Application of the Geographically Weighted Regression (GWR) with the Bi-Square Weighting Function on the Poverty Model in the City/Regency of West Java

Euis Sartika[1,*] Sri Murniati[2]

[1] *Departement of Commerce, Politeknik Negeri Bandung*
[2] *Departement of Refrigeration and Air Conditioning Engineering, Politeknik Negeri Bandung*
[*]*Corresponding author. Email: euis.sartika@polban.ac.id*

ABSTRACT
The Geographically Weighted Regression (GWR) analysis is considered the most appropriate analysis to describe the Poverty model, including location. By employing the GWR analysis, this study is aimed to find out the proper model of poverty in West Java Province, which can describe the geographical characteristics of the location or district/city in West Java with the Kernel Bi-Square Weighting function. The secondary data were taken from 2018 which consist of the response variable of the Poor Percentage (PP) and the independent variables which cover the Open Unemployment Rate (OUR), Human Development Index (HDI), Gross Regional Domestic Product (GRDP), Population Density Level (PDL), Regional Minimum Wage (RMW), Poor Population Percentage aged 15 years and high school (PPHS), and Literacy Rates (LR). These variables are estimated to influence the rate of poverty. Multiple regression (Global) and GWR regression (Local) were applied in the analysis, and the weighting function used was Bi-Square. Whereas the best model was selected using the criteria of the coefficient of determination R2 and the value of AIC. The results showed that the local GWR regression model has a coefficient of determination (R2) of 0.9253, meaning that the independent variables could explain 92.53% of the variation in the Poverty Percentage model. The remaining 7.47% is explained by other factors. Besides, the value of the global regression coefficient of determination is 0.7084. The AIC value for GWR is 352.437, and the AIC value for global regression is 363.227, meaning that the error value for GWR is smaller than the global regression. Thus, it can be concluded that the GWR (local) regression model is considered a better model. The variables that affect the percentage of poor people in the global regression model are the Open Unemployment Rate and the Regional Minimum Wage. Meanwhile, the variables that affect the percentage of poor people for 27 cities/districts of West Java vary.

Keywords: *GWR, poverty, Bi-Square, coefficient of determination, AIC.*

1. INTRODUCTION

Poverty is one of the critical problems faced by the government. In 2018, the government targeted a reduction in the poverty rate by 4.1% to 5%. Unfortunately, the target was not achieved yet. According to BPS, the poverty rate in 2018 was still at 7.25%. Nationally, this decline only reached 13.79% or ranked third after East Java (16.72%) and Central Java (15.06%) [1]. To lower the poverty rate, finding out the factors causing poverty is essential.

The causes of poverty may be explained due to several factors: work productivity, economic growth, income per capita, income inequality, health facilities and services, nutrition and disease outbreaks, infant mortality rates, and educational facilities and curriculum that are less relevant [2]. Many scholars have conducted studies concerning these factors in West Java province using various statistical analyses. The results show that the main factors of poverty in West Java are low human development and open unemployment, while economic growth and others do not significantly affect poverty. However, the study which includes spatial elements in investigating the poverty in West Java in 2018 has not been carried out yet. By including spatial elements in the study, a model describing the characteristics of the various cities/ districts of West Java could be formed.Thus, 27 poverty models adjusted to the number of cities/districts of West Java can be obtained. In conducting the study, there were two types of data. First, the factors used as predictor variables are HDI, GRDP, PDL, LR, OUR, and PPSH. Second, the dependent variable is the percentage of poor population in the

city/district of West Java in 2018. The data were analysed using descriptive analysis, GWR, and classical regression analysis. Then, from the local regression model and global regression, the best model was selected based on the coefficient of determination and AIC. This research aims to examine the factors that influence the poverty level of cities/districts in West Java based on the spatial characteristics of the cities/districts to find out the best poverty model from global and local regression models referring to determination coefficient and AIC analysis. It is expected that the research on the poverty model in West Java using GWR with the bi-square weighting function can provide information for the provincial government or West Java city government regarding poverty alleviation policies in West Java

2. BACKGROUND

The purpose of this research is to examine the factors that influence the poverty level of cities/districts in West Java based on the spatial characteristics of the cities/districts to find out the best poverty model from global and local regression models referring to determination coefficient and AIC analysis. It is expected that the research on the poverty model in West Java using GWR with the bi-square weighting function can provide information for the provincial government or West Java city government regarding poverty alleviation policies in West Java.

2.1 Poverty

The central statistics agency (BPS) measures poverty conditions based on the poverty line (GK). GK is defined as the sum of the food poverty line (GKM) and the non-food poverty line (GKNM) [2]. The poor are the people who are under GK. The temporary poverty rate is the percentage of poor people in the total population. Poverty is influenced by many factors, including productivity for work, economic growth, income per capita, income inequality, health facilities and services, nutrition and disease outbreaks, infant mortality rates, and educational facilities and curriculum that are less relevant [1].

2.2 Global and Local Regression Analysis (GWR)

Spatial regression is an extension of the classical linear method. This development is due to the influence of place or spatial on the analysed data [3]. If there are data with spatial effects, the analysis used is spatial regression analysis. Multiple regression will give less accurate results and inaccurate conclusions if this is used because the assumption of independent error is not met. Based on the data type, spatial modelling can be divided into a point and area approach. Point data shows the locations in the form of points, for example, in the form of points on longitude and latitude. Line data is used to describe something with a long path, not an area, for example, contour lines, road networks, electric rivers, etc. Area data shows a location of an area, such as a country, district, city and so on [4].

One of the statistical methods that can be used to analyse risk factors spatially with a point approach is the spatial model of GWR. This method extends the global regression model framework into a local regression model that allows local parameter estimation. Each regression parameter is estimated at each geographic location point so that the relationship between the response variable (Y) and the explanatory variable (X) varies across locations. The selection of a weighting matrix is one of the main steps in GWR because it will greatly affect the resulting GWR model [5]. The most important thing in the GWR model is the weighting because the weight is the value for each location. Near location has a strong influence in estimation from a far location. Methods used to determine the GWR model's weight are kernel functions, including the gaussian distance and the bi-square function. There are several ways to determine the elements of the weighting matrix in the GWR, one of which is the weighting that adopts the kernel distribution function. Kernel density function is often used in data smoothing by giving weighting according to the optimal bandwidth whose value depends on the condition of the data. The kernel function used in the GWR weighting matrix in this study is a bi-square kernel function that uses the distance between locations in its function.

The form of the regression model with k predictor variables and the number of observations is as follows:

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + \varepsilon_i \quad , \mathrm{i} = 1,2,\ldots,\mathrm{n} \qquad (1)$$

In the global regression, the relationship between the independent and dependent variables is considered the same in each geographic location. The multiple linear regression model assumptions are as follows: homoscedasticity, autocorrelation, multicollinearity, and the error is normally distributed [6]. The model parameter test consists of a partial test and a simultaneous test of the response variables' independent variables.

The GWR model equation parameters for each observation location are different from other locations, a global regression development. The GWR model cannot estimate parameters other than those at the observation site [7]. The GWR model is denoted as follows:

$$Y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) X_{ik} + \varepsilon_i \qquad (2)$$

Description:
Y_I: the response variable at the i-th observation POINT
X_{ik} : the k-th independent variable at point p
(u_i, v_i) **:** coordinates of the i-th observation point
$\beta_o(u_i, v_i)$ **:** intercept model GWR

$\beta_k(u_i, v_i)$: the k-th regression coefficient at the i-th observation point

ε_i : error at the i-th location point

To estimate the parameter (u_i, v_i) at the location, the method of WeightedLleast Squares (WLS) is used [8]. The WLS method gives different weights from all the observation data. Weighting is based on the distance between the locations of the observation data. The parameter estimator of the GWR model is stated as follows:

$$\beta_k(u_i, v_i) = (X^T W(u_i, v_i) X)^{-1} X^T W(u_i, v_i) Y \quad (3)$$

Description :

$$W_i = diag[w_i(u_i, v_i), w_j(u_j, v_j), ..., w_n(u_n, v_n)] \quad (4)$$

W is a different weighting matrix for each location parameter prediction data. The weighting value is highly dependent on the distance between the observation points. The weighting function is used to estimate different parameters at each observation point. The parameter estimate depends not only on the value of the explanatory variable but also on the bandwidth used in calculating the weights. Bandwidth is the radius of the estimation area centred at the i-th observation point. Observation points located in the area have an influence on parameter estimation at the i-th observation point. The selected kernel function is the one that produces the optimum bandwidth by looking at the Akaike Information Criterion (AIC) and Cross-Validation (CV) values. The smaller the AIC and CV values produced by the kernel function, the better the resulting weights [9].

Spatial autocorrelation is an estimate of the correlation between observed values related to the spatial location of the same variable. Spatial autocorrelation measurements for spatial data can be calculated using the Moran's Index [10]. Calculation of Moran's Index is one of the spatial analysis techniques that can be used to determine the presence of spatial autocorrelation between observation locations [11] as follows:

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (5)$$

The commonly used weighting method is the Bi-square kernel which is denoted as:

$$w_{ij} = \begin{cases} \left[1 - \left(\frac{d_{ij}}{b}\right)^2\right]^2 & , if\ d_{ij} < b \\ 0 & , other \end{cases} \quad (6)$$

Euclidean distance is between the -th observation point and the -th observation point [12].

$$d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2} \quad (7)$$

2.3 Bandwidth Determination

Cross-Validation (CV) is one of the criteria to obtain bandwidth optimum. The bandwidth optimum will be inversely proportional to CV. In other words, bandwidth optimum will be obtained when the CV is getting smaller. Bandwidth optimum is essential to get weighting value which is used in calculation process of spatial regression model denoted as:

$$CV(h) = \sum_{i=0}^{n}(y_i - \hat{y}_i(h)^2) \quad (8)$$

2.4 Test Goodness of fit Model GWR

The F test is used for testing the goodness of the GWR model and is denoted as:

$$F = \frac{SS(R)_{OLS} - SS(R)_{GWR}/df_2}{SS(R)_{GWR}/df_2} \quad (9)$$

Description: SS: Sum Square
R: Residual

Hypothesis :

H_o: GWR and OLS models are the same

H_1: GWR model is better than the OLS model.

If the results of the F test are not significant, then there is no difference between the GWR and the OLS model, whereas, if H_o is rejected, then proceed to the model parameter test step.

2.5 Partial GWR model parameter testing

Tests are carried out to find out which parameters significantly affect the response variable. The form of the hypothesis is as follows:

H_o : $\beta_k(u_i, v_i) = 0$ (no independent variable affects the dependent variableβ_k

H_1: $\beta_k(u_i, v_i) \neq 0$; $k = 1,2,\ldots p$ (there is at least one independent variable that affects the dependent variable)

The test statistics used are:

$$t_{count} = \frac{\hat{\beta}_k(u_i, v_i)}{SE[\hat{\beta}_k(u_i, v_i)]} \quad (10)$$

Test criteria: if $|t_{count}| > t_{(1-\alpha/2, -p-1)}$ then H_o is rejected. This means that $\hat{\beta}_k(u_i, v_i) \neq 0$ or, in other words, the local regression coefficient $\hat{\beta}_k(u_i, v_i)$ obtained for the GWR model is significant.

2.6 Best Model Testing

- The coefficient of determination (R2) is used to measure the suitable level of the regression line model, which is denoted as:

$$R_i^2 = \frac{SST_{GWR} - SSR_{GWR}}{SST_{GWR}} \quad (11)$$

Description:
SST = Sum Square Total
SSR = Sum Square Residual

- AIC are methods that can be used to select the best regression model found by Akaike and Schwarz [12]. Both methods are based on the maximum likelihood estimation method. To calculate the AIC values, the following formula is used:

$$AIC = e^{\frac{2k}{n}} \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n} \quad (12)$$

Description:
k = number of parameters estimated in the regression
n = number of observations ; e = 2.718 ; u_i = Residual

According to the AIC methods, the best regression model is a regression model that has the smallest AIC values

3. METHODOLOGY

The data used was secondary data from the publication of the Central Statistics Agency (BPS) of West Java in 2018. The data used as the response variable was the percentage of the poor (PP). Several predictor variables were considered affecting the response variables, namely the Open Unemployment Rate (OUR), Gross Regional Regional Product (GRDP), Human Development Index (HDI), Population Density Level (PDL), Regional Minimum Wage (RMW), percentage of the poor population aged 15 years of high school graduates (PPHS), and Literacy Rate (LR). As the point of observation, the latitude and longitude coordinate data were used to calculate the distance between districts/cities in West Java.

The steps of analysis carried out in this study are as follows:

- Describing the variables as an initial description of poverty in West Java Province along with the factors that are thought to influence it.
- Analyzing the global regression model with the following steps: classic assumption test, global regression model, parameter significance test, simultaneous test, and partial test.
- Analyzing the GWR model with the following steps:
  - calculating the Euclidean distance between the location which lies at the coordinates (u_i, v_i),
  - determining the optimum bandwidth using the CV,
  - calculating the weighting matrix using the Bi-Square kernel function,
  - estimating GWR model parameters using WLS,
  - testing the significance of the GWR parameter,
  - determining the best model using the Coefficient of Determination ($R2$) and AIC.

4. RESULTS

4.1. The variables as an initial description of poverty in West Java Province

The description of poverty data and the factors that influence it in the city/regency of West Java in 2018 is stated as follows:

Table 1. Data Description of Poverty

Variable	Min	Max	Mean	Std.Dev
PP	2.14	12.71	7.942	2.729
OUR	3.68	10.56	7.874	1.830
GRDP	2778.07	215983.05	57040.75	63746.45
HDI	64.62	81.06	70.94	4.817
PDL	78.70	15477.74	3418.31	4833.85
RMW	1558794	3919291	2527033.3	811908.2
PPSH	67	3538	1249.19	992.17
LR	282683	615255	398871.52	84617.29

Several variables that are estimated to affect the poverty rate in cities/districts in West Java have a wide range of values, such as the variables: GRDP, PDL, RM, and LR.

4.2 The steps of the global regression assumption

4.2.1 Normality Assumption Test

Table 2. Normality Test

Tests	Kolmogorov-Smirnov (Sig)	Shapiro-Wilk (Sig)
Unstandardized Residual	0.149	0.499

Based on Shapiro Wilk's sig (probability) value, a value of 0.499 is greater than 0.05, meaning that the residuals are normally distributed.

4.2.2 Multicollinearity Assumption Test

Table 3. Multicollinearity Test

Var	OUR (X1)	GRDP (X2)	HDI (X3)	PDL (X4)	RMW (X5)	PPHS (X6)	LR (X7)
VIF	1.06	2.75	1.11	1.80	2.95	1.29	2.03

Table 3 shows that the VIF value for each independent variable is below 10, which means multicollinearity between the independent variables does not occur.

4.2.3 Autocorrelation Assumption Test

For the autocorrelation test, the Durbin-Watson (DW) value was 2.487, and the values for du and dl based on the table with n=27 and p=6 were 1.86079 and 1.00421, respectively. Because the value of dw=2.487 lies between (4-du)=2.13921 and (4-dl)=2.99579, it means that it cannot be concluded. So the Run test is used, and the P-Value value is 0.235, which is greater than 0.05, which means that there is no autocorrelation.

4.2.4 Heterogeneity Assumption Test

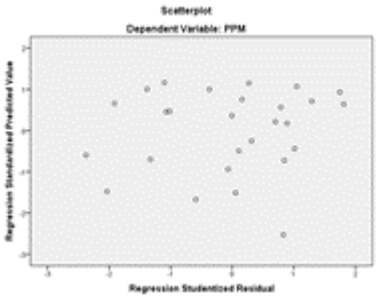

Figure 1 Scatterplot of Percentage of Poverty.

Based on Figure 1, it can be shown that the data points are scattered irregularly and lie on the positive and negative axes. This shows that the data does not contain heteroscedasticity.

4.2.5 Global Regression Model

Table 4 Predictor Variable Estimation

Variable	Coefficient	Std. Error	t_{stat}
CONSTANT	2451.53	1322.65	1.8535
OUR	0.53884	0.233335	2.309291
GRDP	0.000006	0.000023	0.276927
HDI	-0.25198	0.216807	-1.162212
PDL	-0.000064	0.000134	-0.477182
RMW	-0.000198	0.000069	-2.885031
PPHS	-0.033533	0.047365	-0.707971
LR	0.000644	0.000669	0.964043

Based on table 4, it can be shown that the variables that affect the percentage of poor people globally (applicable to all districts/cities in West Java) are OUR and RMW because the smaller $\mathbf{t}_{\mathbf{stat}}$ for OUR is greater than $\mathbf{t}_{\mathbf{table}}$= 2.093 and the $\mathbf{t}_{\mathbf{stat}}$ for the UMR variable is smaller than the value of $\mathbf{t}_{\mathbf{table}}$ = -2,093.

The global regression model formed is stated as follows:

PP= 2451.53+ 0.538838 OUR+ 0.000006 GRDP - 0.251976 HDI -0.000064 PDL - 0.000198 RMW - 0.033533 PPHS + 0.000644 LR

Interpretation : An increase in HDI, PDL, RMW, and PPHS values can reduce the percentage of poor people.

4.3 GWR Model Analysis

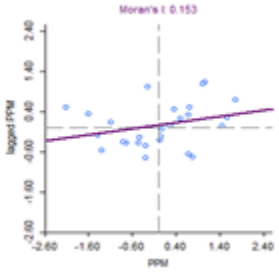

Figure 2 Moran Scatterplot of Percentage of Poor Population.

Based on Figure 2, it can be shown that the Moran Index (I) is 0.153, which is in the range 0< I <1, indicating a positive spatial autocorrelation. Based on the Moran Index significance test with a significance level of 5%, it can be concluded that between cities/districts, one does not have similar values or indicates that the percentage of poor people between cities/districts in West Java is not correlated.

The first step in the GWR analysis is to determine the bandwidth to be used in the Kernel Bi-square weighting function. The bandwidth for the Bi-square kernel weighting function is 10565.891, which gives CV= 373032,196931

The model suitability test is carried out using F. As shown in the following table:

Table 5. GWR Anova

Source	SS	DF	MS	F
Global Residual	64768.734	19		
GWR Improvement	420094.457	14.832	28324.069	
GWR Residual	144674.277	4.168	34708.326	0.81606

Based on the GWR Anova table, it can be shown that the value of F_{count} = 0.816060 is smaller than F_{table} = 2.92. It means that accepting Ho means that the local model (GWR) has almost the same accuracy as the Global model.

The comparison of the accuracy of the global and local regression models (GWR) results is shown in the following table 6.

Table 6. Comparison of GWR model and global regression

	Global Regression	Local Regression (GWR)
R^2	0.708425	0.925309
AIC	363.227719	352.436965

Table 6 explains that the local regression model is relatively better than the global regression model because the Coefficient of Determination (R2) value is greater (0.925309). It means that 92.53% of the percentage of the Poor Population model is influenced by the independent variables, namely OUR, HDI, PDL, RMH, PPHS, and LR. In terms of model accuracy, Local regression is also better because the AIC value of Local Regression is 352.436965, which is smaller than the Global regression model of 363.227719. It indicates that the error in the Local regression model is smaller than the global regression. Based on the GWR output, it can be concluded that the variables which affect the percentage of poor people in cities/districts in West Java are as follows:

Table 7. The Variables Affecting Percentage of Poor Cities/Regencies in West Java Locally

City / District	Significant variables
Bekasi, Sumedang, Kabupaten Bekasi,	OUR
Bekasi, Kabupaten Bekasi, Bogor, Tasikmalaya	RMH
Banjar	PPSH
Indramayu, Kabupaten Bekasi	LR

Table 7 shows that the poverty rate in West Java is divided into four groups of cities/districts in West Java. This grouping is based on the influencing variables.

Table 8. Geographical variability tests of local coefficients

Variable	F	DOF for Criterion	Diff of
Intercept	518.162	1.176	-122.64
OUR	10.858	0.706	-20.80
GRDP	0.9014	0.930	-1.666
HDI	4796.05	1.834	-193.153
PDL	1.522	0.842	-3.536
RMW	4.782	0.962	-13.423
PPHS	1.563	1.101	-4.596
LR	43.109	1.432	-62.544
F_{test}=6.009			

Based on the table of Geographical variability tests of local coefficients, it can be shown that the DIFF of Criterion values are all negative. This means that there are spatial variations between adjacent locations for the variables OUR, GRDP, HDI, PDL, RMW, PPHS, and LR.

There are 27 local regression models, because there are 27 districts/ cities in West Java, one of which is the city of Cirebon:

PP=1007.352 + 0.533 OUR +0.000006 GRDP +0.003 HDI-0.000278 PDL - 0.000203 RMH + 0.012122 PPHS - 0.000169 LR

5. DISCUSSION

Based on the study results, the variables that affect poverty in the cities/districts of West Java globally are OUR and RMW. Meanwhile, the variables that affect poverty locally vary for each city/district in West Java. The best regression model is the GWR model because it has a coefficient of determination of 0.925309 l and an error value or AIC of 352.436965.

6. CONCLUSION

Based on the results of the discussion, the following conclusions can be drawn:

1. Globally, the variables that affect the percentage of poor people in the cities/districts of West Java are OUR and RMW.
2. The local regression model is relatively better than the global regression because it has a greater coefficient of determination, that is 0.925309 compared to the global regression of 0.708425, and the error value (AIC) is smaller, namely 352.436965, compared to the global regression of 363.227719.
3. There are 27 different local regression models according to the number of cities/districts in West Java. Likewise, the variables that affect the Poverty Level for each city/district vary.
4. There are four different city/district area groups in West Java based on the poverty level variables.

REFERENCES

[1] BPS, "Https : // W W W . B . G," *BPS, 2019 (Indonesian Stat.*, p. Jakarta: Badan Pusat Statistik, 2019.

[2] M. Smith, Stephen and Todaro, *Economic Development, 12th Edition (The Pearson Series in Economics)*. 2003.

[3] C. Wang, "The impact of car ownership and public

transport usage on cancer screening coverage : Empirical evidence using a spatial analysis in England," *JTRG*, vol. 56, pp. 15–22, 2016, DOI: 10.1016/j.jtrangeo.2016.08.012.

[4] C. H. Lin and T. H. Wen, "Using geographically weighted regression (GWR) to explore spatial varying relationships of immature mosquitoes and human densities with the incidence of dengue," *Int. J. Environ. Res. Public Health*, vol. 8, no. 7, pp. 2798–2815, 2011, DOI: 10.3390/ijerph8072798.

[5] Gujarati, *Basic Econometrics 4th (fourth) Edition by Gujarati published by McGraw-Hill Inc., US (2003)No Title*, 4th ed. US: McGraw-Hill Inc., US (2003) Hardcover, 2003.

[6] J. Walter, R. Carsten, and L. Jeremy W, *Local and Global Approaches to Spatial Data Analysis in Ecology*. Global Ecology and Biogeography 14, 2005.

[7] C. Brunsdon, A. S. Fotheringham, and M. E. Charlton, "Geographically weighted regression: a method for exploring spatial nonstationarity," *Geogr. Anal.*, vol. 28, no. 4, pp. 281–298, 1996, doi: 10.1111/j.1538-4632.1996.tb00936.x.

[8] A. . Fotheringham, C. Brunsdon, and M. Charlton, *Geographically weighted regression: the analysis of spatially varying relationships*. UK: John Wiley & Son, 2003.

[9] Pfeiffer D and A. Et, *Spatial Analysis in Epidemiology*. New York: Oxford University Press, 2008.

[10] R. Kosfeld, *Spatial Econometrics*. 2007.

[11] M. C. Fotheringham, A. Stewart, Chris Brunson, *Geographically Weighted Regression, The Analysis of Spatially Varying Relationships*. UK: John Willey and Sonc Inc, 2002.

[12] Grasa Aznar Antonio, *Econometric Model Selection A New Approach*. Kluwer Academic Publishers;, 1989.