Handling Multicollinearity on Social Spatial Data Using Geographically Weighted Random Forest

Binti Kurniati [1], Yuliani Setia Dewi [1], Alfian Futuhul Hadi [1]

[1] *Faculty of Mathematics and Natural Sciences Univerisity of Jember, Indonesia*

Abstract **– Crime includes all kinds of harmful acts that violate the laws in force in Indonesia as well as social and religious norms. The crime total is the number of incidents reported to the police, obtained from public reports and events where the perpetrators were caught red-handed by the police. We can use the Poisson model to analyze the data, but the existence of spatial heterogeneity in the data makes the model less accurate. This research investigates the methods when there is spatial heterogeneity in the data by using Geographically weighted regression (GWR), Geographically Weighted Poisson Regression (GWPR) and Geographically Weighted Random Forest (GW-RF). We compare the GWR, GWPR, and GW-RF models for criminal cases in East Java in handling multicollinearity in the data. The results of this study indicate that the GW-RF model is better for modeling criminal cases with the smallest RMSE and MAPE values and an R-Square value close to 1. Based on the three most important variables in each location, they form six groups of regencies/cities in East Java, Indonesia. The variables vary between groups and the poverty severity index is not included in the three most important variables in all locations.**

Keywords **– Crime, GWR, GWPR, GW-RF, multicollinearity, spatial heterogeneity.**

DOI: 10.18421/SAR63-02
https://doi.org/10.18421/SAR63-02

Corresponding author: Yuliani Setia Dewi,
University of Jember, Indonesia
Email: yulidewi.fmipa@unej.ac.id

Received: 24 July 2023.
Revised: 25 August 2023.
Accepted: 01 September 2023.
Published: 26 September 2023.

1. Introduction

Crime is a significant issue in developing nations, including Indonesia. The numerous news articles about criminal acts in print and social media demonstrate how commonplace crime is in Indonesia. In 2020, there were many criminal cases in Indonesia. With 17,642 crimes reported in 2020, East Java Province is one of Indonesia's top three criminal-involved regions [1]. The government has taken several steps to prosecute offenders and prevent crime. Still, more has to be done, specifically by looking at the factors contributing to crime development.

Prior studies on the causes of crime, specifically [2], found that the variables Gross Regional Domestic Product (GRDP) per capita, unemployment rate, population density, and poverty significantly influenced East Java's crime. Using the path analysis approach, [3] evaluated the variables affecting the crime rate in Indonesia in 2018. They found that the population and poverty variables have a substantial impact, whereas the education and unemployment variables have no significant effect. This research was conducted in a single geographic area, assuming the observation sites had similar regional conditions. Due to geographical, sociocultural, and other considerations, each observation location has unique regional properties. This difference in characteristics allows spatial heterogeneity. This spatial heterogeneity occurs when one or more independent variable do not give the same responses at different locations within the research study area [4].

One of the data types that can include geographic heterogeneity is a crime data. Therefore, we use the Geographically model approach for the proper modeling. The GWR method is a local variation of linear regression that computes the parameter estimates using a spatial weight at each location. The research by Anjas and Kencana [5] studied multiple linear regression and GWR to examine pneumonia cases in East Java.

To obtain the optimal GWR model, another study used five independent variables and a fixed bisquare weighting function to model Leptospirosis susceptibility in Bantul Regency [6]. The short birth intervals case in Ethiopia uses the GWR method using five independent variables [7]. Another method for modeling the problem of spatial heterogeneity is Geographically Weighted Poisson Regression (GWPR), a development of Poisson regression. Previous research regarding the GWPR method [8] tested the health services of tuberculosis cases in Vietnam. The results of this study show that the GWPR method adjusts data better than the GLM regression method.

Geographically Weighted Random Forest is another approach to address the issue of spatial heterogeneity (GW-RF). The GW-RF approach uses spatial weighting and is a random forest variation [9]. Analyzing the regional variability of type 2 diabetes mellitus (T2D) prevalence in the United States results that the GW-RF technique has a significant chance of explaining regional variability and forecasting the prevalence of T2D using six key variables [10]. The GW-RF approach to study socioeconomic factors and poverty in China finds that the geographically weighted random forest is better than the random forest based on the R-square, NRMSE, and MAE values [11].

GWR, GWPR, and GW-RF approaches can handle spatial heterogeneity in a data set. This research compares the GWR, GWPR, and GW-RF methodologies to analyze how crime develop in East Java using the year 2020 crime data.

2. Method

We investigate the GWR, GWPR, and GWRF methods for analyzing crime cases in East Java, Indonesia. We use some kernels for GWR and GWPR processes and use an adaptive kernel for the technique based on machine learning (GWRF).

2.1. Data Sources

This study uses secondary data from the Central Statistics Agency of East Java Province 2020 [1]. The observation units used in this study are 29 regencies and nine cities in East Java, Indonesia. This study uses one response variable, the reported crime number (*crime total*) and six predictor variables: the percentage of poor people (X_1), population density (X_2), human development index (X_3), average length of school (X_4), gender ratio (X_5), and poverty severity index (X_6).

2.2. Analysis Method

This research uses the GWR, GWPR, and GW-RF methods and compare them in handling multicollinearity in the data. The following are the steps taken in this research.

1. Performing descriptive statistical analysis.
2. Testing the classical assumption, multicollinearity and spatial heterogeneity. The multicollinearity test based on the VIF value [12]. We use the Breusch-Pagan test to detect spatial heterogeneity with significance level $\alpha = 0.05$.
3. Modeling GWR using the equation

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{m} \beta_k(u_i, v_i)\, X_{ik} + \varepsilon_i \; ; \quad (1)$$

$i = 1,2, \dots, n$

We use Euclidean to calculate the distance between locations and the Cross Validation (CV) method to find the weight. We use several kernels, namely Adaptive Gaussian, Adaptive Bisquare, and Fixed Gaussian [4]. Determining optimum bandwidth based on the smallest Akaike Information Criterion (AIC) value. Estimating the parameters of GWR model uses the Weighted Least Square (WLS) method. We specify the RMSE, R-square (R^2), and MAPE values for measuring the model's goodness of fit.

4. Modeling GWPR [13] using the equation

$$\widehat{\mu_i} = \exp\left(\sum_{j=0}^{k} \beta_j(u_i, v_i)\, X_{ij}\right), \quad (2)$$

$i = 1,2, \dots, n$

The same procedure as GWR model, we conduct to GWPR model.

5. We analyze the model using the GW-RF method with the following steps:

a. Finding the optimum bandwidth and weights uses the adaptive kernel function. Determining the optimum bandwidth value based on the R-square. The adaptive kernel function is based on the number of nearest neighbors in an observation area [14].

b. We calculate the RMSE, R-square (R^2), and MAPE values to measure the goodness of fit.

6. We compare the GWR, GWPR, and GW-RF models by selecting the best model using the RMSE, R-square (R^2), and MAPE values.
7. We find the variables importance and group of locations based on the three highest- variables importance from the best model.

3. Results

Descriptive statistical analysis (Table 1) determines the characteristics of each variable. The total crime variable (Y) has an average of 600, with the highest crime of 1850 cases in Malang Regency and the lowest is 70 cases in Batu City.

Table 1. Descriptive Statistics Results

Variable	Average	Variance	Minimum Value	Maximum Value
Y	600.2	187763.5	70	1850
$X1$	11.02	20.864	3.890	22.780
$X2$	1923	4441299	295	8200
$X3$	71.87	25.465	62.700	82.230
$X4$	7.94	2.340	4.850	11.140
$X5$	97.20	6.493	90.750	101.20
$X6$	0.49	0.079	0.100	1.300

We calculate the value of the Variance Inflation Factor (VIF) to find the multicollinearity between independent variables. A VIF value greater than 10 indicates a high correlation [15].

Table 2. VIF Values

Variable	VIF Values
X_1	7.021614
X_2	3.422645
X_3	23.155702
X_4	24.078785
X_5	1.966448
X_6	3.706591

Table 2 shows symptoms of multicollinearity between independent variables, the VIF values for the variables X_3 and X_4>10.

Heterogeneity test is related to spatial heterogeneity, where there are differences in characteristics between locations. The Breusch-Pagan (BP) tests this heterogeneity. Based on the tests, the $p-value$ is 0.04808, so it rejects H_0, which means there are differences in variance between locations so that spatial heterogeneity occurs.

3.1. Geographically Weighted Regression (GWR) Modeling

The GWR method determines the optimum bandwidth value to find the weight and build the model. This research obtains the best weight by finding the optimum bandwidth by looking at the minimum AIC and Cross Validation (CV) value.

Table 3. Optimum Weight Selection GWR Model

Kernel	AIC	Bandwidth	CV
Adaptive Gaussian	561.4777	0.999945	7105097
Adaptive Bisquare	561.9788	0.999950	7146959
Fixed Gaussian	561.4627	18.38921	7098748

Based on Table 3, the weight with the minimum AIC value is the fixed Gaussian kernel function equal to 561.4627 with an optimum bandwidth of 18.38921. The next step is to estimate the parameters of the GWR model based on the fixed Gaussian kernel function.

Here the examples of the interpretation of the GWR model for Jember Regency based on the predicted model:

$$\hat{y} = -11729.07 - 16.347X_1 + 0.0229X_2 + 122.445X_3 - 347.322X_4 + 66.420X_5 - 452.960X_6$$

Based on this model, it shows that in Jember Regency, the variables of population density (X_2), human development index (X_3), and sex ratio (X_5) have a positive relationship with the crime total. In contrast, the variable percentage of the poor population (X_1), the average length of schooling (X_4), and the poverty severity index (X_6) have a negative relationship with the total crime rate.

3.2. Geographically Weighted Poisson Regression (GWPR) Modeling

The optimum bandwidth and weight selection in the GWPR model is the same as the GWR method, namely, using the minimum AIC and CV values.

Table 4. Optimum Weight Selection GWPR Model

Kernel	AIC	Bandwidth	CV
Adaptive Gaussian	6350.999	37	7457063
Adaptive Bisquare	6427.427	37	8011424
Fixed Gaussian	6426.946	17.47546	7844091

Based on Table 4, the weighting with the minimum AIC value is the adaptive Gaussian kernel function equal to 6350.999 with an optimum bandwidth of 37.

The parameter estimation of the GWPR uses the weight of each location to build the model. By using the GWPR method, we obtain the model as follows for Jember Regency:

$$\hat{\mu} = \exp(-15.6412 - 0.0181X_1 + 0.000017X_2 + 0.2183X_3 - 0.6419X_4 + 0.1157X_5 + 0.7110X_6)$$

This model shows that in Jember Regency, the variable percentage of poor people (X_1) and the average length of schooling (X_4) negatively correlate with the total crime rate in East Java. The variable population density (X_2), human development index (X_3), sex ratio (X_5), and poverty severity index (X_6) have a positive relationship with the total crime rate.

3.3. Geographically Weighted Random Forest (GW-RF) Modeling

Geographically Weighted Random Forest (GW-RF) is one of the machine learning methods to overcome spatial heterogeneity in the data. GW-RF modeling begins with finding the optimum bandwidth and weight for building the GW-RF model. This research uses an adaptive kernel function and selects the optimum bandwidth based on the local GW-RF model's highest R-square (R^2) value. We build the model using the number of trees of 500 and find the optimum bandwidth value is 30.

We compare and select the best model between the GWR, GWPR, and GW-RF to explain the studied problems by comparing the value of the Root Mean Square Error (RMSE), the value of the coefficient of determination (R^2), and the Mean Absolute Percentage Error (MAPE) indicators. A better model has smaller RMSE and MAPE values, which means it has the smaller the error value of a model, the closer the predicted value is to the actual value [16]. A model's R-Square (R^2) value is said to be good if the value is close to 1 [17].

Table 5. Selection of the Best Model

Criteria	GWR Model	GWPR Model	GW-RF Model
RMSE	356.2379	341.2200	80.4159
R^2	0.3059	0.3631	0.9646
MAPE (%)	84.2559	81.8773	19.3948

Based on Table 5, the RMSE value of the GW-RF model is the smallest. The R^2 value of the GW-RF model is closer to 1 compared to the R^2 value of the GWR and GWPR models, and the MAPE value of the GW-RF model is smaller than the GWR and GWPR models. Therefore, GW-RF is the best model for explaining the problems studied in this study.

3.4. Variable Importance

This variable importance determines whether a variable has an important role in this modeling. The higher value of variable importance, the more influential the variable is used [18].

Table 6 gives the grouping of regencies/cities in East Java based on the three highest values of the variable importance for 38 regencies/cities.

Table 6. Group of Locations Based on Three Most Important Variable

Group	Important Variable	Regency/City
1	X_1, X_3 and X_4	Gresik, Bangkalan, and Pamekasan
2	X_1, X_3 and X_5	Tulungagung, Blitar Regency, and Blitar City
3	X_1, X_4 and X_5	Malang
4	X_2, X_3 and X_5	Pacitan, Ponorogo, Trenggalek, Madiun, Magetan, Ngawi, Kediri City, and Madiun City
5	X_2, X_4 and X_5	Bojonegoro
6	X_3, X_4 and X_5	Kediri, Lumajang, Jember, Banyuwangi, Bondowoso, Situbondo, Probolinggo, Pasuruan, Sidoarjo, Mojokerto, Jombang, Nganjuk, Tuban, Lamongan, Sampang, Sumenep, Malang City, Probolinggo City, Pasuruan City, Mojokerto City, Surabaya City, and Batu City

Based on Table 6, there are six groups of locations. The first group consists of three regions, namely Gresik Regency, Bangkalan Regency, and Pamekasan Regency with the important variable: the percentage of the poor population (X_1), human development index (X_3), and the average length of schooling (X_4). The second group has three members, namely Tulungagung Regency, Blitar Regency, and Blitar City with important variables: the percentage of poor people (X_1), human development index (X_3), and sex ratio (X_5). The third and fifth groups are minority groups with one regency/city, namely Malang and Bojonegoro Regency. The fourth group has eight regency/city members with important variable population density (X_2), human development index (X_3), and sex ratio (X_5). The sixth group has the highest number of members, namely 22 regencies/cities with the important variables the human development index (X_3), average length of schooling (X_4), and sex ratio (X_5). Figure 1 exhibits the groups of East Java's 38 regencies/cities based on the variable importance.

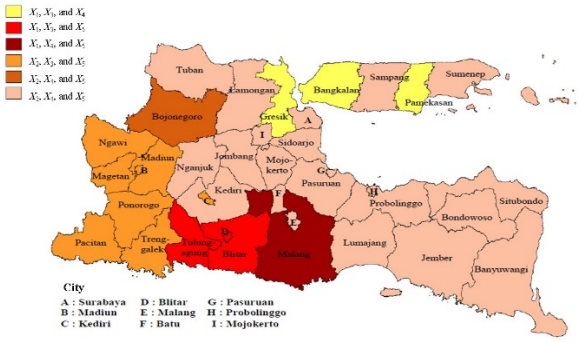

Figure 1. Mapping of Locations Based on Three Most Important Variables

4. Conclusion

The modeling of criminal cases in East Java in this research contains multicollinearity and spatial heterogeneity in the data. We investigate the GWR, GWPR, and GW-RF methods for handling multicollinearity in the spatial data. The GWR and GWPR modeling use Adaptive Gaussian, Adaptive Bisquare, and Fixed Gaussian kernels. We find that the best kernel to find the weight for the GWR model is the Fixed Gaussian kernel, while the GWPR model is the Adaptive Gaussian kernel. GW-RF modeling uses an adaptive kernel based on the nearest neighbor number. A comparison of the methods shows that the GW-RF method is superior to the GWR and GWPR methods in explaining crime cases in East Java, 2020. It has the lowest RMSE and MAPE values and the R-Square value close to 1. There are six groups of regencies/cities based on the three most important variables using the GW-RF model. The important variables vary between locations. The poverty severity index does not include the three most important variables in all locations of East Java Province.

References:

[1]. Lapebesi, R. A., Pramesti, E. N., Ahyandi, M. N., Sari, M. T., & Yuhan, R. J. (2021). Analisis Jalur Faktor-Faktor yang Mempengaruhi Jumlah Kriminalitas di Jawa Timur Tahun 2020. *Jurnal Sains Matematika dan Statistika*, *7*(2), 38-49.

[2]. Purwanti, E. Y., & Widyaningsih, E. (2019). Analisis faktor ekonomi yang mempengaruhi kriminalitas di Jawa Timur. *Jurnal Ekonomi-QU*, *9*(2).

[3]. Putra, A. D., Martha, G. S., Fikram, M., & Yuhan, R. J. (2021). Faktor-Faktor yang Memengaruhi Tingkat Kriminalitas di Indonesia Tahun 2018. *Indonesian Journal of Applied Statistics*, *3*(2), 123-131.

[4]. Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2003). *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons.

[5]. Anjas, M., Sukarsa, I. K. G., & Kencana, I. P. E. N. (2019). Penerapan Metode Geographically Weighted Regression (GWR) pada kasus penyakit Pneumonia di Provinsi Jawa Timur. *E-Jurnal Matematika*, *8*(1), 27-34.

[6]. Widayani, P., Gunawan, T., Danoedoro, P., & Mardihusodo, S. J. (2016). Application of geographically weighted regression for vulnerable area mapping of leptospirosis in Bantul District. *The Indonesian journal of geography*, *48*(2), 168.

[7]. Shifti, D. M., Chojenta, C., Holliday, E. G., & Loxton, D. (2020). Application of geographically weighted regression analysis to assess predictors of short birth interval hot spots in Ethiopia. *PloS one*, *15*(5), e0233790.

[8]. Bui, L. V., Mor, Z., Chemtob, D., Ha, S. T., & Levine, H. (2018). Use of Geographically Weighted Poisson Regression to examine the effect of distance on Tuberculosis incidence: A case study in Nam Dinh, Vietnam. *PloS one*, *13*(11), e0207068.

[9]. Santos, F., Graw, V., & Bonilla, S. (2019). A geographically weighted random forest approach for evaluate forest change drivers in the Northern Ecuadorian Amazon. *PLoS One*, *14*(12), e0226224.

[10]. Quiñones, S., Goyal, A., & Ahmed, Z. U. (2021). Geographically weighted machine learning model for untangling spatial heterogeneity of type 2 diabetes mellitus (T2D) prevalence in the USA. *Scientific reports*, *11*(1), 6955.

[11]. Luo, Y., Yan, J., McClure, S. C., & Li, F. (2022). Socioeconomic and environmental factors of poverty in China using geographically weighted random forest regression model. *Environmental Science and Pollution Research*, 1-13.

[12]. Sriningsih, M., Hatidja, D., & Prang, J. D. (2018). Penanganan multikolinearitas dengan menggunakan analisis regresi komponen utama pada kasus impor beras di Provinsi Sulut. *Jurnal Ilmiah Sains*, *18*(1), 18-24.

[13]. Nakaya, T., Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2005). Geographically weighted Poisson regression for disease association mapping. *Statistics in medicine*, *24*(17), 2695-2717.

[14]. Georganos, S., Grippa, T., Niang Gadiaga, A., Linard, C., Lennert, M., Vanhuysse, S., ... & Kalogirou, S. (2021). Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International*, *36*(2), 121-136.

[15]. Ryan, T. P. (2008). *Modern regression methods* (Vol. 655). John Wiley & Sons.

[16]. Wang, C. H., & Hsu, L. C. (2008). Using genetic algorithms grey theory to forecast high technology industrial output. *Applied mathematics and computation*, *195*(1), 256-263.

[17]. Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*,e623. Doi: 10.7717/peerj-cs.623

[18]. Dewi, C., & Chen, R. C. (2019). Random forest and support vector machine on features selection for regression analysis. *Int. J. Innov. Comput. Inf. Control*, *15*(6), 2027-2037.