Geographically weighted regression model on poverty indicator

I Slamet¹, N F T A Nugroho² and Muslich²
¹Department of Statistics, Universitas Sebelas Maret, Jl. Ir. Sutami no. 36-A, Kentingan, Surakarta, 57126, Indonesia
²Department of Mathematics, Universitas Sebelas Maret, Jl. Ir. Sutami no. 36-A, Kentingan, Surakarta, 57126, Indonesia

E-mail: isnandarslamet@staff.uns.ac.id

Abstract. In this research, we applied geographically weighted regression (GWR) for analyzing the poverty in Central Java. We consider Gaussian Kernel as weighted function. The GWR uses the diagonal matrix resulted from calculating kernel Gaussian function as a weighted function in the regression model. The kernel weights is used to handle spatial effects on the data so that a model can be obtained for each location. The purpose of this paper is to model of poverty percentage data in Central Java province using GWR with Gaussian kernel weighted function and to determine the influencing factors in each regency/city in Central Java province. Based on the research, we obtained geographically weighted regression model with Gaussian kernel weighted function on poverty percentage data in Central Java province. We found that percentage of population working as farmers, population growth rate, percentage of households with regular sanitation, and BPJS beneficiaries are the variables that affect the percentage of poverty in Central Java province. In this research, we found the determination coefficient R² are 68.64%. There are two categories of district which are influenced by different of significance factors.

1. Introduction
Poverty is a problem that has not been resolved completely in Indonesia. The condition of the poor can be known based on the lack of income ability to meet the standard of living (Nugroho[5]). Poverty can be identified by low income ability to meet basic needs in the form of food, health, housing or education. Low income ability is also defined as low purchasing power or the ability to consume.

Central Java province is a province with high poverty level in Indonesia. Currently, the number of poor people in Central Java province amounts to 4.5 million people, or 13.32 percent in 2016. The percentage is higher than the national poverty rate of 11.13 percent. With the increase of the population that happens every year is possible the number of poor people is increasing. So it is necessary to determine an analysis that shows what factors affect poverty in Central Java Province.

In determining a region classified as poor or not, it is necessary to apply appropriate model analysis to each region. In this case a geographically weighted regression model can be used to analyze the data. The geographically weighted regression model is a model used to analyze spatial data resulting in estimation of local model parameters where each location is different from other locations.
(Fotheringham \textit{et al.} [4]). The geographically weighted regression model is part of spatial analysis by weighting based on the position or distance of one observation location with another observation location.

In previous research, Yusuf showed that geographically weighted regression models are more effectively used than global regression models. Sinaga in 2013 examined the geographically weighted regression for poverty data in North Sumatra Province with the weighted function of the kernel Gaussian. In this research will be shown in the form of geographically weighted regression model with Gaussian kernel weighted function on poverty percentage data in Central Java Province.

2. \textbf{Theoretical Basis}

2.1. \textit{Multiple Linear Regression}

Multiple linear regression model is a regression model used to examine the relationship of two or more independent variables to the dependent variable. The method that can be used to estimate the parameters of multiple linear regression model is the least squares method (OLS). A common form of multiple linear regression model with $p-1$ independent variables is presented in the equation below

\[ Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \]

In matrix form the equation can be written as

\[ Y = X\hat{\beta} + \varepsilon \]

where $Y$ is the dependent variable vector ($n \times 1$), $X$ is the independent variable matrix ($n \times p$), $\hat{\beta}$ is vector parameter ($p \times 1$), $\varepsilon$ is an error vector ($p \times 1$) mean zero and variance $\sigma^2 I$. $I$ is the identity matrix.

The parameter estimate of $\hat{\beta}$ in can be expressed by

\[ \hat{\beta} = (X^T X)^{-1} X^T Y \]

The testing statistic $F$ of regression model is

\[ F = \frac{SSR}{SSE} \]

With $SSR = \frac{\sum (Y - \bar{Y})^2}{p}$ and $SSE = \frac{\sum (Y - \bar{Y})^2}{n-p}$. $H_0$ is rejected if $F > F(\alpha; p-1, n-p)$.

Partial test of regression parameter is used to know the effect of regression parameter significantly in model. The testing statistic $t$ of regression model is

\[ t = \frac{b_k}{s(b_k)} \]

$H_0$ is rejected if $|t| > t(\alpha; n-p)$. 

2.2. \textit{Spatial Heterogeneity}

Spatial heterogeneity suggests a diversity of regions. According to Anselin [1], testing of spatial heterogeneity may use the Breusch-Pagan test ($BP$ test) below

\[ BP = \frac{1}{2} \left[ h^T Z(Z^T Z)^{-1} Z^T h \right] \sim \chi^2_k \]

Spatial heterogeneity occurs when the value of $BP > \chi^2_k$.

2.3. \textit{Geographically Weighted Regression}

Geographically weighted regression (GWR) is an expansion of the linear regression model with the assumption of spatial heterogeneity in the error. In geographical weighted regression the dependent variable depends on the location of the region. The weighted regression model is geographically defined as

\[ y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p-1} \beta_k(u_i, v_i)x_{i,k} + \varepsilon_i, i = 1, 2, \ldots, n, \]
where \( y_i \) is the value of the observation of dependent variable at the location \( i \), \( x_{i,k} \) is the observation value of independent variable at the location \( i \), \( \beta_0(u_i, v_i), \beta_1(u_i, v_i), \ldots, \beta_k(u_i, v_i) \) are the parameters of the weighted regression model, \((u_i, v_i)\) denotes the coordinate point (longitude, latitude) destination location \( i \), and \( \varepsilon_i \) is error on the trend location \( i \) (Fotheringham et al. [4]). The GWR model parameter estimation is done using the weighted least square (WLS) method by assigning different weight to each location. With a weighted location in \( i \) against another location \( j \) is \( w_{ij}(u_i, v_i) \) then the sum of squared errors that have been weighted based on

\[
\sum_{i=1}^{n} w_{ij} \varepsilon_i^2(u_i, v_i) \quad y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p-1} \beta_k(u_i, v_i)x_{i,k} + \varepsilon_i, \quad i = 1,2,\ldots,n, \quad (8)
\]

so that obtained GWR model parameter estimation for each location is

\[
\hat{\beta} = (X^T W(u_i, v_i) X)^{-1} X^T W(u_i, v_i) Y \quad (9)
\]

where \( W(u_i, v_i) = diag[w_{i1}, w_{i2}, \ldots, w_{in}] \), \( X \) is independent variable matrix, and \( Y \) is the dependent variable matrix.

The spatial weights are used to estimate the parameters of the GWR model. The following Gaussian kernel function will be used

\[
w_{ij} = \exp\left(-\frac{d_{ij}^2}{2b^2}\right) \quad (10)
\]

where \( d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \). \( d_{ij} \) is the euclidean distance between locations \( i \) and \( j \) while \( b \) is bandwidth. Bandwidth is the circle radius where the point within it is considered still influential in the formation of model parameters. The optimum bandwidth used is the one that produces the minimum cross validation (CV) value, by the formula

\[
CV = \sum_{i=1}^{n} \sum_{j=1}^{n} [y_i - \hat{y}_{j \neq i}(b)]^2 \quad (11)
\]

where \( \hat{y}_{j \neq i}(b) \) is the estimated value of the dependent variable with on-site observations \( i \) omitted from the prediction process (Fotheringham[4]).

2.4. Mapping Models.

Model mapping is done by determining the factors that affect the dependent variable. To determine the factors that influence significantly on the dependent variable used partial test parameters. Partial test is used to know the significance of parameter \( \hat{\beta}(u_i, v_i) \) against the dependent variable partially on the GWR model with test statistic

\[
t = \frac{\hat{\beta}_k(u_i,v_i)}{SE(\hat{\beta}_k(u_i,v_i))} \quad (12)
\]

Where \( \hat{\beta}_k(u_i, v_i) \) is an estimated parameter of the GWR model, and \( SE(\hat{\beta}_k(u_i, v_i)) \) is a standard GWR model error obtained from \( \sqrt{Var(\hat{\beta}_k(u_i, v_i))} \). \( H_0 \) is rejected if \( |t| > t_{\alpha/2,n-p} \). Partial test of parameters is used to determine the factor map affecting poverty in Central Java Province based on district or city location.

3. Research Methodology

3.1. Data Source

The data used in this study is the percentage of poverty and factors affecting poverty in Central Java Province, for the year 2015 taken from Central Bureau of Statistics (BPS [2]). The data consist of 12 independent variables: labor force participation rate \( X_1 \), percentage of population working as farmers
(X_2), and HDI (X_3) as influencing factors terms of employment. Furthermore, population growth rate (X_4) as an influential factor in terms of population.

The percentage of the population who does not have an elementary school certificate (X_5), the percentage of the population graduated from primary school (X_6), the percentage of the population graduated from junior high school (X_7) percentage of the population graduated from senior high school (X_8) as a factor affecting poverty in education. In terms of food needed factors affecting poverty ie minimum wage work (X_9), per capita expenditure (X_10). The percentage of households with regular sanitation (X_11) and BPJS beneficiaries (X_12) as influential factors in terms of health.

3.2. Analysis Method
The research is done by first determining global regression model with stepwise regression method and estimating regression coefficient parameter with ordinary least squares method (OLS). The next step is to test the classical assumption of regression analysis. In addition to the classical assumption test, the next test is to identify the spatial effects on the data by using spatial heterogeneity tests. If the assumptions of spatial heterogeneity are met then we can use geographically weighted regression to overcome the heterogeneity problems contained in the model. Furthermore, the geographical weighted regression coefficient parameter is estimated using the under weighted least square method with the kernel weighting function Gaussian. For deriving the model we used R programing.

4. Analysis and Discussion

4.1. Ordinary Least Square
By using the least square method, we obtained linear regression model for poverty percentage data in Central Java in 2015 which is

\[ Y = 3.736 + 0.084X_2 - 11.622X_4 + 0.120X_{11} + 0.090X_{12} \]  

From the model it can be seen that from 12 independent variables there are only four independent variables that influence significantly. There are percentage of population working as farmer (X_2), population growth rate (X_4) percentage of households with regular sanitation (X_11), and BPJS beneficiaries (X_12).

The normality and multicollinearity assumptions in this model has been fulfilled. From the test of spatial heterogeneity, we obtained BP v value of 11.438 which results greater than the value of \( \chi^2_{(0.1;4)} = 7.779 \) which means there is spatial heterogeneity. For that, we use geographically weighted regression to determine the model.

4.2. Geographically Weighted Regression
The geographically weighted regression model is formed based on four independent variables that have been obtained from the global regression model i.e. (X_2), (X_4), (X_11), (X_12). In determining the estimation value of GWR model parameters depends on the weighting of each location. For that, it needs to be determined first bandwidth optimum obtained from the minimum CV.

The bandwidth value with the gaussian kernel weighted function is 2.7939 with minimum CV 285.3083. Then the bandwidth value is used to determine the weighting matrix of each GWR. Based on the equation \( W(u_i, v_i) = diag[w_{i1}, w_{i2}, ..., w_{in}] \) we have a weighted value for each location. The weighted value is then used to determine the estimated parameters of the GWR model according to the equation (9).

From geographically weighted regression model it can be arranged classification of area in Central Java Province based on factors that affect poverty as in Figure 1.
Figure 1. The division of regions is based on factors affecting the poverty of districts in the province of Central Java.

Figure 1 shows region mapping based on geographically weighted regression model with kernel weighted function Gaussian. With the color difference shows the independent variables that affect the percentage of poverty in the area.

References
[1] Anselin L 1998 *Spatial Econometrics: Methods and Models* (Dordrecht: Kluwer Academic Publishers)
[2] Badan Pusat Statistik 2016 *Provinsi jawa tengah dalam angka tahun 2016* (Jakarta)
[3] Chaudry A, and Wimer C 2016 Poverty is Not Just an Indicator: The Relationship between Income, Poverty, and Child Well-Being *Academic Pediatrics* 16 S23–29.
[4] Fotheringham S, Brunsdon C, and Charlton M 2002 *Geographically Weighted Regression - The Analysis of Spatially Varying Relationships* (Chichester: Wiley)
[5] Nugroho H 1995 *Kemiskinan, Ketimpangan, dan Kesenjangan* (Yogyakarta: Aditya Media)