MODELING OF THE NUMBER OF TUBERCULOSIS CASES IN INDONESIA

Aida Meimela
Statistics of North Sumatera Province
aida.mey@bps.go.id

Diterima: Juli 2020; Disetujui: Desember 2020

Abstract*. One of the health issues listed in the Sustainable Development Goals (SDGs) is to end the tuberculosis epidemic in 2030. Indonesia is the country with the third-highest number of tuberculosis cases in the world after India and China in 2018. Aims of this study to model the number of tuberculosis cases in each province in Indonesia, depending on the characteristics of each region. Geographically Weighted Lasso (GWL) is a method used to overcome the local multicollinearity that appears in the Geographically Weighted Regression (GWR) model. By using this method, each region will have a different regression model according to its respective characteristics. There is local multicollinearity (VIF> 10) in each explanatory variable used. Banten, West Java, South Kalimantan, East Kalimantan, East Nusa Tenggara and Papua Province are provinces where all research variables affect the number of tuberculosis cases. The variable that has the most significant effect on the number of tuberculosis cases in each region in Indonesia is the number of health centers. Therefore, to end the number of tuberculosis cases, the government should increase the number of health centers and improve the health service.*

Keywords: *dependency spatial, lasso, multicollinearity, spatial heterogeneity, tuberculosis.*

Abstraksi. *Salah satu isu kesehatan yang tercantum dalam Sustainable Development Goals (SDG's) adalah mengakhiri epidemi tuberkulosis (TBC) di tahun 2030. Indonesia merupakan negara dengan jumlah kasus TBC tertinggi ketiga di dunia setelah India dan China pada tahun 2018. Tujuan penelitian ini adalah melakukan pemodelan jumlah kasus tuberkulosis di Indonesia sesuai dengan karakteristik wilayah masing-masing. Geographically Weighted Lasso (GWL) merupakan metode yang digunakan untuk mengatasi multikonieritas lokal yang muncul pada model Geographically Weighted Regression (GWR). Adanya multikolineritas lokal (VIF >10) pada setiap variabel penjelas yang digunakan. Banten, Jawa Barat, Kalimantan Selatan, Kalimantan Timur, Nusa Tenggara Timur dan Papua adalah provinsi dimana seluruh variabel penelitian berpengaruh terhadap jumlah kasus tuberkulosis. Variabel yang paling banyak berpengaruh signifikan terhadap jumlah kasus tuberculosis di setiap wilayah di Indonesia adalah jumlah puskesmas. Oleh karena itu, untuk mengakhiri jumlah kasus tuberkulosis pemerintah sebaiknya menambah jumlah puskesmas dan meningkatkan pelayanan kesehatan.*

Kata Kunci: *ketergantungan spasial, keragaman spasial, lasso, multikolinieritas, tuberkulosis.*

BACKGROUND

One of the health issues listed in the Sustainable Development Goals (SDGs) is to end the tuberculosis epidemic in 2030 (Indonesia, 2018). This is because tuberculosis is one of the second-highest causes of death after HIV/ AIDS (Fogel, 2015). World Health Organization (WHO) publication in tuberculosis states that Indonesia is the country with the third-highest number of tuberculosis cases in the world after India and China in 2018 (WHO, 2019).

The data distribution of the number of tuberculosis cases in each province in Indonesia has different patterns depending on the characteristics of each region. In 2018, the provinces with the highest number of tuberculosis cases were West Java, Central Java and East Java. Those three provinces are closely located. This indicates the presence of regional dependencies on the number of tuberculosis cases (Firdaus, 2014).

The previous research was conducted through traditional modeling (linear regression, logistic regression and binomial negative regression) to explain the relationship between tuberculosis prevalence and several factors in general. However, this method is unable to explain the spatial heterogeneity (Sun *et al*., 2015). Therefore, research on the prevalence of tuberculosis taking into account regional factors needs to be done.

In addition, three models are best used if the relationship between predictor variables and response variables is not spatially dependent on the region or stationary (Fotheringham et al., 2002).

Geographically Weighted Lasso (GWL) is used to overcome the local multicollinearity that appears in the *Geographically Weighted Regression* (GWR) model. Geographically Weighted Regression is one method to overcome regional heterogeneity caused by different locations and conditions between regions. By using GWL, each region will have a different regression model according to its respective characteristics.

Previous studies conducted by Wheeler (2009) showed that the estimated error of the GWL model is smaller than the GWR model. Other studies also used the same method in modeling poverty in Java. The results showed that the GWL method is better than the GWR method on spatial data that contains multicollinearity (Setiyorini, 2017). According to those researched, the modeling of the number of tuberculosis cases in Indonesia in this research will be done following the characteristics of each region.

RESEARCH METHOD

The data used in this study are taken from publications published by the Indonesian Ministry of Health and Statistics Indonesia. There are seven explanatory variables in this study. They are the poor population, population density, percentage of households with a per capita floor area < 7,2 m^2 (Indonesia, 2019a, 2019b, 2019c), percentage of districts/ cities that have a clean and healthy behavior policy, the percentage of slum households, the proportion of the population aged > 10 years who smoke every day and the number of health centers (Indonesian Ministry of Health, 2018, 2019a, 2019b). The observation unit is 34 provinces in Indonesia. The response variable is the number of tuberculosis cases in Indonesia modeled by seven explanatory variables.

Dependence and Spatial Heterogeneity

The characteristics of spatial data are the presence of dependencies and spatial heterogeneity (Anselin, 1988). Spatial dependence means that there are similarities between observations that are closely located. Spatial dependencies are measured to see weather observations at one location affect observations at other nearby locations. Meanwhile, spatial heterogeneity is a characteristic difference between one region and another (Fotheringham et al., 2002). Spatial dependency was measured using the Moran I coefficient. Meanwhile, to see the presence or absence of spatial heterogeneity with the Breusch-Pagan test (Anselin, 1988).

Geographically Weighted Lasso (GWL)

Geographically Weighted Lasso (GWL) is a technique that uses the Lasso approach in the GWR model. The GWL model depends on the weight used. The weighting function used in this study is the *Fixed Exponential Kernel* which is written as follows (Fotheringham et al., 2002):

$$w_j\ (u_i, v_i) = \exp\left(\frac{d_{ij}}{h}\right) \qquad (1)$$

where $d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2}$ is the Euclidean distance of the location $((u_i, v_i))$ with the location (u_j, v_j) and h is the fixed or same bandwidth in all locations.

The selection of optimum bandwidth affects the accuracy of the parameter estimation results. One method that can be used is Cross-Validation which is written as follows:

$$CV\ (h) = \sum_{i=1}^{n}[y_i - \hat{y}_{\neq i}(h)]^2 \qquad (2)$$

where $\hat{y}_{\neq i}(h)$ is the estimated value for y_i with bandwidth h. The selection of optimum bandwidth obtained from the iteration process that produces the smallest CV (Fotheringham et al., 2002).

One way to detect the presence of local multicollinearity is to calculate the VIF value which is formulated as follows (Wheeler, 2007):

$$VIF_k\ (i) = \frac{1}{1 - R_k^2\ (i)} \qquad (3)$$

where $R_k^2\ (i)$ is the coefficient of determination when x_k is regressed on the other explanatory variables for each i-th location. VIF values > 10 indicate local multicollinearity. In addition to overcoming the existence of multicollinearity, GWL can also simultaneously select insignificant variables by shrinking the regression coefficient to zero. The solution to GWL is to complete the following formula (Wheeler, 2009):

$$\hat{\beta}^R = arg_\beta \min\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{k=1}^{p} x_{ik}\,\beta_k\right)^2 + \lambda \sum_{k=1}^{p}|\beta_k|\right\} \qquad (4)$$

With the lasso constraint $\sum_{k=1}^{p}|\beta_k| \leq s$ is equivalent to adding the penalty term $\lambda \sum_{k=1}^{p}|\beta_k|$ to the residual sum of a square; hence there is a direct correspondence between the parameters s and λ (Tibshirani, 1996). Efron et al found an algorithm that can solve lasso solutions called the LARS (Least Angle Regression) algorithm (Efron, 2004).

RESULTS AND DISCUSSION

The number of tuberculosis cases in Indonesia was 511.873 cases in 2018. The highest number of tuberculosis cases was West Java Province with 99.398 cases, Central Java with 67.063 cases and East Java with 56.445 cases. The three provinces are located on the island of Java and their locations are close. The following are descriptive statistics of the variables used in this study (Table 1).

From Table 1, the highest standard deviation is variable Y, which means that the number of tuberculosis cases in Indonesia varies greatly. While the lowest standard deviation is the proportion of the population aged > 10 years who smoke every day. This means the proportion of the population aged 10 years and over who currently smoke every day is not too varied in Indonesia.

Spatial dependency test results with the Moran Index obtained Moran Index value of 0,7124564 with a p-value of 6,645384 e-6. With a significance level of 5 percent, it can be concluded that there are spatial dependencies in the number of tuberculosis cases in Indonesia. Meanwhile, the Breusch Pagan test results showed a statistical test value of 19,203 with a p-value of 0,01389. So with a significance level of 5 percent, it can be concluded that there is spatial heterogeneity in the number of tuberculosis cases in Indonesia. From this test, global regression modeling on the number of tuberculosis cases in Indonesia cannot be done.

Table 1.
Minimum, Maximum and Standard Deviation of the Data

Variable	Min	Max	Mean	Standard Deviation
Number of tuberculosis cases	1.421	99.398	15.055	21.395,26
Percentage of the poor population	3,57	27,740	10,81	5,78
Population density	9	15.764	734,68	2.685,57
Percentage of districts/ cities that have a clean and healthy behavior policy	6,90	100	74,64	29,63
Percentage of slum households	1,13	40,01	7,06	6,87
The proportion of the population aged > 10 years who smoke every day	18,80	28,10	23,49	2,60
Percentage of households with a per capita floor area < 7,2 m2	2,58	34,74	10,62	6,29
Number of health centers	56	1.069	293,9	244,54

Source: Data processed, 2020

Table 2.
Local VIF Value

Variable	Min.	Max.	Mean	VIF > 10
Percentage of the poor population	1,96	2.528,50	126,33	23
Population density	1,18	3.853,74	311,11	14
Percentage of districts/ cities that have a clean and healthy behavior policy	1,50	767,33	61,79	14
Percentage of slum households	2,6	637.410	19.020,3	19
The proportion of the population aged > 10 years who smoke every day	1,91	43.986,62	1319,36	27
Percentage of households with a per capita floor area < 7,2 m2	3,4	557.215,7	17.173,1	30
Number of health centers	1,71	126.288,33	3.788,45	22

Source: Data processed, 2020

In Table 2 we can see that VIF > 10 occurs in all explanatory variables. One way to overcome the presence of local multicollinearity is to use the Geographically Weighted Lasso (GWL) method. Where in this modeling the optimum bandwidth obtained is 43. The estimated coefficient values are shown in Table 3.

Table 3.
Parameter Estimates of GWL

Parameter Estimates	Minimum	Maximum	Mean
$\hat{\beta}_0$	0,88	1,62	0,04
$\hat{\beta}_1$	0,18	0	0,07
$\hat{\beta}_2$	0,00	0,14	0,07
$\hat{\beta}_3$	0	0,07	0,03
$\hat{\beta}_4$	-0,17	0	0,02
$\hat{\beta}_5$	0	0,13	0,06
$\hat{\beta}_6$	0	0,26	0,03
$\hat{\beta}_7$	0,68	0,93	0,82

Source: Data processed, 2020

Modeling with GWL shows that in Banten, West Java, South Kalimantan, East Kalimantan, East Nusa Tenggara and Papua Province all significant variables affect the number of tuberculosis cases. Table 4 shows significant variables affecting the number of tuberculosis cases in other provinces. Among all variables, the number of the health center is a significant variable that always affects the number of tuberculosis cases. The number of health center positively affect the number of tuberculosis case in all provinces in Indonesia.

Table 4.
Variables Affect the Number of Tuberculosis Cases in Indonesia

Variable	Province
Percentage of the poor population, Population density, Percentage of districts/ cities that have a clean and healthy behavior policy, Percentage of slum households, the proportion of the population aged > 10 years who smoke every day and Number of health centers.	Nangroe Aceh Darussalam, West Kalimantan, Central Sulawesi and North Maluku.
Percentage of the poor population, Population density, Percentage of districts/ cities that have a clean and healthy behavior policy and Number of health centers.	Bali, Central Kalimantan and West Papua
Percentage of the poor population, Population density, Percentage of districts/ cities that have a clean and healthy behavior policy, the proportion of the population aged > 10 years who smoke every day and Number of health centers.	Bengkulu, Jambi, Bangka Belitung, Maluku, Riau, Southeast Sulawesi, North Sulawesi, West Sumatera and South Sumatera

Variable	Province
Population density and Number of health centers	Central Java, Lampung and West Nusa Tenggara.
Percentage of poor population, Population density and Number of health centers	South Sulawesi
Number of health centers	D.I Yogyakarta, DKI Jakarta, Gorontalo, East Java, North Kalimantan, Riau Islands, West Sulawesi and North Sumatera.

Source: Data processed, 2020

GWL model is more appropriate in modeling the number of tuberculosis cases in Indonesia for data containing spatial heterogeneity and the presence of local multicollinearity. This is also supported by the coefficient of determination (R2). The coefficient of determination shows the percentage of response variables that can be explained by predictor variables. The coefficient of determination of the GWL model is 0,9299. This means that 92,99 percent of the response variable can be explained by the predictor variable.

CONCLUSION

Modeling the number of tuberculosis cases in Indonesia using geographically weighted lasso due to the presence of local multicollinearity. The factors that influence the number of tuberculosis cases are not the same in each region, after all each region has its characteristics. Banten, West Java, South Kalimantan, East Kalimantan, East Nusa Tenggara and Papua Province are provinces where all research variables affect the number of tuberculosis cases. Meanwhile, in other provinces, variables affecting the number of tuberculosis cases varied.

The variable that has the most significant effect on the number of tuberculosis cases in each region is the number of health centers. Therefore, one of the things the government must do in reducing the number of tuberculosis is to increase the number of health facilities or improve services in health facilities.

REFERENCE

Anselin, L. (1988). *Spatial Econometrics: Methods and Models.* (Vol. 85, Issue 411). Kluwer Academic Publisher. https://doi.org/10.2307/2290042

Efron, B. T. H. I. J. (2004). Least Angle Regression. *The Annals of Statistics*, *32*(2), 407–499.

Firdaus, Y. (2014). Pemetaan Penyakit Tuberkulosis di Kota Surabaya Tahun 2014. *Jurnal Ilmiah Keperawatan*, *2*(2), 42–50.

Fogel, N. (2015). Tuberculosis: A disease without boundaries. In *Tuberculosis* (Vol. 95, Issue 5). https://doi.org/10.1016/j.tube.2015.05.017

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. John Wiley & Sons, Inc.

Indonesia, S. (2018). *Indikator Tujuan Pembangunan Berkelanjutan Indonesia Tahun 2018*. Statistics Indonesia.

Indonesia, S. (2019a). *Data dan Informasi Kemiskinan Kabupaten/Kota Tahun 2018*. Statistics Indonesia.

Indonesia, S. (2019b). *Indikator Perumahan dan Kesehatan Lingkungan 2018*. Statistics Indonesia.

Indonesia, S. (2019c). *Statistic Yearbook of Indonesia 2019*. Statistics Indonesia.

Indonesian Ministry of Health. (2018). *Laporan Nasional RISKESDAS (Basic Health Research) 2018*.

Indonesian Ministry of Health. (2019a). *Data dan Informasi Profil Kesehatan Indonesia 2018*. Indonesian Ministry of Health.

Indonesian Ministry of Health. (2019b). *Profil Kesehatan Indonesia 2018*. Indonesian Ministry of Health.

Setiyorini, A. (2017). *Pemodelan Tingkat Kemiskinan Pulau Jawa Dengan Metode Geographically Weighted Lasso*. Padjadjaran Bandung.

Sun, W., Gong, J., Zhou, J., Zhao, Y., Tan, J., Ibrahim, A. N., & Zhou, Y. (2015). A spatial, social and environmental study of tuberculosis in China using statistical and GIS technology. *International Journal of Environmental Research and Public Health*, *12*(2), 1425–1448. https://doi.org/10.3390/ijerph120201425

Tibshirani, R. (1996). Regression Shrinkage and Selection via The Lasso. *Journal of The Royal Statistical Society Series B*, *58*(1), 267–288.

Wheeler, D. C. (2007). Diagnostic tools and a remedial method for collinearity in geographically weighted regression. In *Environment and Planning A* (Vol. 39, Issue 10). https://doi.org/10.1068/a38325

Wheeler, D. C. (2009). Simultaneous coefficient penalization and model selection in geographically weighted regression: The geographically weighted lasso. *Environment and Planning A*, *41*(3), 722–742. https://doi.org/10.1068/a40256

WHO. (2019). *Global Tuberculosis Report 2019*.