MODELING OF MALARIA PREVALENCE IN INDONESIA WITH GEOGRAPHICALLY WEIGHTED REGRESSION

Ita Miranti, Anik Djuraidah, Indahwati
Departement of Statistics, Bogor Agricultural University
Email: itamiranti414@gmail.com

Abstract

Background: *Malaria is a public health problem that can lead to death, especially in high-risk groups i.e. infants, toddlers and pregnant women. This disease is still endemic in most parts of Indonesia. The relation of location factor between regions with the surrounding region was assumed to give the effect of spatial variability in the prevalence of malaria in the region. It would lead to the prevalence of malaria modeling using classical regression methods become less precise due to the assumption of homogeneity of variance was not met. It could be overcome by Geographically Weighted Regression (GWR) modeling*

Methods: *Conducting spatial correlation test between each adjacent observation error. Selecting explanatory variables to be included in the model with stepwise regression. Conducting spatial heterogeneity test on malaria prevalence data. Estimating the parameters of each GWR models and partial test parameters for each province by using a gaussian kernel weighting function and kernel weighting function bisquare. Choosing the best model between GWR models using gaussian kernel weighting function and bisquare kernel weighting function by using R^2 and AIC.*

Results: *GWR modeling is locally modeling therefore the value of the parameter estimators GWR models will be different for each research area .GWR models with gaussian kernel weighting has a R^2 value of 87.82 and AIC value of 143.80 GWR models with bisquare kernel weighting have R^2 value of 90.17% and AIC value of 137.81.*

Conclussion: *GWR modeling using bisquare kernel weighting function on malaria prevalence data by province in Indonesia in 2013 provides better results than a gaussian kernel weighting function based on the value of R^2 and AIC of both models.*

Keywords: *bisquare, gaussian, GWR, malaria prevalence*

1. Introduction

Malaria is a public health problem that can lead to death, especially in high-risk groups ie infants, toddlers and pregnant women. Malaria is caused by plasmodium parasites that are transmitted by the bite of the female Anopheles mosquito. This disease is still endemic in most parts of Indonesia. The prevalence of malaria in 2013 was 6.0 percent. Five provinces with the highest prevalence is Papua (28.6%), East Nusa Tenggara (23.3%), West Papua (19.4%), Central Sulawesi (12.5%), and North Maluku (11.3 %).[1] Malaria prevalence is the number of malaria cases within one year compared to the total population.[1]

The World Health Organization set 25 April as World Malaria Day, The declaration "Towards Indonesia Free Malaria" on May 7, 2008 by President Susilo Bambang Yudhoyono, then the decree of the Minister of Health of the Republic of Indonesia Number 293/MENKES/SK/IV/2009 April 28, 2009 on the Elimination of malaria in Indonesia to create a society that is healthy life, free from malaria transmission gradually until 2030. Elimination stages starting from Thousand Islands (DKI Jakarta), Bali and Batam in 2010. Furthermore, Java, Aceh province, and the province of Riau Islands in 2015. The third phase is the Sumatra (except Aceh and

Riau Islands), West Nusa Tenggara and Sulawesi in 2020. the last is the province of Papua, West Papua, Maluku, East Nusa Tenggara, and North Maluku, in 2030.[2]

Observations at a particular location was influenced by observations in other locations as proposed by Tobler's which states that everything is interconnected to each other, but something close have more influence than something far away.[3] Similarly, the prevalence of malaria cases could have been influenced by the observation location or geographic conditions, including the position of the provinces toward the surrounding provinces. Spatial heterogeneity allegedly occurred in malaria prevalence data would cause data among observations had a difficulty to meet the assumptions of the classical regression i.e., residual variance homogeneity. So, to overcome the spatial approach is to model the prevalence of malaria in the region.

One of the approaches that take into account the spatial location of the observation is Geographically Weighted Regression (GWR). GWR is one method used to overcome the problem of residual variance heterogeneity due to the observation location factor.[4] Selection of weighting function is one determinant of the results of the analysis of GWR.[5] The weighting function is used to construct the GWR models in this study is a gaussian kernel and bisquare kernel weighting function.

The purpose of research is to determine the best model among the GWR models with gaussian kernel weighting function and GWR bisquare with kernel weighting function of malaria prevalence by province in Indonesia in 2013.

2. Research Method

The data used in this research was secondary data published by the Ministry of Health in the form of Health Research Result publications in 2013. The observation units used in this study were 33 provinces in Indonesia. Response variable used in this study was the prevalence of malaria. Based on the availability of data from Riskesdas 2013, the explanatory variables suspected to affect the prevalence of malaria are:

1. The proportion of people who received treatment ACT program
2. The proportion of people who received drug treatment ACT within 24 hours
3. The proportion of people who received drug treatment ACT for 3 days
4. The proportion of people who get effective treatment with ACT
5. The proportion of people treating the disease by themselves
6. The proportion of households using nets
7. The proportion of households using mosquito coils
8. Proportion of households using mosquito netting
9. The proportion of households using repellent
10. Proportion of households using insecticide
11. Proportion of households taking drugs to prevent malaria

Stages of the analysis conducted in this study are as follows:

1. Conducting spatial correlation test between each adjacent observation error with moran index.

   Moran index is used to test the spatial dependency or autocorrelation between observations or location.[3] Used to test the hypothesis that spatial correlation:

   $H_0: I = 0$ (no autocorrelation between locations)
   $H_0: I \neq 0$ (there is autocorrelation between locations)

   Moran index equation is as follows:

$$I = \frac{n\sum_i \sum_{j\neq i} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{(\sum_i \sum_{j\neq i} w_{ij}) \sum_i (y_i - \bar{y})}$$

where w is the spatial weighting matrix element results standardization line, y is the vector of observations and n is the number of observations.

Test statistics:

$$Z_{value} = \frac{I - E(I)}{\sigma(I)}$$

where:

I = moran index
$E(I)$ = expected value of Moran Index
$\sigma(I)$= standard deviation of Moran Index
n = number of observations

Decision making: Reject H_0 if $|Z_{value}| > Z\alpha_{/2}$

2. Selecting explanatory variables to be included in the model with stepwise regression.
3. Conducting spatial heterogeneity test on malaria prevalence data using Breusch – Pagan test.[6] The hypothesis Breusch-Pagan test is:

$$H_0: \sigma^2(u_i, v_i) = \cdots = \sigma^2(u_n, v_n) = \sigma^2$$
$$H_1: \text{at least one } \sigma^2(u_i, v_i) \neq \sigma^2(u_j, v_j) \text{ for } i \neq j, \text{with } i,j = 1,2,\ldots,n$$

Test statistics:

$$BP = \left(\frac{1}{2}\right) h'z(z'z)^{-1}z'h$$

With:

$$h = \left(\frac{e_i^2}{\sigma^2} - 1\right)$$

Test criteria: $BP > \chi^2_{(k)}$, reject H_0

with z is the observation vector y response variable size (n × 1) and standardized for each observation. While ei2 is squared error for the i-th observation and σ^2 is the variance of e_i. BP value will approach chisqure distribution with degrees of freedom k, where k is the number of explanatory variables.

4. Estimating the parameters of each GWR models and partial test parameters for each province by using a gaussian kernel weighting function and kernel weighting function bisquare. General equation GWR models is:[5]

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i)x_{ik} + \varepsilon_i$$

where

y_i : The value of the response variable for observation i-th location
(u_i, v_i) : Stating geographical location coordinates (longitude, latitude) of the i-th observation location
$\beta_0(u_i, v_i)$: Intercept values GWR models

$\beta_k(u_i, v_i)$: Regression coefficient explanatory variables k-th of the i-th observation location

x_{ik} : Observation value explanatory variables k-th on the location of the i-th observation

ε_i : error observations i-th , $\varepsilon_i \sim N(0, \sigma^2)$.

GWR is a locally linear regression model parameters which generate estimators for each location of the observations with Weighted Least Square (WLS), namely:

$$\hat{\beta}(u_i, v_i) = (X^T W(u_i, v_i) X)^{-1} X^T W(u_i, v_i) y$$

where $W(u_i, v_i)$ is the weighting matrix of size nxn. Diagonal elements of the i-th weighting location, which is determined by the distance between the location of the observation, while the diagonal elements other than zero.[7] This study used a spatial gaussian kernel and bisquare kernel weighting function.

Gaussian kernel function is written as follows: $W_{ij} = \exp[-\frac{1}{2}(\frac{d_{ij}}{b})^2]$

Bisquare kernel function is written as follows

$W_{ij} = [1 - (\frac{d_{ij}}{b})^2]^2$ if d_{ij} < b , and W_{ij} = 0 for $d_{ij} \geq b$.

where d_{ij} is the distance from location-i to location-j and b is the bandwidth, which is a smoothing parameter value function whose value is always positive.

Cross validation is one way that can be used as criteria to obtain the optimum value of the width of the window. The optimum window width used is the width value that results in a minimum value of the coefficient of cross validation with a formula:

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

With $\hat{y}_{\neq i}(b)$ is the value of the alleged y_i (fitting value) with the observation in the i-th location removed from the process of prediction.[5] The optimum window obtained by the iteration process to obtain a minimum CV.

To detect global excess GWR models compared with OLS regression for data use cases, can be tested by analysis of variance (ANOVA) as follows:[5]

$$F = \frac{SS_{OLS} - SS_{GWR}/v_1}{SS_{GWR}/\delta_1}$$

SS_{OLS} = sum of squares error of the OLS model
SS_{GWR} = sum of squares error of the GWR model

F value will approach the F distribution with degrees of freedom numerator v_1^2/v_2 and degrees of freedom denominator δ_1^2/δ_2, δ_1 = tr$[(I - S)'(I - S)]^i$, i = 1,2,.., Magnitude v_1 is the value of n-p-1-δ_1 and v_2 is the value of n-p-1-$2\delta_1 + \delta_2$ while S is a hat matrix from the GWR models that transform the vector y estimate $(\hat{y})$ from the value of y observations , F value is small will support the acceptance of the null hypothesis which states that the GWR and OLS models are equally effective in explaining the relationship between variables. With a significance level α, the null hypothesis would be rejected if $F > F_{\alpha(\frac{v_1^2}{v_2^2}, \delta_1^2/\delta_2)}$.

Testing of the model parameters for each location done partially with the aim of knowing any real parameters affect response variable in each location. Hypotheses used are as follows:

$H_0 : \beta_k(u_i, v_i) = 0$
H_1 : at least one $\beta_k(u_i, v_i) \neq 0$

test statistics: $t_k(u_i, v_i) = \frac{\hat{\beta}_k(u_i,v_i)}{Se(\hat{\beta}_k(u_i,v_i))}$

test criteria: H_0 rejected, $|t_k(u_i, v_i)| > t_{\alpha/2}$

With $Se(\hat{\beta}_k(u_i, v_i)) = \sqrt{c_{kk}\sigma^2}$, c_{kk} is a diagonal matrix elements CC' where matrix $C = (X'W(u_i, v_i)X)^{-1}X'W(u_i, v_i)$, σ^2 is the central square of the value of the model GWR, and v is degrees of freedom (n-k-1), k is the number of explanatory variables used.[5]

Choosing the best model between GWR models using gaussian kernel weighting function and bisquare kernel weighting function by using R^2 and AIC.

3. Results and Analysis

Spatial correlation test results performed using Moran index values obtained 0.5445 with p-value (0.01) less than 5% of significance level, thus obtained a rejected decision of H_0, which means that there is spatial autocorrelation in the data of prevalence of malaria in Indonesia. Because the Moran index value is in the range of 0 and 1, it can be concluded that the resulting autocorrelation is positive spatial autocorrelation. Positive autocorrelation indicates that the adjacent provinces have similar values and the data of malaria prevalence in Indonesia tend to clustered.

Stepwise regression is one method to get the best model of a regression analysis. Variables first entry is a variable that has the highest correlation and significant with the response variable, the second incoming variable is a variable that has the highest partial correlation and is still significant. After a certain variable enters into the model, the other variables in the model will be evaluated, if the variable is not significant then it should be removed again. From the 11 explanatory variables included in the model, there are three variables selected based on stepwise i.e the proportion of households who get effective treatment with ACT (X1), proportion of households treating the disease by themselves(X2) and the proportion of households using a mosquito coil (X3).

Testing of spatial heterogeneity using Breusch-Pagan test (BP) resulted in BP value of 21.49 with p-value (0.0008) which is less than 5% of significance level, thus obtained a rejected decision of H_0, which means that there is a spatial heterogeneity in malaria prevalence data by province in Indonesia 2013. The spatial heterogeneity of malaria prevalence shows that every province in Indonesia has its own characteristics, ultimately it requires a local approach to modelling and to address the heterogeneity that occurs in malaria prevalence. One local modeling is a model using geographically weighted regression.

A. Geographically weighted regression model with gaussian kernel weighting function

In the gaussian kernel weighting function the optimum bandwidth that generates minimum cross validation is 1858.801 km which resulted CV = 292.974. The optimum value of the bandwidth is 1858.801 km on gaussian weighting function shows that the distance between provinces is more than or equal to 1858.801 km, considered to have no effect on data observation analyzed.

Table 1. Summary of parameter estimators GWR models with Gaussian kernel weighting function

Parameter	Minimum	Median	Maximum
b0	4.227	5.355	8.971
b1	0.141	0.184	0.287
b2	2.615	3.496	3.950
b3	-0.141	-0.067	-0.046

In Table 1 it is known that the parameter b1 and b2 in the GWR models with a gaussian kernel weighting function has a positive value interval. This suggests that an increase in each variable proportion of people that effective treatment with ACT (X1) and the proportion of people treating the disease by themselves (X2) will increase the prevalence of malaria. While b3 parameter has a negative value hoses which means if there is an increase in variable proportion of households using mosquito coil (X3) will reduce the prevalence of malaria by province in Indonesia in 2013.

ANOVA was used to test whether there are significant differences between the classical regression model with a GWR models. ANOVA test between classical regression model with Gaussian weighting models available GWR obtained value of F (6.472) with a p-value (0.001) smaller than 5% of significance level to reject H0, which means that there are significant differences between the classical regression model with geographically weighted regression with a gaussian kernel weighting. Partial test conducted to determine the parameters of the explanatory variables that significantly affect the prevalence of malaria in each region. Partial test of the GWR models parameters with a gaussian kernel weighting function in 33 provinces in Indonesia to form two groups of regions based on the explanatory variables that significantly affect the prevalence of malaria. In the first group GWR modeling results with gaussian kernel weighting function is the region where the prevalence of malaria is influenced by the proportion of people treating the disease by themselves (X2), there were 16 areas included in this group, namely the Province of Aceh, North Sumatra, West Sumatra, Riau, Jambi, South Sumatra, Bengkulu, Lampung, Bangka Belitung, Riau Islands, Jakarta, West Java, Central Java, Yogyakarta, Banten and West Kalimantan. While 17 other areas were included in the second group that is affected by the variable proportion of people who received effective treatment with ACT (X1), proportion of people treating the disease by themselves (X2) and the proportion of households using a mosquito coil (X3). In general, modelling of malaria prevalence using GWR with gaussian kernel weighting function indicates that there is an explanatory variables that affect malaria prevalence in all provinces in Indonesia namely the proportion of people treating the disease by themselves (X2).

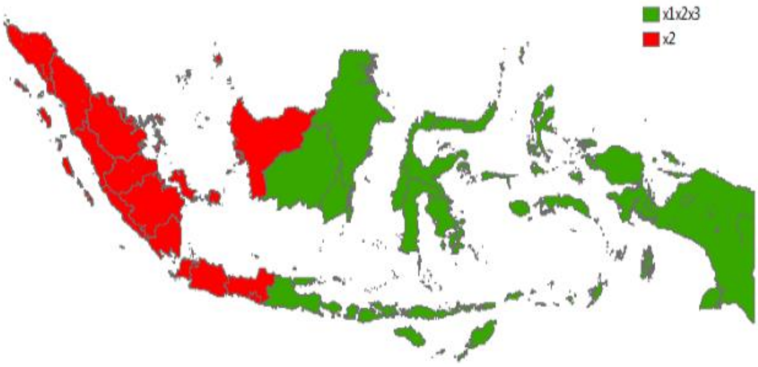

Figure 1. Map of the area group of geographically weighted regression with gaussian kernel weighting function

B. Geographically weighted regression model with bisquare kernel weighting function

In bisquare kernel weighting function the optimum bandwidth that generates minimum cross validation is 2899.979 km which resulted CV = 284.298. The optimum value of the bandwith is 2899.979 km in bisquare weighting function shows that the distance between provinces is more than or equal to 2899.979 km, is considered to have no effect on the observation on data analyzed.

Table 2. Summary of parameter estimators GWR models with bisquare kernel weighting function

Parameter	Minimum	Median	Maximum
b0	3.136	4.864	18.320
b1	0.091	0.162	0.309
b2	1.447	3.840	4.481
b3	-0.301	-0.058	-0.024

In Table 2 it is known that parameter b1 and b2 in the GWR models with a gaussian kernel weighting function has a positive value interval. This suggests that an increase in each variable proportion of people that treatment effective treatment with ACT (X1) and the proportion of people treating the disease by themselves (X2) will increase the prevalence of malaria by province in Indonesia in 2013. While b3 parameter has a negative value hoses which means if there is an increase in variable proportion of households using mosquito coil (X3) will reduce the prevalence of malaria by province in Indonesia in 2013.

ANOVA test between classical regression model with GWR models with weighted bisquare generated value of F (4.3185) with a p-value (0.0289) is smaller than 5% of significance level, so reject H0, which means that there are significant differences between the classical regression model with a GWR models bisquare kernel weighting. Partial test of the GWR models parameters on the kernel weighting function bisquare formed four groups of regions. The difference in the number of groups formed region due to differences in weighting functions are used to build the weighting matrix in the estimation of parameters, thus the weight given to each region for each of the weighting matrix is different and will affect the value of the alleged parameters generated. Group region formed by the significant variables through GWR modeling indicates that there is variation among provinces in the Indonesian region. In the first group of GWR modeling results with bisquare kernel weighting function is an area with malaria prevalence partially at 5% of significance level was not affected by the three explanatory variables, namely Aceh province, so further research is needed to examine the variables that partially affecting the prevalence of malaria. The second group is the region that have been affected by the variable proportion of people treating the disease by themselves (X2), there are 17 areas that fall into this group, namely North Sumatra, West Sumatra, Riau, Jambi, South Sumatra, Bengkulu , Lampung, Bangka Belitung, Riau Islands, Jakarta, West Java, Central Java, Yogyakarta, East Java, Banten, West Kalimantan and Central Kalimantan. The third group is the region that have been affected by three explanatory variables namely the proportion of people that receive effective treatment with ACT (X1), proportion of people treating the disease by themselves (X2) and the proportion of households using a mosquito coil (X3). There are 14 areas that fall into this group, namely the provinces of Bali, West Nusa Tenggara, East Nusa Tenggara, South Kalimantan, East Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku and West Papua. While the Province of Papua into four groups that are affected by the variable proportion of households using a mosquito coil (X3). In general, the malaria prevalence modeling using GWR with bisquare kernel weighting function indicates that the variable proportion of people treating the disease by themselves (X2) effect on malaria prevalence in almost all provinces in Indonesia except the province of Aceh and Papua province.

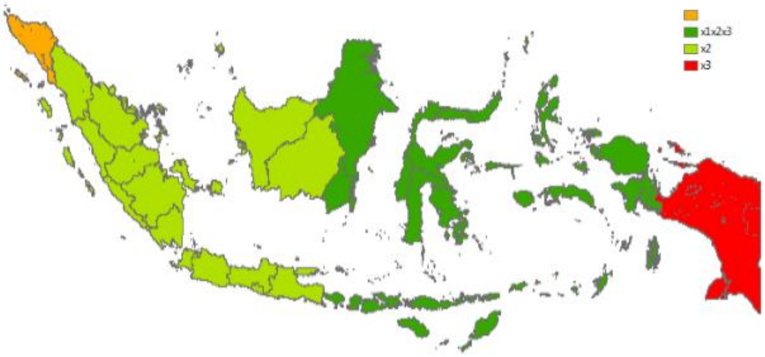

Figure 2. Map of the area group of geographically weighted regression with bisquare kernel weighting function

C. Selection of the best model

Table 3. Comparison of the value of R^2 and AIC values GWR models using gaussian kernel weighting function and bisquare kernel weighting function

Model	R^2	AIC
GWR (gaussian)	87.82%	143.80
GWR (bisquare)	90.17%	137.81

Some of the criteria used in determining the best model is to look at the value of the coefficient of determination (R^2) and the value of the Akaike Information Criterion (AIC). The best model is the model that has the highest R^2 value and the smallest AIC value. Comparison between R2 and AIC for each model is available in Table 3, it is known that based on the value of R^2 and AIC values then the bisquare kernel weighting function a weighting function best in building the GWR models on malaria prevalence data by province in Indonesia in 2013. GWR models with gaussian kernel weighting has a R^2 value of 87.82%, which indicates that 87.82% of the variance of malaria prevalence can be explained by the model, in while the remaining 12.18% is explained by other variables outside the model. GWR models with bisquare kernel weighting have R^2 value of 90.17%, which indicates that 90.17% of the variation of malaria prevalence can be explained by the model, while the remaining 9.83% is explained by other variables outside the model.

4. Conclusion

a. GWR modeling using bisquare kernel weighting function on malaria prevalence data by province in Indonesia in 2013 provides better results than a gaussian kernel weighting function based on the value of R^2 and AIC of both models.

b. Variations variables explanatory affecting the prevalence of malaria in each region showed that the government of each region has the task of prioritizing specific programs in an effort to control malaria in the region, the government through the health personnel area can improve the dissemination and outreach to the community about ways to be effective in prevention mosquito bites and how to obtain effective treatment of malaria in the region.

References

1. KEMENKES, *Riskesdas 2013 dalam Angka,* Jakarta, Kementerian Kesehatan, 2013.
2. KEMENKES, *Buku Saku Menuju Eliminasi Malaria,* Jakarta, Kementerian Kesehatan, 2011.
3. Anselin L., *Spatial Econometrics,* Dallas, University of Texas, 1999.
4. Saefuddin A., Setiabudi N. A., Achsani N. A., On Comparisson between Ordinary Linear Regression and Geographically Weighted Regression: With Application to Indonesian Poverty Data, *European Journal of Scientific Research,* vol/no: 57(2), pp. 275-285, 2011.
5. Fotheringham A. S., Brunsdon C., Chartlon M., *Geographically Weighted Regression, The Analysis of Spatially Varying Relationships,* England, John Wiley & Sons, 2002.
6. Anselin L., *Spatial Econometrics,* Dallas, School of Social Science, 2009.
7. Leung Y., Chang-Lin M., Wen-Xiu Z., Statistical Test for Spatial Nonstationarity Based on The Geographically Weighted Regression Model, *Journal of Environment and Planning A,* vol/no: 32(1), pp. 9–32, 2000.