Advances in Engineering Research, volume 198
International Seminar of Science and Applied Technology (ISSAT 2020)

Comparison of Geographically Weighted Regression Analysis and Global Regression on Modeling the Unemployment Rate in West Java

Euis Sartika[1,*] Anny Suryani[2]

[1]*Jurusan Administrasi Niaga, Politeknik Negeri Bandung, Indonesia*
[2]*Jurusan Akuntansi, Politeknik Negeri Bandung, Indonesia*
[*]*Corresponding author. Email: euis.sartika@polban.ac.id*

ABSTRACT
This study aims to identify the factors Unemployment Rate (UR) in West Java and develop the appropriate model. This study applied the location (spatial) element using Geographically Weighted Regression (GWR). The GWR model was compared with the global regression. The data used in this study are secondary data on 2017 UR for 27 cities/ regencies in West Java. The dependent variable (Y) is the Unemployment Rate (UR), the independent variables include Population Density Level (PDL), Gross Regional Domestic Product(GRDP), Regional Minimum Wage (RMW), Level of Active Labor Participation Rate (ALPR), and Human Development Index (HDI). The results show that the GWR model provides a coefficient of determination (R^2) more significant than the global regression model. The Akaike Information Criteria (AIC) value of the GWR model is smaller than the global regression model, meaning that the local regression model of error value is smaller than the global regression model. In other words, the local regression model is better than the global regression model. The factor affecting UR globally is RMW. There are 27 different combinations of local regression models according to the number of cities/regencies in West Java.

Keywords: *Unemployment Rate, Geographically Weighted Regression, West Java*

1. INTRODUCTION

One of the International Labour Organization (ILO) targets is to reduce unemployment in every country. In measuring the Unemployment Rate (UR), ILO uses the Open Unemployment Rate (UR). Referring to the tool, Indonesia has also applied the tool to measure Open Unemployment Rate. Statistical analysis, i.e., global regression, is employed to find the relationship between UR and its explanatory variables. However, the results show that estimator parameters could be only accepted generally. The unemployment problem in West Java is a very complex problem because it can trigger other social problems. BPS (Central Bureau of Statistics) states that the unemployment rate in West Java is higher than the national figure, which is 5.01 [1]. The high level of the unemployment rate has an impact on the low level of income of the people of West Java. Low levels of education and high levels of unemployment cause a more complex social problem, namely poverty. In 2017, the highest unemployment rate in Indonesia was recorded in West Java province at 7.73%. Ningtias and Rahayu (2017) state that unemployment has a spatial aspect [2]. Spatial effects can appear from one region to another. The reason is that each region has a relationship with one another, whether the relationship is caused by distance, or due to the similarity of culture and characteristics. This is the reason for choosing the spatial method as a statistical analysis that considers the area factor of the data taken. This is caused by conditions at one point or region that are in accordance with conditions at one point or adjacent area [3]. Research conducted by Amalia E, and Kurnia Sari L conducted spatial analysis research on the island of Java with a period 2017 by stating the pattern of distribution of disturbance levels in the western part of Java Island and the variables that influence the area are HDI, RMW, and ALPR [4]. Research conducted by Aini N, Wahyu Utami T, Karim A (2018), modeled the number of unemployed in West Java using the Mixed geographically weighted regression states that there is no difference between the global regression model and the local regression, with the RMW variable influencing globally and the HDI influencing locally [5].

Several studies on the unemployment rate have been carried out using different methods. However, research that uses the spatial effect of West Java city/regency objects in the 2017 period has not been carried out. For this reason, this study was carried out by comparing global regression models and local regression. This comparison needs to be done, because the global regression model provides estimated variable values in general, applicable to all cities/regencies in West Java. While the local regression model, the estimation of the variability is local. So, there are 27 local regression models regarding the unemployment rate according to the number of cities/regencies in West Java. With the knowledge of the factors that affect the local unemployment rate, of course, the district government can prioritize these factors partially.

The purposes of this study are to identify the factors influencing UR in the 27 cities/regencies situated in West Java in 2017. By identifying the factors, it is expected to form the appropriate model of UR using GWR and global regression. The models obtained are then compared based on the AIC value and the coefficient of determination (R Square).

2. BACKGROUND

2.1. Global Regression (Multiple Regression)

The general equation of the linear regression model for n observations and m predictor variables [6], can be written as follows:

$$Y_i = \beta_0 + \sum_{k=1}^{p} \beta_k X_{ik} + \varepsilon_i \quad (1)$$

i=1,2,…,n

Y_i is the response variable in the i-th; intercept $\beta_k$is the value of the X_k predictor variable function on the i-th observation, and ε_i Testing global regression assumptions include residual in the global regression model must have the characteristics: normally distributed, nonmulticollinearity, non-autocorrelation, and nonheteroscedasticity. The global regression model uses Ordinary Least Square (OLS) strongly strict on a few assumptions. If there are one assumptions that are not met, then there are clues to the existence of spatial effects. Normality assumption test to determine whether the residuals are normally distributed. The test used in this study was Shapiro Wilk. Autocorrelation test assumption, aims to determine if there is a correlation between residue Testing is carried out using Durbin-Watson [7]. Multicollinearity means there is a correlation between some or all of the predictor variables. Testing is done by looking at variance inflation factor (VIF). If the VIF value is less than 10 then does not occur multicollinearity [8].

2.1.1. Estimation of Global Regression Parameters

To estimate the global regression model, the OLS (Ordinary Least Square) model is used by minimizing the number of squared errors [9], Then the β estimate is obtained as follow:

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (2)$$

The F test is used to determine the effect of all independent variables on the response variable [10]. The F test is usually expressed in ANOVA. The t test is used to investigate whether the independent variables individually affect the response variable.

2.2. *Local Regression*

A regression method that produces parameter estimators that can predict the response of each location is GWR that can estimate data containing spatial heterogeneity and formulated as follows [9] :

$$y_i = \beta_o(u_i v_i) + \sum_{j=1}^{p} \beta_j(u_i, v_i) + \varepsilon_i \quad (3)$$

$\beta_j(u_i, v_i)$ is the regression coefficient for the j-th predictor variable for each location, (u_i, v_i) is longitude and latitude for the i-th location, and ε_i is the i-observation random variable. The main thing needed to be done in the GWR analysis is finding optimum value bandwidth. The function of bandwidth is to determine the weight of one location against another location [11]. Testing local regression assumptions include analysis of spatial-dependency using the Moran Test and Lagrange Multiplier, spatial diversity test using Breusch-Pagan, determining the optimal window width value by looking at Cross-Validation (CV), determining the goodness of the model using the F test [7]

2.2.1. Estimation of Local Regression Parameters

Estimation of parameters in the GWR model uses Weighted Least Square (WLS), which is a different weighting for each location where the data is taken. [11]. In the GWR model, suppose that the area close to the first observation location has a greater influence on the parameter estimation compared to the distant area. For example the weight for each location (u_i,v_i) is $W_i(u_i v_i)$ with i=1,2,…,n. Minimize the number of squares of the error, then reduce it and make it equal to zero. Then the parameter estimation of the GWR model for each location is as follows :

$$\hat{\beta}(u_i, v_i) = (X^T W(u_i, v_i) X)^{-1} X^T W(u_i, v_i) y \quad (4)$$

Testing of Spatial Effects using the Moran Index is carried out to see whether each variable affects spatial location. If the value of the Moran Index for each variable is greater than the value of E[I], it means that there is positive autocorrelation of the dependent and independent variables. Conversely, if it is less than, it means that there is negative autocorrelation.

$$E[I] = \frac{-1}{n-1} \quad (5)$$

Moran's index is used to calculate the strength of the correlation between observations as a function of distance. Moran's index calculates spatial autocorrelation, positive autocorrelation occurs when neighboring values provide similar values (grouping). Negative autocorrelation occurs when neighboring values give different values (dispersal). Moran's I statistics for spatial autocorrelation are defined as follows [8] :

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij} z_i z_j}{S_0 \sum_{i=1}^{n} z_i^2} \quad (6)$$

n : total number of spatial units indexed by i and j
i dan j : spatial units
z_i: deviation of an attribute for feature I from its mean $(x_i - \bar{X})$
x_i: variable interest
$\bar{X}$: mean of x_i
W_{ij}: the spatial weight between feature i and j
S_0: the aggregate of all the spatial weight

$$S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij} \quad (7)$$

The score for the statistic is computed as :

$$z_j = \frac{I - E[I]}{\sqrt{V - [I]}} \quad (8)$$

Which based on :

$$E[I] = \frac{-1}{n-1} \quad (9)$$

$$V[I] = E[I^2] - E[I]^2 \quad (10)$$

$$E[I^2] = \frac{A-B}{C} \quad (11)$$

$$A = \{(n^2 - 3n + 3)S_1 - nS_2 + 3{S_0}^2\} \quad (12)$$

$$B = D\{n^2 - n)S_1 - w2nS_2 + 6{S_0}^2\} \quad (13)$$

$$C = -(n-1)(n-2)(n-3){S_0}^2 \quad (14)$$

$$D = \frac{\sum_{i=1}^{n} z_i^2}{(\sum_{i=1}^{n} z_i^2)^2} \quad (15)$$

$$S_1 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} (w_{ij} + w_{ji})^2 \quad (16)$$

$$S_2 = \sum_{i=1}^{n}\left(\sum_{j=1}^{n} w_{ij} + \sum_{j=1}^{n} w_{ji}\right)^2 \quad (17)$$

The Lagrange Multiplier Test (LM) is used to select a suitable spatial regression model and can also be used to identify the existence of a spatial model [8].

2.3 Weight and Bandwidth Matrix

The weighted value represents the location of the observation data from one another. In this study using a weighting function, the Gaussian distance function is denoted as follows [11] :

$$W_j(u_i, v_j) = \varphi\left(\frac{d_{ij}}{\sigma h}\right) \quad (18)$$

The Inverse Distance function has a weight of 0, if location j is outside radius b from location I, and is 1 if it is within radius b.

$$d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2} \quad (19)$$

Meanwhile, the Gaussian Kernel functions are:

$$W_j(u_i, v_j) = \exp\left(-\frac{1}{2}\left(\frac{d_{ij}}{h}\right)^2\right) \quad (20)$$

The Gaussian function assigns a weight whose value will decrease according to the Gaussian function as it increases.

2.4 Selection of the Best Model

a. The coefficient of determination (R2) is used to measure the level of fit of the model (goodness of fit) of the regression line, which is denoted :

$$R^2(u_i, v_i) = \frac{JKR_W}{JKT_W} - \frac{\sum_{j=1}^{p}(y_j - \hat{y_j})^2}{\sum_{j=1}^{p}(y_i - \bar{y})^2} \quad (21)$$

b. Akaike's Information Criterion (AIC)

The best model is determined by the smallest AIC value, which is notated as follows [11]:

$$AIC = 2n\log(\hat{\sigma}) + n\log(2\pi) + n + tr(L) \quad (22)$$

$\hat{\sigma}$ = Estimator value of the standard deviation of the error resulting from the maximum estimation of Likelihood.

L= Projection matrix, $\hat{y} = Ly$

2.5 Unemployment

The ILO (International Labor Organization) states that "The rules of the global economy should be aimed at improving the rights, livelihoods, security, and opportunities of people, families and communities around the world". World Commission on the Social Dimension of Globalization, 2004 (ILO: A Fair Globalization: Creating opportunities for all, Report of the World Commission on the Social Dimension of Globalization) [12]. This means that the world labor organization wants the world economy to be focusing on developing a decent standard of living for the community through providing opportunities for all

people to obtain their rights, to realize this (welfare) a person must have a decent income through free and open employment opportunities, this is what is still a major problem, for economic experts and governments around the world to provide a decent life for the people, because of the limited employment opportunities for these individuals or groups of people, this lack of job opportunities results in an economic phenomenon known as unemployment. The data in this study are secondary data from BPS West Java in 2017 with the object 27 cities/regencies of West Java. Response variables in this study are UR (points), and predictor variables are PDL (population/square km), GRDP (billion), RMW (rupiah), ALPR (percent), and HDI (points). The analysis used was global regression analysis and Geographically Weighted Regression (local regression).

Table 1. Descriptive Statistics

Variables	Minimum	Maximum	Mean	Std. Deviation
UR	3,34	10,97	8,0756	1,73772
GRDP	2491,64	228725,92	50198,5278	55611,07620
HDI	63,70	80,31	70,2789	4,88925
RMW	1433901,00	3605271,00	2301867,7037	717062,14439
ALPR	0,25	2,34	1,4367	0,56870
PDL	391,00	15307,00	2965,1481	3789,46659

Based on table 1, it can be shown that there is a sharp difference in the GRDP and PDL values between the minimum and maximum values. This shows that there are differences in income and regional population density between cities/districts in West Java.

The assumptions for the global regression model consist of:

- normality test, using Shapiro Wilk and the unstandardized residual probability value is obtained is 0.255 more than 0.1, which means that the residual is typically distributed normally.
- heteroskedasticity test, using the Scatter diagram, the following results are obtained :

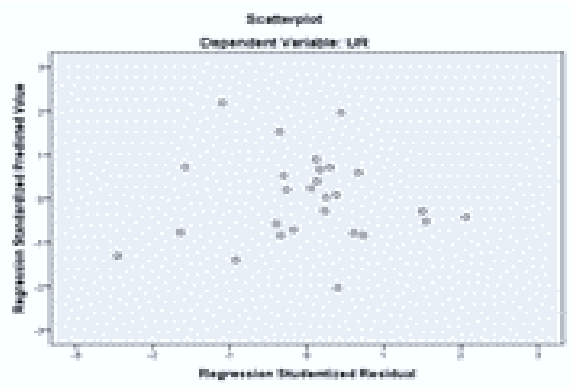

Figure 2. Scatter Plot

Based on Figure 2, it can be shown that the data spreads irregularly (does not form a specific pattern) located above and below the flat axis. This condition shows that the heteroscedasticity assumption is fulfilled.

- multicollinearity test, using the VIF value obtained the following results:

Table 2. VIF of Independent Variables

Variable	GRDP	HDI	RMW	ALPR	PDL
VIF	1,997	2,733	2,154	1,666	1,730

Based on table 2, it can be shown that the VIF value of each independent variable is less than 10, meaning that there is no multicollinearity. Based on table 2, it can be shown that the VIF value of each independent variable is less than 10, meaning that there is no multicollinearity.

- autocorrelation test, using the Durbin Watson value obtained DW = 2.354. Based on the DW table 2, if n = 27 and k = 5, then the value of du = 1,75274 and (4-du)=2,24726. Because DW value is not between du and (4-du) the non-parametric test was done, and the probability value was greater than 0,1 meaning the autocorrelation assumption was fulfilled.

Table 3. Estimated of the Global Regression Model Coefficients

Model	Coefficent	t-Statistic	Probability
Constant	4,683	0,743	0,466
GRDP	6,857E-6	1,028	0,316
HDI	-0.028	-0,311	0,759
RMW	1,528E-6	2,844	0,010
ALPR	0.900	1,510	0,146
PDL	6,075E-5	0,667	0,512

Based on table 3, it can be seen that all of the estimated coefficient values of the global regression model are positive, except for the HDI value. The probability value indicates significant variable, the probability value of the RMW variable is 0.010 less than 0.1, meaning that the RMW variable significantly affects UR. The global regression model is formed as follows:

UR = 4,683+ 6,857E-6GRDP-0.028HDI + 6,075E-5PDL+ 1,528E-6RMW+ 0.900ALPR

The global regression model shows that the coefficient of HDI variable is negative, meaning that each increase in HDI variable by one unit will reduce the UR value by 0,028 units, while the value of the other coefficient of independent variables, namely GRDP,

PDL, RMW, ALPR is positive. The partial parameter estimation test shows that the RMW independent significantly affects the global regression model indicated by the probability value in Table 3, which shows a value of less than 0.1. Meanwhile, the simultaneous estimation test shows that the independent affect the model, as shown by the P-Value. The value indicates in the Anova table of P-Value (F- Stat) = 0.006, less than 0.10. This probability value also shows that the Global regression model formed was moderate.

The next step is to determine the local regression assumption, the first is to check whether there are spatial dependencies.

Table 4. Output Geoda

Test	Value	Probability
Moran"x I (error)	-1.2103	0.22617
Lagrange Multiplier (error)	1,8685	0,17165
Breusch-Pagan test	6,3886	0,27022

Table 4 shows that the local regression assumption for the spatial measure of dependency is Moran's index. Its value is 0.22617, which is greater than 0.10 for the response variable (UR), meaning that there is no spatial effect on the predictor variable. The spatial homogeneity test using Breusch-Pagan produces a value of 0.27022 greater than 0.10, which means it contains an element of heterogeneity. The Lagrange Multiplier value is 0.17165, greater than 0.1, which means that there is no spatial dependence effect on the predictor variable.

Table 5. Index Moran`s I

Variables	Moran Index
UR	-0,099
GRDP	-0,319
HDI	0,223
RMW	0,089
ALPR	-0,176
PDL	-0,205

Based on table 5, it can be shown that the independent variable which has a value greater than E[I] is HDI and RMW, which means that it has positive autocorrelation with the UR variable. It can be concluded that the higher the RMW and HDI values the higher the UR value. Meanwhile, GRDP, ALPR, and PDL have values less than E[I], meaning that there is a negative autocorrelation relationship between these independent variables and UR.

The Lagrange Multiplier test aims to identify spatial dependencies between cities/districts. If the value of the Lagrange Multiplier is greater than 0.1 then there is no spatial dependence [8].

The optimum window width is determined by looking at the smallest Cross-Validation (CV) value using Gaussian. The best bandwidth results are 1.8111, with a CV value of 12. It means that the area around the area within a radius of 1.8111 degrees will be considered to have a location effect. The effect decreases with distance with a minimum CV of 12. The accuracy of the global regression model and local regression is based on the GWR analysis, 27 local regression models were obtained, including the local regression model for the Unemployment Rate of Cirebon city, as follows:

UR = 7.848707+ 0.62008 GRDP + 0.124972 HDI + 0.915516 ALPR

The estimated value limits of the local regression model coefficients for the 27 cities/districts are shown in Table 5.

Table 6. Estimated Value of Local Regression Model Coefficient (GWR)

Variable	Minimum	Maximum	Mean	Std. Deviation
UR	3,34	10,97	8,08	1,74
GRDP	2491,64	228725,92	50198,53	55611,08
HDI	63,70	80,31	70,28	4,89
RMW	1433901	3605271,0	2301867,	717062,14
ALPR	0,25	2,34	1,444	0,57
PDL	391,00	15307,00	2965,15	3789,47

Based on the GWR analysis, three regional groups were obtained based on predictor variables that had a significant effect on each region.

Table 7. Variables that are Significant in the Local Regression Model

Cities / Regencies	**Significant variable**
Karawang, Purwakarta, Bandung Barat,Subang,Bekasi, Bogor, Sumedang, Sukabumi, Bandung, Kabupaten Bekasi, Cimahi, Garut, Kabupaten Bogor, Banjar,Kabupaten Bandung, Cianjur, Kabupaten Sukabumi	RMW, PDL, HDI
Cirebon, Indramayu, Kabupaten Cirebon, Majalengka, Depok	RMW, ALPR, PDL,HDI
Pangandaran,Ciamis, Tasikmalaya, Kab. Tasikmalaya	RMW, GRDP, PDL, HDI

Based on the GWR analysis, three regional groups were obtained based on predictor variables that had a significant effect on each region. Based on table 6, it can be seen that the variables of RMW, PDL, and HDI are significant in all cities/districts in West Java. The ALPR and GRDP variables are quite significant in several cities/districts. This shows that the difference in the value of the minimum wage, the level of population density, and the human development index between cities / districts in West Java affects labor

absorption. Likewise, differences in the number of active workers and regional income in several cities/regencies in West Java affect the unemployment rate in the area. It is hoped that the city/regional government in West Java can narrow this difference. If not, population migration will occur because of these differences. It is feared that the migration of people without skills trigger a new problem.

Table 8. Comparison of Global regression and Local regression

Criteria	Global regression	Local regression
R^2	0.520594	0.678844
AIC	99.592002	95.527444

Table 8 explains that the local regression is relatively better than the local regression because the value of R^2 is 0.678844, greater than global regression, meaning 67,88% of the UR model is influenced by the predictor variables, namely HDI, PDL, ALPR, and RMW. In terms of model accuracy, local regression is also better because the AIC value is smaller than the global regression model, whichis95.527444. In contrast, the value of the AIC Global regression was 99.592002, the global regression error.

Table 9. GWR ANOVA

Source	SS	DF	MS	F_{count}	F_{table}
Global Residual	37.64	21.00			
GWR Improvement	12.42	5.187	2.395		
GWR REsidual	25.21	15.81	1.595	1.50	2,27

Based on GWR Anova, it can be shown that the calculated F_{value} is 1.50 less than F_{table} =2,27, meaning that there is no difference between the global regression model and local regression (GWR).

3. DISCUSSION/CONCLUSION

The factors affecting the Unemployment Rate in West Java are RMW and ALPR. The local regression model (GWR) is the appropriate model to identify the factors of UR. Statistically, GWR has a coefficient of determination (R^2) of 67,88 % higher than the global regression of 52,06%. It means that 67,88% of the Unemployment rate model is influenced by the independent variables, as presented above. On the other hand, the AIC value of the Local regression model of 96,787784 is smaller than the Global regression of 99,592002 meaning that the value of the local regression model error is smaller than the value of the global regression error. Thus, the West Java government, along with the city/regency level, is expected to do some policies to reduce the unemployment rate and migration to higher globally significant variable, but they also show a similar effect locally.

REFERENCES

[1] BPS Jawa Barat, *Provinsi Jawa Barat dalam 2017*, 2018th-08–16th ed. Bandung: 32560.1802, 2017.

[2] I. P. Ningtias and S. P. Rahayu, "Pemodelan Faktor-faktor yang Mempengaruhi Tingkat Pengangguran Terbuka di Provinsi Jawa Timur Tahun 2015 Menggunakan Regresi Spasial," *J. Sains dan Seni ITS*, vol. 6, no. 2, 2017, doi: 10.12962/j23373520.v6i2.24984.

[3] A. Karim, A. Faturohman, S. Suhartono, D. D. Prastyo, and B. Manfaat, "Regression Models for Spatial Data: An Example from Gross Domestic Regional Bruto in Province Central Java," *J. Ekon. Pembang. Kaji. Masal. Ekon. dan Pembang.*, vol. 18, no. 2, p. 213, 2017, doi: 10.23917/jep.v18i2.4660.

[4] E. Amalia and Liza Kurnia Sari, "Analisis Spasial Untuk Mengidentifikasi Tingkat Pengangguran Terbuka Berdasarkan Kabupaten/Kota Di Pulau Jawa Tahun 2017*," *Indones. J. Stat. Its Appl.*, vol. 3 No. 3, pp. 202–215, 2019.

[5] N. Aini *et al.*, "Di Provinsi Jawa Barat Menggunakan Pendekatan Mixed Geographicay Weighted Regression," 2017.

[6] C. Brunsdon, S. Fotheringham, and M. Charlton, "Geographically Weighted Regression as a Statistical Model." pp. 1–12, 2000.

[7] F. F. Solastika Mariani, Wardono, Masrukan, "The Arcview And Geoda Application In Optimization Of Spatial Regression Estimate," *J. Theor. Appl. Inf. Technol.*, vol. 95 no.5, 2017.

[8] J. LeSage and R. K. Pace, *Introduction to Spatial Econometrics*, 1st Editio. New York: Chapman and Hall/CRC, 2009.

[9] M. Fotheringham, Alexander and Brunsdon, Chris and Charlton, "Geographically Weighted Regression : The Analysis of Spatially Varying Relationships," *John Wiley Sons*, vol. 13, 2002.

[10] L. Anselin, *Spatial Econometrics_ Methods and Models _ L*, 1st ed. Netherlands: Springer Netherlands, 1988.

[11] C. Brunsdon, S. Fotheringham, and M. Charlton, "Geographically Weighted Regression-Modelling Spatial Non-Stationarity," *J. R. Stat. Soc. Ser. D (The Stat.*, vol. 47, No. 3, 1998, [Online]. Available: https://www.jstor.org/stable/2988625.

[12] I. Labour Officce Geneva, *Strengthening the ILO's Capacity to Assist Its Member's Efforts to Reach Its*, 97th ed. 2004.