parameter GWR in each location. This happens because each district/city influenced by the relative condition of the district/city surroundings. Summary GWR parameter estimators for the overall observation is shown in Table 4.

TABLE 4
SUMMARRY OF GWR PARAMETER ESTIMATES

Estimates	Min	Mean	Max
b_0	-2.85	-2.62	-2.63
b_1	1.27	1.51	1.75
b_2	-0.00	-0.00	0.00
b_3	0.05	0.06	0.07
b_4	-0.00	0.00	0.00
b_5	0.00	0.00	0.00
b_6	0.00	0.01	0.02
b_7	-0.00	-0.00	-0.00
b_8	0.23	0.26	0.28

Local multicollinearity (covariates are correlated) is done by calculating the local correlation coefficient and local VIF each of the explanatory variables. The local correlation coefficient of variables in each location of the observations contained in Figure 2.

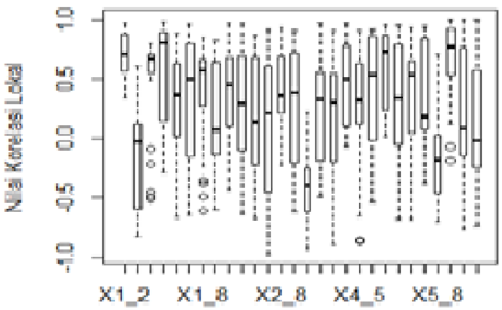

FIGURE 2
THE LOCAL COEFFICIENT CORRELATION OF VARIABLE EXPLANATORY

Table 5 shows that the value of the local VIF variable explanatory ranging from 2.58 until 14.78. With the addition of the weighting matrix in GWR models led to rising multicollinearity compared with the linear regression model. VIF> 5 that is contained in the variables X1, X2, X4, X5, X6, X7 and X8. Based on the value of local VIF can conclude that among independent variables are correlated.

TABLE 5
SUMMARY OF LOCAL VIF VALUE

Variables	Min	Mean	Max	Number of VIF value >5
X_1	3.73	5.47	7.35	16
X_2	6.05	9.68	14.78	27
X_3	2.61	3.67	4.37	0
X_4	3.02	4.56	6.23	14
X_5	5.21	6.04	7.50	27
X_6	2.56	3.68	5.22	1
X_7	4.76	5.98	8.55	22
X_8	3.68	4.47	5.35	5

3.4 Compare the Result Model

The value of R^2 and RMSE for each model are listed in Table 6. According to Table 6 GWL model produces value of R^2 are a maximum of 99.97% and the smallest of RMSE 0.079. This means that the model GWL good to overcome the spatial heterogeneity in the GWR model of LGR value. In addition, the GWL model is able to solve the problem correlated of covariates that can not be handled by GWR model.

TABLE 6
R^2 AND RMSE OF THE RESPONSE VARIABLE FOR THE GWR, GWRR AND GWL MODELS

Model	R^2	RMSE
GWR	97.84%	0.73
GWRR	85.19%	1.90
GWL	99.97%	0.08

3.5 Geographically Weighted Lasso Model

In the GWL modeling, each observation location has different models. A parameter value of shrinkage (si) GWL different models in each location. The advantages lasso compared to ridge regression, multiple regression of coefficient shrink to zero which automatically makes the variable coeffi

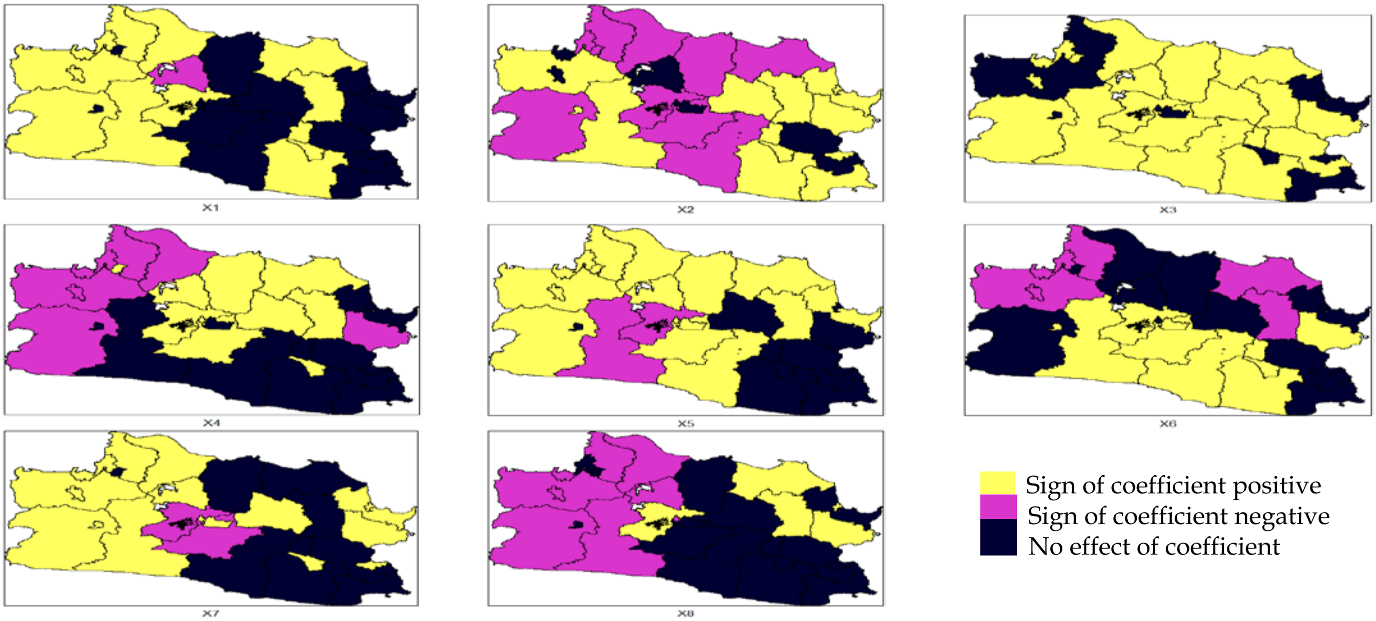
X1
X2
X3
X4
X5
X6
X7
X8
🟨 Sign of coefficient positive
🟪 Sign of coefficient negative
⬛ No effect of coefficient

FIGURE 3
DISTRIBUTION OF EACH INDEPENDENT VARIABLE IN GWL MODEL

cients corresponding to no effect on the model. Results summary of GWL model parameters using R 3.3.0 with shrinkage can be shown in Table 7.

TABLE 7
SUMMARRY OF GWL PARAMETER ESTIMATES, SHRINKAGE AND RESPONSE VARIABLE

Estimates	Min	Mean	Max
b_0	-5.27	0.07	6.20
b_1	-0.21	0.89	4.10
b_2	-0.00	0.00	0.02
b_3	0.00	0.04	0.19
b_4	-0.03	-0.00	0.01
b_5	0.00	0.00	0.00
b_6	-0.05	0.07	0.03
b_7	-0.31	0.16	0.47
b_8	-0.00	0.00	0.00
si	0.00	0.44	1.14
y	0.54	5.46	20.93

For more details distribution is presented maps any real independent variables which give positive and negative effect on the value of LGR in the district/city of West Java province in Fogure 3.

4 CONCLUSION

Based on the purpose of the research and the analyst can be concluded as follows:

1. The value of R^2 GWR , GWRR and GWL respectively were 97.87%, 85.19% and 99.79%. RMSE value GWR, GWRR and GWL are 0.727, 1.903 and 0.079. In GWRR models obtained value $\lambda = 0$, resulting in GWRR models no better to handle local covariates correlated on GWR method.
2. From the value of R^2 and RMSE concluded that GWL method shows better performance than models of GWR and GWRR handling correlated covariates and spatial heterogeneity in the data value LGR districts/cities in West Java Province in 2015. The variable explanatory that influenced for LGR value districts/cities in West Java Province are dominated by variables are the number of medium and large manufacturing, the number of markets and the number of foreign and domestic tourist visitor.

REFERENCES

[1] Anggraini A. 2016. Aplication of Geographically Weighted Regression in case Locally Generated Revenue in Central of Java Province. Bogor[ID] : Bogor Agricultural University.

[2] Anselin. 1988. Spatial Econometrics: Method and Models. Dordrecht (NL) : Kluwer Academic

[3] Anselin L. 1999. Spatial Econometrics. Bruton Center: University of Texas. TX 75083-0688.

[4] Fotheringham AS, Brunsdon C, Charlton M. 2002. Geographically Weighted Regression of Spatially Varying Relationships. England (GB): John Wiley and Sons.

[5] Lesage JP.1997. Spastial Econometrics. Department of Economics. University of Toledo.

[6] Statistic Indonesia (BPS) of West Java Province. Welfare Indicators 2016. Jakarta[ID] : BPS West Java

[7] Statistic Indonesia (BPS) of West Java Province. West java in Figures 2016. Jakarta[ID] : BPS West Java

[8] Wheeler D, Thiefelsdorf M. 2006. Multicolinearity and Correlation among Local Regression Coefficients in Geographically Weighted Regression. J Geograph Syst (2005) 7 : 161-187