IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 7 Issue 1, January 2020
ISSN (Online) 2348 – 7968 | Impact Factor (2019) – 6.248
www.ijiset.com

Modeling and Mapping Incidence Rate of Tuberculosis in Bandung City by Means Spatial Error Model

Salsa Nurdini[1*], I Gede Nyoman Mindra Jaya[2]

[1] Department Statistics, Padjadjaran University, Indonesia

[2] Department Statistics, Padjadjaran University, Indonesia

*email: mindra@unpad.ac.id

Abstract
Tuberculosis is an infectious disease caused by Mycobacterium tuberculosis that most often affects the lungs. Several attempts can be made to reduce the incidence rate of tuberculosis in children by identifying the factors and mapping the incidence rate of tuberculosis in children in Bandung City. Regression analysis is a statistical tool that can be used to model factors affecting the incidence rate of tuberculosis in children. But in this research, the object of research is a spatial object. Therefore the aspects of the region need to be considered. This research is conducted to model and map the incidence rate of tuberculosis in children by considering the spatial aspects. The analysis that can be used to handle spatial autocorrelation is Spatial Error Model.

Keywords: *Tuberculosis, Spatial Autocorrelation, Spatial Error Model*

1. Introduction

Tuberculosis is an infectious disease caused by *Mycobacterium tuberculosis* that most often affect the lungs. Global Tuberculosis Report 2018 presents that tuberculosis is one of the 10 leading causes of death caused by germ infections worldwide [1]. In 2018 in Indonesia there were 511,873 cases of tuberculosis with 54,340 of them are the number of cases of tuberculosis in children [2]. The incidence rate of tuberculosis in children in Bandung City in 2018 was recorded at 143 cases of 100,000 population [3].
The importance of controlling tuberculosis in children in the tuberculosis disease control component is because children aged less than 15 years on average in each country are 20-50% of the total population. Tuberculosis transmission in children reflects the continuing transmission of tuberculosis in the population [4]. The ongoing transmission of tuberculosis in population is the reason for the need to be controlled by identifying the factors that influence the incidence rate of tuberculosis in children in Bandung City.
Tuberculosis transmission can be influenced by several factors such as population factors and environmental factors. Population factors include gender, age, nutritional status, immunization status and socioeconomic conditions. While environmental risk factors include occupancy density, house floor, ventilation, lighting, humidity, temperature and altitude [5]. In addition, aspects of the sub-district area in Bandung City have their respective backgrounds and characteristics that can be related to one another [6].
One way to control the incidence of tuberculosis in children is by identifying the factors that influence the rate of tuberculosis in children. This problem can be handled by regression analysis. However, the presence of spatial dependency needs to be considered because each location in Bandung has its own characteristics. Therefore, the method that can accommodate these spatial effects is the spatial error model.

2. Method

The data used in this study is the incidence rate of tuberculosis in children in 30 sub-districts of Bandung City in 2018, with the variables used include:
Y: incidence rate of tuberculosis in children (case rate/100,000 population)
X_1: population density (person / km^2)

X_2: percentage of child malnutrition status
X_3: incidence rate of tuberculosis in adult (case rate/100,000 population)

2.1. Multiple Linear Regression Analysis

Regression analysis is a method generally used to explain the relationship between response variables and predictor variables. The term regression was first introduced by Francis Galton in 1886 [5]. Regression models with several predictor variables can be formulated as in the following:

$$y_i = \beta_0 + \sum_{k=1}^{K} \beta_k X_{ik} + \varepsilon_i \qquad (1)$$

i = 1,…, N and k = 1,…, K
y_i : the i-th observation value of response variable
β_0 : intercept
$\beta_1, \ldots \beta_K$: slop coefficient
X_{ik} : the k-th independent variables for observation i
ε_i : error at location i, assumed ε_i ~ IID (0, σ^2)
These basic assumptions are known as classic assumptions which consist of:

1. $E(\varepsilon_i) = 0$, for i = 1,…, N
2. $Var\ (\varepsilon_i) = \sigma^2$, for i = 1,…, N
3. $Cov\ (\varepsilon_i, \varepsilon_j) = 0$, for $i \neq j$

Estimating the parameters of the linear regression model uses the Ordinary Least Square (OLS) method by minimizing the sum of squares error. Estimating parameters of the model is obtained from the following equation:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T y} \qquad (2)$$

2.2. Spatial Weight Matrices

In spatial analysis, dependency can be seen by forming a spatial weighting matrix. Spatial weight matrix (**W**) illustrates the relationship between regions and is obtained based on distance or neighbor information with the dimension $N \times N$.
In this study, a spatial weighting matrix will be formed with a distance matrix. The closer the area will be given a greater weight because the farther the distance will give greater variation. The distance between locations is determined by the distance of the two locations. This value is obtained from the inverse value of the distance defined as follows:

$$c_{ij} = \frac{1}{d_{ij}}\ ;\ d_{ij} \leq D \qquad (3)$$

The value of D is the specified distance limit, while d_{ij} is the distance between region i and j. Then the spatial weight matrix will be standardized by the following equation:

$$w_{ij} = \frac{c_{ij}}{\sum_{i=1}^{N} c_{ij}} \qquad (4)$$

2.3. Spatial Dependence

The data in this study consisted of several observation units. Each location that has the characteristics of each of these can be related with one another, or often called spatial dependence. This dependence shows the relationship between what happens in one location with another location. For $H_0 : I = 0$, this autocorrelation can be tested as follows:

$$Z(I) = \frac{I - E(I)}{\sigma(I)} \qquad (5)$$

where:

$$E(I) = \frac{-1}{(n-1)}$$

$$\sigma(I) = \sqrt{\frac{n^2 \sum_{ij} w_{ij}^{\ 2} + 3\left(\sum_{ij} w_{ij}\right)^2 - n \sum_i \left(\sum_j w_{ij}\right)^2}{(n^2-1)\left(\sum_{ij} w_{ij}\right)^2}}$$

2.4. Spatial Error Model

Spatial Error Model appears when the error term at a location correlates with the error term at the surrounding location or in other words there is a spatial correlation between errors. In the Spatial Error Model, the form of error at location *i* is a function of error at location *j* where *j* is a location located around location *i*. Spatial Error Model is stated as follows:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{u} \quad (6)$$

$$\mathbf{u} = \lambda\mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon} \quad (7)$$

with:

$\boldsymbol{y}$: observation value at location *i* of response variable
$\boldsymbol{\beta}$: regression coefficient parameters
$\boldsymbol{X}$: observation value at location *i* and of the *k*-th predictor variable
$\boldsymbol{u}$: residual vector
ε : residual vector that are normally distributed with a mean 0 and variance σ^2
λ : spatial coefficient parameter
$\mathbf{W}$: spatial weight matrix

To see whether the modeling with the Spatial Error Model is a good model, Lagrange multiplier testing will be performed. The coefficient determination (R^2) needs to be calculated because it is the most commonly used quantity to measure how far the ability of the model in explaining the variation of response variables in research.

3. Result and Discussion

3.1. Data description

The following is a map of the incidence rate of tuberculosis in children in Bandung City in 2018:

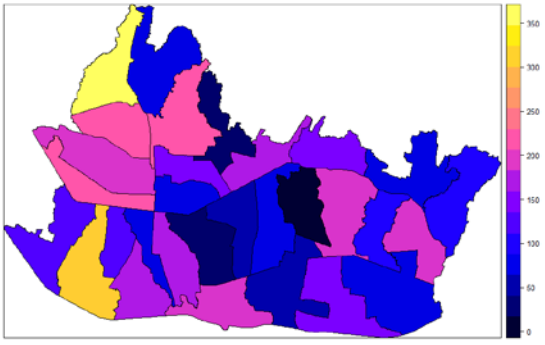

Figure 1 Map of Incidence Rate of Tuberculosis Among Children in Bandung City in 2018

Based on Figure 1 it can be seen that the locations that have the darkest color has a low incidence rate are Antapani District, Cibeunying Kaler District and Lengkong District. While locations that have the brightest color has a high incidence rate are Sukasari District and Babakan Ciparay District.

3.2. Multiple Linear Regression

The following are the results of modeling using the Multiple Linear Regression method:

Table 1 Estimate Parameters of Multiple Linear Regression Model

Parameter	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-7.983	39.28	-0.203	0.8405
β_1	-0.00075	0.1612	-0.466	0.6454
β_2	-0.1033	44.44	-0.233	0.8180
β_3	1.778	0.4688	3.793	0.0008

With the result coefficient determination (R^2) of 89.99%, the regression model obtained is as follows:

$$\hat{y} = -7.983 - 0.00075X_1 - 0.1033X_2 + 1.778X_3$$

In this section, modeling the incidence rate of tuberculosis in children in Bandung City using the Multiple Linear Regression method. The model assumption test is first performed to answer whether a regression model is valid or not is used as an explanation for the influence between predictor variables on the response variable. Testing of residuals using Kolmogorov-Smirnov concluded that residuals were normally distributed, testing using Durbin Watson concluded that there were no autocorrelations, testing spatial heterogeneity with the Breusch Pagan and based on the VIF value obtained, each variable had a VIF value of less than 10, which means that there was no multicollinearity problem.

Testing using Durbin Watson gives a p-value of 0.0074 while α=0.05 so that from the test results it can be concluded that there is autocorrelation. Based on VIF values obtained, each variable at each location has VIF value of less than 10, which means there is no local multicollinearity problem. Testing using Breusch Pagan concluded that there is not spatial heterogeneity and the residuals are normally distributed. Therefore, analysis using Spatial Error Model is appropriate.

3.3. Spatial Error Model Analysis

Before testing spatial autocorrelation with the value of Moran's I, first form a spatial weight matrix with a distance matrix measuring 30 x 30. Based on the calculation of Moran's I obtained a p-value of 0.0065 which is smaller than α=0.05, it is concluded that there is a spatial autocorrelation.

After that the Lagrange Multiplier test is performed to test the effect of spatial dependence. LM test results are as follows:

Table 2 Lagrange Multiplier Test

Model	LM	p-value
Lagrange Multiplier (Lag)	0.26842	0.6004
Lagrange Multiplier (Error)	0.66005	0.00415

Based on Table 2 it can be concluded that the modeling will be done with a spatial error model. At this stage the parameters are estimated and tested. The estimation results and parameter testing for the Spatial Error Model are presented in Table 3 as follows:

Table 3 Estimate Parameters of Spatial Error Model

Parameter	Estimate	Std. Error	t value	Pr(>|t|)
$\hat{\lambda}$	0.24249	32.4537	0.102	0.0188
$\hat{\beta}_1$	-0.00103	0.00149	-0.6899	0.4903
$\hat{\beta}_2$	-2.177539	41.5068	-0.0525	0.9582
$\hat{\beta}_3$	1.7464	0.43176	4.0449	0.00052

Based on Table 3 it can be seen that the variables that have a positive influence is incidence rate of tuberculosis in adult, while the other variables have negative influences. The incidence rate of tuberculosis in adult variable in adults (X_3) has a significant result so these factors affect the incidence rate of tuberculosis in children in Bandung City.

In addition, the spatial coefficient parameter (λ) also has significant results, so that the incidence rate of tuberculosis of children in a location is not only influenced by the number of tuberculosis incidence rates in adults in that location but is also influenced by the error of other locations that are nearby and have the same characteristics. The Spatial Error Model formed is:

$$\hat{y} = -0.00103X_1 - 2.177539X_2 + 1.7464X_3 + u$$

$$u = 0.24249 \sum_{i=1}^{N} w_{ij}\,\hat{u} + \varepsilon$$

After the model above is formed, the next step is to test the assumptions in the Spatial Error Model. Based on the assumption test results obtained that all assumptions are fulfilled, identical, independent, and normal distribution.

From the result model, 90.98% the predictor variables are able to explain the incidence rate of tuberculosis in children in Bandung City and the remaining are explained by other variables outside the model.

IV. Conclusion

Based on the results and discussion, it can be concluded that modeling using Spatial Error Model gives better results and incidence rate of tuberculosis of children in a location is not only influenced by the number of tuberculosis incidence rates in adults in that location but is also influenced by the error of other locations that are nearby and have the same

characteristics. This is also supported by the R^2 value of the Spatial Error Model (90.98%) which is greater than multiple linear regression (89.99%).

Acknowledgments

We thank to the Rector Universitas Padjadjaran for funding this research through the ALG program of Prof. Dr. Budi Nurani Ruchjana, M.S. (ALG 2019 No. 3330/UN6.D/LT/2019)

Reference

[1] World Health Organization. (2018). *Global Tuberculosis Report 2018.* Geneva: World Health Organization.
[2] Kementrian Kesehatan Republik Indonesia. (2019). *Profil Kesehatan Indonesia 2018.* Jakarta: Kementrian Kesehatan Republik Indonesia.
[3] Dinas Kesehatan Kota Bandung. (2017). *Profil Kesehatan Kota Bandung 2017.* Bandung: Dinas Kesehatan Kota Bandung.
[4] Kementrian Kesehatan Republik Indonesia. (2016). *Tuberkulosis: Temukan Obati Sampai Sembuh.* Jakarta: Kementrian Kesehatan Republik Indonesia.
[5] Ruswanto, B. (2010). *Analisis Spasial Sebaran Kasus Tuberkulosis Paru Ditinjau dari Faktor Lingkungan*
[6] *Dalam dan Luar Rumah di Kabupaten Pekalongan.* Semarang: Universitas Diponegoro.
[7] Anselin, L. (1988). *Spatial Econometrics: Methods and Models.* Santa Barbara: Departments of Geography and Economics University of California.
[8] Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. West Sussex: Wiley.