Proceeding of ICSA 2019, p: 9-17
ISBN 978-979-19256-3-1 (PDF)

Identification of Factors Affecting Tuberculosis in West Java using Spatial Modeling

Yusma Yanti[1*], Septian Rahardiantoro[2]

[1]*Computer Science Department, Pakuan University,Bogor, Indonesia*
[2]*Department of Statistics, IPB University,Bogor, Indonesia*

[*)]Corresponding author: ***yusmayanti.fn@gmail.com***

Abstract

Tuberculosis (TB) is an infectious disease caused by the bacillus Mycobacterium tuberculosis. In 2017 WHO records there are 1.7 billion TB sufferers in the world. Whereas in the same year TB sufferers in Indonesia reached 421 thousand cases and 10 thousand of them were in the province of West Java. In this study, the factors that suspected to influence TB include poverty, population density and malnutrition were analyzed by looking at the spatial aspects. In addition to these factors, smoking and consuming alcoholic beverages can also trigger TB. The method used was Spatial Autoregressive Model (SARM), Spatial Error Model (SEM), and Generalized Spatial Model (GSM), then the best model is chosen based on the best criteria of lagrange multiplayer test. The result indicated that SEM performed better than others, with the following significant variables were malnutrition and unemployment factor.

***Keywords*:** Generalized Spatial Model, Spatial Autoregressive Model, Spatial Error Model, Tuberculosis

Introduction

Tuberculosis (TB) is an infectious disease caused by Mycobacterium tuberculosis bacillus (Zumla *et al*, 2013). This disease included in the dangerous disease because in 2017 the WHO recorded 1.7 billion TB sufferers over the world. This study focused on TB cases that occurred in West Java Province in 2017, because at that time West Java Province contributed a lot of TB cases in Indonesia. The number of TB patients in West Java can be said to be high, so it is very important to know the trigger factors of the cases. Several factors that suspected can influence TB were: poor and niserable, population density and malnutrition. In addition to these factors, smoking and consuming liquor can also include the factor that can effect TB cases. Moreover, the last factor that probably also effect the TB cases was very close to the population who do not work or better known as unemployment factor.

It is to be more interesting cases to see the another point of view in the modeling of TB cases. This study applied spatial effect in the modeling process, to see influenced factor spatially. It is suspected that the level of TB sufferers in West Java is also affected by directly neighboring location, because the spread of TB also through the air (Zumla *et al*, 2013). When a TB patient sneezes or speaks and transmits in the air, and is inhaled, the individual can be infected. But TB will not be spread through household appliances. TB is also spread with intense and prolonged association with sufferers. The amount of influence given by regions with high levels of cases is very diverse. This can be analyzed using a spatial approach.

In this paper, the analysis taken only deal with spatial dependences. The method applied were Spatial Autoregressive Model (SARM), Spatial Error Model (SEM), and Generalized Spatial Model (GSM). The best model was chosen based on the criteria of Lagrange multiplayer test. Therefore, the limitation of this paper is not consider the heterogeneity as critical aspect concerned. This study applied R software to analyze all of procedures needed (Anselin, 2007)

Materials

In this study secondary data were obtained from Badan Pusat Statistik in the publication of Jawa Barat Province in Figures (BPS, 2018). There are 27 district or city in West Java Province that to be the observation in this study. The response variable used in this study was number of TB cases in each district or city in West Java 2017. We proposed to use several predictor variables that suspected to influenced of TB cases besides the neighborhood location, such as: poor and miserable, malnutrition, unemployment, and population density. Table 1 shows a list of predictor variables completed with scale of these variables.

Table 1. List of predictor variables and its scale

No	**Predictor Variable Name**	**Scale**
1	Malnutrition (A2)	Person
2	Poor & Miserable (B1)	Person
3	Unemployment (B3)	Person
4	Population Density (D1)	Density

Methods

The following list, contains the procedure used to analyze the data in this study.

1. Data exploration on the response variable, number of TB cases in West Java, through mapping many cases for each district or city. In this step, we simplified the number of TB cases to be four categories based on amount of the value, Low: TB cases < 1000, Medium: 1000 ≤ TB cases < 2000, High: 2000 ≤ TB cases < 3000, Very High: TB cases ≥ 3000.
2. Identification of neighborhood criteria. In this study, we used the neighborhood criteria based on the distance coverage (Sander *et all*, 2010). Euclidian distance was chosen to apply in tis research.
3. Calculation the Moran Index that indicates the value of spatial autocorrelation as follows.

$$I = \frac{N}{\sum_{ij} w_{ij}} \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2}$$

where N is the number of spatial units indexed by i and j, y is the variable of interest (response variable), $\bar{y}$ is mean of y, w_{ij} is i-th row and j-th column element matrix of spatial weights.

4. Modeling process using regression analysis (Jerrett *et all*, 2010). In this step also evaluated normality of residuals, homogeneity of variance, also residuals randomness.
5. Performing the Lagrange Multiplayer test for identifying preferable model of SARM, SEM, or GSM (Anselin, 1988). The following formula describes the Lagrange Multiplayer test in the detecting spatial term.

   SARM: $H_0\colon \rho = 0;\ H_1\colon \rho \neq 0$

   $LM_{sar} = \frac{\left(\boldsymbol{\varepsilon}'\boldsymbol{Wy}/(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/N)\right)^2}{((\boldsymbol{WX\beta})'(\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')(\boldsymbol{WX\beta})/\sigma^2)+tr(\boldsymbol{W}^2+\boldsymbol{W}'\boldsymbol{W})}$; reject H_0 if $LM_{sar} > \chi^2_{(1)}$

   SEM: $H_0\colon \lambda = 0;\ H_1\colon \lambda \neq 0$

   $LM_{sem} = \frac{\left(\boldsymbol{\varepsilon}'\boldsymbol{W\varepsilon}/(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/N)\right)^2}{tr(\boldsymbol{W}^2+\boldsymbol{W}'\boldsymbol{W})}$; reject H_0 if $LM_{sem} > \chi^2_{(1)}$

   GSM: H_0: ρ and or $\lambda = 0$; $H_1\colon \rho$ and or $\lambda \neq 0$

   $LM_{gsm} = \frac{1}{E}\left[R_y^2 T - 2R_y R_\varepsilon T + (D+T)\right];$

   $R_y = \frac{1}{\hat{\sigma}^2}(\boldsymbol{\varepsilon}'\boldsymbol{Wy});\ R_y = \frac{1}{\hat{\sigma}^2}(\boldsymbol{\varepsilon}'\boldsymbol{W\varepsilon}); D = ((\boldsymbol{WX\beta})'(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')(\boldsymbol{WX\beta})/$
   $\sigma^2)$

   $T = tr(\boldsymbol{W}^2 + \boldsymbol{W}'\boldsymbol{W}); E = (D+T)T - T^2$

   reject H_0 if $LM_{gsm} > \chi^2_{(2)}$
6. Modeling SARM, SEM, and GSM based on the result of Lagrange Multiplier test. The modeling process follow the formulas below (Anselin, 1988):

   SARM: $\boldsymbol{y} = \rho\boldsymbol{Wy} + \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$

SEM: $\boldsymbol{y} = \boldsymbol{X\beta} + \lambda \boldsymbol{Wu} + \boldsymbol{\varepsilon}$

GSM: $\boldsymbol{y} = \rho \boldsymbol{Wy} + \boldsymbol{X\beta} + \lambda \boldsymbol{Wu} + \boldsymbol{\varepsilon}$

where $\boldsymbol{y}$ denotes $N \times 1$ vector of response variable, $\boldsymbol{X}$ denotes $N \times p$ matrices of predictor variables, $\boldsymbol{\beta}$ denotes $p \times 1$ vector of model coefficient, $\boldsymbol{\varepsilon}$ denotes $N \times 1$ vector of error of model, $\boldsymbol{W}$ denotes $N \times N$ matrices of neighborhood, $\boldsymbol{u}$ denotes $N \times 1$ vector of random spatial effect, ρ and λ denote the coefficient of autoregressive and error term respectively.

7. Selection of the best method based on the minimum AIC value and interpretation of results.

Results and Discussion

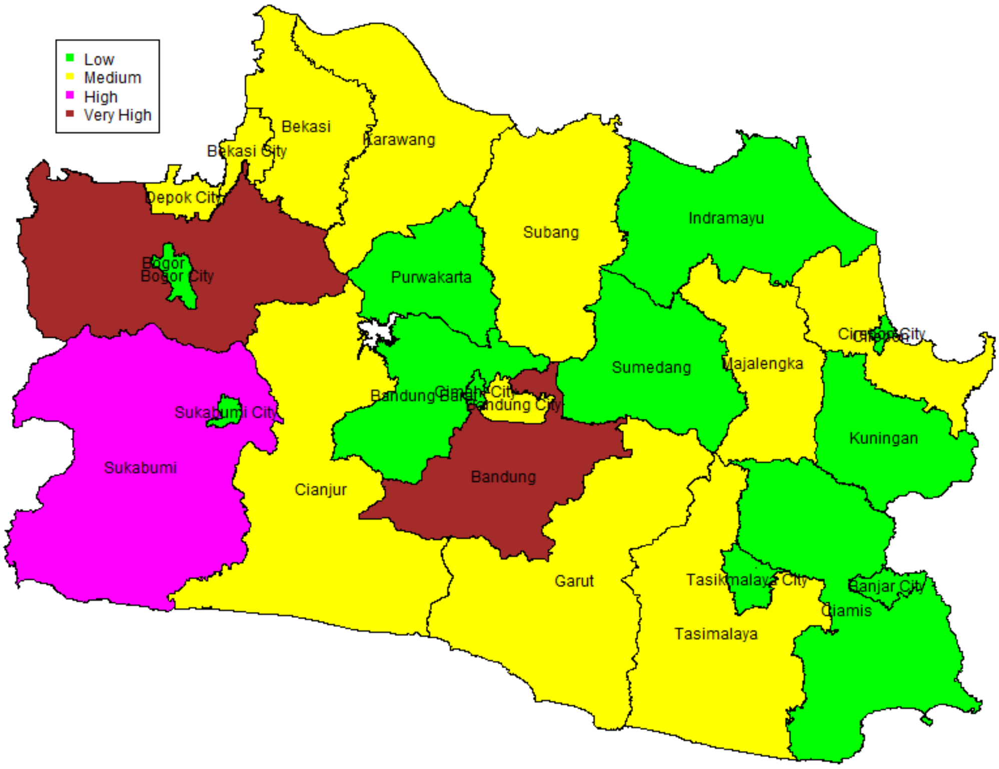

Figure 1. Spread of TB in West Java Province

Figure 1 presents an overview of the spread of TB disease in West Java for each Regency and City. Based on Figure 1, Bogor District and Bandung District have the highest amount of TB cases in West Java. Then, Sukabumi District is neighboring directly with Bogor followed with high amount of TB cases. The spread of TB is relatively more in the western region and more relative to the district. If observed, there are fewer cases in urban areas. There are something interesting from this, the districts that have high value of TB cases has opposite condition with its city, the Bogor City, Bandung City, and Sukabumi City have lower level of TB cases in West Java. It can be say that number of TB cases occurred in the district was higher of the value in the similar city.

To find out the level of closeness of the region with the area of TB spread, the Index of Moran (IM) was calculated. Neighboring determination method based on a distance of 1.5 units based on Euclidian distance. Regions that have distance ≤ 1.5 means that these regions (districs) to be neighbor, and vice versa. Then, the weighting $\boldsymbol{W}$ matrix can be obtained by using standardized matrix. The result of IM was 0.01600711, with p-value = 0.007274 > 0.05, therefore the spatial autocorrelation was significant in this data. Because of that, it is possible to continue the data analysis using spatial aspect modeling.

The next step was creating the initial model a regression analysis. Table 2 presents the summary of regression modeling result of this data. Based on this table, the result of regression model was suitable for this data, because it has $R^2 = 62.28\%$ and significant in F test. While, the normal assumption based on Anderson Darling test was not met, but other regression assumptions such as homogeneous and mutually random variety are fulfilled. Furthermore, this model has AIC value 426.525.

Table 2. R output of regression analysis summary and assumption testing

Regression analysis R summary output

```
Residuals:
    Min      1Q  Median      3Q     Max
-624.81 -332.68  -78.69   81.25 2040.45

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.326e+02  2.159e+02   2.004   0.0576 .
A2           5.753e-01  2.173e-01   2.647   0.0147 *
B1          -2.735e-04  4.230e-04  -0.647   0.5246
B3           6.369e-03  2.757e-03   2.310   0.0306 *
D1          -4.862e-03  2.514e-02  -0.193   0.8484
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1

Residual  standard  error:  578.2  on  22  degrees  of
freedom
Multiple R-squared:  0.6228,    Adjusted R-squared:
0.5543
F-statistic: 9.083 on 4 and 22 DF,  p-value: 0.0001724
```

Assumptions testing R output

Normality test

```
Anderson-Darling normality test

data:  err.regklasik
A = 1.5682, p-value = 0.0003757
```

Homogeneity Variance test

```
studentized Breusch-Pagan test

data:  reg.klasik
BP = 0.962, df = 4, p-value = 0.9155
```

Randomness residual test

```
Runs Test for Randomness

data:  err.regklasik
runs = 16, m = 14, n = 13, p-value =
0.5681
alternative hypothesis: true number of
runs is not equal the expected number
sample estimates:
median(x)
-78.68803
```

Then, Lagrange Multiplier test was performed to determine the spatial influence of each region. Table 3 shows the summary of Lagrange Multiplier test based on R output. The results obtained were not significant, but this study still continued the analysis to the next step, to perform the model with lower AIC value.

Table 3. Summary output of Lagrange Multiplier Test

SARM	SEM	GSM

`LMlag = 0.14703, df = 1,`
`p-value = 0.7014`
	`LMerr = 0.59379, df = 1,`
	`p-value = 0.441`
		`SARMA = 1.0042, df = 2,`
		`p-value = 0.6053`

Although the Lagrange Multiplier test was not significant in all spatial term schemes, we decided still to continue spatial modeling analysis to perform the model that has better AIC value compared with ordinary regression analysis. Table 4 resumes the result of spatial modeling using SARM, SEM, and GSM. The lowest AIC value was performed by SEM method, AIC = 425.81. Because of that, the chosen model was SEM with significant variables

(p-value < 0.05) were malnutrition and unemployed variable. Therefore, based on this result, malnutrition and unemployed variable were significant to effect the TB cases in West Java 2017 by consider the spatial term.

Table 4. Comparison of estimating coefficient and AIC values of each method

Output	Estimated Cofficient					AIC
SARM		Estimate	Std. Error	z value	Pr(>|z|)	428.29
	(Intercept)	8.1045e+02	8.0790e+02	1.0032	0.31579	
	A2	5.5821e-01	1.9694e-01	2.8345	0.00459	
	B1	-2.3955e-04	3.8354e-04	-0.6246	0.53226	
	B3	6.8360e-03	2.6566e-03	2.5733	0.01007	
	D1	-1.1898e-03	2.0548e-02	-0.0579	0.95383	
SEM		Estimate	Std. Error	z value	Pr(>|z|)	425.81
	(Intercept)	3.5755e+02	1.0651e+02	3.3571	0.0007878	
	A2	5.1533e-01	2.0164e-01	2.5557	0.0105975	
	B1	-1.7313e-04	3.5216e-04	-0.4916	0.6229862	
	B3	7.4759e-03	2.1114e-03	3.5407	0.0003991	
	D1	1.5496e-03	2.1395e-02	0.0724	0.9422620	
GSM		Estimate	Std. Error	z value	Pr(>|z|)	427.79
	(Intercept)	4.2327e+02	6.2864e+02	0.6733	0.500752	
	A2	5.1223e-01	2.0761e-01	2.4673	0.013613	
	B1	-1.6362e-04	3.6748e-04	-0.4453	0.656137	
	B3	7.6289e-03	2.5641e-03	2.9752	0.002928	
	D1	2.6613e-03	2.5324e-02	0.1051	0.916304	

Conclusions

Based on the analysis above, we performed the better model using spatial consideration based on AIC criteria. Although the Lagrange Multiplier test indicated there were no significant spatial parameter, while SEM model seemed promising to give better result than ordinary regression analysis. The significant variables were malnutrition and unemployment factor.

References

Anselin L. 1988. *Spatial Econometrics: Methods and Models*. Academic Publishers. Dordrecht.

Anselin L. 2007. *Spatial Regression Analysis in R A Workbook*. Center for Spatially Integrated Social Science.

BPS-Statistics of West Java Province. 2018. *Jawa Barat Province in Figure*. Adiatama, CV

Jerrett M, Gale S, Kontgis C. 2010. Spatial Modeling in Environmental and Public Health Research. *Int. J. Environ. Res. Public Health*. 7. pp 1302-1329

Sander HA, Ghosh D, Riper Rv, Manson SM. 2010. How do you measure distance in spatial models? An example using open-space valuation. *Environment and Planning B: Planning and Design*. 37. pp 874-894

Zumla A, Raviglione M, Hafner R, Reyn CFv. 2013. Current Concepts Tuberculosis. *The New England Journal of Medicine*. 368:8. pp 745-755.