Geographically weighted logistic regression modeling on stunting cases in Indonesia

F K Alam¹, Y Widyaningsih¹* and S Nurrohmah¹

¹Department of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Indonesia, Kampus Baru UI, Depok, 16424, Indonesia

*yekti@sci.ui.ac.id

Abstract. Stunting is a condition of failure to thrive in children as a result of chronic malnutrition, so the child is too short at his/her age. Stunting harms children's growth and affects the quality of human resources in the future. To reduce the prevalence of stunting in Indonesia, the government determined priority areas for handling stunting cases in Indonesia. This study aims to determine the variables that affect the status of priority areas for handling stunting in Indonesia. The model used in this study is Geographically Weighted Logistic Regression (GWLR) as a development of logistic regression model that considers spatial effect. This study used Maximum Likelihood Estimation (MLE) method to estimate the parameter model. The spatial weighting function used in this study is the Fixed Gaussian and Fixed Bisquare kernel weighting functions. The response and predictor variables in this study contain missing values, so Classification and Regression Tree (CART) method used to handle the missing values. The results showed that the best GWLR model on stunting cases modeling in Indonesia is the GWLR model with Fixed Bisquare kernel weighting function with AIC value of 622.806477 and model classification accuracy of 0.7257.

Introduction

Stunting is a chronic nutritional problem experienced by children under-five years in the world. Stunting is a condition which a child is shorter compared to the normal in his/her age. A lack of balanced nutrition causes stunting during the golden period. The golden period starts when the child is still in the womb until the child is two years old [1]. Stunting has bad effect on children's growth and the quality of human resources in the future. Based on the Indonesia Toddler Nutrition Status Survey (SSGBI) in 2019, the prevalence of stunting in Indonesia is 27.67%. This prevalence is still higher than the prevalence of stunting set by WHO, 20%. To reduce the prevalence of stunting in Indonesia, the government determined districts/cities as priority areas. Based on this situation, the Indonesian government needs to understand the factors that influence the status of priority areas for handling stunting cases. This study will construct a model that can analyze the relationship between the priority status of stunting cases handling in Indonesia and the variables that influence it.

The model used in this study is the Geographically Weighted Logistic Regression model, a development model of the logistic regression model that considers spatial influence. The spatial influence is represented through a weighting matrix at each observation location to produce an estimation of the local model parameters for each observation location. This model has been widely used in several studies [2, 3]. The advantage of applying the Geographically Weighted Logistic Regression model is the parameter model can be different for each region because the estimation model is specific for each region, whereas the global model represented by only one formula may not be adequate represent local variations [4].
There are missing values in several variables used in this study. Missing values can be interpreted as observations or missing or unavailable information about a research object on certain variables [5]. This study uses the imputation method in dealing with the missing values. The imputation method used is Classification and Regression Tree (CART) which is a classification technique using binary recursive partitioning algorithms [6].

**Method**

1.1. **Spatial Autocorrelation**

Spatial autocorrelation or spatial dependency is correlation between observational values related to spatial locations on the same variable. The existence of spatial autocorrelation shows that observations at one location are affected by observations at other locations that are close to each other [7]. The amount of spatial autocorrelation is calculated using the Moran’s Index, which is formulated in the following equation [8]

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \quad (1)$$

The Moran’s I test used to see whether there is spatial autocorrelation with the following hypothesis [9].

$H_0: I = 0$ (there is not spatial autocorrelation)

$H_1: I \neq 0$ (there is spatial autocorrelation)

with the test statistic is [10]

$$Z(I) = \frac{I - E(I)}{\sqrt{Var(I)}} \quad (2)$$

where $x_i$ represents the value of variable $x$ at the $i$ location for $i = 1, 2, \ldots, n$, $x_j$ represents the value of variable $x$ at the $j$ location for $j = 1, 2, \ldots, n$, $\bar{x}$ represents the average value of the variable $x$, and $w_{ij}$ represents the weight between the $i$-th location and the $j$-th location. The decision rule of the hypothesis is to reject $H_0$ if $|Z(I)| > Z_{\alpha}$, the value of Normal distribution table value with the level of significance $\alpha$. If $H_0$ is rejected, there is a spatial autocorrelation.

1.2. **Geographically Weighted Logistic Regression**

Geographically Weighted Logistic Regression (GWLR) is the development of logistic regression model which considers spatial or location factor. The model of the GWLR formulated as follows.

$$\pi(x_i) = \frac{\exp(\beta_0(u_i, v_i) + \beta_1(u_i, v_i)x_{i1} + \cdots + \beta_p(u_i, v_i)x_{ip})}{1 + \exp(\beta_0(u_i, v_i) + \beta_1(u_i, v_i)x_{i1} + \cdots + \beta_p(u_i, v_i)x_{ip})} ; i = 1, 2, \ldots, n \quad (3)$$

with logit transformation, the equation (3) formed as follows.

$$g(x_i) = \ln \left( \frac{\pi(x_i)}{1 - \pi(x_i)} \right) = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_k, v_k)x_{ik} \quad (4)$$

where $\pi(x_i)$ is the probability of the response variable ($y = 1$) given the value of $x$ at the $i$ location and $\beta_k(u_i, v_i)$ is the $k$-variable regression parameter for each $i$-location based on the latitude and longitude coordinates ($u_i, v_i$). Before constructing the GWLR model, the initial step taken is to determine the geographical location (latitude and longitude) of each district/city in Indonesia. The
latitude and longitude coordinates represent the center of geographical location \((u_i, v_i)\). The next step is to calculate the distance between the observation locations based on the latitude and longitude of each district/city in Indonesia. The concept of distance used in this study is the Euclidean distance with the following formula

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

where \(d_{ij}\) represents the Euclidean distance between observation \(i\) and \(j\), and the variables \(u\) and \(v\) are the geographical location (coordinates) of each object of observation. The next step is to calculate the optimum bandwidth. Bandwidth is the radius of a circle where the points within the radius of the circle considered to affect the parameters of the \(i\)-location model [11]. The method used to select the optimum bandwidth in this study is the Akaike Information Criterion (AIC) method. The optimum bandwidth has a minimum AIC value. The AIC value formulated as follows [12]

\[
AIC = 2p - 2\ln{(L)}
\]  

where \(L = L(\hat{\theta})\) is the maximum value of the GWLR likelihood function and \(p\) is the number of parameters in the GWLR model. After the value of optimum bandwidth is obtained, the next step is to form a weighting matrix. The spatial weighting matrix is used to estimate the parameters at each observation location so each district/city has a different model. The weighting value obtained using the spatial weighting function [13]. The spatial weighting function used to form the weighting matrix in this study is the Fixed Bisquare kernel function and the Fixed Gaussian kernel function. Table 1 shows the formula of spatial weighting function which used in this study.

| Fixed Gaussian | Fixed Bisquare |
|----------------|---------------|
| \(w_{ij} = \exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right]\) | \(w_{ij} = \begin{cases} \exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right], & \text{if } d_{ij} < b \\ 0, & \text{others} \end{cases}\) |

where \(d_{ij}\) is the Euclidean distance between observations \(i\) and \(j\), and \(b\) is the value of fixed optimum bandwidth or bandwidth that has the same value for each location. The next step is to estimate the GWLR model parameters. The method used to estimate the parameters of the GWLR model is Maximum Likelihood Estimation (MLE) method. The first step of the MLE method is to form the likelihood function with the Bernoulli distribution response variable shown in equation (7).

\[
L(\beta(u_i, v_i)) = \prod_{i=1}^{n} P(Y = y_i) \\
= \prod_{i=1}^{n} [(\pi(x_i))^y_i(1 - \pi(x_i))^{1-y_i}] \\
= \prod_{i=1}^{n} \{[1 + \exp \sum_{k=0}^{p} \beta_k(u_i, v_i)x_{ik}]^{-1} \exp \sum_{k=0}^{p} y_i \beta_k(u_i, v_i)x_{ik}] \}. 
\]  

The next step is to form the \(\ln\) likelihood function of the equation (7)

\[
\ln(L(\beta(u_i, v_i))) = \sum_{k=0}^{p} \sum_{i=1}^{n} y_i \beta_k(u_i, v_i)x_{ik} - \sum_{i=0}^{n} \ln \left( [1 + \exp \sum_{k=0}^{p} \beta_k(u_i, v_i)x_{ik}] \right) .
\]  

Spatial factor are weighting factors in the GWLR model. This factor has a different value for each location in the GWLR model [14]. The weighting is given to the \(\ln\) likelihood function to obtain the
The GWLR model. Supposed that the weighting factors for each location \((u_i, v_i)\) is \(w_i(u_i, v_i)\), then the weighted likelihood function is obtained by equation (9)

\[
\ln \left( L^*(\beta(u_i, v_i)) \right) = \sum_{k=0}^{p} \sum_{i=1}^{n} w_i(u_i, v_i)y_i\beta_k(u_i, v_i)x_{ik} - \sum_{i=0}^{n} w_i(u_i, v_i)\ln \left[ 1 + \exp \sum_{k=0}^{p} \beta_k(u_i, v_i)x_{ik} \right].
\]

The parameter estimation is obtained through the first partial derivative of equation (9) of the parameter \(\beta_k(u_i, v_i), k = 0,1,2, ..., p\) then equated to zero, so the following results obtained.

\[
\frac{\partial \ln(L^*(\beta(u_i, v_i)))}{\partial \beta_0(u_i, v_i)} = \sum_{i=1}^{n} w_i(u_i, v_i)y_i - \sum_{i=1}^{n} \pi(x_i)w_i(u_i, v_i) = 0
\]

\[
\frac{\partial \ln(L^*(\beta(u_i, v_i)))}{\partial \beta_1(u_i, v_i)} = \sum_{i=1}^{n} w_i(u_i, v_i)x_{1i} - \sum_{i=1}^{n} x_{1i}\pi(x_i)w_i(u_i, v_i) = 0
\]

\[
\frac{\partial \ln(L^*(\beta(u_i, v_i)))}{\partial \beta_p(u_i, v_i)} = \sum_{i=1}^{n} w_i(u_i, v_i)x_{pi} - \sum_{i=1}^{n} x_{ip}\pi(x_p)w_i(u_i, v_i) = 0.
\]

Based on equation (10), the function that is not closed-form so numerical approaches are needed to get the GWLR model parameter estimator such as the Newton-Raphson iteration method [15].

### 1.3. The Determination of Model Quality

This study uses Information Criterion (AIC) and the accuracy of the model classification to determined the model quality. Akaike Information Criterion (AIC) is a relative measure of the goodness of fit of a statistical model. If two models are compared, the model with the smallest AIC value is the best. The AIC value can be obtained by the following formula [12].

\[
\text{AIC} = 2p - 2\ln (\mathcal{L})
\]

where \(\mathcal{L} = \mathcal{L}(\hat{\theta})\) is the maximum value of the GWLR likelihood function and \(p\) is the number of parameters in the GWLR model.

Accuracy is a measure of model evaluation that shows how much percentage of data classified into the correct class by the classification model. Accuracy information can be obtained through confusion matrix as explained in Table 2.

| Observation | Prediction |
|-------------|------------|
| +           | True Positive (TP) | False Negative (FN) |
| -           | False Positive (FP) | True Negative (TN) |

The accuracy formula based on confusion matrix is [16].

\[
\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}.
\]
1.4. Classification and Regression Tree

Classification and Regression Tree (CART) is a classification technique using binary recursive partitioning algorithms and is one of the methods of a decision tree. Decision tree is a prediction model that uses hierarchical tree structures [6]. Variables that contain missing values will be used as response variables, and other variables will be used as predictor variables. If the response variable is categorical, the decision tree that will be generated is the classification tree. And if the response variable is numeric, the decision tree that will be generated is a regression tree [17]. The first step of using the Classification and Regression Tree (CART) method is constructing the decision tree. Constructing the decision tree uses observations with complete data. Observations with missing values will not be included in the decision tree constructing process. After the decision tree constructed, the next step is predicting the missing values. Prediction is made by entering other variable information that does not contain missing values in the observations into the decision tree that has been constructed.

Result and Discussion

1.5. Databases

The data used in this study is a secondary data obtained from the Public Health Development Index (IPKM) book in 2018 [18] and the poverty data in 2018 [19]. This study uses data at the district/city level in Indonesia. There are 514 districts/cities in Indonesia. In this study, the response variable used is the status of priority areas for stunting handling \((y)\) with the category 0 being the non-priority areas for handling stunting cases (stunting prevalence is less than the average of stunting prevalence in Indonesia) and the category 1 is the priority area for stunting case handling (stunting prevalence is higher than the average of stunting prevalence in Indonesia). The average of stunting prevalence in Indonesia used in this study is 32.01%. The predictor variables are the prevalence of chronic energy deficiency in women \((x_1)\), percentage of poor population \((x_2)\), access to proper sanitation \((x_3)\), access to clean water \((x_4)\), the percentage of handwashing habits \((x_5)\), the prevalence of diarrhoea in children under five \((x_6)\), the prevalence of acute respiratory infections in children under five \((x_7)\), and complete basic immunization coverage \((x_8)\). There are missing values in the data used in this study. Table 3 shows the number of missing values in each variable used in this study.

| Variable | Complete Observation | Number of Missing Values | Percentage of Missing Values |
|----------|----------------------|--------------------------|-----------------------------|
| \(y\)    | 513                  | 1                        | 0.19%                       |
| \(x_1\)  | 513                  | 1                        | 0.19%                       |
| \(x_2\)  | 514                  | 0                        | 0%                          |
| \(x_3\)  | 514                  | 0                        | 0%                          |
| \(x_4\)  | 512                  | 2                        | 0.38%                       |
| \(x_5\)  | 514                  | 0                        | 0%                          |
| \(x_6\)  | 511                  | 3                        | 0.57%                       |
| \(x_7\)  | 498                  | 16                       | 3.11%                       |
| \(x_8\)  | 509                  | 5                        | 0.95%                       |

The missing values will be imputed using Classification and Regression. Missing value in variable \(y\) will be imputed using classification tree. Otherwise, missing values in the other variable \((x_1, x_4, x_6, x_7, \text{ and } x_8)\) will be imputed using regression tree.
1.6. Moran's I Test
Firstly, Moran's I test for every variable is carried out to know if there is spatial autocorrelation in the data. Table 4 shows the results of Moran's I test.

| Variable | Moran's I   | P-value    |
|----------|-------------|------------|
| y        | 0.005406    | 0.008497   |
| x₁       | < 2.2e⁻¹⁶   | < 2.2e⁻¹⁶  |
| x₂       | < 2.2e⁻¹⁶   | < 2.2e⁻¹⁶  |
| x₃       | 1.212e⁻⁸    | 1.212e⁻⁸   |
| x₄       | 3.946e⁻¹⁵   | 4.289e⁻¹⁴  |
| x₅       | < 2.2e⁻¹⁶   | < 2.2e⁻¹⁶  |
| x₆       | 0.002115    | 0.00276    |
| x₇       | 2.218e⁻¹⁰   | 5.086e⁻¹⁰  |
| x₈       | < 2.2e⁻¹⁶   | < 2.2e⁻¹⁶  |

Based on Table 4, because the p-value of the response variable and the eight predictor variables are less than the 5% significance level, H₀ is rejected so the local approach is a suitable model for analyzing the relationship between the status of priority areas for handling stunting and the eight predictor variables.

1.7. Geographically Weighted Logistic Regression
The model used in this study is GWLR model as formulated below.

\[
g(x_i) = \ln \left( \frac{\pi(x_i)}{1 - \pi(x_i)} \right) = \beta_0(u_i, v_i) + \beta_1(u_i, v_i)x_{i1} + \beta_2(u_i, v_i)x_{i2} + \beta_3(u_i, v_i)x_{i3} + \beta_4(u_i, v_i)x_{i4} + \beta_5(u_i, v_i)x_{i5} + \beta_6(u_i, v_i)x_{i6} + \beta_7(u_i, v_i)x_{i7} + \beta_8(u_i, v_i)x_{i8} \tag{13}\]

GWLR model processing in this study was carried out by GWR4 software version 4.0.90. GWR4 is a software to calibrate the GWLR model where diversity between variables can be explored geographically. Table 5 shows the summary of the GWLR model with the Fixed Gaussian kernel weighting function and the Fixed Bisquare kernel weighting function.

| Model                | Optimum Bandwidth | AIC         | Accuracy |
|----------------------|-------------------|-------------|----------|
| GWLR (Fixed Bisquare)| 4372.904          | 622.806477  | 0.7257   |
| GWLR (Fixed Gaussian)| 1542.791          | 622.865125  | 0.7276   |

Based on Table 5, it can be concluded that the GWLR model with the Fixed Bisquare kernel weighting function is the best model for analyzing the relationship between the response variable and the predictor variables in this study.
The existence of spatial influences results in the relationship between predictor variables and response variables vary so that not all predictor variables analyzed have a significant effect on a location. Table 6 shows groups of predictor variables that significantly influence the response variable.

Table 6. Groups of predictor variables that significantly influence the response variable

| Group | Predictor Variables That Significantly Influence | The Number of District/City |
|-------|--------------------------------------------------|-----------------------------|
| 1     | -                                                | 22                          |
| 2     | $x_3$                                            | 55                          |
| 3     | $x_3$ and $x_8$                                  | 91                          |
| 4     | $x_3$, $x_4$, and $x_8$                          | 298                         |
| 5     | $x_2$, $x_3$, $x_4$, and $x_8$                   | 48                          |

Note:
- $x_2$: percentage of poor population
- $x_3$: access to proper sanitation
- $x_4$: access to clean water
- $x_8$: complete basic immunization coverage

Based on Table 6, there are five area groups based on significant predictor variables. In group 1, there are no predictor variables in this study that significantly influence the response variable. This indicates that in group 1, the number of priority areas for stunting handling is influenced by other variables that not in the model. Group 1 is the group with the fewest number of districts/cities, consisting of 22 districts/cities. Group 2 is the group with $x_3$ as the significant variable. Group 3 is the group with $x_3$ and $x_8$ as the significant variables. Group 4 is the group with the most number of districts/cities. The significant variables in group 4 are $x_3$, $x_4$, and $x_8$. And the last group is group 5. Significant variables in this area group are $x_2$, $x_3$, $x_4$, and $x_8$. Figure 1 shows the distribution of groups of predictor variables that significantly influence the response variable.

Figure 1. The distribution map of significant predictor variables groups

As seen on Figure 1, group 1 tends to be in the eastern part of Indonesia, precisely on Papua. Group 2 tends to be on Papua and Maluku. Group 3 tends to spread in the central and eastern parts of Indonesia. Group 4 tends to be spread in western Indonesia, especially in Java. Moreover, group 5 is spread only on Sumatera.

Conclusion
This study used Classification and Regression Tree (CART) method to deal with the missing values. Based on Moran’s I test results, the Geographically Weighted Logistic Regression model is an appropriate model to analyze the relationship between the status of priority areas for stunting treatment and the variables that influence it. Based on the AIC value and the accuracy value, the GWLR model with the Fixed Bisquare kernel weighting function is the best model for analyzing the relationship between the status of the priority area for stunting handling and the variables that influence it. Based
on the analysis of the best GWLR model, there are five groups of area. Group 1 tends to be in the eastern part of Indonesia, precisely in Papua. In group 1, none of the predictor variables in this study has a significant effect on the response variable. Group 2 tends to be in Papua and Maluku. The significant variable in this area group is $x_3$. Group 3 tends to be scattered in the central and eastern parts of Indonesia. The variables that are significant in this area group are $x_3$ and $x_8$. Group 4 tends to be scattered in the western part of Indonesia, especially in Java. The variables that are significant in this area group are $x_3$, $x_4$, and $x_8$. Moreover, group 5 spreads only in Sumatera. The variables that are significant in this area group are $x_2$, $x_3$, $x_4$, and $x_8$.

**Acknowledgments**

This study was funded by Directorate of Research and Development of Universitas Indonesia (DRPM UI) as a grant of Publikasi Terindeks Internasional (PUTI) Prosiding 2020 No. NKB-1028/UN2.RST/HKP.05.00/2020.

**References**

[1] Nelson C A 1999 Neural plasticity and human development *Curr. Dir. in Psychol. Sci.*, 8 42–45

[2] Albuquerque P H M, Medina F, and Da Silva A R 2017 Geographically weighted logistics regression applied to credit scoring models Rev. Contab. e Financas, 28 93–112

[3] Wu L, Deng F, Xie Z, Hu S, Shen S, Shi J and Liu D 2016 Spatial analysis of severe fever with Thrombocytopenia Syndrome virus in China using a geographically weighted logistic regression model *Int. J. Environ. Res. Public Health*, 13 1125

[4] Atkinson P M, German P M, Sear P M and Clark P A 2003 Exploring the relations between riverbank erosion and geomorphological controls using geographically weighted logistic regression *Geogr. Anal.*, 35 58–82

[5] Buhi E R, Goodson P and Neilands T B 2008 Out of sight, not out of mind: Strategies for handling missing data *A. J. Health Behav.*, 32 83–92

[6] Rokach L and Maimon O 2015 *Data Mining with Decision Trees: Theory and Applications* (New Jersey: World Scientific)

[7] Mathur M 2015 Spatial autocorrelation analysis in plant population: An overview *J. Appl. Nat. Sci.*, 7 501–513

[8] Moran P A P 1950 Notes on continuous stochastic phenomena *Biometrika*, 37 17–23

[9] Viton P A 2010 Notes on spatial econometric models *City Reg. Plan.*, 870 9–10

[10] Wong D W S and Lee J 2005 Statistical analysis of geographic information with ArcView GIS and ArcGIS *Int. J. Geogr. Inf. Sci.*, 11 1–3

[11] Purhadi, Rifada M and Wulandari S P 2012 Geographically weighted ordinal logistic regression model *Int. J. Math Comp.*, 16 116–126

[12] Akaike H 1998 *Information Theory and Extension of The Maximum Likelihood Principle* (New York: Springer)

[13] Nakaya T, Fotheringham A S, Brunsdon C and Charlton M 2005 Geographically weighted Poisson regression for disease association mapping *Stat. Med.*, 24 2695–2717

[14] Runadi T, Widyansih Y and Lestari D 2020 Modeling total crime and the affecting factors in Central Java using geographically weighted regression *J. Phys. Conf. Ser.*, 1442 012026

[15] Agresti A 1990 *Categorical Data Analysis* (New York: John Wiley and Sons, Inc.)

[16] Novakovic D J, Veljovic A, Ilic S S, Papic Z and Tomovic M 2017 Evaluation of classification models in machine learning *Theory Appl. Math Comp. Sci.*, 7 39–46

[17] Breiman L 1984 *Classification and Regression Trees* (New York: Routledge)

[18] Kemenkes RI 2018 *Buku Indeks Pembangunan Kesehatan Masyarakat Tahun 2018* (Jakarta: Kementerian Kesehatan RI)

[19] BPS 2019 *Data Statistik dan Informasi Kemiskinan Kabupaten/Kota Tahun 2018* (Jakarta: Badan Pusat Statistik)