Distance weight of GWR-Kriging model for stunting cases in East Java

D Ardianti*, H Pramoedyo and N Nurjannah
Department of Statistics Faculty of Mathematics and Natural Sciences, Brawijaya University, Malang, Indonesia

*debyardianti@student.ub.ac.id

Abstract. The chosen of distance weights is needed to form an accurate Geographically Weighting Regression model. There are 3 types of distance weights namely Gaussian kernel, Bisquare kernel and Tricube kernel. The weighting in GWR describes the closeness relation between locations. For data that has spatial heterogeneity, GWR models are more suitable models than OLS models. This study was conducted to obtain the best distance weighting based on minimum cross-validation method. Using secondary data from the Health Department in East Java with 34 districts for observation, the dependent variable is stunting and five independent variables that influence stunting cases. Based on the result, GWR models with fixed gaussian models produces a better accuracy in higher $R^2$ values compared to OLS models. The predicted map of the spread stunting cases conducted by interpolation GWR Kriging using exponential semivariogram.

1. Introduction

When data collected from different spatial locations that indicate the dependence between measurement data and location is called spatial data [1]. Spatial heterogeneity is one of the assumptions on spatial data. The GWR method is a method that is suitable for modelling data that contains spatial heterogeneity compared to multiple regression methods. Geographically Weighted Regression (GWR) is a method that is often used in spatial modelling with a point approach to linear regression models which will be weighted regression models [2]. The accuracy of the weighting method is needed because the weighted value will represent the location of the origin of the data [3]. Several studies have been conducted using the GWR model, namely [4–6].

The GWR model is used to predict at each observation location, and observations outside the study sample are not well predicted by the GWR model [7]. To solve this problem, the Kriging method will be used to predict the predictive value and the mapping. Kriging is a geostatistical technique used to predict and interpolate data at unsampled locations [8]. In ordinary kriging interpolation semivariogram is used which is a function that describes, models, and calculates spatial autocorrelation between data from a variable and functions as a measure of variance [9]. Several studies have been conducted using the Kriging, namely [10][11].

Stunting is a condition in which toddlers have less length or height compared to age [12]. Toddler stunting is a chronic nutritional problem caused by various factors such as socio-economic conditions, maternal nutrition during pregnancy, illness in infants, and lack of nutritional intake in infants. Based
on RISKESDAS data in East Java, the prevalence of stunting in East Java is on average 32.81%, this average value has decreased by 2.99% over the last 5 years.

Research related to stunting modelling is very important to do because with this study each region can find out what factors influence the incidence of stunting and detect pockets of areas that have a high stunting rate. This research will use stunting data and GWR Kriging interpolation method with weighted matrix in the form of distance.

2. Data and methods
The data used are secondary data obtained from the publication book Riskesdas and Health Profile of East Java and Malang Regency in 2018, namely stunting prevalence data (Y), K1 visit coverage data (X1), FE3 tablet consumption data (X2), Exclusive breastfeeding data (X3), complete immunization coverage data (X4) and hygiene and healthy lifestyle data (X5). The observation unit is 38 district/cities in East Java. In multiple linear regression analysis produces a small $R^2$ value, namely 0.291 and in the research data there is spatial heterogeneity so that the GWR method is used and the spatial interpolation of GWR kriging using a distance and area weighted matrix with the following stages:

I. Doing Multiple Regression Analysis
II. Testing spatial heterogeneity
III. Doing GWR analysis,
IV. Calculate euclid distance.
   i. Choose the best weighting based on cross-validation method
   ii. Estimating the parameters of the GWR model.
   iii. Testing the parameters of the GWR model.
V. Doing Kriging Analysis,
   i. Prepare a map of the spatial data location, which is a digitized map containing spatial data attribute information.
   ii. Modeling the exponential semivariogram for stunting interpolation
   iii. Make interpolation of stunting contour map

3. Regression analysis, spatial heterogeneity, GWR, and Kriging

3.1. Regression analysis
The general form of the multiple linear regression model is as follows [13]:

$$ y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \epsilon_i, \quad i = 1, 2, ..., n ; k = 1, 2, ..., p $$

(1)

where, $y_i$ is the observed value of the i-th predictor variable, $x_{ik}$ is the k-th predictor variable's observed value, $\beta_0$ is the regression model intercept value, $\beta_k$ is the k-th predictor variable regression coefficient and $\epsilon_i$ is the i-error.

Estimation of parameters from multiple regression models using the Ordinary Least Square (OLS) method, with the following equation [13]:

$$ \hat{\beta}(u_i, v_i) = [X'X]^{-1}X'Y $$

(2)

3.2. Spatial heterogeneity
Spatial heterogeneity testing used the Breusch-Pagan test as follows [14]:

$$ H_0: \sigma_1^2 = \sigma_2^2 = ... = \sigma_n^2 = \sigma^2 $$

$$ H_1: \text{there are at least one } i \text{ where } \sigma_i^2 \neq \sigma^2 $$

If $H_0$ True, Test Statistic of Breusch-Pagan:

$$ BP = \frac{1}{2} f'Z(Z'Z)^{-1}Z'f + \left( \frac{1}{T} \right) \left[ \frac{e'We}{\sigma^2} \right]^2 \sim \chi^2_{(p+1)} $$

(3)
where: \[ f_i = \left( \frac{e_i}{\sigma^2} - 1 \right) \]

3.3. Geographically Weighted Regression (GWR)

Geographically Weighted Regression (GWR) is a spatial analysis using points which is also the development of linear regression analysis by considering the location (spatial). The GWR model is as follows [2]:

\[ y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i)x_{ik} + e_i ; \quad i = 1, 2, \ldots, n \quad (4) \]

where, \( y_i \) is the observed value of the i-th predictor variable, \( x_{ik} \) is the k-th predictor variable’s observed value, \( \beta_0(u_i, v_i) \) is the regression model intercept value, \( \beta_k(u_i, v_i) \) is the k-th predictor variable regression coefficient and \( e_i \) is the i-error.

3.3.1. Estimates parameter of GWR model.

Estimation parameters of the Geographically Weighted Regression (GWR) model uses the Weighted Least Square (WLS) method, which means giving different weights for each location. The following is the parameter estimation for the GWR model [2]:

\[ \hat{\beta}(u_i, v_i) = [X'W(u_i, v_i)X]^{-1}X'W(u_i, v_i)y \quad (5) \]

From (5) equation, the parameter coefficient of the GWR model for each location has different values.

3.3.2. Weighting matrix.

In this research, we use the fixed kernel weighting, which is a function with the same bandwidth value at each location[2]. Bandwidth tends to be narrow if the observation point is spread across the observation location, whereas if the observation point tends to be far from the ith observation point, the bandwidth will be wider. Following are the formulas for the Fixed Kernel[3]:

a. Fixed Gaussian Kernel

\[ w_{ij} = \exp \left[ -\frac{1}{2} \left( \frac{d_{ij}}{b} \right)^2 \right] \quad (6) \]

b. Fixed Bisquare Kernel

\[ w_{ij} = \begin{cases} 1 - \left( \frac{d_{ij}}{b} \right)^2, & \text{if } d_{ij} < b \\ 0, & \text{if } d_{ij} \geq b \end{cases} \quad (7) \]

c. Fixed Tricube Kernel

\[ w_{ij} = \begin{cases} 1 - \left( \frac{d_{ij}}{b} \right)^3, & \text{if } d_{ij} < b \\ 0, & \text{if } d_{ij} \geq b \end{cases} \quad (8) \]

with,

\[ d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2} \quad (9) \]

dan \( b \) is optimum bandwidth with Cross Validation (CV) method

\[ CV(b) = \sum_{i=1}^{n}(y_i - \hat{y}_i(b))^2 \quad (10) \]

3.3.3. Testing GWR Models

a. Simultaneous testing to determine the effect of predictor variable together against the response variable [2]

\[ H_0: \beta_k(u_i, v_i) = \beta_k \]

\[ H_1: \text{at least one } \beta_k(u_i, v_i) \neq \beta_k \]
If \( F > F_{\text{table}} \), reject \( H_0 \).

b. Partial testing to determine which predictor variables that influence the response variable for each observation location[3], using the t test statistics based on the hypothesis:

\[
H_0: \beta_k(u_i, v_i) = 0 \\
H_1: \text{at least one } \beta_k(u_i, v_i) \neq 0
\]

Test of statistic:

\[
t = \frac{\hat{\beta}_k(u_i, v_i)}{s.e(\hat{\beta}_k(u_i, v_i))} \quad (11)
\]

Reject \( H_0 \), if the test statistic \(|t| > t_{\text{table}}\).

3.4. Ordinary Kriging

In this research, to make interpolation kriging we use Exponential Semivariogram, with equation as follows [15]:

\[
\gamma(h) = C_0 + C \left[ 1 - \exp \left( \frac{3h}{\alpha} \right) \right] \quad (12)
\]

4. Results and discussions

4.1. Identification of spatial heterogeneity for stunting data

The results of Breusich Pagan test for distance weighting are presented in Table 1.

| BP         | P-Value |
|------------|---------|
| Distance Weight | 11,484  | 0.004259 |

Based on Table 1, \( P \text{Value} = 0.004259 < \alpha = 0.05 \), reject \( H_0 \), so there is spatial heterogeneity in the data of stunting.

4.2. GWR model

4.2.1. Choose the best distance weighting. Based on Table 2, the minimum Cross Validation (CV) value of 14.0822 for the Fixed Gaussian kernel weighting function. So the weighting used is the Fixed Gaussian weighting function with the resulting bandwidth value of 647589.2. After determining the most optimum bandwidth value, then determining the weighting matrix using the Fixed Gaussian kernel function. The weighting matrix obtained at each location will be used to form a model so that each location will have a different model.

| Model          | CV       | Bandwidth |
|----------------|----------|-----------|
| Fixed Gaussian | 14,0822  | 647589.2  |
| Fixed Bisquare | 15,81945 | 40377     |
| Fixed Tricube  | 14,0822  | 647689.2  |

4.2.2. Estimates parameter OLS and GWR model

After obtaining the optimum Bandwidth value, then forming a weighting matrix used to determine the estimated GWR parameter at each research location. The estimates parameter presented in Table 3.
Table 3. Estimates parameter OLS and GWR model with Fixed Gaussian Weighting.

| Parameter | Global Model (OLS) | Fixed Gaussian Model (GWR) |
|-----------|--------------------|----------------------------|
| $\beta_0$ | 2.448              | Min 1.992                  |
|           |                    | Max 2.699                  |
| $\beta_1$ | -0.080             | Min -0.085                 |
|           |                    | Max -0.075                 |
| $\beta_2$ | 0.060              | Min 0.061                  |
|           |                    | Max 0.062                  |
| $\beta_3$ | 0.032              | Min 0.032                  |
|           |                    | Max 0.033                  |
| $\beta_4$ | 0.001              | Min 0.001                  |
|           |                    | Max 0.003                  |
| $\beta_5$ | -0.004             | Min -0.005                 |
|           |                    | Max -0.004                 |
| $R^2$     | 0.291              | 0.599                      |

The minimum and maximum parameter estimates of the GWR model in Table 3 show that the minimum and maximum limits are variable $X_1$ with interval $-0.085 \leq \beta_1 \leq -0.075$. The minimum limit is located at Banyuwangi and the maximum limit is located at Bondowoso, which means that the greatest effect of the variable coverage visiting of pregnant women on the stunting variable is in Bondowoso and the smallest effect is in Banyuwangi. While the independent variable that has the greatest influence on the stunting variable is the consumption of FE tablets with an interval of $0.032 \leq \beta_2 \leq 0.033$.

4.2.3. Testing the parameter of GWR model

a. Simultaneous testing

Table 4. ANOVA GWR model.

| Source            | Df | Sum Sq | Mean Sq | F Value |
|-------------------|----|---------|---------|---------|
| Global Regression | 6  | 8,7369  |         |         |
| GWR Improvements  | 0.45 | 0.3915  | 0.8746  |         |
| GWR Residuals     | 27,55 | 8.6243  | 0.31306 | 2.7937  |

Based on the table above, the statistical value of the test $F_{value} = 2.7937 > F_{(0.15;0.451;27,55)} = 2.195$.

So reject $H_0$, the conclusion is Fixed Gaussian Kernel weighting affects the estimation parameter of the GWR model.

b. Partial Testing

The partial test shows the predictor variable with fixed gaussian weighting affect the response variable at each location, as described in Table 5.
Table 5. Grouping of district based on significant variable.

| District                        | Significant Variable                      |
|---------------------------------|------------------------------------------|
| Blitar, Bojonegoro, Bondowoso,  | Coverage Visiting of Pregnant Woman,     |
| Gresik, Jember, Jombang, Kediri,| Consumption FE Tablet                     |
| Lamongan, Lumajang, Madiun,     |                                          |
| Magetan, Malang, Mojokerto,     |                                          |
| Nganjuk, Ngawi, Pacitan,        |                                          |
| Pasuruan, Ponorogo, Probolinggo,|                                          |
| Sidoarjo, Trenggalek, Tuban,   |                                          |
| Tulungagung, Batu City, Blitar  |                                          |
| City, Kediri City, Madiun City,|                                          |
| Malang City, Mojokerto City,    |                                          |
| Pasuruan City, Probolinggo City,|                                          |
| Surabaya City                   |                                          |
| Banyuwangi, Situbondo           | Consumption FE Tablet                     |

4.3. Interpolation GWR-Kriging

Mapping with kriging interpolation uses coordinate point data and Ypredicted value data for each GWR model data. The semivariogram used in this study is the exponential semivariogram. The results of forecasting on the GWR Kriging interpolation will be presented in the form of a forecast map to make it more informative and more useful.

Based on the interpolation map in Figure 1, it can be seen that the stunting forecasting value using GWR Kriging interpolation ranges from 31% - 36.9%. High stunting forecasting values are presented on the red map and the low are presented on the green map. It can be seen that only a small proportion of districts in East Java have a high prevalence value of stunting, this is illustrated in the red and outraged maps in Ponorogo and Ngawi Districts.

For example, Banyuwangi Regency located on the green map is classified as 1 or low stunting prevalence with a value between 31.1% - 33.1%. The map also shows that Malang City is affected by Malang Regency and Batuhal City. It can be seen from the location of the points which are close together and are on the same colored map, namely green which is in the low prevalence classification.

![Figure 1. Interpolation GWR Kriging.](image)
5. Conclusion
The GWR model with a distance weighting, namely the Fixed Gaussian Kernel, is better used to model stunting data in East Java because it has $R^2 = 0.599$ greater than the OLS model that has $R^2 = 0.291$, which means that the independent variables in the research simultaneously have an effect on stunting by 59.9%. and the remaining 40.1% is influenced by other variables outside the research variables. From the partial parameter testing, the prevalence of stunting in 32 district affected by Coverage Visiting of Pregnant Woman & Consumption FE Tablet and 2 district affected by Consumption FE Tablet. The results of forecasting the interpolation of GWR Kriging with exponential semivariogram on the high prevalence of stunting in the red areas and the low ones in green with a value range between 31% to 36.9%.

Acknowledgement
We would like the first thank to University of Brawijaya which have supported this research, and the second thank to Health Department of East Java for providing data resources and thank to anonymous reviewer of this paper.

References
[1] Ver Hoef J M, Cressie N A C and Glenn-Lewin D C 1993 Spatial models for spatial statistics: some unification. J. Veg. Sci. 4 441–52
[2] Fotheringham A S, Brunsdon C and Charlton M 2003 Geographically weighted regression: the analysis of spatially varying relationships (John Wiley & Sons)
[3] Chasco C 2007 Modeling spatial variations in household disposable income with Geographically Weighted Regression
[4] Hidayah R N, Wulandari S P and Pramono S 2014 Pemodelan Proporsi Kasus Penyakit Infeksi Saluran Pernapasan Akut (ISPA) bagian Atas pada Balita di Kabupaten Gresik dengan Geographically Weighted Regression. J. Sains dan Seni ITS 3 D146–51
[5] Firdial L 2010 Pemodelan Angka Harapan Hidup di Propinsi Jawa Timur dan Jawa Tengah dengan Metode Geographically Weighted Regression (GWR)
[6] Pratiwi L P S, Hanief S and Suniantara I K P 2018 Pemodelan Angka Putus Sekolah Usia Pendidikan Dasar Dengan Metode Spasial Geographically Weighted Regression Proceeding Seminar Nasional Sistem Informasi dan Teknologi Informasi vol 1 pp 621–5
[7] Walter J and Jeremy W L 2005 Local and global approaches to spatial data analysis in ecology Spatial autocorrelation 97–8
[8] Armstrong M 1998 Basic linear geostatistics (Springer Science & Business Media)
[9] Gaetan C and Guyon X 2010 Spatial statistics and modeling vol 90 (Springer)
[10] Widhita P J A 2008 Penaksiran Kandungan Cadangan Bauksit di Daerah Mempawah Menggunakan Ordinary Kriging dengan Semivariogram Anisotropik Univ. Indones. Depok
[11] Sukarsa K G D E and Dharmawan K 2015 Interpolasi Spasial Dengan Metode Ordinary Kriging Menggunakan Semivariogram Isotropik Pada Data Spasial (Studi Kasus: Curah Hujan di Kabupaten Karangasem) E-Jurnal Mat. 4 26–30
[12] Kementerian Kesehatan RI 2019 Situasi Balita Pendek (Stunting) di Indonesia (Jakarta: Kemenkes RI)
[13] Neter J, Kutner M H, Nachtsheim C J and Wasserman W 1996 Applied linear statistical models
[14] Anselin L 2013 Spatial econometrics: methods and models vol 4 (Springer Science & Business Media)
[15] Rozalia G, Yasin H and Ispriyanti D 2016 Penerapan Metode Ordinary Kriging Pada Pendugaan Kadar NO2 Di Udara (Studi Kasus: Pencemaran Udara Di Kota Semarang) J. Gaussian 5 113–21