Modelling mixed geographically weighted poisson regression for tuberculosis disease in Surabaya

Hani Khaulasari
Department Of Mathematics Education, IKIP Widya Darma, Surabaya, Indonesia
Email : hanikhaulasari@gmail.com

Abstract
Tuberculosis is one of the deadly infectious diseases caused by mycobacterium tuberculosis. Patients with Tuberculosis will experience pain and organ damage leading to death. Surabaya is one of the cities with the most densely populated settlements and the highest number of Tuberculosis sufferers in East Java. In this study, we want to get the factors of the spread of the number of tuberculosis sufferers in Surabaya using Mixed Geographically Weighted Poisson Regression (MGWPR) method. The advantage of the MGWPR method is to pay attention to the differences in each location where it is estimated to affect the spread of Tuberculosis locally and globally. The results show that the Gaussian kernel fixed weighting MGWPR model forms four groups with different factors. Global factors that influence Tuberculosis sufferers are the number of HIV/AIDS sufferers, the percentage of households applying PHBS, the ratio of health education, The percentage of the population receives Tuberculosis information, the number of medical personnel, the number of the population not completed elementary and The number of population graduated from high school while the influenced locally variables on the spread of Tuberculosis was the percentage of healthy homes. Model MGWPR is better than model GWPR and Regression Poisson, because model MGWPR get AIC value is the smallest and $R^2$ is the biggest.

1. Introduction
Tuberculosis is a deadly infectious disease caused by the bacterium Mycobacterium Tuberculosis. The bacteria enter the human body through the air that is inhaled into the lungs, then the germs spread from the lungs to other parts of the body through the circulatory system, the lymphatic system, the respiratory tract (bronchus) or other body parts. TB bacteria is a bacilli bacterium that is so strong that it requires relatively long treatment. Patients will feel pain in the organs that are attacked by Tuberculosis bacteria and organ damage leading to death. The terrible impact of Tuberculosis is not only felt by TB sufferers but is also felt by all people and even countries. This disease is commonly found in densely populated settlements with unhealthy environments, poor sanitation, lack of ventilation and solar lighting. Surabaya City is one of the cities with densely populated settlements and the highest number of Tuberculosis sufferers in East Java [1]. Tuberculosis is a rare case of disease and data on the number of sufferers of Tuberculosis are data in the form of count. Appropriate methods for analyzing cases of tuberculosis, one of them is Poisson Regression. Research on Poisson regression has been conducted by [2]and [3]The results of these studies provide conclusions that there are spatial effects that affect the model. One of the spatial methods that can be used is Geographically Weighted Poisson Regression. Research on Geographically Weighted Poisson Regression has been carried out by [4], [5], [6], [7]. The study provides results that the GWPR method is better than the Poisson regression method and the GWPR method can overcome the heterogeneity of errors that occur
because of spatial effects (location). In previous studies of the GWR model, it was found that in addition to local factors there are also global factors that influence the model. Research on Mixed Geographically Weighted Poisson Regression was conducted by [8] to model the number of hypertension sufferers in East Java and [9] to model the number of very poor households in Kulonprogo Regency in 2011. The research resulted that the Mixed Geographically Weighted Poisson Regression model gave better results than the GWR and Poisson Regression models based on AIC values. Therefore, this study uses the Mixed Geographically Weighted Poisson Regression method for the analysis of the number of Tuberculosis sufferers in Surabaya. The aim is to obtain global and local factors that affect the number of tuberculosis sufferers in Surabaya, get the number of groupings and compare the best models between MGWPR, GWPR and Poisson Regression. In this study using a Fixed Gaussian kernel function weighter.

2. Mixed Geographically Weighted Poisson Regression

Mixed Geographically Weighted Poisson Regression (MGWPR) method is the development of the Geographically Weighted Poisson Regression (GWPR) method. This method is used when there are certain predictor variables that affect the response variable globally and there are also predictor variables that affect locally [9]. The MGWPR model is as follows

\[
\begin{align*}
\mu_i &= \exp \left( \sum_{j=0}^{k} \beta_j(u_i, v_i)x_{ij} + \sum_{p=k+1}^{p} \gamma_p x_{ip} \right) = \exp(X_i^T \beta(U_i, V_i) + X_i^T \gamma) \\
\end{align*}
\]

(1)

Parameter estimates for the MGWPR model can be obtained using the Maximum Likelihood Estimation (MLE) method. Model suitability test is obtained by using the Maximum Likelihood Ratio Test (MLRT) method. The suitability test of the Mixed Geographically Weighted Poisson Regression model uses a comparison value of the Poisson regression model and the Mixed Geographically Weighted Poisson Regression model.

\[
H_0 : \beta_j(u_i, v_i), \gamma_p = (\beta_j, \gamma_p), j=1,2,...,p*; i=1,2,...,n
\]

(2)

\[
H_1 : \text{there is at least one } (\beta_j(u_i, v_i), \gamma_p) \neq (\beta_j, \gamma_p), j=1,2,...,p*; i=1,2,...,n
\]

Statistic Test : \( F_{\text{test}} = \frac{D(\beta^2)}{/v_1} / \frac{D(\beta(u_i, v_i), \gamma)/v_2}{2} \) \( F_{\text{test}} > F_{\alpha,1,2} \)

(3)

Reject \( H_0 \) if \( F_{\text{test}} > F_{\alpha,1,2} \)

In partial testing the parameters of the Mixed Geographically Weighted Poisson Regression model are partial tests globally and locally. For testing locally the hypothesis is as follows.

\[
H_0 : \beta_j(u_i, v_i) = 0
\]

(4)

\[
H_1 : \beta_j(u_i, v_i) \neq 0
\]

Statistic Tes : \( Z_{\text{test}} = \frac{\beta_j(u_i, v_i)}{SE(\beta_j(u_i, v_i))} \)

(5)

Reject \( H_0 \) if \( |Z_{\text{test}}| \geq Z_{\alpha/2} \)

For testing globally the hypothesis is as follows.

\[
H_0 : \gamma_p = 0
\]

(6)

\[
H_1 : \gamma_p \neq 0
\]

Statistic Tes : \( Z_{\text{test}} = \frac{\gamma_p}{SE(\gamma_p)} \)

(7)
Reject $H_0$ if $|Z_{test}| \geq Z_{\alpha/2}$

3. Methodology
The data used in this research are collected from a health office of Surabaya and publications of BPS Surabaya. The Tuberculosis data contains 12 attributes and 31 sub-district of Surabaya observations, with one response attribute is the number of Tuberculosis sufferers in Surabaya 2018 and 11 predictor attributes that are thought to influence Tuberculosis (The number of HIV/AIDS sufferers, total population density, the number of poor people, percentage of healthy homes, percentage of households applying PHBS, health education ratio, the percentage of the population that receives Tuberculosis information, the number of medical personnel, the number of health facilities, the number of population not graduated from elementary school, the number of population graduated from high school). We applied Tuberculosis data with Mixed Geographically Weighted Poisson Regression strategy. We get global and local influential factors with parameter significance test Mixed GWPR. We compared the performance of the proposed method with poisson regression and GWPR as a benchmark. All the computation were implemented using GWR4. The AIC dan $R^2$ was used as the performance measurement. First Stage is standardized variable, then exploring data and modelling regression poisson, GWPR and MGWPR. MGWPR method can be used overcome outlier.

4. Data Experimental Result
The results of this study will discuss about data exploration by describing the data distribution of Tuberculosis in Surabaya then proceed to MGPWR modeling through the stages of poisson regression modeling and GWPR modeling. the last step is comparing the goodness of the model with the criteria of AIC and $R^2$.

4.1. Description Data

![The Distribution pattern of Tuberculosis Disease In Surabaya](image)

**Figure 1.** The Distribution pattern of Tuberculosis Disease In Surabaya
Based on Figure 1, it can be described that the Surabaya area with a high rate of Tuberculosis distribution or prone to Tuberculosis are Semampir, Kenjeran, Tambaksari, Krembangan, Tegembangan, Sawal, Sawahan, Wonokromo, Gubeng, Sukolilo and Rungkut. Subdistrict groups with high levels of tuberculosis must receive attention and guidance from the government and all parties.

4.2. Modelling Mixed Geographically Weighted Poisson Regression
The stages in modeling are perform multicollinearity diagnostic checks and correlation test. On the results of multicollinearity diagnostic checks on 11 predictor variables, it can be concluded that there is no multicollinearity on 11 predictor variables and eleven of these variables are feasible to be used in modeling.

Based on the results of the correlation test between response variables with 11 predictor variables, it can be concluded that there is a relationship between the number of people with Tuberculosis (Y) with eleven predictor variables. In the results of the global regression model test (poisson regression) a violation of homogeneity and independent assumptions occurred. This indicates the presence of location (spatial) effects that affect the model.

The spatial aspects of the spatial dependency analysis is performed using the Moran's I Test and spatial heterogeneity with the Pagan Breush test. In the test of dependency and spatial heterogeneity it is concluded that the errors are correlated and have Spatial heterogeneity. therefore, it was developed by the GWPR method. In the GWPR model test results, it can be concluded that the factors influence the spread of Tuberculosis in Surabaya not only affect locally but also affect globally. Therefore, the Mixed GWPR method was developed.

Table 1. Parameter Significance Mixed GWPR Test

| Parameter | Estimation | Standard Error | Zhitung |
|-----------|------------|----------------|---------|
| Intercept | 5.16026    | 0.014636       | 352.5734|
| Global Parameter | 0.046100 | 0.015795 | 2.918573* |
| The number of HIV/AIDS sufferers (B1) | -0.148509 | 0.028895 | 4.139625* |
| The percentage of households applying PHBS (B2) | 0.102192 | 0.018424 | 5.602156* |
| The ratio of health education (B3) | 0.0232541 | 0.014438 | 2.253878* |
| The percentage of the population receives Tuberculosis information (B4) | 0.106684 | 0.023244 | 4.589694* |
| The number of medical personnel (B5) | 0.028034 | 0.016356 | 1.714016 |
| The number of population not graduated from elementery school (B6) | 0.017222 | 0.011913 | 1.454969 |
| The number of population graduated from high school (B7) | -0.063165 | 0.016895 | -3.73870* |
| Local Parameter | 0.022347 | 0.014332 | 1.559302 |

*) Significant to 5% significance level
Figure 2. Sub-district Grouping Mixed GWPR Models

Table 1 is the results of the Mixed GWPR model test on the spread of Tuberculosis in Surabaya. The results showed that the significant influence globally variables were The number of HIV/AIDS sufferers (X₁), Percentage of households applying PHBS (X₅), the ratio of health education (X₆), The percentage of the population receives Tuberculosis information (X₇), the number of medical personnel (X₈), The number of population not graduated from elementary school (X₁₀) and The number of population graduated from high school(X₁₁) while the influenced locally variables on the spread of Tuberculosis was the percentage of healthy homes (X₄). This influential variable is based on the calculated Z test statistic value greater than Z table with a significance level of 5% that is equal to 1.96.

Figure 2 shows that the grouping of the Mixed GWPR model forms four groups with different local characteristic factors. The first group consists of Asemrowo, Bubutan, Hamlet fern, Jambangan, Karang Pilang, Krembangan, Customs Cantian, Sawahan, Semampir, Simokerto, Sukomanunggal, Tandes, Tegalsari, Wiyung, Sambikerep, Lakarsantri, Kenjeran, Benowo, Pakal with local factor is Percentage of healthy homes (X₄). The second group consisted of Gayungan, tile, gubeng, tambaksari, wonocolo with local factors are Total population density (X₂) and Percentage of healthy homes (X₄). The third group consists of Gunung anyar and rungkut with local factors are Total population density (X₂), The number of poor people (X₃) and Percentage of healthy homes (X₄). The fourth group Mulyorejo, Sukolilo, Tenggilis Mejoyo, Bulak with local factors are Total population density (X₂), Percentage of healthy homes (X₄) and The number of health facilities (X₉).

| Model                  | AIC         | R²   |
|------------------------|-------------|------|
| Regresi Global (Poisson) | 91,748106  | 45%  |
| GWPR                   | 84,963683   | 52%  |
| MGWPR                  | 75,954786   | 70%  |

Comparisons between Poisson, GWPR, Mixed GWPR regression models are performed for find out which model is better applied to the criteria of the model goodness used is AIC and R² the best model. Model MGWPR get AIC value is the smallest AIC and R² is the biggest. Results the comparison shown in Table 2 gives the conclusion that the Mixed model GWPR with Fixed Gaussian
kernel function weighting is the best model for analysis of the number of spread of Tuberculosis in Surabaya in 2018.

5. Conclusion
In this study, the empirical results showed that the MGWPR model the fix Gaussian kernel form four groups with different local factors. Global factors that influence Tuberculosis sufferers are the number of HIV / AIDS sufferers (X₁), the percentage of households applying PHBS (X₅), the ratio of health education (X₆), The percentage of the population receives Tuberculosis information (X₇), the number of medical personnel (X₈), The number of population not graduated from elementary school (X₁₀) and The number of population graduated from high school (X₁₁), the influenced locally variables on the spread of Tuberculosis was the percentage of healthy homes (X₄). Model MGWPR is better than model GWPR and Regression Poisson, because model MGWPR get AIC value is the smallest and R² is the biggest. In further research, it is better to compare the merits of the model with different kernel functions.

Acknowledgments
This research was supported and funded by Ministry of Research and Technology Education Higher for lecturer research grant assistance beginners (PDP) 2019, Thank you very much.

6. References
[1] Dinkes Kota Surabaya 2016 Profil Kesehatan Kota Surabaya Tahun 2015 Surabaya
[2] S N Aulele 2012 Pemodelan Jumlah Kematian Bayi di Provinsi Maluku Tahun 2010 dengan Menggunakan Regresi Poisson BAREKENG J. Ilmu Mat. Dan Terapan. 6, 2, 23–27
[3] D P Firdaus 2017 Pemodelan Jumlah Kasus Tuberkulosis di Provinsi Jawa Barat Menggunakan Regresi Binomial Negatif dan Regresi Nonlinier (Surabaya : Thesis Institut Teknologi Sepuluh Nopember)
[4] T Nakaya, A S Fotheringham, C Brunsdon and M Charlton 2005 Geographically weighted Poisson regression for disease association mapping Stat. Med. 24, 17,2695–2717.
[5] A Hadayeghi, A S Shalaby and B N Persaud 2010 Development of planning level transportation safety tools using Geographically Weighted Poisson Regression Accid. Anal. Prev. 42, 2, 676–688
[6] R D Lestari, S P Wulandari and Purhadi 2014 Pemodelan Faktor-Faktor yang Mempengaruhi Jumlah Kasus Penyakit Tuberkulosis di Jawa Timur dengan Pendekatan Generalized Poisson Regression dan Geographically Weighted Poisson Regression. Sains Dan Seni ITS. 3, 2, D188–D193
[7] Y Ji, X Ma, M Yang, Y Jin and L Gao 2018 Exploring spatially varying influences on metro-bikeshare transfer: A geographically weighted poisson regression approach Sustainability. 10, 5,1526
[8] E D Priyana and Purhadi 2014 Pemodelan Jumlah Penderita Hipertensi di Propinsi Jawa Timur dengan Mixed Geographically Weighted Poisson Regression Sains Dan Seni ITS. 3, 2, D260–D265
[9] H Nurcahyani 2014 Pemodelan Spasial Kemiskinan dengan Mixed Geographically Weighted Poisson Regression dan Flexibly Shaped Spatial Scan Statistic (Studi Kasus: Jumlah Rumah Tangga Sangat Miskin di Kabupaten Kulonprogo, Bantul, dan Gunungkidul Tahun 2011) (Surabaya : Tesis Institut Teknologi Sepuluh November)