Identifying Determinants of Child Malnutrition Using Spatial Regression Analysis

K M Zurnila1*, M Saiful1 and M Selvi1
1Statistics Department, Faculty of Mathematics and Natural Science, Syiah Kuala University, Banda Aceh, Indonesia

*Corresponding author: kesumaku@yahoo.com

Abstract. As a public health problem, the nutritional status of childrens under five years old is a sensitive indicator of a country's health status. Malnutrition affected by many factors, including spatial dependence factors. This factor indicates the value of observations from a region affected by the value of observations in other areas. It assumed that child health factor, maternal health, environmental health, and region influenced the prevalence of malnutrition among under-five children. The objective of the study was to determine factors that associated with malnutrition among under-five children in Sumatera Island and the spatial regression model for the case. The study used Basic Health Research (Riskesdas) and Public Health Development Index (IPKM) survey 2013 from 125 districts on Sumatera Island. Spatial regression analysis and Spatial Autoregressive Model (SAR) applied in this study to obtain the significant determinants of child malnutrition. The results of this study revealed that the lack of chronic energy in women of childbearing age, the proportion of family planning users, households that having clean and healthy life behaviour (PHBS), and access to clean water were the significant factors of child malnutrition.

Keywords: malnutrition, Sumatera Island, spatial dependence, Spatial Autoregressive Model (SAR)

1. Introduction
Nutrition problem is a public health problem that can occur in all age groups. The most age group that should be concerned is 0 to 5 years old age group (infants and toddlers). The underweight condition of children under five causes them having a higher risk factor in the face of growth disorders than the adult age group. The disruption of growth during this period cannot be fulfilled in the future and will negatively affect the quality of future generations [1].

Nutrition status of children under five measured by measuring height and weight. These indicators based on body weight, height and ages mostly used to assess the population nutritional status. The nutritional based on the anthropometry status of children fewer than five could be divided into three indices, namely stunting (length or height not appropriate for age), wasting (un suitable weight to height) and underweight (lack of weight against age). In this study, the nutritional status used was malnutrition or underweight based on body weight index by age [1].

By 2015, World Health Organization (WHO) found 250 from 667 million children in the world suffering from malnutrition. Public health problem is considered serious if malnutrition prevalence less between 20-29% and considered very high prevalence if ≥ 30%. Prevalence is the number of incidence of disease within one year compared with the population. The Nutrition Status monitoring survey 2016 [2] conducted at 514 districts in 34 provinces of Indonesia showed that malnutrition prevalence in
Indonesia was 17.8%. Based on WHO criteria, child malnutrition in Indonesia is a public health problem that fall into the medium category. While on the Sumatra Island, there are three provinces that have child malnutrition prevalence less than the national prevalence figures namely Aceh, North Sumatra and Riau Province [1, 3].

Generally, statistical analysis that describes the cause and effect relationship between one dependent variable with one or several independent variables is linear regression analysis. One type of data used in linear regression analysis is cross-sectional data. Cross-sectional data can also be referred as spatial data or area data, which is data that not only states the value of an observation, but also indicates the location where the data resides. There is possibility of spatial dependence in the data meaning that observations in a region can influence by observations in other areas. As Tobler states that everything is related to one another, but something adjacent is more influential [4]. Linear regression analysis ignores the existence of this spatial dependence so that the data analysis for spatial data is a spatial regression analysis. The fundamental component of the spatial model is the spatial weighted matrix reflecting the relationship between one region and another. This area aspect is important to be studied because inter-region certainly has different characteristics [4, 5].

To determine the factors that influence child malnutrition on Sumatra Island by considering aspects of the region, the authors conducted a study to identify child malnutrition using spatial regression analysis.

2. Spatial Regression Analysis

The first stage of the study was to determine spatial association among areas or districts. Spatial associations or spatial autocorrelation indicate an inter-regional relationship to observational values. There are several methods used to detect spatial autocorrelation, they are spatial weighted matrix, Global and Local Moran Index and Moran’s spotted plot [6].

Spatial weighted matrix ($W$) is a matrix that shows the inter-regional relationships in the observation. The adjacent area called as neighbour. The general form of the $W$ matrix is as equation (1) follows:

$$
W = \begin{pmatrix}
    w_{11} & \cdots & w_{1n} \\
    \vdots & \ddots & \vdots \\
    w_{n1} & \cdots & w_{nn}
\end{pmatrix}
$$  \hspace{1cm} (1)

Where $w_{ij}$ denotes the spatial weights for the region-$i$ and region-$j$, with $i = 1, 2 \ldots n$ and $j = 1, 2 \ldots n$. Each element in row $i$ is 0 for a region which is not a neighbour of region $i$, and 1 for a region which is a neighbour of region $i$. Based on that definition, it is not allowed region $i$ to be a neighbour against itself. Therefore, the diagonal element in the matrix $W$ is 0 [7].

There are several methods to express the relationship structure of the region using a spatial weighted matrix ($W$). One of the methods is the k-nearest neighbours method used for areas case. The number of nearby areas (neighbours) selected using the distance approach. Furthermore, the $W$ matrix is standardized by row normalized weights method. If the weights in each row are normalized, then each row must have a value which if summed to 1. This can show in equation (2)

$$
\sum_{j=1}^{n} w_{ij} = 1, \ i = 1, \ldots, n
$$  \hspace{1cm} (2)

Any normalization of the weighing line, $w_{ij}$ can be interpreted as the fraction of all spatial influences on region $i$ caused by region $j$. In the nearest neighbours K method, the normalization of spatial weights using weights of $\frac{1}{k}$ is performed [7].

One of the most important to do for spatial analysis is determining global and local moran index. The Moran Index is used to test spatial dependencies that indicate the circumstances in which the observation value of a single location depends on the value of neighbouring observations territories at adjacent locations. The Global Moran Index can be obtained through the following equations (3):
where \( I \) is Moran Index value, \( n \) is the number of observations, \( \bar{y} \) is the average value of \( y_i \) from \( n \) location, \( y_i \) is observation value of location \( i \), \( y_j \) is observation value of location \( j \) and \( w_{ij} \) is a spatial weighted matrix element. 

Furthermore, the Local Moran Index is useful for detecting hotspots or coldspots in area data. The local Moran index with a spatial weighted matrix defined as equations (4) follows:

\[
I_i = \frac{(y_i - \bar{y})}{\sqrt{\frac{1}{n-1} \sum_{j=1}^{n} w_{ij} (y_j - \bar{y})}} \sqrt{\frac{1}{n-1} \sum_{j=1}^{n} (y_j - \bar{y})^2} 
\]

Visual exploration analysis capable of detecting spatial autocorrelation is Moran's scatter plot. The resulting output is not the original data but the data that has been standardized in Z-score. In visualization, the standardized observed value of a region on the x-axis and the observed value of the average observation of its neighbours on the y-axis. There are four quadrants in the Moran scatter plot, namely [5]:

i. Quadrant I. Indicates a positive autocorrelation (high-high).
ii. Quadrant II. Indicates a negative autocorrelation (low-high) / coldspot.
iii. Quadrant III. Indicates a positive autocorrelation (low-low).
iv. Quadrant IV. Indicates a negative autocorrelation (high-low) / hotspot.

The Moran's I test is a spatial dependency test on the OLS linear regression model's error. Moran's I statistics can write as equations (5):

\[
I = \frac{\epsilon' W \epsilon}{\epsilon' \epsilon} 
\]

where \( \epsilon \) is a \( n \times 1 \) error vector error of the regression model with OLS estimation method with \( N \) of observation and \( W \) is a standardized spatial weighted matrix of size \( n \times n \). If there is a spatial dependence, then it can be continued in the manufacture of spatial regression models [8].

A method used to model a data that contains location or spatial information is spatial regression model. Spatial Autoregressive Models (SAR) is a spatial regression model whose dependent variables are spatially correlated meaning that this model has dependence on one observation in a region with observations in its neighbouring region. The general SAR model [6] is as equations (6)

\[
y = \rho W y + X \beta + \epsilon \\
\epsilon \sim N(0, \sigma^2 I)
\]

Where \( y \) is the vector of dependent variable with size \( n \times 1 \). \( X \) is the matrix of independent variables with size \( n \times (k + 1) \); \( \beta \) is the vector coefficient of regression parameters with size \( (k + 1) \times 1 \); \( \rho \) is the spatial parameter coefficient lag dependent variable. Besides, \( W \) is the spatial weighted matrix by size \( n \times n \); \( \epsilon \) is the error vector of size \( n \times 1 \); \( n \) is the number of observations or location and \( k \) is a number of independent variables \((k=1, 2, \ldots, l)\). Lastly, \( I \) is the identity matrix by size \( n \times n \).

The spatial autoregressive lag parameter (\( \rho \)) indicates the degree of correlation of the spatial component of an area against the other region around it. Estimator of \( \beta \) and \( \rho \) parameters for SAR model using maximum likelihood method [8]. Hence, equations (7-9) is shown as follows;

\[
\hat{\beta} = (X'X)^{-1}X'(I - \rho W)y
\]
One of the basic principles of Maximum Likelihood estimator is asymptotic normality, meaning that the larger the size of $n$ the curve will be closer to the normal distribution. Statistical test significance of the parameters used is as equations (10)

$$
\hat{\theta} = \frac{1}{n} \left( Y'yW'y - \rho W'y \right)
$$

Where $s \cdot b(\hat{\theta})$ is asymptotic standar error. Hypothesis for partial parameter test is:

$H_0 : \hat{\theta} = 0$ (The regression coefficient is not feasible to use)

$H_1 : \hat{\theta} \neq 0$ (The regression coefficient is feasible to use)

Where $\hat{\theta}$ is spatial parameter regression ($\beta$, $\lambda$, and $\rho$). If $Z_{\text{stat}} > Z(\alpha/2)$ or $p\text{-value} < \alpha/2$, then the decision reject $H_0$, meaning regression coefficient feasible to be used on the model [9].
4. Result And Discussion
The prevalence of child malnutrition in Sumatra Island is still quite alarming. The reasons there are still some districts or cities that have a prevalence rate above 30%. That is, public health problems on Sumatra Island are quite high. The spread of child malnutrition prevalence in eight provinces in Sumatera Island can be seen in Figure 1.

![Distribution of child malnutrition’s prevalence in Sumatera Island.](image)

Figure 1 shows that the province of Aceh is the province with the highest number of child malnutrition’s prevalence East Aceh District is at 35.51% while the district with lowest prevalence of child malnutrition were Banda Aceh City at 8.17% in Aceh Province.

Before a spatial weighted matrix is established, keep in mind the neighbours of each district on the Sumatra Island. The method used to determine nearest neighbour in this study is $k$-nearest neighbor method. Given the wide coverage of the Island of Sumatra, the nearest number of $k$ selected is $k = 4.0$. This means that each district will have four nearest neighbours with a spatial weighted matrix that formed is not symmetrical with the size 125 x 125.

The Global Moran Index’s value for child malnutrition prevalence in Sumatera Island was 0.297937 (P-value = 1.72 x 10^{-7}). These results indicate that there was a positive spatial autocorrelation or adjacent districts having similarities in the prevalence of child malnutrition. In another word, we can say that the neighbouring regions influenced the prevalence of child malnutrition in the district of Sumatra Island.

If globally it is known that there is a positive spatial autocorrelation on the data of malnutrition prevalence and underfive nutrition in infants on the island of Sumatra, it is necessary to analyze locally using Local Moran Index. The aim is to know which districts are significant to have spatial autocorrelation or which districts affected by the neighboring regions.

Moran Scatter Plot for child malnutrition prevalence in Sumatera Island can be seen in Figure 2. There are 13 districts with positive autocorrelation (quadrant I), 7 districts with negative autocorrelation (quadrant III), and no districts in quadrant II (hotspot) and quadrant IV (coldspot). The districts in quadrant I had positive autocorrelation, then Pidie Jaya District, Nias Districts, Mandailing Natal
District, South Nias District, Padang Lawas Utara District, Padang Lawas Regency, North Nias District, West Nias District, Padang Sidempuan City, Indragiri Hulu and Indragiri Hilir District had a high prevalence of malnutrition and were surrounded by districts (neighbouring regions) which also had high prevalence rates.

While the districts in quadrant III had negative autocorrelation, meaning that the district of Empat Lawang, Kota Pagar Alam, Seluma District, Bintan District, Natuna District, Anambas Islands District and Tanjung Pinang City had low malnutrition’s prevalence in children under five and surrounded by districts (neighbouring regions) which also had low prevalence rates.

Before performing spatial regression analysis, spatial dependence test is performed. Before testing spatial dependence to determine the spatial regression model, a spatial dependence test was performed on the error of linear regression using Moran Index. The value of Moran Index (I) obtained is 0.103858 (P-value = 0.01768). By using α = 0.05, it can be concluded that there is positive spatial autocorrelation in linear regression error so that need to be continued at making of spatial regression model.

One of the assumptions that must be met before creating a spatial regression model is that the range of errors in the linear regression model should be homogeneous. Testing homogeneity of various errors was done by Breusch-Pagan (BP) test statistic. BP test statistic value is 16.081 (P-value = 0.2448). Using α = 0.05, it can be concluded that the homogeneous error is so that the assumption of homogeneity of various errors is met.

To determine which model is appropriated with the data, it is necessary to test Lagrange Multiplier. The result of Lagrange Multiplier test can be seen in Table 2.
Table 2. Lagrange Multiplier Test.

| Statistic               | Value  | P-value |
|-------------------------|--------|---------|
| Lagrange Multiplier (SAR) | 5.4921 | 0.0191* |
| Lagrange Multiplier (SEM) | 3.0231 | 0.08208 |
| Lagrange Multiplier (SARMA) | 5.5067 | 0.06371 |

Table 2 showed the results of the Lagrange Multiplier test against three spatial regression models. The test results show that SAR model is significant at $\alpha = 0.05$ meaning that there is spatial dependence on lag or dependent variable. Therefore, the spatial regression model used in this study is Spatial Autoregressive Model (SAR). The result of spatial regression estimation tabulated in Table 3.

Table 3. The result of spatial regression parameter estimation of SAR model

| Main Factor        | Variable | $\beta$  | Z-value  | P-value     |
|--------------------|----------|----------|----------|-------------|
| Intersept ($\beta_0$) | 47.157   | 5.0523   | 4.36 x 10^{-7}*** |
| $\rho$             | 0.24335  | 2.626    | 0.0086** |
| Child Health       | $X_1$   | -0.131   | -1.4368  | 0.151       |
|                    | $X_2$   | -0.011   | -0.3322  | 0.739       |
|                    | $X_3$   | 0.016    | 0.5699   | 0.569       |
|                    | $X_4$   | 0.025    | 1.026    | 0.305       |
|                    | $X_5$   | -0.067   | -1.592   | 0.111       |
|                    | $X_6$   | 0.039    | 1.9573   | 0.338       |
|                    | $X_7$   | 0.036    | 1.4497   | 0.147       |
| Maternal Health    | $X_8$   | -0.015   | -0.3987  | 0.690       |
|                    | $X_9$   | 0.139    | 2.5151   | 0.012*      |
|                    | $X_{10}$ | -0.370   | -4.7213  | 2.34 x 10^{-6}*** |
| Environmental Health| $X_{11}$ | -0.136   | -2.4117  | 0.016*      |
|                    | $X_{12}$ | -0.254   | -2.9968  | 0.002**     |
|                    | $X_{13}$ | -0.044   | -1.2589  | 0.208       |

Note: level of significance *(0.05), ** (0.01), *** (0.000)
Based on the results in Table 3, the SAR model for child malnutrition cases on the Sumatra Island written as equation (11):

$$\hat{y}_i = 47.157 + 0.24335\sum_{j=1, j\neq i}^{125} w_{ij} y_j + 0.139x_9 - 0.370x_{10} - 0.136x_{11} - 0.254x_{12}$$  \hspace{1cm} (11)

The informations that can explained in the equation above is $\hat{y}_i$ as the prevalence of children under-five malnutrition in district $i$, where $i=1, 2, \ldots, 125$. $y_j$ as the prevalence of children under-five malnutrition in district $j$ which is the neighbor of region $i$, where $j=1, 2, \ldots, 125$ and $j \neq i$. In addition, $x_9$ as the prevalence of chronic energy deficiency in women of childbearing age; $x_{10}$ as the proportion of family Planning (KB) users; $x_{11}$ is the percentage of households are living clean and healthy (PHBS); $x_{12}$ as percentage of access to clean water; $w_{ij}$ is the spatial weighted matrix by size 125 x 125 and lastly $\rho = 0.24355$.

The result of estimation and testing of parameters of SAR model shows that there are four significant independent variables at $\alpha = 0.05$ that is prevalence of chronic energy deficiency in women of child-bearing age ($x_9$), proportion of family planning ($x_{10}$), percentage of households are living clean and healthy (PHBS) ($x_{11}$), and percentage access to clean water ($x_{12}$).

Based on equation stated, the coefficient value $\beta_0$ is 47.157. This means that if the prevalence of chronic energy deficiency in women of child-bearing age ($x_9$), the proportion of family planning ($x_{10}$), percentage of households with PHBS ($x_{11}$), and coverage of access to clean water ($x_{12}$) is 0, and no spatial dependence in the model, then the prevalence of malnutrition and under nutrition in children will be 47.157%. This suggests that the prevalence of child malnutrition will be very high although there is no influence of the independent variables in the model.

Furthermore, the $\beta_9$ coefficient value of 0.139 has significance if the prevalence of chronic energy deficiency in women of reproductive age increases 1%, hence the prevalence of child malnutrition will increase by 0.139%, while other variables are considered constant. This result is in accordance with the statement of Ministry of Health of Republic Indonesia (2016) that women who are lacking energy or nutrition are at risk of delivering babies with low birth weight or malnutrition as well.

The value of $\beta_{10}$ coefficient of -0.370 has a meaning if the proportion of family planning users increased by 1%, it will reduce the prevalence of child malnutrition by 0.370%. This indicates that the family planning program has succeeded in reducing the prevalence of child malnutrition in Sumatera Island. If the family planning program can be run with the maximum proportion, then the child nutritional problems can be prevented.

The coefficient of $\beta_{11}$ is valued at -0.136, meaning that if the percentage of households is living clean and healthy PHBS increases by 1%, then the prevalence of child malnutrition will decrease by 0.136%. If the number of households with PHBS increases, the prevalence of child malnutrition can be reduced. It also shows that clean and healthy living behavior for family members is very important to prevent various health problems.

The coefficient value of $\beta_{12}$ is -0.254, meaning if coverage of access to clean water increased by 1% it will decrease the prevalence of child malnutrition by 0.254%. These results prove that access to clean water is quite important given the large number of malnutrition prevalence rates in under-fives children on the Sumatra Island. Therefore, it is necessary to make every district have sufficient access to clean water so that malnutrition problem in children can be discharged.

Furthermore the coefficient $\rho$ on the SAR model is significant at $\alpha = 0.05$ with a value of 0.24335. This means that if a district surrounded by four neighbouring regions, then the influence of these areas is 0.24335 multiplied by the average prevalence of child malnutrition in neighboring areas.

Through a pre-established SAR model, 22 spatial regression models established for each district or city with its neighbouring regions. For example, the SAR model for Pidie Jaya District in Aceh Province are as equation (12) and (13):
\[ \hat{y}_{16} = 47.157 + 0.24335 \left[ 0.25(y_9 + y_{10} + y_{16} + y_7) \right] + 0.139x_9 - 0.370x_{10} - 0.136x_{11} - 0.254x_{12} \]  
\[ (12) \]

\[ \hat{y}_{16} = 47.157 + 0.06y_9 + 0.06y_{10} + 0.06y_{16} + 0.06y_7 + 0.139x_9 - 0.370x_{10} - 0.136x_{11} - 0.254x_{12} \]  
\[ (13) \]

SAR model for Pidie Jaya regency can be interpreted that if the prevalence of malnutrition and under-five nutrition in Pidie (Y9) increased 1%, the prevalence of malnutrition and under-nutrition in under-five children in Pidie Jaya district will increase by 0.06%, with assume another variable is constant. Likewise, for the influence of Bireun (Y10), Aceh Jaya (Y16), and Aceh Barat (Y7) District towards Pidie Jaya District can be interpreted in the same way. While the interpretation of the factors that influence the prevalence of child malnutrition the same as the previous explanation.

5. Conclusion
Based on the results of the study, the conclusions of this study are:

i. Factors affecting malnutrition and nutritional status among under-fives in Sumatera Island are prevalence of chronic energy deficiency in women of childbearing age, proportion of family planning users’ percentage of clean and healthy households and percentage of access to clean water.

ii. There are 22 spatial regression models for each district or city with neighbouring regions using SAR model. For Pidie District (Y9) for example, we conclude that if the prevalence of malnutrition among infants increase by 1%, the prevalence of malnutrition in children under five in Pidie Jaya District will increase by 0.06%, assuming other variables are constant. Likewise, for the influence of Bireun District (Y10), Aceh Jaya (Y16), and Aceh Barat (Y7) on Pidie Jaya District, it can be interpreted in the same way.

References

[1]. Republic of Indonesia Ministry of Health. 2016. Indonesian Health Profile 2015. Republic of Indonesia Ministry of Health

[2]. Republic of Indonesia Ministry of Health. 2016. Handbook on Nutrition Status Monitoring and Nutrition Performance Indicators in 2015. Republic of Indonesia Ministry of Health.

[3]. WHO. 2010. Nutrition Landscape Information System (NLIS): Country Profile Indicators. World Health Organization, Geneva.

[4]. Tobler W. 1979. Philosophy in Geography, Reidel, Dordrecht. Cellular Geography, inS. Gale and G. Olsson (eds 379-386).

[5]. Anselin L. 1995. Local Indicators of Spatial Association. Study Paper 9331 Regional Study Institute West Virginia.

[6]. Ward M. D and Gleditsch, K. S. 2008. Spatial Regression Model. Sage Publication, Inc, United States.

[7]. Sarrias, Mauricio. 2016. Lecture 1: Introduction to Spatial Econometrics. Universidad Catolica del Norte, Chile.

[8]. Anselin L and Bera A.K. 1998. Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics. Dekker, New York.

[9]. Anselin L. 2003. Spatial Econometrics: A Companion to Theoretical Econometrics. Blackwell Publishing Ltd.