JURNAL MATEMATIKA, STATISTIKA DAN KOMPUTASI e-ISSN: 2614-8811
Published by Departement of Mathematics, Hasanuddin University, Indonesia
https://journal.unhas.ac.id/index.php/jmsk/index
Vol. 19, No. 2, January 2023, pp. 342-350
DOI:10.20956/j.v19i2.23481

Modeling of COVID-19 Cases in Indonesia with the Method of Geographically Weighted Regression

Samsul Arifin [1]*, Erna Tri Herdiani[2]*

**Department of Statistics, Hasanuddin University, Makassar, Indonesia, 90245*

E-mail: *samsul.arrr@unhas.ac.id[1], herdiani.erna@unhas.ac.id[2]*

Received: 12 October 2022; Accepted: 9 November 2022; Published: 5 January 2023

Abstract

The COVID-19 pandemic has spread to all corners of the world, including Indonesia. Various factors affect the spread of COVID-19 cases in an area so that the government and the community can make prevention and control efforts so that this pandemic does not spread. This study aims to model the number of COVID-19 cases in Indonesia using the Geographically Weighted Regression (GWR) method, which develops a linear regression model. The GWR model uses weights based on the location of each observation so that the model is obtained for that location. Determine the weighting on the bandwidth. Optimum bandwidth selection is obtained by minimizing the value of Cross-Validation (CV). The GWR model using a fixed bisquare kernel weighting function has an optimum bandwidth of 0.999948 with a minimum CV value of 397.076.128 with a coefficient of determination (R^2) of 85.1 %. The results show that the number of positive cases positively correlates with the number of patients who died from COVID-19. In contrast, the number of recovered patients negatively correlates with the number of patients who died from COVID-19.

Keywords: COVID-19 Pandemic, Geographically Weighted Regression, Cross-Validation, Coefficient of Determination

1. INTRODUCTION

The Indonesian Ministry of Health stated that Coronavirus is a large family of viruses that cause disease in humans and animals. In humans, it usually causes respiratory tract infections, ranging from the common cold to serious diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). Coronavirus is known to have originated in the city of Wuhan, China, in December 2019 [14], later named Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-COV2), and caused Coronavirus Disease-2019 (COVID-19). This new type of virus even causes a pandemic status, which includes spreading the virus in a wide geographical area to all corners of the world, including Indonesia. The spread of this

virus can be said to be very fast, and no one has escaped exposure to this pandemic. The following is a graph of positive cases of COVID-19 in Indonesia until the end of 2021.

As of December 31, 2021, the baseline confirmed positive was 4,262,697 cases, with details of as many as 3,743 patients being treated, 4,114,744 patients recovered, and 144,210 patients died. As many as 80% of positive cases of COVID-19 in Indonesia came from the island of Java, DKI Jakarta Province, which has the most COVID-19 cases, with 218,684,767 confirmed. Many experts have also researched the spread of COVID-19, including Ogundokun et al. [9], predicting confirmed cases of COVID-19 in Nigeria. Furthermore, Fitrial et al. [4] modeled the number of COVID-19 cases in Indonesia using the Poisson and negative binomial regression approaches. Sameni [12] carried out mathematical modeling of epidemic diseases in COVID-19 cases, and Fajar [3] carried out parametric modeling of the growth curve of the COVID-19 epidemic in Indonesia.

The spread of COVID-19 has spread to every province in Indonesia. A speedy distribution from one location to another indicates a spatial effect in the modeling. One method used in spatial analysis is Geographically Weighted Regression (GWR). This technique brings the framework from a simple regression model to a weighted regression model [5]. Ma, Xue, and Hu [7] conducted economic data modeling using geographically weighted regression models. Song et al. [13] Conducted GWR modeling to see the properties of the soil. Putra et al. [11] conducted a geographically weighted regression modeling with a kernel function on the open unemployment rate in East Java. The linear regression model only produces parameter estimators that apply globally, while in the GWR model, local parameter estimators are generated for each observation location [10]. Based on this, this research will focus on modeling the number of COVID-19 cases in Indonesia using Geographically Weighted Regression (GWR).

2. METHODOLOGY

The data used in this study is the number of positive cases of COVID-19 in Indonesia, sourced from https://www.kaggle.com. The method used in this study is the Geographically Weighted Regression method.

2.1 Linear Regression Analysis

Linear regression analysis is a method used to express the relationship between a dependent variable or response (Y) with several independent variables or predictors (X). Multiple linear regression model for k predictor variables and the number of observations as many as n can be written in equation (2.1) as follows:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i \qquad 2.1$$

Where: Y_i is the dependent variable, $\beta_0, \beta_1, \ldots, \beta_p$ are Unknown parameters $X_{1i}, X_{2i}, \ldots, X_{pi}$ is the Predictor variable, and ε_i Is the error term for observation $i(i = 1,2, \ldots, n)$.

In the multiple linear regression model, if there are n observations, the model can be arranged in the form of a matrix:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{p1} \\ 1 & X_{21} & X_{22} & \cdots & X_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Where: $\boldsymbol{Y}$ is the response variable vector of size $nx1$, $\boldsymbol{X}$ is the independent variable matrix with nxp size, $\boldsymbol{\beta}$ is the parameter vector $px1$, and $\boldsymbol{\varepsilon}$ is the error vector $nx1$.

The parameter $\boldsymbol{\beta}$ is estimated using the ordinary least square (OLS) method by minimizing $\boldsymbol{\varepsilon'\varepsilon}$ by decreasing $\boldsymbol{\varepsilon'\varepsilon} = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})$ to $\boldsymbol{\beta}$. The derivative result is then equated with zero so that the Estimator $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$ is obtained.

2.2 Geographically Weighted Regression

Geographically Weighted Regression (GWR) is a development of linear regression which applies locally to GWR. This model calculates the parameters at each observation location so that each location will have a different regression model [8]. The model form of GWR is as in equation (2.2) as follows:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i)x_{ik} + \varepsilon_i \qquad 2.2$$

Where: y_i is Dependent variable at location i for $i = 1,2,\dots,n$, x_{ik} is an independent variable at the location, (u_i, v_i) is Longitude latitude coordinates of the i-th point in a geographic location, $\beta_k(u_i, v_i)$ Is the regression coefficient at each location, ε_i Is Error that is assumed to be identical, independent, and generally distributed with zero mean and constant variance σ^2.

With the Weighted Least Square (WLS) method, the parameter estimator at location i is formulated as equation (2.3) as follows:

$$\hat{\beta}_i = (X'W(i)X)^{-1}X'W(i)y \qquad 2.3$$

where: $W(i) = diag[w_1(i), w_2(i), \dots, w_n(i)$. $W(i)$ is a diagonal matrix of size nxn, which is the spatial weighting matrix whose values for the diagonal elements are determined by the proximity of the i-th location to other locations.

2.3 Geographically Weighted Regression

In the GWR model, a weighting matrix shows the close relationship between locations. In the GWR model, there are several weighting functions [5], such as

a. $W_j(u_i, v_i) = 1$, for all i and j. This weighted GWR model will produce an ordinary regression model, where each data in all locations is given the same weight, namely 1, regardless of its location or distance from other locations.
b. $W_j(u_i, v_i) = exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{h}\right)^2\right]$, where d_{ij} The distance from location i to location j and h is the optimum bandwidth fixed and the same at all locations. This function is called the fixed Gaussian kernel function.
c. $W_j(u_i, v_i) = \left[1 - \left(\frac{d_{ij}}{h}\right)^2\right]^2$, if $d_{ij} < h$, and $W_j(u_i, v_i) = 0$ for $d_{ij} \geq h$. This function follows the form of a weighted kernel and is commonly called a fixed Bisquare kernel function.
d. $W_j(u_i, v_i) = exp\left[-\frac{R_{ij}}{h}\right]$, where R_{ij} Is the rank (rank) distance from location-i to location-j for $j = 1,2,\dots,n$. The closest distance will result in the value of $W_j(u_i, v_i)$ Approaching one will decrease with increasing distance from location i to location j.

In this study, the weighting function is the fixed gaussian kernel function and the fixed Bisquare kernel function.

2.4 Optimum Bandwidth Selection

The selection of the optimum bandwidth affects the accuracy of the parameter estimation results [2]. The method used is to use Cross Validation which is written in equation (2.4) as follows:

$$CV = \sum_{i=1}^{n}(y_i - \hat{y}_{\neq i}(h))^2 \quad 2.4$$

where; y_i Is the Observation value of the response variable and $\hat{y}_{\neq i}$ Estimator where every observation at location i is removed from the estimation process. The selection of the optimum bandwidth is obtained from an iteration process that produces a minimum CV value.

2.5 Spatial Dependency and Spatial Heterogeneity

Spatial data is data with spatial dependencies and spatial diversity characteristics. One way to determine the existence of spatial dependencies between locations is to perform a spatial autocorrelation test using the Moran index statistic [6]. The spatial dependency measurement tool uses the Moran Index, which can be written in equation (2.5) as follows:

$$Z = \frac{I - E(I)}{\sqrt{Var(I)}} \quad 2.5$$

Where: Z is the Moran index test statistic value, and I is the Moran index value.

The spatial diversity that is a requirement in GWR modeling is spatial heterogeneity. o detect the presence or absence of spatial heterogeneity in the model, the Breusch-Pagan test [1] is written in equation (2.6) as follows:

$$BP = \frac{1}{2}b^T A(A^T A)^{-1}A^T b \quad 2.6$$

Where: b is Vector of size 1xn with $b_i = \frac{e_i^2}{\sigma^2} - 1$, e^i Residual at observation i, σ^2 is the variance of residual, and A is the Vector of the normalized response variable for each observation

3. RESULT AND DISCUSSION

Data on the development of COVID-19 cases until the end of December 2021 is given in Table 3.1.

Table 3.1. Characteristics of COVID-19 Cases in Indonesia

Positives	Recovered	Died
4.262.697	4.114.744	144.210

COVID-19 has spread to every province in Indonesia until December 2021. There were 4,262,697 positive cases recorded, 4,114,744 patients recovered, and 144,210 patients died.

3.1 Linear Regression Model

The Linear regression model is a global model using OLS estimation. The estimation results from the global regression are given in Table 3.2

Table 3.2. Estimation Results of Linear Regression Model

Variable	Estimation	T	P-Value	R-Square
Intercept	469.002	0,908	0,371	
Positives (x_1)	0,838	7,543	0,000	84,2 %
Recovered (x_2)	-0,813	-7,290	0,000	

Based on the results of the F test, the statistical value of F = 82.44 with a p-value of 0.000 means that the predictor variables are simultaneously significant to the response variable. In comparison, the partial test results in the table above show that the number of positive cases and recovered patients are significant to the number of COVID-19 patients who died in Indonesia. Based on this global regression model, spatial dependencies and spatial heterogeneity tests will be carried out, which are assumptions in the GWR model. The test results using the Moran and Breusch-Pagan indexes are given in Table 3.3.

Table 3.3. Dependency Test and Spatial Heterogeneity

Test	P-Value	Result
Moran Index	0.000	Significant
Breusch-Pagan	0,000	Significant

The value of the Moran and Breusch-Pagan Index test results in a significant p-value of =5%. So it can be concluded that there are spatial dependencies and spatial heterogeneity. With this assumption fulfilled, global modeling cannot be used because it will cause inefficient parameter estimation, so local modeling through GWR is needed by considering spatial effects on the data.

3.2 GWR Model with Fixed Gaussian Kernel Function

Optimum bandwidth selection is obtained by minimizing the value of Cross-Validation (CV). By using the fixed gaussian kernel function, the CV calculation results are given in Table 3.4.

Table 3.4. Optimum Bandwidth Selection

Bandwidth	CV
0,381966	668.472.235
0,618034	571.600.803
0,763932	509.913.843
0,854102	458.185.252
0,909830	453.140.816
⋮	⋮
0,999948	397.076.128

The optimum bandwidth is 0.999948, with a minimum CV value of 397.076.128. GWR modeling uses a weighting matrix for each location based on the optimal bandwidth. The summary of parameter estimates in GWR modeling is given in Table 3.5.

Table 3.5. GWR Model with Fixed Gaussian Kernel Function

Variable	Minimum	Median	Maximum	R-Square
Intercept	407,075	418,868	511,810	
Positives (x_1)	0,820	0,855	0,860	84,4 %
Recovered (x_2)	-0,834	-0,830	-0,795	

The results show that the number of positive cases positively correlates with the number of patients who died from COVID-19. In contrast, the number of recovered patients negatively correlates with the number of patients who died from COVID-19.

3.3 GWR Model with Fixed Bisquare Kernel Function

Optimum bandwidth selection is obtained by minimizing the value of Cross-Validation (CV). By using the fixed bisquare kernel function, the CV calculation results are given in Table 3.6.

Table 3.6. Optimum Bandwidth Selection

Bandwidth	CV
0,381966	1.389.418.745
0,618034	1.005.790.334
0,763932	834.023.618
0,854102	723.189.072
0,909830	710.104.218
⋮	⋮
0,999955	582.119.244

The optimum bandwidth is 0,999955, with a minimum CV value of 582.119.244. GWR modeling uses a weighting matrix for each location based on the optimal bandwidth. The summary of parameter estimates in GWR modeling is given in Table 3.7.

Table 3.7. GWR Model with Fixed Gaussian Kernel Function

Variable	Minimum	Median	Maximum	R-Square
Intercept	223,089	316,459	603,022	
Positives (x_1)	0,719	0,885	0,914	85,1 %
Recovered (x_2)	-0,889	-0,859	-0,691	

The results show that the number of positive cases positively correlates with the number of patients who died from COVID-19. In contrast, the number of recovered patients negatively correlates with the number of patients who died from COVID-19.

Modeling the number of COVID-19 cases in Indonesia using the fixed bisquare kernel weighting function is better because it has a higher coefficient of determination of 85.1%. The local model for each region can be seen in table 3.8.

Table 3.8. Local model for regions in Indonesia

No	Province	Intercept	β_1	β_2	R^2
1.	Aceh	223.089	0.914	-0.889	0.848
2.	Bali	346.805	0.875	-0.849	0.852
3.	Banten	259.175	0.906	-0.880	0.848
4.	Bengkulu	240.758	0.910	-0.885	0.848
5.	DI Yogyakarta	291.139	0.896	-0.871	0.849
6.	DKI Jakarta	263.292	0.904	-0.879	0.848
7.	Gorontalo	435.996	0.799	-0.773	0.863
8.	Jambi	241.756	0.910	-0.884	0.848
9.	Jawa Barat	268.463	0.903	-0.878	0.848

10.	Jawa Tengah	288.546	0.897	-0.871	0.849
11.	Jawa Timur	314.949	0.888	-0.863	0.850
12.	Kalimantan Barat	292.146	0.891	-0.866	0.852
13.	Kalimantan Selatan	346.582	0.870	-0.845	0.855
14.	Kalimantan Tengah	317.970	0.882	-0.857	0.853
15.	Kalimantan Timur	356.430	0.858	-0.832	0.858
16.	Kalimantan Utara	349.090	0.858	-0.832	0.859
17.	Kepulauan Bangka Belitung	260.025	0.904	-0.879	0.849
18.	Kepulauan Riau	267.767	0.898	-0.873	0.851
19.	Lampung	252.850	0.907	-0.882	0.848
20.	Maluku	548.900	0.756	-0.729	0.857
21.	Maluku Utara	512.967	0.757	-0.730	0.859
22.	Nusa Tenggara Barat	375.583	0.860	-0.834	0.856
23.	Nusa Tenggara Timur	459.371	0.827	-0.801	0.859
24.	Papua	603.022	0.719	-0.692	0.852
25.	Papua Barat	567.521	0.733	-0.706	0.854
26.	Riau	237.818	0.910	-0.885	0.849
27.	Sulawesi Barat	393.867	0.837	-0.811	0.862
28.	Sulawesi Selatan	414.150	0.832	-0.806	0.861
29.	Sulawesi Tengah	423.126	0.816	-0.790	0.863
30.	Sulawesi Tenggara	453.464	0.815	-0.789	0.861
31.	Sulawesi Utara	468.108	0.777	-0.750	0.861
32.	Sumatera Barat	233.604	0.912	-0.887	0.848
33.	Sumatera Selatan	248.249	0.908	-0.883	0.848
34.	Sumatera Utara	228.849	0.913	-0.888	0.848

Each location has a different model. For example, the local model for South Sulawesi Province is:

$$y_{sulsel} = 414.150 + 0.832 - 0.806$$

The equation explains that if the independent variable is fixed and the number of positive cases increases by one unit, the number of patients dying will increase by 0.839. If the number of recovered patients decreased by one unit, the number of patients who died would decrease by 0.814. The coefficient of determination in South Sulawesi Province is 86.1%. This shows that the variance of the independent variables used in the model can explain 86.1% of the variable number of patients dying, and the rest is explained by other variables that have not been included in this study.

4. CONCLUSION

The COVID-19 case originated in Wuhan, China, and continues to spread to all corners of the world, including Indonesia. This virus causes mild respiratory infections, such as the flu. However, this virus can also cause severe respiratory infections, such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). As of December 31, 2021, there were 4,262,697 positive cases in Indonesia, with details of 3,743 patients being treated, 4,114,744 patients recovered, and 144,210 patients died. The GWR model using a fixed bisquare kernel weighting function has an optimum bandwidth of 0.999948 with a minimum CV value of 397.076.128 with a coefficient of determination (R^2) of 85.1 %. The results show that the number

of positive cases positively correlates with the number of patients who died from COVID-19. In contrast, the number of recovered patients negatively correlates with the number of patients who died from COVID-19.

REFERENCES

[1] Anselin, L.,1988. *Spatial Econometrics: methods and models*. Dordrecht. Kluwer Academic Publishers, 1988.
[2] Brundson, C., Fotheringham, A.S., dan Charlton, M.E., 1996. Geographically Weighted Regression: Method for Exploring Spatial Nonstationarity. *Geographical Analysis*, 28(4), 281-298.
[3] Fajar, M., 2021. Pemodelan Parametrik Kurva Pertumbuhan Epidemi Covid-19 Di Indonesia. *Jurnal Matematika dan Sains*: ISSN 0854-5154.
[4] Fitrial, N.H., Fatikhurrizqi, A.., 2020. Pemodelan Jumlah Kasus Covid-19 Di Indonesia Dengan Pendekatan Regresi Poisson Dan Regresi Binomial Negatif. *Seminar Nasional Official Statistics* 2020: *Pemodelan Statistika tentang COVID-19*.
[5] Fotheringham, A.S., Brundson, C., Charlton, M.E., 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationship*. John Wiley and Sons Ltd.
[6] Goodchild, M.F., 1989. *Spatial Autocorrelation*. Norwich: Geobooks, Norwich.
[7] Ma, Z., Xue, Y., Hu, G., 2020. Geographically Weighted Regression Analysis for Spatial Economics Data: A Bayesian Recourse. *International Regional Science Review* XX(X).
[8] Mahdy, Ilham,F., 2020. Pemodelan Jumlah Kasus Covid-19 Di Jawa Barat Menggunakan Geographically Weighted Regression. *Seminar Nasional Official Statistics 2020: Statistics in the New Normal: A Challenge of Big Data and Officials Statistics*. Departemen Statistika, Universitas Padjadjaran.
[9] Ogundokun, R., O. Lukman, A., F. Golam B.M. Kibria, F.,G. Awotunde, J., B. Aladeitan, B., 2020. Predictive Modelling of COVID-19 Confirmed Cases In Nigeria. *Infectious Disease Modelling 5.*
[10] Purhadi & Yasin,. 2008. Mixed Geographically Weighted Regression Model (Case Study: the Percentage of Poor Households in Mojokerto 2008. *European Journal of Scientific Research*, 188-196.
[11] Putra, R., Tyas, S.W., Fadhlurrahman, M.G., 2022. Geographically Weighted Regression with The Best Kernel Function on Open Unemployment Rate Data in East Java Province. *ENTHUSIASTIC: International Journal Of Statistics And Data Science*. Volume 2, April 2022.
[12] Sameni, R., 2020. *Mathematical Modeling of Epidemic Diseases; A Case Study of the COVID-19 Coronavirus*. Quantitative Biology: Cornell University.
[13] Song, Y., Shen, Z., Wul, P., Rossel, R.A.V., 2021. Wavelet Geographically Weighted Regression For Spectroscopic Modelling Of Soil Properties. *Scientific Reports:* https://doi.org/10.1038/s41598-021-96772-z
[14] Susilo A., Rumende C.M., Pitoyo C.W., Santoso W.D., Yulianti M, Herikurniawan H., et al., 2020. Coronavirus Disease 2019: Tinjauan Literatur Terkini. *Jurnal Penyakit Dalam Indonesia*, Vol. 7(1), Hal. 45–67.
[15] Teguh, R., Sahay, A.S., Adjib. F.F., 2020. Pemodelan penyebaran Infeksi COVID-19 di Kalimantan. *Jurnal Teknologi Informasi*.