International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Poverty Data Modeling in North Sumatera Province Using Geographically Weighted Regression (GWR) Method

Kristina Pestaria Sinaga

Faculty of Mathematics and Natural Science, University of Sumatera Utara
Jalan Bioteknologi No. 1, Kampus Padang Bulan 20155, Medan, Indonesia

Abstract: ***In regular regression equation, a response variable is connected with some predictor variables in one main output, which is parameter measurement. This parameter explains relationships of every predictor variable with response variable. However, when it is applied to spatial data, this model is not always valid because the location difference can result in different model estimation. One of the analyses that recommend spatial condition is locally linear regression called Geographically Weighted Regression (GWR). The basic idea from this GWR model is the consideration of geographical aspect or location as weight in estimating the model parameter. Model parameter estimation of GWR is obtained using Weight Least Square (WLS) by giving different weights to every location where the data is obtained. In many analyses of GWR, also in this research, the weight used is Gauss Kernel, which needs bandwidth value as distance parameter that still affects each location. Bandwidth optimum can be obtained by minimizing cross validation value. In this research, the researcher aims to compare the results of global regression model with GWR model in predicting poverty percentage. The data used as a case study are data from 33 cities/regencies in North Sumatera province.***

Keywords: GWR, WLS, Kernel Function, Poverty

1. Introduction

Poverty is one of fundamental issues that have been government's concern in any country all over the world. In Indonesia, poverty is still one of the biggest problems. Both central government and local governments has tried to implement policies and programs to overcome poverty but there seem a lot of things that have not been accomplished. One of important aspects to overcome poverty is determining the poverty measurement value. Reliable measurement can be a very important element in policy making regarding the lives of the poor [3].

To know the number, spread, and condition of poverty in certain area, a perfect poverty measurement is needed to achieve effectiveness in overcoming poverty through policies and programs. BPS also develops a certain method to obtain a criterion that operationally can be used to determine the number of poor households. This method is used in *Pendataan Sosial Ekonomi* (PSE) (Social Economical Census) in 2005 by using 14 variable indicators to determine the poverty status [2, 3]. However in reality, the method to determine poverty rate, according to this notion, is still global; in other words, it applies to all locations being observed. In fact, the condition of one location is not always the same with the condition of other locations, may be due to geographical factor (spatial variation), social cultural background, and other things that surround the location. Therefore, the model to determine the global poverty rate does not fit to be used for its spatial heterogeneity. One of the effects emerging from spatial heterogeneity is spatial varied regression parameter. In global regression, it is assumed that the predictive value of regression parameter will be constant, which means the regression parameter is the same for every point in the research area. If spatial heterogeneity happens to regression parameter, then global regression becomes less capable in explaining the real data phenomena. This research aims to model poverty in North Sumatera province in 2013 with Fixed Gaussian Kernel weight and to test the GWR model parameter.

2. Literature Review

2.1. Linear Regression

The method that is often used to declare the relationship between response variable and predictor variable is regression method. Linier regression model for p predictor variable and the n number of observation in matrix equation is [11,12]:

$$Y=\begin{pmatrix}Y_1\\Y_2\\\vdots\\Y_n\end{pmatrix}=\begin{pmatrix}\sum X_1 & \sum X_{11} & \dots & \sum X_{1p}\\ \sum X_2 & \sum X_{21} & & \sum X_{2p}\\ \vdots & \vdots & & \vdots\\ \sum X_3 & \sum X_{31} & & \sum X_{3p}\end{pmatrix}\begin{pmatrix}\beta_0\\\beta_1\\\vdots\\\beta_n\end{pmatrix}+\begin{pmatrix}\varepsilon_0\\\varepsilon_1\\\vdots\\\varepsilon_n\end{pmatrix} \quad (1)$$

Equation 1 is the general form of regression equation in matrix symbol. In this general form, Y is a response vector $n\times 1$, X states that predictor matrix with measurement $n\times(k+1)$, β is parameter vector with measurement $(k+1)\times 1$ and ε is error vector with measurement $n\times\ 1$.

Model (2) is also called global regression model because global regression model assumes that the relationship between response variable with predictor variable is constant, so the parameter of which the estimation value is the same in all places where the data taken [4, 11]. Ordinary global regression equation is usually defined using parameter estimation method *Ordinary Least Square* (OLS) [13]. For n observation with p independent variable, the regression model can be noted as below:

$$y=\beta_0+\sum_{k=1}^{p}\beta_k x_{ik}+\varepsilon_i \quad (2)$$

with $i = 1, 2, \ldots, n$; $\beta_1, \beta_2, \ldots, \beta_p$ is model parameter and $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ is error assumed identical independent that has normal distribution with zero mean and constant variants. In this model, the relationship between independent variable with dependent variable is considered constant in each geographical location [9, 11]. The estimator of model parameter can be obtained below:

$$\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right)^T = (X^T X)^{-1} X^T Y \qquad (3)$$

The testing statistic F_{value} of regression model is [11]

$$F_{value} = \frac{SSR}{SST} \qquad (4)$$

H_0 is rejected if $|F_{value}| > |F_{table\,(\alpha,p,n-p-1)}|$.

The coefficient value of determination can be formulated using analysis of variance table [11]:

$$R^2 = \frac{SSR}{SST} \qquad (5)$$

Spatial test is done to obtain significant parameter that can be used on model. Statistical test t_{value} of regression model [7, 8] is as below:

$$t_{value} = \frac{\hat{\beta}_k}{S(\hat{\beta}_k)} \qquad (6)$$

The parameter is significant to the model if $|t_{value}| > \left|t_{table\,(\frac{\alpha}{2},n-p-1)}\right|$.

2.2. Geographically Weighted Regression (GWR)

Geographically Weighted Regression (GWR) is a development technique from global regression model to weighted regression model [4, 6, 12]. Response variable depends on the area location. GWR model can be formulated as below.

$$\hat{y}_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i)\, x_{ik} + \varepsilon_i \qquad (7)$$

in which:

y_i : coordinate point (longitude, latitude) of *i*
$\beta_k(u_i, v_i)$: regression coefficient of predictor converter *k* for each location (u_i, v_i)
(u_i, v_i) : longitude and latitude for location *i*
x_{ik} : observation value of predictor *k* in observation *i*
ε_i : random observation changer *i*

In hypothesis test, there are a few assumptions used in GWR model, such as:

1. Error forms $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are assumed identical independent and have normal distributions with zero means and constant variants $(\varepsilon_i \sim IIDN(0, \sigma^2))$.
2. If $\hat{y}_i$ is an estimator of y_i in location *i*, then in all locations $(i = 1, 2, \ldots, n)$, $\hat{y}_i$ is non-bias estimator for $E(y_i)$ or it can be written as $E(\hat{y}_i) = E(y_i)$ for all *i*.

2.3. Making GWR Model

Spatial weight is a weight that explains data locations. Close location and medium distance location are given big weight while far location is given small weight. Kernel function is a way to determine the size of weight in each different location on GWR model [6]. The weight functions can be written as below:

1. Gaussian
$$w_j(u_i, v_i) = exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{h}\right)^2\right]$$
2. Adaptive Gaussian
$$w_j(u_i, v_i) = exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{h}\right)^2\right]$$
3. Bisquare
$$w_j(u_i, v_i) = \begin{cases}\left(1 - \left(\frac{d_{ij}}{h}\right)^2\right)^2, & for\ \ d_{ij} \le h \\ 0, & d_{ij} > h\end{cases}$$
4. Adaptive Bisquare
$$w_j(u_i, v_i) = \begin{cases}\left(1 - \left(\frac{d_{ij}}{h}\right)^2\right)^2, & for\ \ d_{ij} \le h_i \\ 0, & d_{ij} > h_i\end{cases}$$
5. Tricube
$$w_j(u_i, v_i) = \begin{cases}\left(1 - \left(\frac{d_{ij}}{h}\right)^3\right)^3, & for\ \ d_{ij} \le h \\ 0, & d_{ij} > h\end{cases}$$
6. Adaptive Tricube
$$w_j(u_i, v_i) = \begin{cases}\left(1 - \left(\frac{d_{ij}}{h}\right)^3\right)^3, & for\ \ d_{ij} \le h_i \\ 0, & d_{ij} > h_i\end{cases}$$

With $d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2}$, the euclidean distance between locations (u_i, v_i) to location (u_j, v_j) and h is non-negative parameter usually known and called as smoothing parameter or bandwidth. If the weight used is kernel function, then the choice of bandwidth is very important because bandwidth is a balance controller between curves towards data and data smoothness [7, 13]. The method used to choose optimum bandwidth is Cross Validation (CV). This method is noted as below:

$$CV(h) = \sum_{i=1}^{n}\left(y_i(h) - \hat{y}_{\neq i}(h)\right)^2 \qquad (8)$$

with:

$y_i(h)$: fitting value y_i in which observation on location (u_i, v_i) is omitted from estimation process.
$\hat{y}_{\neq i}(h)$: fitting value y_i in which observation on location (u_i, v_i) is included in the estimation process.
n : sample total number.

2.4. Model Hypothesis Testing (GWR)

The goodness of fit test of GWR model is done using the following hypothesis:

H_0 : $\beta_k(u_i, v_i) = \beta_k$ (there is no difference between OLS and GWR)

H_1 : at least there is one $\beta_k(u_i, v_i) \neq \beta_k$ (there is difference between OLS and GWR

Test statistics:

$$F_{value} = \frac{\frac{(RSS_{OLS} - RSS_{GWR})}{v}}{\frac{RSS_{GWR}}{\delta_i}} \quad (9)$$

Rejection location: reject $H_0, F_{hitung} > F_{tabel\,(\alpha, df_1, df_2)}$.

Parameter significance test in each location is done by testing spatial parameter. This testing is done to know the significance of (u_i, v_i) to response variable in partial on GWR model. The hypothesis is as the following:

H_0 : $\beta_k(u_i, v_i) = 0$

H_1 : at least there is one $\beta_k(u_i, v_i) \neq 0$

The parameter estimation $\beta_k(u_i, v_i)$ will follow multivariate normal distribution.

Statistic test:

$$T = \frac{\hat{\beta}_k(u_i, v_i)}{\hat{\sigma}\sqrt{g_{kk}}} \quad (10)$$

Reject H_0 if $|T| > t_{\left(\frac{\alpha}{2}, db\right)}$ which means parameter $\beta_k(u_i, v_i)$ is significant to the model.

2.5. Poverty

Poverty is a multi-dimensional problem that interconnects many parties. Poverty in Indonesia is followed by discrepancy among citizens and local developments, indicated by, for example, poor education and health state and low income and purchasing power, and reflected from the low Human Development Index. The citizens categorized as poor if their income rate is below the poverty line and included in official poverty rate in Indonesia.

3. Methodology

3.1 Data Source

The data from this research is secondary data coming from the National Social Economical Survey (SUSENAS) in 2013.

3.2 Research Variable

Variables used in this research are:

1. Response variable, that is poverty percentage
2. Predictor variable, that is participation rate of work (X_1), citizens who study in elementary school (X_2), junior high school (X_3), senior high school (X_4), households of which the main power source is not State Electricity Company (PLN) (X_5), households in which the water and sanitation source is just ordinary well (X_6), households that join JAMKESMAS (health insurance) (X_7), and households of which the main fuel is kerosene (X_8).

Besides the variables above, two geographical variables of the location of cities and regencies in North Sumatera, spatial coordinate (longitude and latitude), are also used. The research units being investigated here are 26 regencies and 7 cities in North Sumatra province.

3.3 Analysis Method

The analysis stages used to achieve the goal of research are:

1. Analyzing global regression model to determine the poverty rate in North Sumatera with the following steps:
   - Modeling response variable (Y) with predictor variable (X)
   - Testing linier regression models' goodness of fit all at once
   - Testing model parameter partially
   - Testing assumptions needed in regression
2. Analyzing GWR model to determine the poverty rate in North Sumatera province with the following steps:
   - Determining optimum bandwidth value for every regions based on CV value.
   - Determining the eucledian distance among observation locations based on geographical position. Euclidian distance between location i that occurs in coordinate (u_i, v_i) and location j that occurs in coordinate (u_i, v_i).
   - Determining weight using Gausian Kernel function.
   - Estimating GWR model parameter using Weight Least Square method
   - Testing the goodness of fit of GWR
   - Testing model parameter

The software used by the researcher in finishing the model is Minitab 16 and GWR4.

4. Analysis and Discussion

Before using GWR for data analysis, global regression model should be formed first; that is the best regression model between poverty and the influencing factors. From several combination models of predictor variable, the last model is assumed by asserting four predictor variables and intercept. The first regression model produced is:

$$\hat{Y} = 37.1 - 0.392X_1 - 0.315X_2 + 0.245X_7 + 0.167X_8$$

The equation model above is pretty suitable to be used with $R^2 = 66.8\,\%$. The table test of model parameter value above is available in Table 1.

Table 1: Model Parameter Test of Global Regression Model

Source	Coefficient	SE Coefficient	T	P value	Conclusion
Intercept	53.88	15.33	3.52	0.002	Significant
X_1	-0.379	0.115	-3.29	0.003	Significant
X_2	-0.315	0.147	-2.14	0.041	Significant
X_7	0.199	0.089	2.23	0.035	Significant
X_8	0.126	0.056	2.24	0.034	Significant

Note: Processed Using Minitab 16

Table 2: ANOVA

Source	DF	SS	MS	F	P
Regression	4	849.83	212.46	14.01	0.000
Residual Error	28	421.77	15.06		
Total	32	1271.60			

Note: Processed Using Minitab 16

The first analysis step in GWR is determining the bandwidth used in Gauss Kernel weight function. Determining optimum Bandwidth (h) with *Cross Validation* criteria results in h with value 0.381966 and minimum CV value 7.756. h value and distance among locations will be used in making weight matrix. The following is weight matrix for Nias regency:

$$W(u_i, v_i) = diag[w_1(u_1, v_1) \quad w_2(u_1, v_1) \quad \dots \quad w_{33}(u_1, v_1)]$$
$$= diag[1 \quad 0.9419 \quad \dots \quad 1]$$

The weight matrix is used to estimate parameter in location (u_i, v_i). To estimate parameter in location (u_2, v_2), weight matrix $W(u_2, v_2)$ can be searched in by the same step as the step above; it goes for also for the last observation weight matrix $W(u_{33}, v_{33})$. The equation solving can be done using GWR4 Software so that the parameter estimation is obtained in all locations $(u_i, v_i), i = 1, 2, \dots, 33$.

Table 3: GWR Model Parameter Estimator

Source	Model GWR					
	Nilai		SE		T	
	Min	Maks	Min	Maks	Min	Maks
Intercept	-29.027	218.27	4.614	23.713	-2.916	10.433
X_1	-3.495	0.377	0.054	0.494	-8.370	2.963
X_2	-0.476	6.145	0.089	2.730	-3.560	4.157
X_7	-3.039	0.536	0.438	0.042	-5.130	5.407
X_8	-0.809	0.736	0.035	0.230	-33.52	10.94
R-sqr	99.69 %					
R-adj	95.97 %					
Bandwith	0.382					
Iteration	14					

Note: Processed Using GWR4

ANOVA which shows that GWR model and OLS model explain the relationship among changer is equally good or rejected is Table 4. Table 4 shows that using GWR can result in less residuals.

Table 4: ANOVA

Source	**SS**	**DF**	**MS**	**F**
Global Residuals	421.769	28.000		
GWR Improvement	417.833	24.539	17.027	
GWR Residuals	3.936	3.461	1.137	14.969589

Note: Processed Using GWR4

The $F_{value} = 14.97 > F_{table} = 2.61$ shows that null hypothesis which mentions that the reliability rate of 95% global regression is equally good with GWR, rejected. From the hypothesis testing, it can be concluded that there is spatial influence among poverty rate with the affecting variables if GWR analysis is used.

It can be concluded that the poverty rate in cities or regencies in North Sumatra is better if it is explained by clarifier changer in geographically coefficient way, compared to using global regression with constant coefficient in all cities or regencies. The estimator value is shown in the Table 5.

Table 5: Estimator Values of GWR Model

Regencies/Cities	β_0	β_1	β_2	β_7	β_8	$\hat{Y}_{GWR}$
Nias	195.846	-3.021	3.986	-2.192	0.720	17.418
Mandailing Natal	17.584	-0.110	-0.469	0.170	0.196	9.620
Tapanuli Selatan	0.078	0.153	0.039	0.017	-0.052	11.397
Tapanuli Tengah	7.953	0.129	-0.476	0.046	0.113	15.407
Tapanuli Utara	28.998	-0.209	-0.228	0.034	0.083	10.504
Toba Samosir	18.857	-0.149	-0.090	0.197	0.041	9.853
Labuhan Batu	21.407	-0.181	-0.125	0.207	0.058	8.892
Asahan	20.677	-0.172	-0.113	0.203	0.052	11.455
Simalungun	20.049	-0.163	0.103	0.197	0.043	10.262
Dairi	26.755	0.186	-0.313	0.064	0.095	9.314
Karo	-10.293	0.230	-0.131	0.492	-0.222	9.732
Deli Serdang	-8.925	0.197	-0.080	0.463	-0.193	5.098
Langkat	-10.317	0.233	-0.139	0.488	-0.221	10.481
Nias Selatan	217.620	-3.467	5.655	-2.873	0.716	18.877
Humbang Hasundutan	36.781	-0.311	-0.277	0.013	0.131	10.780
Pakpak Barat	36.851	-0.326	-0.287	0.018	0.148	11.202
Samosir	32.910	-0.275	-0.264	0.032	0.124	13.780
Serdang Bedagai	13.531	-0.152	-0.062	0.293	0.166	9.304
Batu Bara	-0.499	0.070	-0.180	0.483	0.038	11.911
Padang lawas Utara	-0.010	0.162	0.014	-0.017	-0.043	10.017
Padang Lawas	0.152	0.166	0.013	-0.025	-0.043	8.719
Labuhan Batu Selatan	3.746	0.193	0.033	-0.184	-0.102	12.266
Labuhan Batu Utara	20.463	-0.167	-0.110	0.199	0.046	10.935
Nias Utara	194.753	-2.996	3.783	-2.122	0.736	30.933
Nias Barat	218.265	-3.495	6.145	-3.039	0.665	29.644
Sibolga	28.439	-0.212	-0.286	0.086	0.108	12.833
Tanjung Balai	21.378	-0.182	-0.125	0.208	0.060	14.917
Pematang Siantar	20.049	-0.163	-0.103	0.197	0.043	10.374
Tebing Tinggi	29.023	0.377	0.489	0.536	-0.809	11.739
Medan	-9.632	0.218	-0.118	0.474	-0.207	9.541
Binjai	12.172	-0.125	-0.076	0.290	0.156	7.371
Padang Sidempuan	0.404	0.165	-0.015	-0.026	-0.047	9.214
Gunung Sitoli	195.846	-3.021	3.986	-2.192	0.720	17.418

5. Conclusion

GWR can result in different parameter for each geographical location. GWR can then show the significant difference of poverty rate for each regencies and cities in North Sumatera province.

References

[1] BPS, "Sumatera Utara In Figures 2014," Indonesia, Jakarta. 2014.
[2] BPS, "Welfare Indicators," Indonesia. 2011.
[3] BPS, "Measurement and Analysis of Poverty," Indonesia. 2013
[4] A.S. Fotheringham, Brundson, C., and Charlton, M., "Geographically Weighted Regression: The Analysis of Spatially Varying Relationships," Ritsumeikan University: Departement of Geography. 2002.
[5] N. Tomoki, M. Charlton, P. Lewis, C. Brundson, J. Yao and Fotheringham, A.S., "GWR4 User Manual," Ritsumeikan University : Department of Geography. 2014.
[6] C.L. Mei, "Geographically Weighted Regression Technique for Spatial Data Analysis," School of Science Xi'an Jiaotong University. 2005.
[7] N. Draper and H. Smith, "Applied Regression Analysis Third Edition," New York, Wiley. 1998.
[8] R. K. Sembiring, "Regression Analysis Second Edition," Bandung Institute of Technology, Armico, Bandung, Indonesia. 1995.
[9] J. Supranto, "Econometrics Second Edition," Ghalia Indonesia. 2004.
[10] S. Asep, Nur, A.S. and Noer, A.A., "On Comparison between Ordinary Linear Regression and Geographically Weighted Regression: With Application to Indonesia Poverty, European Journal of Scientific Research" Bogor Agricultural University, Indonesia. 2011, 275-285 (In Bahasa Indonesia).
[11] R.S. Indriya, D.R.S. Saputro dan Purnami W, "Model Geographically weighted Regression Penderita Diare Di Provinsi Jawa Tengah Dengan Fungsi Pembobot Kernel Bisquare [Journal]," Yogyakarta: FMIPA UNY. 2013 (In Bahasa Indonesia).
[12] S. Astutik, N.W. Ni Wayan dan Kurniawan, D. 2007. "On Estimation Geographically Weighted Regression on Heterokedastisitas Spatial," University of Brawijaya : Malang, Indonesia.
[13] Sugiyanto, "Geographically Weighted Regression Technique for Spatial Data Analysis (Case Studies: Poverty Data in Papua Province) [Master Thesis in Bahasa Indonesia]," Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. 2013.

Author Profile

Kristina Pestaria Sinaga received a BSc in Mathematics (2013) from University of Sumatera Utara. She is now in her second year of graduate studies at University of Sumatera Utara, concentrating in operations research. Kristina's research interests include Regression Analysis, Graph, Linear and Non Linear Programming, ARIMA, and Nonparametric Statistics.