International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Semi-parametric Geographically Weighted Regression Modelling using Linear Model of Coregionalization

Zakiyah Mar'ah[a]*, Anik Djuraidah[b], Aji Hamim Wigena[b]

[a,b]*Departement of Statistics, Faculty of Mathematics and Natural Science, Bogor Agricultural University, Indonesia*

[a]*Email: zakiyah1192@gmail.com*

[b]*Email: anikdjuraidah@gmail.com, Email: ajiwigena@ymail.com*

Abstract

Geographically Weighted Regression is a weighted analysis regression for local or spatially varying parameters, therefore each location has different regression parameters. In its application, one often finds a condition that needs some global parameters. Geographically Weighted Regression that has local and global parameters is called Semi-parametric Geographically Weighted Regression. This study modelled Semi-parametric Geographically Weighted Regression using Linear Model of Coregionalization to assist spesification of local and global parameters. Linear Model of Coregionalization represented spatial variability proportion at different spatial distances and spatial dependence of parameters. High spatial dependence variables were as local parameters while the other variables were as global parameters. The data used was poverty data in North Sulawesi Province. The results of Geographically Weighted Regression and Semi-parametric Geographically Weighted Regression models were compared based on Akaike Information Criterion Corrected and Mean Square Prediction Error. It showed that Semi-parametric Geographically Weighted Regression model was better than Geographically Weighted Regression.

Keywords: Geographically Weighted Regression; Semi-Parametric Geographically Weighted Regression; Linear Model of Coregionalization.

---

* Corresponding author.

1. Introduction

Geographically Weighted Regression (GWR) is a weighted analysis regression for local or spatially varying parameters, therefore each locations has different regression parameters [1]. In its application, one often finds a condition that needs some global parameters. GWR that has local and global parameters is called Mixed Geographically Weighted Regression or Semi-parametric Geographically Weighted Regression [1-2]. There are some methods in spesification of local and global parameters. Fotheringham and his colleagues in [3] adopted a stepwise procedure that all possible combinations of global and local parameters were tested and the optimum mixed/semi-parametric model was selected based on the smallest AICc value. Mei and his colleagues in [4] used a spatial variability test (F-Test) in determining local parameters. Pongoh in [5] specified local and global parameters based on the confidence interval of GWR coefficients. In addition, the spesification of local and global parameters can also be obtained using geostatistical approach called Linear Model of Coregionalization (LMC). Goulard and Voltz in [6] stated that LMC is a useful tool for describing spatial relationships among variables. Ribeiro and his colleagues in [7] used LMC on Semi-parametric Geographically Weighted Poisson Regression.

This study aimed to develop Semi-parametric Geographically Weighted Regression (SGWR) model which its parameter spesification was based on LMC and also to show the proportion of spatial variability on different spatial distances and spatial dependence of parameters. The data used was poverty data in North Sulawesi Province which the parameter spesification was based on confidence interval of GWR coefficients in the previous study [5]. SGWR model with confidence interval and SGWR model with LMC were compared based on Akaike Information Criterion Corrected and Mean Square Prediction Error.

2. Literature Reviews

2.1 Geographically Weighted Regression

Geographically Weighted Regression model *(1)* is a development of a global regression model used to model and analyzes parameters that have spatial variability, hence each location has different regression parameter values [1].

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

for $i = 1, 2, ..., n$, $k = 1, 2, ..., p$ and (u_i,v_i) is the i-th coordinate points (longitude, latitude). Parameter estimation of GWR is obtained using weighted least square method *(2)* by giving different weighting in each location, hence data from observation close to location-i has a higher weighting value than far observation.

$$\hat{\boldsymbol{\beta}}(u_i, v_i) = \left(\boldsymbol{X}'\boldsymbol{W}(u_i, v_i)\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}(u_i, v_i)\boldsymbol{y} \qquad (2)$$

$\boldsymbol{W}(u_i,v_i)$ is the spatial weighting matrix of location-i which its diagonal elements are determined by the distance

of the location-*i* with the other location.

2.2 Semi-parametric Geographically Weighted Regression

Semi-parametric Geographically Weighted Regression model *(3)* has geographically varying and constant coefficients in the same model [1]. Nakaya and his colleagues in [2] stated that the parameter estimation procedure in SGWR combines a parametric and non-parametric methods.

$$y_i = \sum_{j=1}^{k} \alpha_j x_{ij} + \sum_{l=k+1}^{p} \beta_l (u_i, v_i) x_{il} + \varepsilon_i \qquad (3)$$

for j = 1, 2, ..., k, l = 1, 2, ..., p, α_j is a global (constant) parameter and $\beta_i(u_i,v_i)$ is a local (geographically varying) parameter. Local parameter estimation in SGWR uses the same method as GWR estimation, i.e. weighted least square and global parameter estimation is obtained using ordinary least square [1].

2.3 Linear Model of Coregionalization

Linear Model of Coregionalization consists of semivariogram *(4)* and cross-semivariogram *(5)* of two or more variables. Each variable is characterized by semivariogram meanwhile each pair of variables is characterized by cross-semivariogram.

$$\hat{\gamma}_X(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{(i,j)=1}^{N(\mathbf{h})} \left(x_i - x_j\right)^2 \qquad (4)$$

$$\hat{\gamma}_{XY}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{(i,j)=1}^{N(\mathbf{h})} \left(x_i - x_j\right)\left(y_i - y_j\right) \qquad (5)$$

for i,j = 1, 2, ..., k, $N(\mathbf{h})$ is observation total on each spatial distance, $\mathbf{h}$ is vector of distance, x and y are observation variables. LMC is formed by nested structure models of semivariogram Linear combination [8]. In this study, LMC was consisted of nugget effect and two basic semivariogram models *(6)*.

$$\gamma_{ij}(\mathbf{h}) = c_{ij,0} g_0(h_0) + c_{ij,1} g_1(h_1) + c_{ij,2} g_2(h_2) \qquad (6)$$

which $c_{ij,s}$ is the s-th structure of semivariogram coefficient (sill), s = 0, 1, 2,

g_0 is the nugget effect $g(h) = \begin{cases} 0, & \text{if } h = 0 \\ 1, & \text{otherwise} \end{cases}$,

g_1 is the first basic semivariogram using Spherical function $g(h) = \begin{cases} 1.5\left(\frac{h}{a}\right) - 0.5\left(\frac{h}{a}\right)^3, & \text{if } h \le a \\ 1, & \text{otherwise} \end{cases}$,

g_2 is the second basic semivariogram using Gaussian function $g(h) = 1 - \exp\left(\frac{-3h^2}{a^2}\right)$ and a is range.

Each semivariogram and cross-semivariogram are built from the same models. Matrix form of Eq. *(6)* is shown in Eq. *(7)*.

$$\boldsymbol{\Gamma}(h) = \begin{bmatrix} \gamma_{11}(h) & \cdots & \gamma_{1p}(h) \\ \vdots & \ddots & \vdots \\ \gamma_{np}(h) & \cdots & \gamma_{pp}(h) \end{bmatrix} = \sum_{s=0}^{2} \boldsymbol{C}_s * g_s(h_s) = \sum_{s=0}^{2} \begin{bmatrix} c_{11,s} & \cdots & c_{1p,s} \\ \vdots & \ddots & \vdots \\ c_{p1,s} & \cdots & c_{pp,s} \end{bmatrix} * g_s(h_s) \qquad (7)$$

which $\mathbf{C_s}$ is the positive definit of coregionalization matrix, the diagonals of $\boldsymbol{\Gamma}(h)$ are semivariogram values and the off-diagonals are cross-semivariogram values.

3. Data and Methodology

3.1 Data

The data were from the integrated database of poor families for the July 2012 Social Protection Program in 159 sub-districts at North Sulawesi Province. Data was taken from TNP2K (National Team for Accelerating Poverty Reduction) website [5]. The response variable is the percentage of low welfare status (Y) and the explanatory variables are percentage of female household heads per sub-districts (X_1), percentage of children not attending school (X_2), percentage of people with disabilities (X_3), percentage of people with chronic disease (X_4), percentage of unemployed individuals (X_5), percentage of households having their own buildings (X_6), percentage of households using protected drinking water sources (X_7), percentage of households using electricity/PLN (X_8), percentage of households using gas cooking fuel/LPG/electricity (X_9), percentage of households using their own latrines (X_{10}) and percentage of households using public septic tank (X_{11}).

3.2 Methodology

Data analysis was performed using statistical software R Studio with the steps:

a. Describing data by showing the locations that have the lowest and highest percentage of poverty rate.

b. Testing the model assumptions such as the error normality using Kolmogorov-Smirnov (KS) test and multicolinearity by calculating Variance Inflation Factor (VIF) at explanatory variables then testing the spatial effects by calculating Moran's Index (I) and Breusch-Pagan (BP) test.

c. Modelling GWR on significant explanatory variables.

d. Estimating LMC and calculating the proportion of spatial variability and spatial dependence of variables.

e. Modelling SGWR which its parameter specifications were performed using LMC.

f. Comparing the GWR to SGWR based on the AICc and MSPE values.

4. Result and Discussion

4.1 Data Description

The poverty rate in North Sulawesi Province was quite high. The average was around 40% to 60%. Figure 1 shows that the highest and lowest percentage of poverty was in Sangihe Island District, the highest percentage was in South Tabukan Subdistrict (79.79%), meanwhile the lowest percentage was in East Tahuna Subdistrict (5.83%).

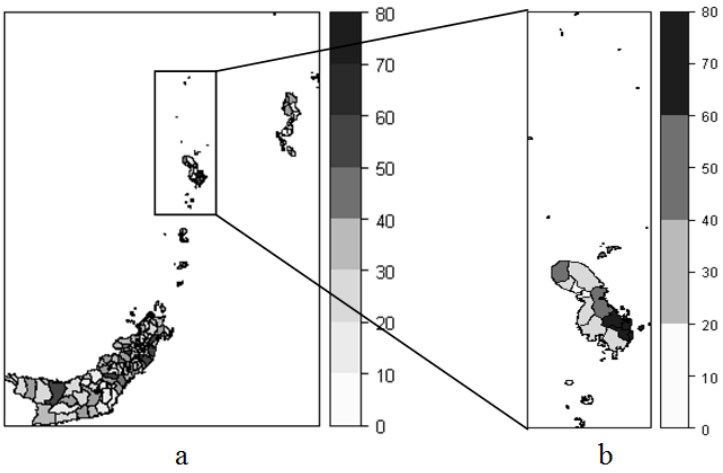

Figure 1: Map of Poverty Percentage Distribution in North Sulawesi Province (a) and Sangihe Island District (b)

4.2 Geographically Weighted Regression Modelling

The data used must satisfy the model assumptions such as error normality, multicolinearity, autocorrelation and heterogeneity. The error of data was normally distributed (KS = 0.067, p-value = 0.085) and there was no multicolinearity (VIF < 5). In addition, since the data was geographical data that has spatial effects, the autocorrelation and heterogeneity tests were performed using spatial effect tests. There was no spatial autocorrelation (I = -0.01, p-value = 0.77) and the spatial variability was different at each observation location (BP = 39.60, p-value = 4.18×10^{-5}), it causes different characteristics of each subdistrict locations hence a local approach is needed in overcoming the variability occured by using GWR models.

The explanatory variables used were significant variables affecting the response variable, i.e. X_1, X_3, X_5, X_8 and X_{10}. GWR model was performed using the selected kernel weighting based on the smallest AICc value, i.e adaptive Kernel Bisquare function weighting. GWR model with adaptive Kernel Bisquare function obtained the coefficient of determination 99.44% and the AICc value 662.06.

4.3 Model Linear of Coregionalization Estimation

The first step in establishing the LMC was to select the nested structure of semivariogram and cross-semivariogram [6]. In this study, the Nug(0) + Sph(54) + Gau(161) model was selected because of the smallest weighted mean square of error (1.66). Nug(0) is the nugget effect, Sph(54) is an observation location with distance 0-54 km modelled using Spherical function and Gau(161) is the observation location with distance 54-161 km modelled using Gaussian function. Figure 2 shows the LMC estimation (the straight lines), semivariogram and cross-semivariogram are shown by the black dots. Variable Y and X_5 have negative spatial relationship (cross-semivariogram value < 0), means that the spatial variability of Y and X_5 got smaller when the distance got larger, meanwhile the other variables have positive spatial relationship. The coregionalization matrices obtained were different at each spatial distance. In this study, the estimation of coregionalization matrices was performed using least square method then setting any negative eigen values to zero in obtaining the positive definite matrices, this method was suggested by Pebesma in [9].

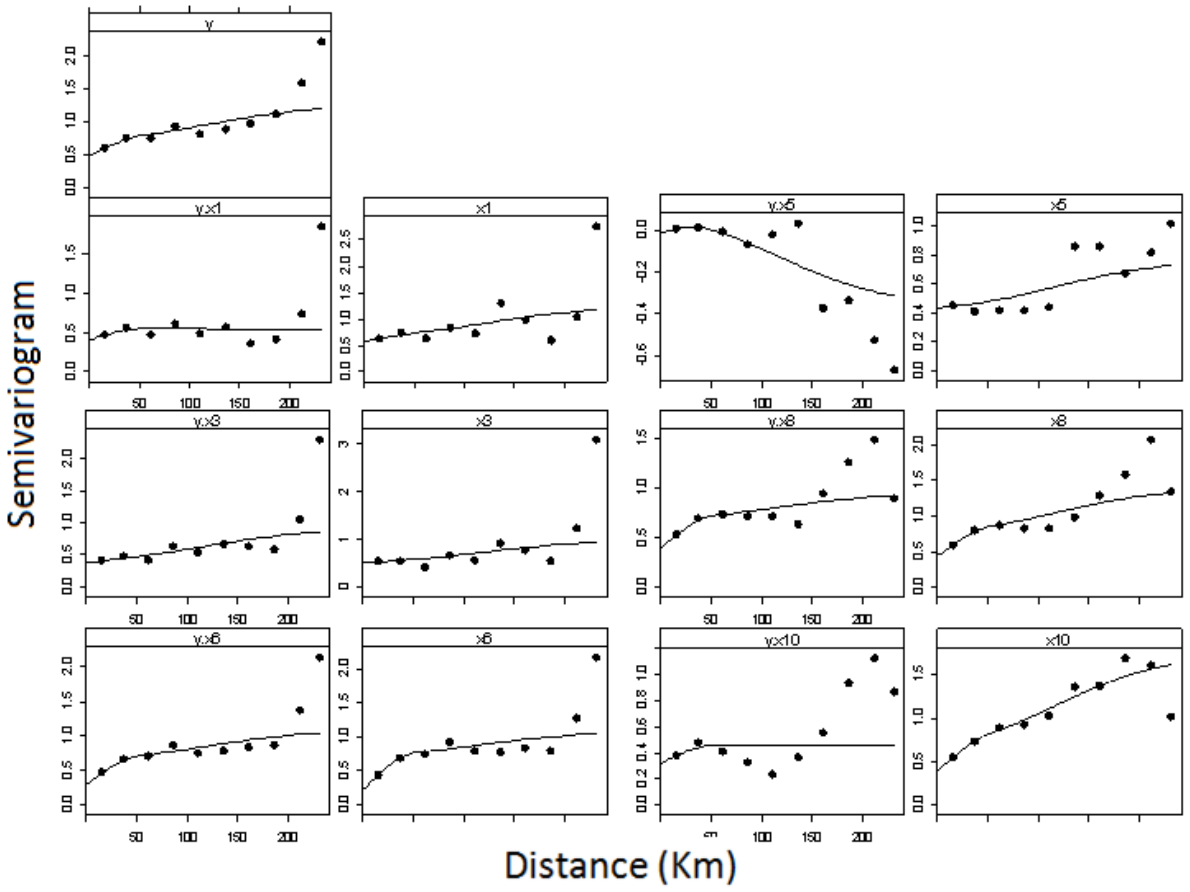

Figure 2: Linear Model of Coregionalization Nug(0) + Sph(54) + Gau(161)

Table 2 shows the relative contribution estimations of each spatial scales to overall variability, 32.37% of the overall variability was explained by nugget effects, 20.44% was explained by the variability occurring in small scale (0-54 km) or in adjacent sub-districts and 47.18% was explained by the variability in large scale (within 0-161 km). The highest variability proportion of X_6 was in small scale (0-54 km) means that the variability

proportions of households having their own buildings were explained by the variability located in adjacent sub-districts, whereas the other variables was explained by the variability within 0-161 km. Table 3 shows the variability proportions between pairs of variable, the variability proportions among Y and X_1, X_8 and X_{10} were dominated by high nugget effects, meanwhile the variability proportions among Y and X_5 and X_6 were dominated by low nugget effects. Ribeiro and his colleagues in [7] stated that nugget effect represents unmeasured factors, measurement errors and the absence of information over small distances.

The size of non-nugget effects (small scale + large scale) represents the spatial dependence that can be modelled, variable that has the highest proportion spatial dependence was X_6 (81.64%) then X_8 (71.64%) and X_{10} (77.52%), they were determined as local parameters in SGWR and the other variables were determined as global parameters. Spesification local and global parameter using LMC was more measurable and objective because the spesification was based on spatial variability, as for the results of LMC and confidence interval of GWR coefficients used by Pongoh in [5] were different, using confidence interval of GWR coefficients obtained X_5, X_6 and X_8 as local parameters and X_1, X_3 and X_{10} as global parameters, meanwhile using LMC obtained X_6, X_8 and X_{10} as local parameters and X_1, X_3 and X_5 as global parameters.

Table 2: Proportion (%) of Spatial Variability

Variable	Nugget	Small Scale	Large Scale	Overall
Overall	32.37	20.44	47.18	100
Y	37.01	21.06	41.93	100
X_1	38.78	15.67	45.55	100
X_3	43.19	7.22	49.58	100
X_5	44.72	7.41	47.86	100
X_6	18.36	42.21	39.43	100
X_8	28.36	27.59	44.05	100
X_{10}	22.48	19.00	58.52	100

Tabel 3: Proportion (%) of Spatial Variability between Y and X

Variable	Nugget	Small Scale	Large Scale
Y, X_1	69.37	27.83	2.80
Y, X_3	41.07	8.27	50.66
Y, X_5	3.52	9.43	87.05
Y, X_6	27.42	33.21	39.38
Y, X_8	41.68	29.30	29.02
Y, X_{10}	70.22	29.64	0.14

4.4 Semi-parametric Geographically Weighted Regression Modelling

LMC obtained that X_6, X_8 and X_{10} have high spatial dependence proportion (81.64%, 71.64% and 77.52% respectively). SGWR model using adaptive kernel Bisquare which X_6, X_8 and X_{10} as local parameters and and X_1, X_3 and X_5 as global parameters obtained AICc value 631.2. In previous study by Pongoh in [5], X_5, X_6 and X_8 as local parameters and X_1, X_3 and X_{10} as global parameters performed using SGWR with fixed kernel Bisquare weighting obtained AICc values 744.2. Models with adaptive kernel function weighting produced smaller AICc and MSPE values than models with fixed kernel function weighting. It was represented in Table 4. Table 4 also shows the comparison between GWR and SGWR models, the comparison between GWR and SGWR models based on AICc value obtained M4, which parameter spesification was determined using LMC,

as the best model, meanwhile the comparison among SGWR models based on MSPE value also obtained M4 as the best model. Therefore, SGWR model using adaptive kernel Bisquare which X_6, X_8 and X_{10} as local parameters was better than GWR model to be applied on poverty data in North Sulawesi Province.

Table 4: Comparison between GWR and SGWR Models

Model Spesification	Model	AICc	MSPE
GWR fixed kernel Bisquare	M0	746.05	
GWR adaptive kernel Bisquare	M1	662.06	
SGWR fixed kernel Bisquare (X_6, X_8, X_{10} local)	M2	743.5	5.25
SGWR fixed kernel Bisquare (X_5, X_6, X_8 local)	M3	744.2	5.28
SGWR adaptive kernel Bisquare (X_6, X_8, X_{10} local)	M4	631.2	1.69
SGWR adaptive kernel Bisquare (X_5, X_6, X_8 local)	M5	633.2	1.67

5. Conclusion

Linear Model of Coregionalization is a tool that can help Geographically Weighted Regression models in showing the spatial variability and also can assist parameter specification of Semi-parametric Geographically Weighted Regression. The results in this study obtained that the variability proportion in poverty data in North Sulawesi Province was explained by variability occurring within 161 km. It was obtained that X_6 (percentage of households have their own buildings), X_8 (percentage of households using electricity/PLN and X_{10} (percentage of households using their own latrines) as the local or geographically varying parameters. Based on AICc and MSPE values, the SGWR model was better than GWR model to be applied on data of poverty rate in North Sulawesi Province. In addition, GWR and SGWR using adaptive kernel function weighting obtained models with the smallest AICc.

Acknowledgements

The authors thank to **F. Pongoh** over the data provided.

References

[1] A.S. Fotheringham, C. Brundson, M Chaltron. Geographically Weighted Reggression: The Analysis of Spatially Varying Relationships. England: John Wiley & Sons Ltd, 2002.

[2] T. Nakaya, A.S. Fotheringham, M. Charlton, C. Brunsdon. “Geographically Weighted Poisson Regression for Disease Association Mapping.” Statist. Med, vol. 24, pp. 2695-2717, 2005.

[3] A.S. Fotheringham, M. Charlton, C. Brunsdon, T. Nakaya. “Model Selection Issues in Geographically Weighted Regression,” in Proceedings of the 8th International Conference on GeoComputation, University of Michigan, USA, 2005.

[4] C.L. Mei, N. Wang, W.X. Zhang. “Testing The Importance of The Explanatory Variables in A Mixed

Geographically Weighted Regression Model." Environment and planning, vol. 38, pp. 587-598, 2006.

[5] F. Pongoh. "Regresi Terboboti Geografis dan Regresi Terboboti Geografis Campuran: Studi Kasus Status Kesejahteraan Rendah di Sulawesi Utara." M.Si. thesis, Bogor Agricultural University, Indonesia, 2015.

[6] M. Goulard, M. Voltz. (1992). "Linear Coregionalization Model: Tools for Estimation and Choice of Cross-Variogram Matrix." Mathematical Geology. 24(3), pp. 269-286.

[7] M.C. Ribeiro, A.J. Sousa, M.J. Pereira. "A Coregionalization Model Can Assist Spesification of Geographically Weighted Poisson Regression: Application to an Ecological Study." Spatial and Spatio-temporal Epidemiology, vol. 17, pp. 1-13, 2016.

[8] E.H. Isaaks, R.M. Srivastava. An Introduction to Applied Geostatistics. New York: Oxford University Press, 1989.

[9] E.J. Pebesma. "Multivariable Geostatistics in S: The Gstat Package." Comput Geosci, vol. 30, pp. 683-691, 2004.