ISSN Online 3025-1605
Journal of Statistical Methods and Data Science
Volume 01 Number 2 December 2023
https://jurnalmipa.unri.ac.id/jsmds

IMPLEMENTATION OF GEOGRAPHICALLY WEIGHTED LASSO (GWL) IN ANALYZING RICE PRODUCTION FACTORS IN INDONESIA

Reka Agustia Astari[1*], Megawati[2], Setyo Wahyudi[3]
[1,2,3] Department of Statistics, IPB University, Jawa Barat, 16680, Indonesia

***e-mail**: rekaagustiaastari@apps.ipb.ac.id

Article Info:
Received: 25-03-2024
Accepted: 04-04-2024
Available Online: 05-04-2024

Keywords:
global collinearity
geographically weighted regression
geographically weighted lasso
spatial heterogeneity

Abstract: *Geographically Weighted Lasso (GWL) is a combination of two regression methods, namely Geographically Weighted Regression (GWR) and Least Absolute Shrinkage Selection Operator (LASSO). Both methods have their own uses. GWR is a regression that takes into account the geographical location aspect because the spatial heterogeneity test is not met. LASSO is a regression method to overcome multicollinearity in the data. The two problems are simultaneously contained in one regression model, namely the GWL method. This study will analyze the factors that affect rice production in 34 provinces in Indonesia by applying and interpreting the results of the Geographically Weighted Lasso method. The results of the analysis show that the coefficient of determination of the GWL model is 0.9703 so it can be concluded that the explanatory variables in this study can that the global level of rice production in each province in Indonesia is 97.03%.*

1. INTRODUCTION

Statistical methods are often used as a tool to determine the relationship between variables by forming a model that is appropriate in describing the characteristics of the data. As in linear regression models that are able to describe the relationship between explanatory variables and response variables. Looking at the relationship between variables in spatial data can be done with spatial statistics methods. Spatial data is geographically oriented data and has a certain coordinate system as its reference base, so that it can be presented in a map (Yulita, 2016).

The problem that is often found in spatial data is the variety that is not always homogeneous at each observation location or called spatial heterogeneity. Spatial heterogeneity can be caused by several things such as differences in geographical conditions, socio-culture, and economic policies that vary in each location. This will be a problem if spatial data is still analyzed using the Least Squares Method (LSM) in estimating its parameters, because it can cause the variance of the estimates to be large. To overcome this problem, a method is needed that is able to overcome the heterogeneity of variance in spatial data to form a more efficient model (Yulita, 2016).

Several previous studies on Geographically Weighted Regression (GWR) have been conducted including those conducted by Setiyorini et al. (2017) in their research on poverty in Java concluded that the Geographically Weighted Lasso (GWL) method is better than the GWR method on spatial data containing multicolinierity. Furthermore, the geographical layout of a region will produce different modeling, this is because differences in geographical location will affect the potential owned or used by a region (Pamungkas et al., 2016).

GWR modeling can overcome the problem of heterogeneity by exploring spatial diversity (Fotheringham et al, 2002) (Bangun & Meimela, 2020). In addition, according to Wheeler (2009) that there are problems that usually arise in GWR, namely local collinearity (local multicollinearity) in the estimated coefficients, which can increase the variance of the estimated regression coefficients, to overcome this Wheeler (2009) proposed the GWL method which is a development of GWR by applying the lasso technique in its estimation so that the estimated results obtained become more stable. Based on this, GWL will be applied to modeling rice production factors in Indonesia.

According to the Badan Pusat Statistik (2020), rice production in Indonesia was 54.60 million tons of dry mill rice in 2019, a decrease of 4.60 million tons (7.76%) compared to 2018. In 2019, a relatively large increase in rice production occurred in the provinces of West Kalimantan, DI Yogyakarta, and South Kalimantan. A relatively large decrease in 2019 occurred in the provinces of South Sulawesi, Central Java, East Java, West Java, and South Sumatra. So that rice production data with its two main factors, namely harvest area and productivity, is an interesting material to be studied more deeply (Astuti et al., 2023). This research will explain the application of the GWL model to determine the factors that affect production in each province in Indonesia. Based on the literature study, this research will use explanatory variables in the form of rice production, harvest area, amount of rainfall and farmers.

2. LITERATURE REVIEW

2.1. Spatial Dependence and Diversity

The existence of location elements in data results in spatial effects so that analysis using global regression cannot be done because the Gauss-Markov assumption is violated (Anselin, 1988). Examination of spatial effects is done through testing for spatial dependence and spatial heterogeneity. Spatial dependence is measured by spatial autocorrelation statistics that describe the similarity between adjacent observations. Spatial dependency testing is done by finding the Moran index. Septiyanto & Tusianti (2020) explain that a positive Moran's I value indicates the presence of spatial auto correlation with a clustering pattern in an area. Fotheringham in Ramdhani (2016) spatial heterogeneity is a condition in which the global regression model is unable to explain between variables because of the diversity of characteristics between regions, spatial heterogeneity testing is carried out through Breusch-Pagan testing (Anselin, 1988).

2.2. Geographically Weighted Regression (GWR)

Geographically Weighted Regression (GWR) is a development of the global regression method by taking into account location aspects and fulfilling spatial diversity assumptions. Each location produces different parameter estimation coefficients. The GWR model depends on the weighting used. According to Fotheringham et al. (2002) and Wheeler (2009), GWR Model at location $i = 1,2, \dots, n$, namely:

$$y(i) = \boldsymbol{X}(i)\boldsymbol{\beta}(i) + \varepsilon(i) \quad (1)$$

with $y(i)$ response variable at location $i, \boldsymbol{X}(i)$ explanatory variable in location to $\boldsymbol{i}, \boldsymbol{\beta}(i)$ regression coefficient is located to $i, \varepsilon(i)$ error at location to i. The estimation of the regression coefficient at location *i* is:

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

with $X = [X^T(1); \dots X^T(n)]^{-1}$ explanatory variable matrix $W(i) = diag[w_1(i), \dots, w_n(i)]$ The diagonal of the weighting matrix calculated for each location i and y is the response variable, and $\hat{\beta}(i) = (\hat{\beta}_{i0}, \hat{\beta}_{i1}, \hat{\beta}_{ip})^T$ local regression coefficient at location *i* for the explanatory variable *p*.

2.3. Geographically Weighted Lasso (GWL)

Wheeler (2009) applied the Least Absolute Shrinkage Selection Operator (LASSO) technique to the GWR Model, which was then called the GWL model to overcome the problem of spatial heterogeneity and local multicollinearity. GWL is a method used to overcome regional diversity caused by different locations and conditions between regions as well as the emergence of local multicollinearity (Wang & Zuo, 2020). GWL is also a construction model between GWR and LASSO (Yuliana & Saputro, 2017). In solving GWL, the Least Angle Regression (LARS) algorithm (Efron, et al. 2004) is used which is modified by adding a weighting matrix to the variables, with the following algorithm:

a. Estimating optimal kernel bandwidth with cross validation (CV):
   Calculate the weighting matrix *W* of size *n x n* from the distance matrix in equation (3), with $[W_1(i), \cdots, W_n(i)]$ is a diagonal matrix $W(i)$ which is defined in equation (2).
b. For each location to $i, i = 1, \cdots, n$:
   1) Forming $\boldsymbol{W}^{\frac{1}{2}}(i) = \sqrt{diag(\boldsymbol{W}(i))}$
   2) Forming $X_w = \boldsymbol{W}^{\frac{1}{2}}(i)X$ dan $X_y = \boldsymbol{W}^{\frac{1}{2}}(i)y$ using the square root of the kernel weighting $W(i)$ at location *i*.
   3) Call the algorithm X_w, y_w, save a series of Lasso solutions, which minimize the error y_i, and save these soluions.
c. Stop when there is only a small change in the φ estimate, then save the φ estimate.

2.4. Spatial Weighting Function

In Wheeler (2009), the weighting matrix $W(i)$ with kernel exponential function weights between location j and location i. The location is calculated by:

$$w_j(i) = \exp\left(-\frac{d_{ij}}{\varphi}\right) \quad (3)$$

Where d_{ij} is the distance between location point i and location j obtained from the euclidean distance, while φ is the kernel bandwidth at location i. The optimal bandwidth selection is obtained by minimizing the CV value at all locations.

2.5. Cross Validation (CV)

Cross validation (CV) or rotational estimation is a model validation technique to assess how the results of statistical analysis will generalize to an independent data set. This technique is primarily used to perform model predictions and estimate how accurate a predictive model is when run in practice. One of the techniques of cross validation is k-fold cross validation, which breaks the data into k parts of the data set with the same size. The use of k fold cross validation is to eliminate bias in the data. Training and testing are performed k times (Bramer, 2007).

Fotheringham et al. (2002), the CV formula is as follows:

$$CV\ (h) = \sum_{i=1}^{n} [y_i - \hat{y}_{\neq i}(h)]^2 \quad (4)$$

where $\hat{y}_{\neq i}(h)$ is the expected value for $y_i$with the i observation omitted from the prediction process, and the optimum bandwidth (h) will be obtained by iterating until the minimum CV is obtained. Therefore, the estimated response variable is predicted by $\hat{y}(i) = X(i)\hat{\beta}(i)$ (Wheleer, 2009).

3. METHODOLOGY

3.1. Data and Sources

The data used in the modeling is secondary data from the publication of the Ministry of Agriculture of the Republic of Indonesia and the publication of the Badan Pusat Statistika. The data is rice production data taken from 34 provinces in Indonesia in 2019. The variables used are distinguished by two variables, namely dependent and independent variables. The dependent variable used is rice production in each province in Indonesia in 2019 (tons). The independent variables used are factors that are thought to affect rice production in Indonesia. The variables used are listed in the following table:

Table 1. Research variable

Variable	Description
Y	Rice Production (Ton)
X_1	Harvested Area (Ha)
X_2	Total Rainfall (mm)
X_3	Fertilizer (ton)
X_4	Farmers (people)

3.2. Data Analysis

The rice production data contains spatial effects that have heterogenous versity at each observation location, so to explore the spatial diversity, GWR modeling will be carried out, with the following model:

$$y(i) = \beta_0(i) + \sum_{k=1}^{4} \beta_k(i) + \varepsilon(i), \text{ dengan } k = 1,2,\cdots 4 \; dan \; i = 1,2,\cdots 4$$

where $y(i)$ is the response variable at the ith location for rice production data, $X_k(i)$ is the kth explanatory variable at the i location, , $\beta_k(i)$ is the local parameter coefficient for the i location and the remainder $\varepsilon(i) \sim N(0, I\sigma^2)$. In addition, because it uses several explanatory variables that can cause local multicollinearity, each location has different parameter coefficients, so the method of estimating the coefficients of the local model formed is done by GWL, namely by selecting variables at each location. Variables that have coefficients equal to zero will be selected for certain areas, so that the estimated results obtained become more stable. The stages of the analysis are as follows:

a. Perform multicollinearity and autocorrelation tests
b. Perform spatial heterogeneity test with Breusch Pagan test.
c. Determining the distance matrix with euclidean distance.
d. Determining the optimum bandwidth based on minimum CV with exponential kernel.

e. Determining the local collinearity of the local model formed.
f. Selecting the coefficient of rice production model with GWL.
g. Determining the goodness of the GWL model based on the comparison seen from the RMSE value.

Stages of analysis in modeling and selection using the modified LARS algorithm (Efron et al, 2004).

4. RESULT AND DISCUSSION

4.1. Multicollinearity Test

Descriptive Analysis of Rice Production data in Indonesia in 2019 is shown in table 2 below:

Table 2. Descriptive Analysis

Variable	Mean	Stdev	Max	Min
Rice Production (Ton)	1.606.001	2.674.447	9.655.654	1.151
Harvested Area (Ha)	314.055	470.167	1.702.426	356
Total Rainfall (mm)	1.950,22	689	4.072,70	637,60
Fertilizer (ton)	270.119	520.621	2.637.877	50
Farmers (people)	937.309	1.226.304	6.054.066	14.098

Based on the table above, it is found that the average rice production in Indonesia in 2019 is 1,606,001 tons with a standard deviation of 2,674,447 and the average harvest area of rice production in Indonesia reaches 314,055 ha. Regression parameter estimates on rice production in Indonesia are shown in table 3 below:

Table 3. Regression Parameter Estimation

Variable	VIF
Harvested Area (Ha)	7.318031
Total Rainfall (mm)	1.059607
Fertilizer (ton)	6.239974
Farmers (people)	1.957322

The multicollinearity test results obtained using the VIF method show that there is no multicollinearity because there are no variables that affect each other with the VIF value of each variable less than 10.

4.2 The Autocorrelation test is performed with the Breusch Godfrey test.

The hypothesis for testing autocorrelation is as follows:
H_0 : There is no autocorrelation.
H_1 : There is autocorrelation.

Table 4. Autocorrelation with Breusch Godfrey

Breusch Godfrey	*p-value*	conclusion
1.7508	0.1858	Accept H_0

Breusch-Godfrey test results with a p-value of 0.1858. H_0 is accepted because the p-value of 0.1858 > 0.05, so it can be concluded that there is no autocorrelation in rice production data in Indonesia.

4.3 Testing for spatial heterogencity

The hypothesis for testing spatial heterogeneity using the Breusch-Pagan test is as follows:
H_0: there is no heterogeneity between locations
H_1 : there is heterogeneity between locations

Table 5. Autocorrelation with Breusch-Pagan

Breusch-Pagan	***p-value***	**conclusion**
13.965	0.007408	Reject H_0

Based on the results of the spatial heterogeneity test, the p-value is 0.002288 with a significance level of α = 5% then H_0 is not accepted, so it can be concluded that there is an effect of spatial heterogeneity in the residuals. Before analyzing the GWR model, first determine the optimum bandwidth using the minimum CV value by determining the fixed kernel gaussian function at each location. The optimum bandwidth value and CV score are presented in table 6 below:

Table 6. Autocorrelation with Breusch-Pagan

Bandwidth Optimum	***CV Score***
0,381966	$4{,}356178 \times 10^{12}$
0.618034	$4{,}450313 \times 10^{12}$
0.236068	$3{,}916639 \times 10^{12}$
0.145898	$3{,}495054 \times 10^{12}$
0.090169	$5{,}497643 \times 10^{12}$
0.180339	$3{,}560044 \times 10^{12}$
0.144263	$3{,}503679 \times 10^{12}$
0.158367	$3{,}455303 \times 10^{12}$
0.166760	$3{,}468337 \times 10^{12}$
0.159146	$3{,}45538 \times 10^{12}$
0.158499	$3{,}455297 \times 10^{12}$
0.158549	$3{,}455297 \times 10^{12}$
0.158580	$3{,}455297 \times 10^{12}$
0.158540	$3{,}455297 \times 10^{12}$

After obtaining the bandwidth value, then determine the weighting matrix for each location by calculating the euclidean distanceat each location. So that the optimum bandwidth value is 0.158540 and CV score is 3.455297 × 1012.

Parameter estimates of response variables that have a significant effect on rice production factors in each province in Indonesia are presented in table 7:

Table 7. Parameter significance results by province

No.	Province	Significant variables
1.	Aceh, Sumatera Utara, Sumatera Selatan, Lampung, Kepulauan Bangka Belitung, Kepulauan Riau, DKI Jakarta, Jawa Barat, Jawa Tengah, DI Yogyakarta, Jawa Timur, Banten, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku, Maluku Utara, Papua Barat, Papua	Harvested Area and Fertilizer
2.	Sumatera Barat, Riau, Jambi, Bengkulu, Bali, Nusa Tenggara Barat, Nusa Tenggara Timur	Harvested Area

Based on the results obtained significant variables that affect rice production in each province in Indonesia is the variable area of harvest. Table 8 shows the RMSE and R^2 values of the Lasso and GWL models.

Table 8. Parameter significance results by Province

Model	RMSE	R^2
Lasso	51388.64	0.9587223
GWL	453818.4	0.9703338

5. CONCLUSIONS

The results obtained from the comparison of Lasso and GWL by using the RSME and R^2 values for modeling rice production in Indonesia is the GWL model because it has a smaller RSME value compared to Lasso, It also shows that the coefficient of determination of the GWL model is 0.9703 so it can be concluded that the explanatory variables in the study can be said that the global level of rice production in each province in Indonesia is 97.03%. Furthermore, the independent variables that affect rice production data in Indonesia are divided into 2 groups based on the province, namely the harvest area and fertilizer groups and only the harvest area alone, namely in the provinces of Sumatera Barat, Riau, Jambi, Bengkulu, Bali, Nusa Tenggara Barat, Nusa Tenggara Timur.

ACKNOWLEDGMENT

Thank you and especially to our parents, siblings and beloved family for all the prayers, efforts, sacrifices, and love that have been given. Thank you for being a witness to the process of my brother's life. Furthermore, thank you to the lecturers and teaching staff at IPB University in general and the Statistics department in particular. As well as proud friends of Batch 59. Last but not least, to myself. Thank you for struggling so far.

REFERENCES

Anselin L. (1988). Spatial Econometrics : Methods and Models, Dordrecht: Kluwer Academic Publisher.

Astuti, F., Bekti, R. D., Arianita, A., Keliat, B., & Inya, T. (2023). Vol . 16 No . 1 September 2023 ISSN : 1979-8415 Analisis Produksi Padi Di Indonesia Menggunakan Model Regresi Robust Estimasi M , Estimasi S Dan Estimasi MM ISSN : 1979-8415. 16(1), 33–40.

Badan Pusat Statistik. (2020). Statistik Indonesia Tahun 2020. Jakarta : Badan Pusat Statistik.

Bangun, R. H. B., & Meimela, A. (2020). Pemetaan Kemiskinan Melalui Pendekatan Geographically Weighted Lasso. *Jurnal Ekonomi Indonesia*, *9*(3), 233–246. https://doi.org/10.52813/jei.v9i3.58

Efron B, Hastie T, Johnstone I, Tibshirani R. (2004). *Least Angle Regression. The Annals of Statistics* 32(2): 407-451.

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

Lestari, S. S. S., Meimela, A., & Revildy, W. D. (2021). Analisis Faktor Tingkat Pengangguran Terbuka Dengan Metode Geographically Weighted Lasso. *Seminar Nasional Official Statistics*, *2020*(1), 1286–1293. https://doi.org/10.34123/semnasoffstat.v2020i1.693.

Pamungkas, R. A., Yasin, H., & Rahmawati, R. (2016). Perbandingan Model GWR Dengan Fixed dan Adaptive Bandwith Untuk Presentase Penduduk Miskin Di Jawa Tengah. *Jurnal Gaussian, 5*(3), 535–544. http://ejournal-s1.undip.ac.id/index.php/gaussian.

Septiyanto, W. G., & Tusianti, E. (2020). Analisis spasial tingkat pengangguran terbuka di Provinsi Jawa Barat. Jurnal Ekonomi Indonesia, 9 (2), 119-131.

Setiyorini, A., Suprijadi, J., & Handoko, B. (2017). Implementations of geographically weighted lasso in spatial data with multicollinearity (Case study: Poverty modeling of Java Island). AIP Conference Proceedings, 1827(1), 020003. doi: https://doi.org/10.1063/1.4979419.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via The Lasso. Journal of the Royal Statistical Society B 58(1): 267-288.

Wang, J., & Zuo, R. (2020). Assessing geochemical anomalies using geographically weighted lasso. Applied Geochemistry, 119 , 104668.

Wheeler DC. (2009). Simultaneous Coefficient Penalization and Model Selection in Geographically Weighted Regression: The Geographically Weighted Lasso. Journal of Environment and Planning A 41 (3): 722-742.

Yuliana, & Saputro, D. R. S. (2017). Algoritme least angle regression untuk model geographically weighted least absolute shrinkage and selection operator. Paper pre- sented at Seminar Matematika dan Pendidikan Matematika UNY ,pp.139-144,Yogyakarta, November 11, 2017.

Yulita, T. (2016). Pemodelan Geographically Weighted Ridge Regression dan Geographically Weighted Lasso pada Data Spasial Dengan Multikolinieritas. https://repository.ipb.ac.id/handle/123456789/80158.