Geographically weighted models for modelling the prevalence of tuberculosis in Java

S Noorcintanami*, Y Widyaningsih, and S Abdullah
Departemen Matematika, Universitas Indonesia. Depok 16424, Indonesia
* yekti@sci.ui.ac.id

Abstract. Indonesia’s position as one of the high burden countries for the infectious disease, tuberculosis (TB), has caused TB to be a major health problem in Indonesia. As means to control the number of TB cases, it becomes important for the government to identify factors associated with it. Commonly, multiple linear regression models are used to evaluate the linear relationship between the identified factors and the number of TB cases. Unfortunately, this model does not have the ability to expose the spatial variation in the data. Therefore, this study proposes to implement a spatial model: a model that takes the geographical location in the model. This research examined two types of geographically weighted models (GWM): geographically weighted regression (GWR) and mixed geographically weighted regression (MGWR). These spatial models assign weights to observations based on its’ geographical location. These two models were constructed to evaluate the relationship between the prevalence of TB in regency/city in Java in 2017 and the factors associated with it: population size, success rate of TB treatment, percentage of toddlers receiving BCG vaccine, percentage of HIV patient, percentage of household with adequate sanitation, percentage of poor people and the number of public health centre per one hundred thousand people. Akaike’s Information Criterion (AIC) and adjusted $R^2$ were used to assess the model performances. We found that the GWR model fits the data better than MGWR, as it has a smaller AIC value (1558.67) and a higher adjusted $R^2$ (0.754). It is also found that BCG vaccine is important to reduce the prevalence of TB, as the percentage of toddlers receiving BCG vaccine is negatively associated with it. Among the examined areas, Jakarta is the area with the highest association between the percentage of toddlers receiving BCG vaccine and the prevalence of TB.

1. Introduction
Tuberculosis (TB) is an infectious disease caused by mycobacterium tuberculosis bacteria. Indonesia’s position as one of the high burden countries for TB [1] has caused TB to be a major health problem in Indonesia. Hence, it becomes the responsibility for the Indonesian government to tackle this issue, which makes it important for the government to understand the possible factors associated with the prevalence of TB. The term prevalence of TB used throughout this study is measured by the number of TB cases.

The factors affecting the prevalence of TB considered in this research include the factors related to the individual, such as an immunosuppressive condition (HIV), factors related to lifestyle and living conditions including sanitation level, housing density, the effectiveness of BCG vaccine, alcohol and cigarettes intake, and lastly, the quantity and quality of public health centres [2]. A common approach to evaluate the linear relationship between those factors and the prevalence of TB is a multiple linear regression model. Such approach, however, fails to expose the spatial variation in the data [3]. This could be attributed to the fact that a multiple linear regression model is a global model, where all observations are used to estimate parameters that represent the whole study area. This approach makes
it a major drawback, considering the parameter estimates may not be representative to specific areas of observation [3].

Different areas might have different characteristics (e.g. its population socio-economic-demographic condition) and -probably- policies regarding health care system. It is reasonable to assume that these factors might affect, or at least associated with, the prevalence of such disease in the area. Thus, one area might have different association between the factors and the TB prevalence compared to the other areas. Therefore, we propose to use a spatial model where the location (area) is also accounted for in the model. A spatial model takes the geographical location of the observations into account for the model construction process. This research models the linear relationship between of the prevalence of TB and the possible factors associated with it. The spatial model used for this research is geographically weighted models. As the name suggests, weights are assigned to observations based on their geographical locations. The assignment process of the weights is based on Tobler’s First Law of Geography [4], suggesting that nearer observations are more related to each other than that of observations located farther away. The Geographically Weighted Models considered in this research includes Geographically Weighted Regression (GWR) and Mixed Geographically Weighted Regression (MGWR), which will further be discussed in the next section of this paper.

Geographically Weighted Models has been applied to several research including Astuti’s research on Dengue Fever in Surakarta [4] Runadi’s research on crime in Central Java [5], and Zhang’s research on the price of houses in Nanjing [6]. Previous studies on the Geographically Weighted Models have not dealt with the case of TB and Java as the study area. Much of the research has rather been descriptive and have failed to elaborate the reasons of what might cause the parameter estimates to vary spatially in detail. In this research, we aim to find which spatial model (GWR or MGWR) that performs better in representing the prevalence of TB based on the factors associated with it. To achieve this aim, we used AIC and adjusted R² values as the model performances indicator. Furthermore, we aim to see how the relationships are between the prevalence of TB and the possible factors associated with it and identify which factors are affecting it the prevalence of TB significantly in specific locations.

This study focuses on Java island as the study area, with the observations per regency/city. This is due to Java being the most populated island in Indonesia. The scope of this research is limited to the year 2017 using secondary data taken from Indonesian Health Department, with the assumption that there are no measurement errors in the collection of the data. The performance evaluation of both models is limited to the AIC and adjusted R² values.

2. Method

The dataset used in this research was built upon the data collected by the Health Department of each province, for Jakarta [7], West Java [8], Central Java [9], Yogyakarta [10] and East Java [11]. The data collected are based on public health facility records and health surveys conducted by the local government. This study focuses on modelling the prevalence of TB in Java island in 2017, consisting of 105 regencies/cities, excluding Banten Province due to the data unavailability. The dataset consists of 105 observations, each representing one specific regency/city in Java Island. An observation unit includes the data on the prevalence of TB, the possible factors associated with it, as well as the longitude and latitude coordinate of a regency/city. The prevalence of TB and the possible factors associated with it are measured as of per year.

The demographic aspect is represented by the population size variable. The success rate of TB treatment was chosen to measure how effective the treatments are for TB control measures, which is obtained by calculating the percentage of patients who have completed the treatment process and are cured, out of the people who are treated. The percentage of toddlers receiving BCG vaccine may measure how prepared each location is for TB prevention measures, where the percentage is measured out of the number of toddlers. The percentage of HIV patients are measured out of the population size to represent how the immunosuppressive condition affects the prevalence of TB. The percentage of households with adequate sanitation represents measures the sanitation level and the housing density, as it is associated with the spread of TB [2]. From the socio-economic aspect, the percentage of poor people is measured out of the population size. Lastly, the public health facility aspect is measured using the data on the number of public health centres, which was converted into the number of public health centre per 100000
people as we assume that it might provide a better insight. The variables considered for this model are shown in Table 1 below.

**Table 1. Variables in the model**

| No | Variables                                      | Type of Variable | Definition                                                                                           |
|----|-----------------------------------------------|------------------|-----------------------------------------------------------------------------------------------------|
| 1  | Prevalence of TB (y)                         | Numeric          | The number of recorded TB cases                                                                     |
| 2  | Population size (x₁)                         | Numeric          | Population size (person)                                                                            |
| 3  | Success Rate (x₂)                            | Numeric          | The percentage of the success rate of TB treatment out of the number of TB patients (%)              |
| 4  | Percentage of toddlers receiving BCG vaccine (x₃) | Numeric          | The percentage of toddlers receiving BCG vaccine out of the number of toddlers (%)                    |
| 5  | Percentage of HIV patients (x₄)              | Numeric          | The percentage of HIV patients out of the population size (%)                                       |
| 6  | Percentage of households with adequate sanitation (x₅) | Numeric          | The percentage of households with adequate sanitation out of the number of houses (%)                |
| 7  | Percentage of poor people (x₆)               | Numeric          | The percentage of poor people out of the population size (%)                                       |
| 8  | Number of public health centre per 100000 people (x₇) | Numeric          | The ratio of the number of public health centre over 100000 people                                   |

2.1. Geographically Weighted Regression (GWR)

In a multiple linear regression model, the assumptions that needs to be satisfied are the error term being normally distributed with a mean zero and a constant variance (referring to homoscedasticity) and the error must be independent between observations. However, when dealing with a spatial data in a GWR model, it is acceptable when the homoscedasticity assumption cannot be satisfied. A GWR model can be written in the general form below [3]:

\[
y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i)x_{ik} + \epsilon_i, \ i = 1, 2, \ldots, n
\]  

where \( y_i \) represents the prevalence of TB at location \( i \) and \( \beta_k(u_i, v_i) \) represents the parameter estimates for the \( k \)th explanatory variable at location coordinate \( (u_i, v_i) \), indicating how the respective independent variables are associated with the prevalence of TB at a specific location. The vector of parameter estimates of the GWR model obtained by minimizing the weighted sum square error is given by [3]:

\[
\hat{\beta}(u_i, v_i) = (X'W(u_i, v_i)X)^{-1}X'W(u_i, v_i)y
\]  

where \( \hat{\beta}(u_i, v_i) = (\hat{\beta}_0, \ldots, \hat{\beta}_p) \) at location \( i \), \( X = (x_1, \ldots, x_n)' \) with \( x_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip}) \), and \( W(u_i, v_i) \) is a spatial weight matrix of location \( i \). The matrix \( W(u_i, v_i) \) introduced in equation (2) highlights the consideration of spatial effect to the model. It is a diagonal matrix representing the weights assigned for every location relative to the location \( i \) [3], which is given by:
For example, \( w_{in} \) is the weight for the location \( n \) relative to the location \( i \). These weights can be calculated using a kernel function, which obeys Tobler’s First Law of Geography. This research used the bi-square kernel function, which is given by:

\[
\begin{align*}
    w_{ij} &= \left\{
    \begin{array}{ll}
        1 - \frac{d_{ij}^2}{b^2}, & |d_{ij}| < b \\
        0, & \text{otherwise}
    \end{array}
\right.
\end{align*}
\]

where \( w_{ij} \) is the weight of location \( j \) relative to location \( i \), \( d_{ij} \) is the Euclidean distance between the observation \( i \) and \( j \), and \( b \) represents the bandwidth [12].

The kernel function requires two inputs, which are distance between the observation \( i \) and \( j \), as well as the bandwidth. Bandwidth acts as a radius calculated from the center of location \( i \), where the observations outside the scope of the bandwidth are considered not affecting the estimation of parameters at location \( i \). The two types of bandwidth include fixed and adaptive bandwidth. Fixed bandwidth refers to a condition where only one bandwidth value holds for every area, which is better used when the observations are regularly spaced across the area [3]. This research adopts an adaptive bandwidth, meaning each location has a different bandwidth, making sure that the number of observations included for the estimation of parameters at every location are equal. An adaptive bandwidth was chosen considering the location of observations that are compact in West Java, however sparse in the East, which may result in a little amount of observation included in the sparse area if a fixed bandwidth was used.

2.2. Mixed Geographically Weighted Regression (MGWR)

In the construction of MGWR, not all parameter estimations are estimated locally. Some of the explanatory variables in MGWR will treated as global variables, where the parameter estimates apply for all the locations. The rest of the explanatory variables are treated as local variables, where the local parameters only apply at each of the location points. An MGWR model is also referred to as a combination of the multiple linear regression model (global model) and the GWR model (local model) [6]. The general form of the MGWR model can be written as follows:

\[
y_i = \sum_{k=1}^{g} \alpha_k x_{ik} + \beta_{i0} + \sum_{k=g+1}^{p} \beta_{ik} x_{ik} + \epsilon_i, \quad i = 1, 2, \ldots, n
\]

where \( g \) is the number of global variables, \( \alpha_k \) is the parameter estimate for the \( k^{th} \) global explanatory variable, \( \beta_{ik} \) is the \( k^{th} \) parameter estimate at location \( i \), \( x_{ik} \) is the \( k^{th} \) explanatory variable at location \( i \). It is clearly shown that the first summation refers to the parameters that are constant over space (global) and the second summation refers to the parameters that vary locally (local). MGWR can also be written in matrix form below [3]:

\[
y = X_i \beta_i(u_i, v_i) + X_g \beta_g + \epsilon
\]

where \( X_g \) is the matrix of global explanatory variables, \( X_i \) is the matrix of local explanatory variables, \( \beta_g \) is the vector of global parameters, \( \beta_i \) is the vector of local parameters [3]. The global parameter estimates can be estimated with the formula \( \hat{\beta}_g = [X'_g(I - S)'(I - S)X_g]^{-1}X'_g(I - S)'(I - S) \).
The selection of global variables was done using the Monte Carlo test to check for spatial variability of the parameters in a GWR model [12]. This was done by testing the null hypothesis that the relationship between the dependent and independent variables are constant (no spatial variation) against the alternative hypothesis where the spatial variation exists. $H_0$ will be rejected if $p$-value < $\alpha$, which implies that there is enough evidence to conclude that there is a spatial variation in the parameter in test, hence that variable will be treated as local and the estimation of parameters will be done at every location points [12].

2.3. Model performance Indicator
2.3.1. Akaike’s Information Criterion (AIC)

The performance of the models in this research was measured using AIC and $R^2$. The AIC used for this model is as follows [3]:

$$AIC = 2n\ln(\hat{\sigma}) - n\ln(2\pi) + n + tr(S)$$  \hspace{1cm} (7)

where $n$ is the number of observations, $\pi$ is 3.14, and $\hat{\sigma}$ is the estimated residual standard deviation, given by $\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$, where $I$ is an identity matrix and $S$ is a hat matrix composed by the vectors $r_i$ which can be written as follows $r_i = X'(X'WX(u_i, v_i))^{-1}X'W(u_i, v_i)$. Compared to other models in comparison, the model with the smallest AIC value is considered the relatively best model.

2.3.2. Adjusted $R^2$

This research uses the adjusted $R^2$ value to evaluate the performance of the models, which takes the number of observation and explanatory variables in the model into account. It explains the percentage variation of the dependent variable that can be explained by the explanatory variables in the model. Hence, a higher adjusted $R^2$ value indicates a better model performance as more variation of the dependent variable is accounted for [13]. The adjusted $R^2$ can be calculated with the following formula:

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}$$ \hspace{1cm} (8)

where $R^2$ is the coefficient of determination (ranging from 0 to 1), $n$ is the number of observation and $p$ is the number of explanatory variables [13].

3. Result and Discussion
3.1. Geographically Weighted Regression (GWR)

With a total of 7 variables and 105 observations, each representing a regency/city in Java, the general form of GWR for this data is given by:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{7} \beta_k(u_i, v_i)x_{ik} + \epsilon_i \ , \ i = 1, 2, \ldots, 105. \hspace{1cm} (9)$$

Assessing the model performance, the GWR model has an AIC of 1558.676 and adjusted $R^2$ of 0.7436. The adjusted $R^2$ value for this model suggests that 74.36% of the variation of the prevalence of TB in Java in 2017 can be explained with the explanatory variables in the model. There are a total of 105 models, each representing respective locations. The equation below is an example of the model calibrated at the location Bandung City ($i = 1$):

$$\hat{y}_1 = 1537.285 + 0.000619x_{11} - 2.576x_{12} - 8.159x_{13} + 3733.595x_{14} - 2.254x_{15} - 15.639x_{16} - 62.492x_{17}.$$  

Based on our findings, population size is proven to be a significant factor affecting the prevalence of TB (based on t-test in [14]) at every location. The population size is positively associated with the
prevalence of TB, with the varying value of parameter estimates in every regency/city is possibly due to the demographic difference in these areas. For example, there might be a higher number of citizens classified into an age group with the highest risk for TB at locations with a stronger relationship compared to the rest. In contrast, the success rate of TB treatment, percentage of poor people and percentage of households with adequate sanitations are proven to be the non-significant factors affecting the prevalence of TB in all locations, which reveals the need for further investigations by the government regarding the non-significant effect of these factors on the prevalence of TB in Java. On the other hand, the percentage of toddlers receiving BCG vaccine, the percentage of HIV patients, and the number of public health centre per 100000 people are proven to affect the prevalence of TB significantly at some locations, however do not affect significantly at some others. Nevertheless, these factors are still considered as the points of interest for this research.

The number of toddlers receiving BCG vaccines is considered a good indicator that measures the scope of immunization and how prepared each regency/city for TB control measures. Figure 1 illustrates the effect of the percentage of toddlers receiving BCG vaccine to the prevalence of TB. In Figure 1, blue represents the negative value of the parameter estimates, while grey represents the non-significance of the parameter estimates. The darker the blue colour, the stronger the negative relationship is between the percentage of toddlers receiving BCG vaccine to the prevalence of TB.

Figure 1 implies that the effect of the percentage of toddlers receiving BCG vaccine is only significant in parts of Jakarta, regencies/cities in West Java and a part of Central Java. It is evident that the percentage of toddlers receiving BCG vaccines is negatively associated with prevalence of TB in these areas, however, shows a non-significant effect in Central Java, Yogyakarta and East Java. Based on our findings, the strongest negative relationship is shown in parts of Jakarta, where to every 1% increase of the number of toddlers receiving BCG vaccine may cause the prevalence of TB to decrease by approximately 8 cases in Jakarta. In contrast, the weakest negative relationship is shown at Banyumas Regency (Central Java) with the estimate of a decrease by 6 cases. The evidence thus supports the idea that the increase of 1% in Jakarta affects the decrease of prevalence more than in Banyumas. The potential cause of the variety of the effect in both areas is the effectiveness of BCG immunization programmes that performs better in Jakarta. Another potential cause of this difference is parents’ awareness regarding the importance of BCG immunization that is higher in a big city like Jakarta compared to other areas.

3.2. Mixed Geographically Weighted Regression (MGWR)

The overall performance of the MGWR model for this data can be shown through the AIC and adjusted R² values. The AIC value for this model is 1574 and the adjusted R² is 0.688. This suggests that 68.8% of the variation of the prevalence of TB in Java in 2017 can be explained with the explanatory variables in the MGWR model. The construction of MGWR model begins with the selection of global variables using Monte Carlo test as discussed in the previous section, which assess the spatial variability of the parameters in a GWR model. Table 2 shows the result of spatial variability test using the Monte Carlo test.
The variables marked with * in Table 2, which includes success rate of TB treatment ($x_2$), percentage of HIV patients ($x_4$), percentage of households with adequate sanitation ($x_5$), and percentage of poor people ($x_6$), are the variables representing the factors with no spatial variability. Hence, these variables are treated as global variables, while the rest of the variables are treated as local variables. The global parameter estimates obtained applies for every application. Focusing on the success rate of TB, based on our findings, it is negatively associated with the prevalence of TB at all regency/city, with an increase by 1% will decrease the prevalence of TB by 3 cases. This model assumes that the success rate has the same effect to the prevalence of TB in every location. The equation below shows the MGWR model calibrated at Bandung City ($i = 1$), where the terms written in bold are the global variables, meaning these parameter estimates are constant over the study area.

$$\hat{y}_1 = 1504.161 - 2.979x_{12} + 3587.471x_{14} - 0.96x_{15} - 8.842x_{16} + 0.0006x_{11} - 8.68x_{13} - 76.20x_{17}.$$  

Figure 2 shows the relationship between the percentage of toddlers receiving BCG vaccine and the prevalence of TB.

Figure 2 shows that the result the percentage of toddlers receiving BCG vaccines vary spatially and is negatively associated with the prevalence of TB in Regencies/Cities in Jakarta, West Java and Central Java, which is aligned with prior hypothesis of this research. However, the positive relationship shown in a part of Central Java, Yogyakarta and East Java does not comply with the prior hypothesis of this study that it is negatively associated. This rather contradictory result may be due to the possibility that an increase of the percentage of toddlers receiving BCG vaccine by 1% will not effectively be able to decrease the prevalence of TB in these areas. Similar to the GWR model, the differences between these values might be caused by the difference between the effectiveness of the BCG immunization programme by the government in every location.

Table 3 summarizes the model performance results of Geographically Weighted Models, based on the AIC and $R^2$ values.
Table 3. Model Performance results

| Model | AIC      | R²   |
|-------|----------|------|
| GWR   | 1558.676 | 0.7543 |
| MGWR  | 1574     | 0.688  |

4. Conclusion
The conclusion drawn for this research are limited to this data and these results therefore need to be interpreted with caution. In real-life situations, there might be some other factors affecting the model however not accounted in the model, such as people’s awareness on the BCG vaccine. GWR has appeared to be the model that performs relatively better than MGWR, having the smallest AIC value and largest R² value. Based on our findings, the significant factors affecting the prevalence of TB are population size, the percentage of toddlers receiving BCG vaccine, the percentage of HIV patients, and the number of public health center per 100000 people. One of the conclusions that can be drawn is that BCG vaccine is important to reduce the prevalence of TB, as it is negatively associated with the prevalence of TB in all areas. Among the examined areas, Jakarta is the area with the highest association between the percentage of toddlers receiving BCG vaccine and the prevalence of TB, in contrast, Banyumas Regency (Central Java) has the lowest association. Suggestions made towards the government is to focus on developing strategies to increase awareness and improve socialization of the importance of BCG vaccine. Another possible area of future research would be to investigate the TB issue in more specific areas (a smaller scale) and conduct a significance test for the MGWR model.

Acknowledgment
This research was funded by Directorate of Research and Development of Universitas Indonesia (DRPM UI) as a grant of Publikasi Terindeks Internasional (PUTI) Prosiding 2020 No. NKB-1028/UN2.RST/HKP.05.00/2020.

References
[1] World Health Organization 2017 Global tuberculosis report 2017. World Health Organization.
[2] Narasimhan P, Wood J, MacIntyre C R, and Mathai D 2013 Hindawi Publishing Corp: Pulmonary Medicine, 2013 828939
[3] Fotheringham A, Brunsdon C and Charlton M 2002 Geographically Weighted Regression: The Analysis of Spatially Varying Relationships (England: Wiley)
[4] Astuti H N, Saputro D R S, Susanti Y 2017 J. Phys. Conf., 855 012007
[5] Runadi T, Widyaningsih Y and Lestari D 2020 J. Phys. Conf., 1442 012026
[6] Zhang S, Wang L and Lu F 2019 ISPRS Int. J. Geo-Inf. 2019 8 431
[7] Dinas Kesehatan Provinsi DKI Jakarta 2017 Profil Kesehatan Provinsi DKI Jakarta (Jakarta: Dinas Kesehatan Provinsi DKI Jakarta)
[8] Dinas Kesehatan Provinsi Jawa Barat 2017 Profil Kesehatan Provinsi Jawa Barat (Jawa Barat: Dinas Kesehatan Provinsi Jawa Barat)
[9] Dinas Kesehatan Provinsi Jawa Tengah 2017 Profil Kesehatan Provinsi Jawa Tengah (Jawa Tengah: Dinas Kesehatan Provinsi Jawa Tengah)
[10] Dinas Kesehatan Provinsi Yogyakarta 2017 Profil Kesehatan Provinsi Yogyakarta (Yogyakarta: Dinas Kesehatan Provinsi Yogyakarta)
[11] Dinas Kesehatan Provinsi Jawa Timur 2017 Profil Kesehatan Provinsi Jawa Timur (Jawa Timur: Dinas Kesehatan Provinsi Jawa Timur)
[12] Lu B, Charlton M, Harris P and Fotheringham A 2014 Int. J. GIS 17 85-101
[13] Mendenhall W and Sincich T 2012 A Second Course in Statistics: Regression Analysis (Boston, MA: Prentice Hall, Pearson)
[14] Utami T W, Rohman A, and Prahatama A 2016 Media Statistika 9 133-147.