Modeling total crime and the affecting factors in Central Java using geographically weighted regression

T Runadi, Y Widyaningsih and D Lestari
Department of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Indonesia, Kampus UI Depok, Depok 16424, Indonesia

Corresponding author's e-mail: dian.lestari@sci.ui.ac.id

Abstract. Analysing the relationship between number of crime cases and affecting factors became an interesting research topic over the last ten years. The total number of crime in Indonesia did not show a consistent decrease. In order to upgrade people safeness quality, the government need to know the factors influence people committing crime acts. Rather than using classical regression analysis, geographically weighted regression (GWR) was preferable since it gave a better representative model by effectively resolve spatial non-stationary problem which is generally exist in spatial data of social phenomenon. Spatial non-stationary is a situation when the relationship between variables are significantly different in each location of observation point, so that classic regression analysis will result a misleading interpretation in some location. GWR handled the spatial non-stationary problem by generating a single model in each observation point which allow different relationship to exist at different point in space. This study used number of crime cases (y) as the dependent variable and the factors which affect the number of crime cases as independent variables that consist of the number of illiterates (x1), the number of unemployed (x2), the number of poor population (x3), population density (x4), the number of victims of drug (x5). This study used secondary data collected by POLRI, BPS and Indonesian Ministry of Social Affairs in Central Java during 2015. GWR generated model for 35 city/regency in Central Java.

Keywords: bisquare, crime, Gaussian, geographically weighted regression (GWR).

1. Introduction
Crime is defined as an act of someone who may be threatened under the criminal code or any other law in Indonesia [1]. By the data compiled by the Indonesian National Police (POLRI), within the period 2013–2015, the number of crime acts in Indonesia does not indicate a consistent decline but tend to fluctuate. The number of crime in Indonesia in 2013 occurred is 342,084 cases and decreased to 325,317 in 2014 and raised again to 352,936 cases by 2015. During 2015, there is at least one crime every 1 minute 29 seconds [2]. This indicates that there is still a need to improve the quality of crime prevention in Indonesia.

In an order to improve the quality of security, the government needs to understand what factors that significantly affect the number of crimes in a region [3]. The relationship between number of crime and its factors can be analyzed by linear regression method. Linear regression analysis is a statistical technique for investigating and modeling the relationship between variables [4]. Linear regression analysis can be applied to a spatial data (data containing only attribute information) as well as spatial data (data containing attribute and location information). In spatial data, regression analysis will provide a good parameter estimate if the relationship between independent and dependent variables is stationary, which means that the relationships between variables are fixed in each region of study area. The relationship between variables in social phenomena data will tend to be non-stationary, or it can be said that relationships among variables tend to differ in each region [5].
Table 1. Dependent and independent variables

| Symbol | Variable | Definition |
|--------|----------|------------|
| Y      | Number of crime act | The number of infringement actions that can be punishable by Penal Code or the law that reported by people to the police, or the perpetrator is caught by the police. |
| X<sub>1</sub> | Number of illiterate | Number of residents over 15 years old who do not have the ability to read and write simple sentences in Latin, Arabic letters, and other letters. |
| X<sub>2</sub> | Number of unemployment | Number of people looking for work, people preparing for a business, people who find it impossible to get a job, people who already have jobs but have not started work. |
| X<sub>3</sub> | Number of poverty | The number of people who have an average monthly per capita expenditure below the poverty line. The poverty line in Central Java in 2015 is 297,851 rupiah/capita/month. |
| X<sub>4</sub> | Population density | The number of residents in a region divided by area. |
| X<sub>5</sub> | Number of drug and alcohol victims | The number of people who use narcotics, psychotropic substances, and other addictive substances includes alcohols without doctor’s supervision or certain medical purposes. |

This condition is also known as spatial heterogeneity. Spatial heterogeneity may exist as an impact of diversity of geographical, social, cultural, economic, or political conditions in each region. If spatial heterogeneity occurs, geographically weighted regression (GWR) is a suitable method to analyze the relationship between independent and dependent variables.

The objective of this study is analyzing significant factors that influence number of crime case in regencies/cities in Central Java using GWR.

2. Data and methods

This study analyzed secondary data that obtained from publication book by Badan Pusat Statistik [1–2]. The independent variable (Y) was the number of crime case that occurred in each regencies/cities in Central Java during 2015. While the dependent variables was the factors that suspected influencing the number of crime case that described in table 1.

2.1. Spatial data and spatial heterogeneity

Spatial heterogeneity is a condition where the relationships among variables are vary at each location. The existence of spatial heterogeneity can be tested using Breuch-Pagan Test [6]. The procedure of Breuch-Pagan Test is explained as follow:

Hypothesis:

H<sub>0</sub>: Spatial heterogeneity not exist
H<sub>1</sub>: Spatial heterogeneity exists.

Test Statistic:

\[
BP = \left( \frac{1}{n} \right) f^T Z (Z^T Z)^{-1} Z^T f
\]

(1)

Reject H<sub>0</sub> criteria: \( BP > \chi^2_{A,p} \)

The value of \( \chi^2_{A,p} \) is obtained from the Chi Square Table Information:

- \( f \): vector with size \((n\times1)\) which its elements are \( f_i = \frac{\epsilon_i^2}{\sigma^2} - 1, \ i = 1, 2, \ldots, n. \)
- \( Z \): standardized matrix \( X \)
- \( n \): number of observations
- \( p \): number of independent variables
- \( \epsilon_i \): error of \( i^{th} \) observation of linear regression model
- \( \sigma^2 \): variance of errors.
2.2. Geographically weighted regression (GWR)

The GWR method is used in spatial data when spatial heterogeneity is detected. The GWR model is as follows:

\[ y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i)X_{ik} + \epsilon_i, \quad i = 1, 2, \ldots, n \]  

Information:
- \( y_i \): dependent variable in \( i^{th} \) location
- \( X_{ik} \): \( k^{th} \) independent variable in \( i^{th} \) location
- \( (u_i, v_i) \): coordinate of \( i^{th} \) location
- \( \beta_0(u_i, v_i) \): constant of GWR model in \( i^{th} \) location
- \( \beta_k(u_i, v_i) \): parameter of \( k^{th} \) independent variable of GWR model in \( i^{th} \) location
- \( \epsilon_i \): error in \( i^{th} \) location.

The parameter estimation on the GWR model used the following formula [3].

\[ \hat{\beta}(u_i, v_i) = [X'W(u_i, v_i)X]^{-1}X'W(u_i, v_i)Y \]  

Information:
- \( \hat{\beta}(u_i, v_i) \)
- \( Y \) = \[
\begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_{n-1} \\
y_n
\end{bmatrix}
\]
- \( X \) = \[
\begin{bmatrix}
x_{11} & \cdots & x_{1(p-1)} & x_{1p} \\
x_{21} & \cdots & x_{2(p-1)} & x_{2p} \\
\vdots & \ddots & \vdots & \vdots \\
x_{(n-1)1} & \cdots & x_{(n-1)(p-1)} & x_{(n-1)p} \\
x_{n1} & \cdots & x_{n(p-1)} & x_{np}
\end{bmatrix}
\]
- \( W(u_i, v_i) \)
- \[
\begin{bmatrix}
w_1(u_i, v_i) & 0 & \cdots & 0 & 0 \\
0 & w_2(u_i, v_i) & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & w_{n-1}(u_i, v_i) & 0 \\
0 & 0 & \cdots & 0 & w_n(u_i, v_i)
\end{bmatrix}
\]

The weight value \( w_j(u_i, v_i) \) is obtained using the spatial weighting function \( (K) \). The spatial weighting function will give a weight value between 0 and 1 based on the distance between the point of the model constructed with the observation point. When the point of building of the model is equal to the observation point then the spatial weighting function will give the maximum weight value that is 1. The further the observation point with the point of constructing the model, the weight value will be smaller [7]. Let \( d_{ij} \) denote the Euclidean distance of the observing points \( i \) and \( j \) with the formula

\[ d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2} \]

then the weight of the \( j^{th} \) observation point for the model at the \( i^{th} \) location is \( w_j(u_i, v_i) = K(d_{ij}) \). This study used fix Gaussian kernel function which the formula is:

\[ K(d_{ij}) = \exp \left[ -\frac{1}{2} \left( \frac{d_{ij}}{b} \right)^2 \right] \]  

Information:
- \( d_{ij} \): distance of \( i^{th} \) location to \( j^{th} \) location.
- \( b \): bandwidth optimum [8].

3. Results and discussion

The GWR method produces different models in each observation location. The existence of spatial heterogeneity leads to the relationship between independent variables and dependent variable may vary in each location so that not all independent variables have a significant influence to dependent variable on a certain location.
Table 2. Groups of independent variable, which significantly influence dependent variable

| Group | Independent variables *) | Number of Regency/City | Regency/City |
|-------|--------------------------|------------------------|--------------|
| 0     | -                        | 4                      | (Reg) Cilacap, (Reg) Tegal, (Reg) Brebes, (City) Tegal |
|       |                          |                        | (Reg) Banyumas, (Reg) Purwalingga, (Reg) Banjarnegara, |
|       |                          |                        | (Reg) Kebumen, (Reg) Pekalongan, (Reg) Pemalang, |
|       |                          |                        | (City) Pekalongan. |
| 1     | $X_2$                    | 7                      | (Reg) Purbalingga, (Reg) Tegal, |
|       |                          |                        | (Reg) Banyumas, (Reg) Purbalingga, (Reg) Banjarnegara, |
|       |                          |                        | (Reg) Kebumen, (Reg) Pekalongan, (Reg) Pemalang, |
|       |                          |                        | (City) Pekalongan. |
| 2     | $X_2$, $X_3$, $X_4$      | 10                     | (Reg) Purworejo, (Reg) Wonosobo, (Reg) Magelang, (Reg) |
|       |                          |                        | Semarang, (Reg) Temanggung, (Reg) Kendal, (City) |
|       |                          |                        | Magelang, (City) Salatiga, (City) Semarang, (Reg) Batang |
|       |                          |                        | (Reg) Boyolali, (Reg) Klaten, (Reg) Demak |
| 3     | $X_2$, $X_3$, $X_4$      | 3                      | (Reg) Sukoharjo, (Reg) Wonogiri, (Reg) Karanganyar, |
|       |                          |                        | (Reg) Sragen, (Reg) Grobogan, (Reg) Pati, (Reg) Kudus, |
|       |                          |                        | (Reg) Jepara, (City) Surakarta |
| 4     | $X_2$, $X_3$             | 9                      | (Reg) Blora, (Reg) Rembang |
| 5     | $X_4$                    | 2                      | * $X_1$: Number of illiterate |
|       |                          |                        | $X_2$: Number of unemployment |
|       |                          |                        | $X_3$: Number of poverty |
|       |                          |                        | $X_4$: Population density |
|       |                          |                        | $X_5$: Number of drug and alcohol victims |

Figure 1. Distribution of groups of independent variables which significantly influence dependent variable.

Figure 1 shows the distribution of independent variables that have a significant influence on the dependent variable with a significant level 5%. There are six location groups based on significant independent variables as described in table 2. Group 0 is a group of districts where no independent variables significantly affect the dependent variable. From figure 1 it can be seen that group 0 is located in the west part of Central Java Province. There are four locations included in this group that is Cilacap Regency, Tegal Regency, Brebes Regency, and Tegal City. This indicates that in those regions, the number of crime is influenced by other factors not analyzed in this study.
4. Conclusions
Analyzing the relationship of the number of crimes and its factors using the GWR method has resulted different regression model parameter values in each regency/city in Central Java. The significant test of estimated parameters shows that there are six groups of independent variables (crime factors), which significantly influence the dependents variable (number of crime case). The number of illiterate does not significantly affect the number of crimes in all regency/city.

References
[1] BPS 2015 Statistik Politik dan Keamanan Jawa Tengah 2015 (Semarang: Badan Pusat Statistik Provinsi Jawa Tengah)
[2] BPS 2016 Statistik Kriminal 2016 (Jakarta: Badan Pusat Statistik)
[3] Miller J M 2009 21st Century Criminology: A Reference Handbook (Thousand Oaks: SAGE Publications, Inc.)
[4] Montgomery D C, Peck E A and Vining G G 2012 Introduction to Linear Regression Analysis 5th ed (Hoboken: John Wiley & Sons, Inc.)
[5] Fotheringham A S, Brunsdon C and Charlton M 2002 Geographically Weighted Regression: The Analysis of Spatially Varying Relationships (Chichester: John Wiley & Sons, Inc.)
[6] Gollini I, Lu B, Charlton M, Brunsdon C and Harris P 2015 J. Stat. Softw. 63 1–50
[7] Brunsdon C, Fotheringham S and Charlton M 1998 J. R. Stat. Soc. Ser. D. (The Statistician) 47 431–43
[8] Anselin L 1988 Spatial Econometrics: Methods and Models. Studies in Operational Regional Science (Dordrecht: Kluwer Academic Publishers)