CAUCHY –Jurnal Matematika Murni dan Aplikasi
Volume 7(2) (2022), Pages 240-248
p-ISSN: 2086-0382; e-ISSN: 2477-3344

# Spatial Autoregressive to Model Tuberculosis Cases in Central Java Province in 2019

**Hasrat Ifolala Zebua[1], I Gede Nyoman Mindra Jaya[2*]**

[1] Post-graduate program in Applied Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Indonesia
[2] Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Indonesia

*Corresponding Author
Email: mindra@unpad.ac.id*,
hasrat20001@mail.unpad.ac.id

**ABSTRACT**

Tuberculosis is an infectious disease caused by the bacterium Mycobacterium tuberculosis. Central Java is one of the three provinces with the highest tuberculosis cases in Indonesia. Some of the risk factors used in this research are the spatial lag of the number of tuberculosis cases representing the agent component, the morbidity rate representing the host component, population density, proper sanitation, and proper drinking water which represent environmental components. This study aims to model the tuberculosis cases in Central Java Province using the Spatial Autoregressive (SAR) model. The SAR model is a regression model where the response variable has a spatial correlation. The estimation method usually used in SAR model is maximum likelihood. The Moran's I on the number of tuberculosis cases in Central Java shows a positive spatial autocorrelation. The model was chosen based on the LM test and AIC. The best model is the SAR model. The results show that the greater the number of tuberculosis cases is influenced by the number of tuberculosis cases in the neighbouring areas. Proper sanitation has a negative effect, on the contrary, the dense population has a positive effect on the number of tuberculosis cases in the province of Central Java.

**Keywords**: Maximum Likelihood; SAR; Spatial; Tuberculosis

## INTRODUCTION

Tuberculosis is an infectious disease caused by infection with the bacterium Mycobacterium tuberculosis [1]. Tuberculosis can be transmitted from human to human through the splash of the saliva of tuberculosis sufferers which spreads into the air when coughing or sneezing. Single cough or sneeze can produce up to 3000 splashes of saliva [2]. An infection occurs when other people breathe air containing these droplets. This disease usually affects the lungs, but can also affect other parts of the body. According to WHO [3], in 2019 around 10 million people were infected with tuberculosis and 11,4 million of them died.

In 2016, 45 percent of the estimated tuberculosis cases were in Southeast Asia

Submitted: September 29, 2021 Reviewed: December 09, 2021 Accepted: December 14, 2021
DOI: http://dx.doi.org/10.18860/ca.v7i1.13451

and Indonesia is one of them. Indonesia is the third country with the highest tuberculosis cases in the world after China and India with an estimated 845 thousand cases. During tuberculosis endemic in Indonesia, the spread of tuberculosis is still very high. Central Java is one of the three provinces with the highest tuberculosis cases in Indonesia, with 73,171 cases in 2019 [4].

Tuberculosis prevention and control efforts have been carried out such as the Bacille Calmette Guerin (BCG) vaccine in infants [5], increasing the number of case finding and treatment success in healthcare facilities. The biggest challenge in controlling tuberculosis is that there are many missing cases (unreported) which will further increase the transmission process due to patient ignorance. From an epidemiological perspective, the incidence of tuberculosis is an interaction between three components, namely the agent, host, and environment [1]. From the agent's side, intensive interactions between patients and other people can facilitate the transmission. The duration of contact time or the intensity of contact with people with tuberculosis can cause a person to be more easily exposed [6]. The host side (i.e., a person's susceptibility to tuberculosis) is strongly influenced by his body's resistance. As stated by Pangaribuan et al [7] the factors that are related from the host side are age, gender, race, socioeconomic, living habits, marital status, occupation, heredity, nutrition, and immunity. In terms of the environment, Arinil, et al [8] stated that environmental factors such as the physical environment of the house (occupancy density, ventilation, sanitation) and weather climate (temperature, humidity) were closely related to tuberculosis.

Prevention of tuberculosis transmission certainly requires control of these three components. Identifying area with a high risk of tuberculosis transmission is important to know to see the inter-regional linkages. Analytical tool for spatial data to model the number of cases of tuberculosis that occur is needed. In spatial epidemiology, the use of maps as a visualization method is needed to see the distribution of disease by geographic area [9]. Several risk factors used in this research are the spatial lag of the number of tuberculosis cases representing the agent component, the morbidity rate representing the host component, population density, proper sanitation, and proper drinking water representing the environmental component. One of the spatial models that can be used is the Spatial Autoregressive (SAR) model. The SAR model is a regression model with a spatial correlation in the response variable. Using cross-section data, this model combines a linear regression model with a spatial lag of the response variable [10]. The use of the SAR model in infectious diseases, especially tuberculosis in Central Java, is due to agent factors that have high mobility from district to other districts, besides that there are other factors, namely host and environment that can be used as covariates. The estimation method usually used in the SAR model is maximum likelihood. Therefore, the purpose of this study was to model the SAR of the number of tuberculosis cases in Central Java Province using the maximum likelihood approach. In this case, we will use a standardized weighting matrix using the queen contiguity.

## METHODS

### Data and Variables

This study used data from the publication of the Health Profile of the Central Java Province [4] and the Central Bureau of Statistics of the Central Java Province [11]. The units of analysis are 35 districts/cities in Central Java. The variables used in this study are shown in Table 1:

**Table 1.** Variables and Data Sources

| Notation | Variable | Source |
|---|---|---|
| $y$ | Number of confirmed cases of tuberculosis | Health Profile of the Central Java Province |
| $x_1$ | Proper sanitation | |
| $x_2$ | Eligible drinking water facilities | |
| $x_3$ | Population Density | Central Bureau of Statistics of the Central Java Province |
| $x_4$ | Morbidity Rate | |

**Moran's I**

Moran's I is the value of the test statistic used to determine whether there is spatial autocorrelation or spatial dependence in the data. Moran's I has global and local measures. Moran's I values range from -1 and 1. The global measure is used to measure the overall autocorrelation and the local is used to identify the autocorrelation on each unit. Global Moran's I can be seen in the following formula [12]:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (1)$$

with,

n : number of spatial units

$\bar{y}$ : mean of n locations

$y_i$ : observation variable at location *i*

$y_j$ : observation variable at location *j*

$w_{ij}$ : elements of the spatial weight matrix W

and the local Moran's I can be seen in the following formula:

$$I_i = \frac{(y_i - \bar{y})}{\sum_{k=1}^{n}(y_k - \bar{y})^2/n} \sum_{j=1}^{n}(y_j - \bar{y}) \quad (2)$$

The null hypothesis for autocorrelation is $I = E(I)$ no spatial dependence. The formula of test statistics can be written as follows:

$$Z(I) = \frac{I - E(I)}{\sqrt{VAR(I)}} \sim N(0,1) \quad (3)$$

with,

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

$$VAR(I) = \frac{n^2 S_1 - nS_2 + 3S_0^2}{(n^2-1)S_0^2} - [E(I)]^2$$

$S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$ ; $S_1 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(w_{ij} + w_{ji})^2$; $S_2 = \sum_{i=1}^{n}(\sum_{j}^{n} w_{ij} + \sum_{j}^{n} w_{ji})^2$.

**Lagrange Multiplier (LM) Test**

The Lagrange Multiplier (LM) test is used to determine whether there is a spatial dependence are not. There are two types of LM tests that have been developed, namely the spatial dependence of the dependent variable and the spatial error dependence. LM test statistics on the spatial dependence ($LM_{LAG}$) of the dependent variable are as follows [13]:

$$LM_{LAG} = \frac{[(e^T W_A y)/(e^T e/n)]^2}{\left[(W_A X\hat{\beta})^2 M(W_A X\hat{\beta})/(e^T e/n)\right] + [tr(W_A^T W_A + W_A^2)]} \sim \chi^2_{(1-\alpha);df=1} \quad (4)$$

If the test statistic value is greater than the chi-square value (reject $H_0$), then the model

made is the Spatial Autoregressive (SAR) model. LM test statistics for the dependence of spatial error ($LM_{ERR}$) can be seen in the following formula:

$$LM_{ERR} = \frac{[(e^T W_A e)/(e^T e/n)]^2}{tr(W_A^T W_A + W_A^2)} \sim \chi^2_{(1-\alpha);df=1} \quad (5)$$

If the test statistic value is greater than the chi-square value (reject $H_0$), then the model made is the Spatial Error Model (SEM). Meanwhile, if the $LM_{LAG}$ and $LM_{ERR}$ values are both significant, the best model can be chosen by comparing the Akaike Information Criterion (AIC). Model with the smaller AIC value is the best model.

**Spatial Autoregressive (SAR) Model**

The SAR model is a combination of a linear regression model with a spatial lag of the response variable using cross-section data. In general, the SAR model can be written as follows [14]:

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}; \qquad \boldsymbol{\varepsilon} \sim \text{MVN}(0, \sigma_\varepsilon^2 \text{I}_\text{n}) \quad (6)$$

with:

**y** : continuous response variable
ρ : autoregressive coefficient
**W** : spatial weight matrix
**β** : intercept and regression coefficient
**X** : predictor variable
**ε** : error

This model assumes that the autoregressive process is only found in the response variable.

Maximum Likelihood is one of the most commonly used estimators because it can provide the Best Linear Unbiased Estimation (BLUE) and overcome endogeneity in the SAR model. The estimated parameters using the Maximum Likelihood method are as follows:

$$\widehat{\boldsymbol{\beta}}_{ML} = \underbrace{(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T y}}_{\widehat{\boldsymbol{\beta}}_{OLS}} - \rho \underbrace{(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T W_\rho y}}_{\widehat{\boldsymbol{\beta}}_{L}} = \widehat{\boldsymbol{\beta}}_{OLS} - \rho\widehat{\boldsymbol{\beta}}_{L} \quad (7)$$

where $\widehat{\boldsymbol{\beta}}_L$ is an estimator of regression parameters that depends on the spatial autocorrelation ρ and the weight matrix (**W**). However, this form cannot be solved directly because the value of ρ is unknown. To be able to estimate the regression parameters, it can be done using a concentrated log-likelihood function ($L_c$) which is a function of the MLE residual which is defined as follows:

$$\ln L_c(\rho) = C - \frac{n}{2}\ln\left[\frac{1}{n}(\boldsymbol{e_0} - \rho\boldsymbol{e_L})^T(\boldsymbol{e_0} - \rho\boldsymbol{e_L})\right] + \ln|\boldsymbol{I} - \rho\boldsymbol{W}_\rho| \quad (8)$$

The formula cannot be solved analytically so a numerical method is needed to find the estimated ρ parameter of the equation.

## RESULTS AND DISCUSSION

The case notification rate (CNR) of tuberculosis in Central Java Province is 211, which means that there are 211 cases of tuberculosis being treated and reported among 100,000 residents in Central Java. Judging from the number of cases of tuberculosis in Central Java there were as many as 73,171 cases during 2019. Tuberculosis cases are a disease that is spread throughout the district in Central Java Province. The lowest number of cases was in Karanganyar Regency with 514 cases and Cilacap Regency with the highest number of 4,703 cases. However, when viewed from the district/city CNR,

the highest CNR is Tegal City at 832.5 per 100,000 population and the lowest CNR is Temanggung Regency at 45.72 per 100,000 [4]. The map of the distribution of tuberculosis cases in Central Java can be seen in Figure 1.

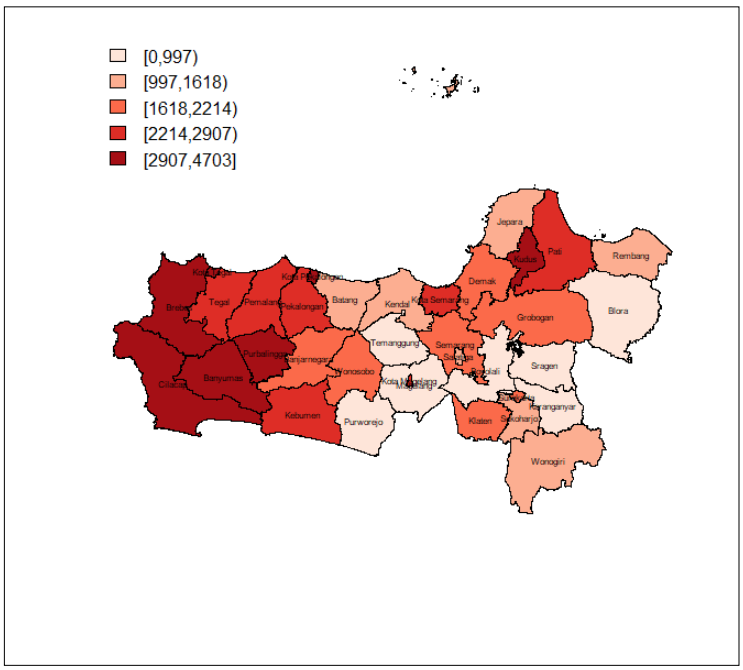


**Figure 1.** Map of tuberculosis case quantile in Central Java Province in 2019

Figure 1 shows that areas with a large number of cases (in solid red) tend to be close to areas with a large number of cases. Areas with a small number of cases (in faded red) also tend to be adjacent to areas with a small number of cases. This indicates that there is a spatial dependence between regions. Before conducting SAR modeling, it is necessary to test the classical assumptions of multiple linear regression and Moran's I tests.

**Table 2.** The statistic test result of classic assumptions and Moran's I

| **Statistic test** | **p-value** |
|---|---|
| Normality (Shapiro-Wilk) | 0.9821 |
| Nonautocorrelation (Durbin-Watson) | 0.4893 |
| Nonmulticolinierity (VIF) | 1.138817(X1); 1.049796(X2); 1.065710(X3); 1.074787(X4) |
| Homoscedasticity (Breusch-pagan) | 0.9749 |
| Moran's I | 0.000 (I= 0.4991) |

Table 2 shows that all assumptions in multiple linear regression have been fulfilled. The results of the Moran's I value using a spatial weight matrix based on queen contiguity on the number of tuberculosis cases in the province of Central Java is 0.499 with a p-value <0.05 (reject $H_0$) which means there is a positive spatial autocorrelation. High tuberculosis cases areas will be surrounded by high areas as well, and low tuberculosis cases areas will be surrounded by low areas as well.

For local Moran's I values will produce different values for each location. There are several significant areas in local Moran's I which is divided into two parts. First, the high-high areas include Cilacap, Banyumas, Brebes, Tegal, Puralingga, and Tegal City. This indicates that districts/cities in high-high areas have a high number of tuberculosis cases and the surrounding areas are also high, which is possible due to the transmission

of tuberculosis to the surrounding area. Second, in the low-low area, there are Magelang, Boyolali, Sragen, and Karanganyar. This shows that in the low-low area the number of tuberculosis cases is low and the surrounding area is also low. For other regions, the local Moran's I is not significant. Moran's I local results can be seen in Figure 2:

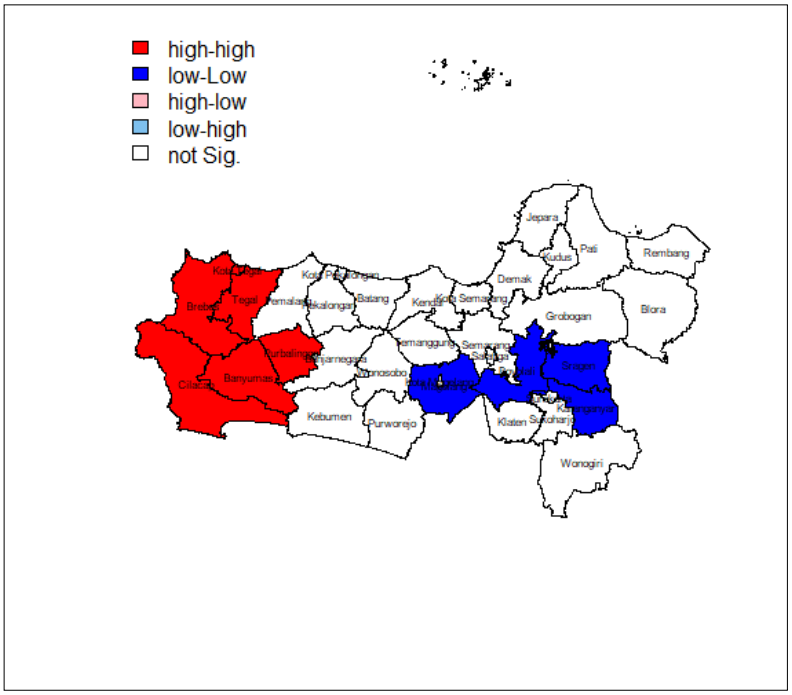


**Figure 2.** Local Moran's I of tuberculosis case in Central Java Province in 2019

The selection of the spatial model was carried out through the Lagrange Multiplier (LM) test as an initial identification. LM test is used to determine the spatial dependence more specifically whether the dependency on a response variable (lag), dependency on other variables that are not studied (error), or both (lag and error). The results of the LM test carried out can be seen in Table 3.

**Table 3.** LM test results

| Model | LM-test | Value | p-value | AIC |
|---|---|---|---|---|
| SAR | $LM_{LAG}$ | 12,635 | 0.000378 | 52.72579 |
| SEM | $LM_{ERR}$ | 8,973 | 0.002739 | 53.10059 |

From Table 2 it can be seen that the spatial dependence in lag and error is significant because the p-value is smaller than alpha (0.05). The SAR model will be applied in this study due to the AIC value is smaller than the SEM model. The SAR model is a spatial regression model that involves spatial lag in the response variable. The estimation results of the SAR model using the Maximum Likelihood method and using a spatial weight matrix based on Queen Contiguity can be seen in Table 4.

**Table 4.** SAR parameter estimation results

| | Estimate | Std. Error | z-value | p-value |
|---|---|---|---|---|
| (Intercept) | 3.383100 | 1.346100 | 2.51318 | 0.01196 |
| Proper sanitation ($x_1$) | -0.010300 | 0.005600 | -1.85760 | 0.06322 |
| Eligible drinking water facilities ($x_2$) | 0.006170 | 0.005800 | 1.06111 | 0.28864 |
| Population Density ($x_3$) | 0.000094 | 0.000029 | 3.20521 | 0.00134 |
| Morbidity Rate ($x_4$) | -0.003020 | 0.020255 | -0.14910 | 0.88147 |

| | Estimate | Std. Error | z-value | p-value |
|---|---|---|---|---|
| Rho: 0.58787, LR test value: 11.396, p-value: 0.000<br>Asymptotic standard error: 0.14246<br>z-value: 4.1267, p-value: 0.000<br>Wald statistic: 17.03, p-value: 0.000<br>AIC: 52.726 | | | | |

The spatial lag variable (rho) has a positive and significant coefficient in influencing the number of tuberculosis cases in Central Java Province. This means that the greater the number of tuberculosis cases is influenced by a large number of tuberculosis cases in the surrounding area. This is in accordance with the research of Mindra, et al [15] in the city of Bandung. The variable of proper sanitation has a negative regression coefficient value and significantly influences the number of tuberculosis cases in Central Java Province. This means that the more families that have access to proper sanitation (healthy latrines), the number of tuberculosis cases will decrease with the assumption that the other variables are constant. The population density variable has a positive and significant regression coefficient value in influencing the number of tuberculosis cases in Central Java Province. This means that the denser the population of an area, the number of tuberculosis cases will increase with the assumption that the other variables are constant. Variables eligible drinking water facilities and morbidity rates have no significant effect on tuberculosis cases in Central Java Province.

In the SAR model, the covariate impact can be categorized in three types, namely direct impact, indirect impact, and total impact which can be seen in Table 5.

**Table 5.** Direct and indirect impact measure of SAR model

| Variable | Direct | Indirect | Total |
|---|---|---|---|
| Proper sanitation ($x_1$) | -0.0116 | -0.0135 | -0.0251 |
| Eligible drinking water facilities ($x_2$) | 0.0069 | 0.0080 | 0.0149 |
| Population Density ($x_3$) | 0.0001 | 0.0001 | 0.0002 |
| Morbidity Rate ($x_4$) | -0.0034 | -0.0039 | -0.0073 |

The direct impact is an impact that occurs locally in an area as a result of changes in predictor variables. The indirect impact is a spillover effect, which is the impact that occurs when the predictor variable is in the surrounding area. The total impact is a change that occurs in an area as a result of changes in the area and its surroundings.

To find out whether the obtained SAR model is good, it is necessary to carry out a diagnostic check, including assumptions of normality, non-autocorrelation, and homogeneity. The results of the diagnostic check performed in Table 6 show that these assumptions have been fulfilled.

**Table 6.** Diagnostic check of SAR Model

| Statistic test | p-value |
|---|---|
| Normality (Shapiro-Wilk) | 0.2812 |
| LM test for residual autocorrelation | 0.4840 |
| Homoscedasticity (Breusch-pagan) | 0.7637 |

Based on Table 6, it can be seen that the p-value for the assumption of normality using the Shapiro-Wilk test is 0.2821 which is greater than alpha (0.05), which means the residuals are normally distributed. In the autocorrelation test, it was found that the p-value was also greater than alpha (0.05), which means that the residuals meet the non-autocorrelation assumption. Likewise, in the homoscedasticity test using the

Breusch-Pagan test, it was found that the p-value was also greater than alpha (0.05) so that the assumption of homogeneity was also fulfilled.

## CONCLUSIONS

Tuberculosis cases in Central Java Province showed a positive spatial autocorrelation. It supports the hypothesis that tuberculosis cases are spatially dependent and that a spatial econometrics model should be considered. The best spatial econometrics model was chosen based on the LM test and the AIC. The best model is the Spatial Autoregressive (SAR) model. The estimation results of the SAR show that the number of tuberculosis cases is influenced by a large number of tuberculosis cases in the neighbouring areas. Proper sanitation (Ownership of healthy latrines) has a negative effect on the number of tuberculosis cases, on the other hand, the dense population has a positive influence on the number of tuberculosis cases in the province of Central Java.

## REFERENCES

[1] K. RI, "Pusat Data dan Informasi Kementrian Kesehatan RI Tuberkolosis," Kementrian Kesehatan Republik Indonesia, Jakarta, 2018.

[2] K. RI, "Pedoman Nasional Pengendalian Tuberkolosis," Jakarta, 2014.

[3] W. H. Organization, "Global Tuberculosis Report 2020," Geneva, 2020.

[4] D. K. P. J. Tengah, "Profil Kesehatan Provinsi Jawa Tengah Tahun 2019," Semarang, 2020.

[5] M. Dara, C. D. Acosta, V. Rusovich, J. P. Zellweger, R. Centis and G. B. Migliori, "Bacille Calmette–Guérin vaccination: the current situation in Europe," *European Respiratory Journal,* vol. 43, no. 1, pp. 24-35, 2014.

[6] T. D. Kristini and R. Hamidah, "Potensi Penularan Tuberculosis Paru pada Anggota Keluarga Penderita," *Jurnal Kesehatan Lingkungan Indonesia,* vol. 15, no. 1, 2020.

[7] L. Pangaribuan, Kristina, D. Pangaribuan, T. Tejayanti and D. B. Lolong, "Faktor-Faktor yang Mempengaruhi Kejadian Tuberkulosis pada Umur 15 Tahun ke Atas di Indonesia (Analisis Data Survei Prevalensi Tuberkulosis (SPTB) di Indonesia 2013-2014)," vol. 23, no. 1, 2020.

[8] A. Haq, U. F. Achmadi and D. Susanna, "Analisis Spasial (Topografi) Tuberkulosis Paru di Kota Pariaman, Bukittinggi, dan Dumai Tahun 2010-2016," *Jurnal Ekologi Kesehatan,* vol. 18, no. 3, p. 149 – 158, 2020.

[9] M. Souris, Epidemiology and Geography Principles, Methods and Tools of Spatial Analysis, Great Britain and the United States: ISTE Ltd and John Wiley & Sons, Inc, 2019.

[10] J. LeSage and R. K. Pace, Introduction to Spatial Econometrics, London: Chapman and Hall/CRC, 2009.

[11] B. P. S. P. J. Tengah, "Provinsi Jawa Tengah Dalam Angka," Badan Pusat Statistik, Semarang, 2020.

[12] G. Grekousis, Spatial Analysis Methods and Practice, Cambridge: Cambridge University Press, 2020.

[13] L. Anselin, "Lagrange Multiplier Test Diagnostics for Spatial Dependence and Spatial

Heterogeneity," *Geographical Analysis,,* vol. 20, pp. 1-17, 2010.

[14] I. G. N. M. Jaya and Y. Andriyana, Analisis Data Spasial Perspektif Bayesian, Sumedang: Alqaprint Jatinangor, 2020.

[15] I. G. N. Jaya and e. al, "Metode Bayesian dalam Penaksiran Model Spasial Autoregressive (SAR) (Studi Kasus Pemodelan Penyakit TB Paru di Kota Bandung)," *Jurnal Euclid,* vol. 4, no. 2, 2017.