Modelling space of spread Dengue Hemorrhagic Fever (DHF) in Central Java use spatial durbin model

Dwi Ispriyanti¹, Alan Prahutama¹, Arkadina PN Taryono¹

¹Department of Statistics, Faculty of Science and Mathematics, Diponegoro University
Jl. Prof. Soedharto, SH, Tembalang, Semarang 50275, Indonesia

E-mail: dwiispriyanti@yahoo.com

Abstract. Dengue Hemorrhagic Fever is one of the major public health problems in Indonesia. From year to year, DHF causes Extraordinary Event in most parts of Indonesia, especially Central Java. Central Java consists of 35 districts or cities where each region is close to each other. Spatial regression is an analysis that suspects the influence of independent variables on the dependent variables with the influences of the region inside. In spatial regression modeling, there are spatial autoregressive model (SAR), spatial error model (SEM) and spatial autoregressive moving average (SARMA). Spatial Durbin model is the development of SAR where the dependent and independent variable have spatial influence. In this research dependent variable used is number of DHF sufferers. The independent variables observed are population density, number of hospitals, residents and health centers, and mean years of schooling. From the multiple regression model test, the variables that significantly affect the spread of DHF disease are the population and mean years of schooling. By using queen contiguity and rook contiguity, the best model produced is the SDM model with queen contiguity because it has the smallest AIC value of 494.12. Factors that generally affect the spread of DHF in Central Java Province are the number of population and the average length of school.

Keywords: DHF, Spatial Durbin Model, queen contiguity, rook contiguity

1. Introduction

Dengue fever is an acute febrile illness caused by dengue virus and is spread through an Aedes aegypti mosquito that has been infected with the dengue virus. Dengue virus incubation period in humans (intrinsic incubation) ranges from 3 to 14 days before symptoms appear, average clinical symptoms appear on the fourth until the seventh day, while the extrinsic incubation period (inside the mosquito body) lasts about 8-10 days. It is the most important in terms of morbidity, mortality, economic burden as well as disability life-years lost ([1], [2]). Dengue Hemorrhagic Fever (DHF) is one of the major public health problems in Indonesia. The number of sufferers and the extent of the spreading area is increasing along with the increasing mobility and population density. From year to year, DHF causes Extraordinary Event in most parts of Indonesia, including in Central Java. Many factors influence the spread of DHF diseases including host factors (vulnerability), environmental factors, demographic conditions, geographical location and mosquito species. Central Java consists of 35 districts / cities where each region is close to each other. Statistics method used to analyze any factors that influence the spread of dengue disease by involving the region in it is spatial regression. Most previous studies, where heterogeneities in socioeco-nomic variables were associated with differential risks of...
infection, considered significantly larger spatial scales (including whole countries). However some studies have also shown association between dengue incidence and socioeconomic conditions at household level ([3],[4]). Others, Delmelle et al [5] has been research a spatial model of socioeconomic and environmental determinant of dengue fever in Cali, Colombia. The results presented in this paper make two important contributions. First, both environmental factors impacting the spread and distribution of the vector and pathogen as well as socioeconomic factors shaping social vulnerability need to be considered in disease risk assessments. Second, the contribution of these factors varies spatially, thus demanding approaches that are able to capture this variability. It was may not have been identified by traditional regression analysis. This further underlines the importance of small-scale level studies.

In spatial regression modeling, there are spatial autoregressive model (SAR), spatial error model (SEM) and spatial autoregressive moving average (SARMA). Spatial durbin model is the development of spatial autoregressive model where spatial influence is not only in the dependent variable but also in independent variable [6].

Spatial weights are the most important element in spatial regression modeling. Spatial weights are usually written in matrix form. There are 6 types of spatial weights i.e. linear contiguity, rook contiguity, bishop contiguity, double linear contiguity, double rook contiguity and queen contiguity. In this research, the method that is going to used by researchers is Spatial Durbin Model (SDM) with weighting Queen contiguity and Rook contiguity. The dependent variable is the number of DHF patients in Central Java, while the independent variables are population, population density, number of hospitals, number of health centers and mean years of schooling in Central Java.

2. Literature Review

2.1 Multiple Linear Regression

Multiple linear regression is a development of a simple linear regression model with more than one independent variable [7]. The multiple linear regression equation with $k$ independent variables as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

The least squares method (Ordinary Least Square) was used to estimate multiple regression parameters. The least squares method can be achieved by deriving the $L$ function partially to $\beta$ and equating the result with zero, so that the error is obtained as small as possible. Obtained least squares estimator of $\beta$ as follows:

$$\hat{\beta} = (X^T X)^{-1} (X^T y)$$

According to Montgomery and Runger (2011) [7], the significance test of regression is intended to determine whether there is a linear relationship between the dependent variable and the independent variable.

Hypotesis:

- $H_0$: $\beta_1=\beta_2=\ldots=\beta_k=0$
- $H_1$: $\beta_j \neq 0$, atleast one $j$

Test of statistics:

$$F = \frac{SSR/k}{SSE/(n-k-1)}$$

Decision: $H_0$ refused if value of $F > F_{a,k,n-k-1}$.

Individual testing is used to examine whether there is the influence of each independent variable on the regression model or not [7]

Hypotesis:

- $H_0$: $\beta_j = 0$ (Parameters are not significant)
- $H_1$: $\beta_j \neq 0$ (Parameters are significant)

Test of statistics:
\[ t_{count} = \frac{\beta_j}{Se(\beta_j)} \]

Decision: \( H_0 \) refused if value of \( |t_{count}| > t_{0.01,n-k-1} \)

Regression model needs assumption, such as normal distribution of residual, non-autocorrelation and nonmulticollinierity. Visually, the normality test can be detected by viewing the spread of data (dots) on the diagonal axis of the P-P plot graph. If the data spreads around the diagonal line and follows the direction of the diagonal line, the regression model will meet the assumption of normality.

The formal test used to determine whether the study sample is a normal distribution type or not is Kolmogorov Smirnov Goodness of Fit Test on each variable. The data consists of a random sample of \( X_1, X_2, ..., X_n \) of size \( n \) with unknown distribution of the type denoted \( F(x) \).

Hypotesis:
\[
H_0: F(x) = F^*(x) \quad \text{(sample of data is normally distributed)}
\]
\[
H_1: F(x) \neq F^*(x) \quad \text{(sample of data is not normally distributed)}
\]

Test of statistics:
\[ T = \sup_x |F^*(x) - S(x)| \]

Where: \( S(x) \) is Cumulative frequency distribution function and \( F^*(x) \) is Cumulative probability of normal distribution. Decision: \( H_0 \) is rejected if \( T > q_{1-\alpha} \) where \( q_{1-\alpha} \) is the critical value obtained from the Kolmogorov-Smirnov table.

For checking non-autocorrelation assumption we can used Durbin Watson test. It is used to detect the presence of autocorrelation between errors \[8\].

Hypothesis
\[
H_0: \text{There is no autocorrelation between errors}
\]
\[
H_1: \text{There is an autocorrelation between errors}
\]

Test of statistics:
\[ d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \]

Decision: Receiving \( H_0 \) if \( d > d_U \) or \( d < (4-d_U) \). Accepting \( H_0 \) if \( d < d_L \) or \( d > (4-d_L) \). Without decision if \( d_L \leq d \leq d_U \) or \( (4-d_U) \leq d \leq (4-d_L) \). Where \( d_U \) is the upper limit and \( d_L \) is the lower limit obtained from the Durbin-Watson table.

To checking multicollinierity assumption we can used VIF value. According to Kleinbaum et al (2014) \[9\], tolerance and variance inflation factor (VIF) is a common method used to calculate the linear relationship between independent variables in multiple regression. If the VIF value of a variable exceeds 10, then it is said that the variable is highly correlated. The test statistics for VIF as follows:
\[ VIF_j = \frac{1}{1-R_j^2} \quad j=1,2,...,k \]

\( R_j^2 \) is The coefficient value of determination on \( j \)-variable

2.2 Spatial Regression

Spatial regression is a model formed from a classical regression that gets a spatial influence (location) within it. According to Lesage (1999) \[10\], a general model of spatial regression can be written as follows:
\[ y = \rho W_1 y + X\beta + u \]
\[ u = \lambda W_2 u + \epsilon, \epsilon \sim N(0, \sigma^2 I_n) \]

where: \( \rho \) is spatial coefficient lag of the dependent variable; \( W_i, W_2 \) are Spatial weighting matrix of size \( n \times n \); \( \lambda \) is coefficient of spatial error auto-regression; \( u \) is An error vector that has a spatial effect of size \( n \times 1 \); \( \epsilon \) is Error vector with size \( n \times 1 \)

From the general model equation of spatial regression, it can form several other models as follows:

1. If \( \rho = 0 \) and \( \lambda = 0 \) then it is called classical linear regression model with the equation formed:
\[ y = X\beta + \epsilon \]
2. If \( \rho \neq 0 \) and \( \lambda = 0 \) are called \textbf{Spatial Autoregressive Model (SAR)} with the equation formed:

\[
y = \rho W_1 y + X \beta + \varepsilon
\]

3. If \( \rho = 0 \) and \( \lambda \neq 0 \) it is called \textbf{Spatial Error Model (SEM)} with the equation formed:

\[
y = X \beta + u \\
u = \lambda W_2 u + \varepsilon
\]

4. If \( \rho \neq 0 \) and \( \lambda \neq 0 \) it is called \textbf{Spatial Autoregressive Moving Average (SARMA)}, the equation is formed:

\[
y = \rho W_1 y + X \beta + u \\
u = \lambda W_2 u + \varepsilon
\]

2.3 \textit{Spatial Effect Test}

To determine whether there is autocorrelation or spatial dependence between locations, then spatial autocorrelation test is done by Moran's method [11].

**Hypothesis:**

\( H_0: \) There is spatial autocorrelation between locations  \\
\( H_1: \) There is no spatial autocorrelation between locations

**Test of statistics:**

\[
Z_{count} = \frac{I - E(I)}{\sqrt{\text{var}(I)}}
\]

Where:

\[
I = \frac{n}{S_0} \frac{(x - \bar{x})^T W (x - \bar{x})}{(x - \bar{x})^T (x - \bar{x})} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{with} \quad S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} ; \quad E(I) = -\frac{1}{n-1}
\]

\[
\text{var}(I) = \frac{n^2 S_1 - n S_2 + 3 S_0^2}{(n^2 - 1) S_0^2} \quad [E(I)]^2 ; \quad S_2 = \sum_{i=1}^{n} \left( \sum_{j=1}^{n} w_{ij} + \sum_{j=1}^{n} w_{ji} \right)^2 ; \quad S_1 = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (w_{ij} + w_{ji})^2
\]

To determine whether there is spatial heterogeneity, spatial heterogeneity testing is performed using Breusch-Pagan Test (Anselin, 1988).

**Hypothesis:**

\( H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2 = \sigma^2 \)  \\
\( H_1: \) Minimally has one \( \sigma_i^2 \neq \sigma^2 \)

**Formula of Breusch-Pagan test:**

\[
BP = \frac{1}{2} \left[ f^T Z (Z^T Z)^{-1} Z^T f \right] \sim \chi^2_p
\]

\[
f_i = \left( \frac{\hat{e}_i^2}{\hat{\sigma}^2} - 1 \right)
\]

Where \( \sigma^2 \) is variance of the model to be tested; \( \hat{e}_i \) is Error for i-th observation; \( Z \) is an \( n \times (k + 1) \) matrix that contains a constant vector. Decision: Reject \( H_0 \) if \( BP > \chi^2_{\alpha,p} \)

2.4 \textit{Spatial Autoregressive Model (SAR)}

Spatial autoregressive or spatial lag is a model that combines a simple regression model with spatial lag on the dependent variable [11]. The SAR equation as follows:

\[
y = \rho W_1 y + X \beta + \varepsilon, \varepsilon \sim N(0, \sigma^2 I_n)
\]

Obtained estimation of SAR model parameters as follows:

\[
\hat{\beta} = (X^T X)^{-1} X^T (y - \rho W_1 y)
\]
\[ \hat{\sigma}^2 = \frac{\varepsilon^T \varepsilon}{n} \]
\[ \hat{\beta} = (y^T W_1^T W_1 y)^{-1} (y^T W_1^T y - \beta^T X^T W_1 y) \]

2.5 Spatial Durbin Model (SDM)

The SDM model is a spatial regression model that not only has spatial lag on the dependent variable but also has spatial lag on the independent variable [10]. According to LeSage & Pace [6], the SDM model has the following form of equation:

\[ y = \rho W y + \alpha 1_n + X\beta + WX\theta + \varepsilon, \varepsilon \sim N(0, \sigma^2 I_n) \]

Or can be written as follows:

\[ y = \rho W y + \Delta + \varepsilon \]

Where:

\[ Z = \begin{bmatrix} 1_n & X & WX \end{bmatrix} \]

\[ \delta = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \]

\[ a = \text{The vector of constant parameter (n x 1)} \]

\[ \theta = \text{The vector of the independent variable spatial lag parameter (k x 1)} \]

\[ 1_n = \text{The vector containing the number 1 (n x 1)} \]

Obtained estimation of SDM model parameters as follows:

\[ \hat{\delta} = (Z^T Z)^{-1} Z^T (I_n - \rho W) y \]
\[ = (Z^T Z)^{-1} Z^T y - \rho (Z^T Z)^{-1} Z^T W y \]

\[ \hat{\sigma}^2 = \frac{\varepsilon^T \varepsilon}{n} \]
\[ \hat{\beta} = (y^T W^T W y)^{-1} (y^T W^T y - \delta^T Z^T W y) \]

2.6 Spatial Weights

According to Viton (2010), spatial weights are the relationship of neighbouring between areas with one another region. Spatial weights are usually written in the following matrix form:

\[ W = \begin{bmatrix} w_{11} & w_{12} & w_{13} & \cdots & w_{1n} \\ w_{21} & w_{22} & w_{23} & \cdots & w_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & w_{n3} & \cdots & w_{nn} \end{bmatrix} \]

There are 6 types of Spatial Weights [10]:

1. Linear Contiguity: Define \( W_{ij} = 1 \) for entities that share a common edge to the immediate right or left of the region of interest.
2. Rook Contiguity: Define \( W_{ij} = 1 \) for regions that share a common side with the region of interest.
3. Bishop Contiguity: Define \( W_{ij} = 1 \) for entities that share a common vertex with the region of interest.
4. Double Linear Contiguity: For two entities to the immediate right or left of the region of interest.
5. Double Rook Contiguity: For two entities to the right, left, north and south of the region of interest define \( W_{ij} = 1 \).
6. Queen Contiguity: For entities that share a common side or vertex with the region of interest define \( W_{ij} = 1 \).
2.7 The Best Model Selection use Akaike’s Information Criterion (AIC)
The AIC method is a method that can be used to select the best regression model based on the Maximum Likelihood Estimation method. The best model is the model that has the smallest AIC value. According to Hu (2007), the general formula of AIC is shown in the following equation:

$$AIC = -2 \log(L(\hat{\theta}|y)) + 2k$$

Where: $L(\hat{\theta}|y)$ is the function of likelihood parameter in estimation; $k$ is The number of parameters in the estimation.

3. Result And Discussion
3.1. Multiple Regression Model
The model formed on multiple regression analysis as follows:

$$\hat{y} = -1088,894 + 0,001x_3 + 122,482x_5$$

Based on the results, it can be concluded that the test of regression significance multiple regression model has been appropriate to describe the relationship between dependent and independent variables. In the regression coefficient test individually, variables $X_3$ and $X_5$ have a significant effect on the regression model. Normality assumptions, non autocorrelation and multicollinearity are met.

3.2. Spatial Weights (W)
Spatial Weights Matrix can be described with a connectivity graph that can be seen in this following picture:

![Figure 1. Connectivity Chart of Queen Contiguity](image1)

![Figure 2. Connectivity Chart of Rook Contiguity](image2)

Figure 1 and 2 inform the connectivity between regions. The graph color shows the number of neighboring/neighborhood relationships, while the high graph shows the frequency/number of areas.

3.3. Spatial Durbin Model (SDM)
3.3.1. SDM Model with Queen Contiguity. Obtained initial model of SDM as follows:

$$\hat{y}_i = 0,61347 \sum_{j=1}^{n} w_{ij}y_j - 536,42 + 0,00067379x_{i3} + 125,78x_{i5} - 0,00027639 \sum_{j=1}^{n} w_{ij}x_{j3} - 80,857 \sum_{j=1}^{n} w_{ij}x_{j5}$$

Where $w_{ij}$ is queen contiguity.

From the output, there is value $F_{count} = 11.1435$ where the value is greater than $F_{table} = 2.545386$. It can be concluded that the SDM model has been appropriate to describe the relationship between the dependent variable and the independent variable.
Based on Table 1, it can be seen that the parameters ρ, β3 and β5 have a significant effect on the model because it has a p-value less than 0.05. The absence of a lag of independent variable with significant weighting causes the parameter estimation results using SDM to be insignificant but on the identification of the Moran’s I value, it identifies the spatial dependence on the independent variable. Based on the output, it can be seen that the sig value. (K-S) = 0.195 > α = 0.05 and from the Q-Q plot chart, it can be seen that the plots are spreading around the normal line. So it can be inferred to receive H0 which means that the residual follows the normal distribution.

From the output, it is obtained that the value of Breusch-Pagan Test of 14.2548 where the value is more than (χ^2 = 9.488) and p-value equals to 0.00625 > α = 0.05. So, it can be concluded that there is spatial heterogeneity.

3.3.2. SDM Model with Rook Contiguity

Obtained initial model of SDM with Rook Contiguity as follows:

\[ \hat{y}_t = 0.5861 \sum_{j=1}^{n} w_{ij} y_j - 533.87 + 0.00067062 x_{t3} + 126.1 x_{t5} - 0.00027226 \sum_{j=1}^{n} w_{ij} x_{j3} - 79.955 \sum_{j=1}^{n} w_{ij} x_{j5} \]

Where \( w_{ij} \) is rook contiguity.

From the output, it is obtained that value F = 10.5336 where the value is greater than F table is 2.545386. It can be concluded that the SDM model has been appropriate to describe the relationship between the dependent variable and the independent variable.

| Parameter | coefficient | p-value |
|-----------|-------------|---------|
| Rho       | 0.5861      | 0.00085043 |
| Intercept | -533.87     | 0.4255469  |
| β3        | 0.00067062  | 3.18 x 10^10 |
| β5        | 126.1       | 0.0008614   |
| θ3        | -0.00027226 | 0.2789902   |
| θ5        | -79.955     | 0.3394002   |

Based on Table 2, it can be seen that the parameters ρ, β3 and β5 have a significant effect on the model because it has a p-value less than 0.05. The absence of a lag of independent variable with significant weighting causes the parameter estimation results using SDM to be insignificant but on the identification of the Moran’s I value, it identifies the spatial dependence on the independent variable. Based on the output, it can be seen that the sig value. (K-S) = 0.195 > α = 0.05 and from the Q-Q plot chart, it can be seen that the plots are spreading around the normal line. So, it can be inferred to receive H0 which means that the residual follows the normal distribution.
From the output, it is obtained that the value of Breusch-Pagan Test of 14.2531 where the value is more than ($\chi^2_{0.05,4} = 9.488$) and p-value equals to $0.00653 > \alpha = 0.05$. So, it can be concluded that there is spatial heterogeneity.

3.4. The Selection Of The Best Model

The selection of the best model uses AIC value criteria. Here is a comparison of AIC values between SDM models with weighted queen and rook.

| No | Model          | AIC  |
|----|----------------|------|
| 1  | SDM Queen      | 494.12 |
| 2  | SDM Rook       | 495.16 |

From Table 3, it is found that the smallest AIC value is SDM with queen contiguity, so it can be concluded that SDM with queen contiguity is better in analyzing the factors that influence the spread of DHF disease.

4. Conclusion

From the results of the analysis and discussion that has been done in the previous chapter, it can be drawn some conclusions:
1. Factors that generally affect the spread of DHF in Central Java Province are the number of population and the average length of school.
2. The spread of dengue disease in Central Java as well as several factors that influence it has spatial effects so that modeling for SAR and HR can be done.
3. Based on the value of Akaike Information Criterion (AIC), modeling with SAR is better than the human resources in determining the model of factors that affect the spread of DHF in Central Java Province.

References

[1] LaBeaud AD 2008 Why arboviruses can be neglected tropical diseases PloS Negl Tropical Disease 2 247.
[2] Hotez PJ, Molyneux DH, Fenwick A, Savioli L and Takeuchi T 2008 A Global Fund to Fight Neglected Tropical Diseases: is the G8Hokkaido Toyako 2008 Summit ready? PLoS Negl. Trop. Dis. 2 220.
[3] Waterman SH, Novak RJ, Sather GE, Bailey RE, Rios I and Gubler DJ 1985 Dengue transmission in two Puerto Rican communities in 1982. Am. J. Trop. Med. Hyg. 34 625–632.
[4] Reiter P, Lathrop S, Bunning M, Biggerstaff B, Singer D, Tiwari T, Baber L, Amador M, Thirion J, Hayes J, Seca C, Mendez J, Ramirez B, Robinson J, Rawlings J, Vorndam V, Waterman S, Gubler D, Clark G, and Hayes E 2003 Texas lifestyle limits transmission of dengue virus. Emergency Infection Disease 9 86-89.
[5] Delmelle E, Hagenlocher M, Kienberger S and Casas I 2016 A Spatial model socioeconomic and environmental determinants of dengue fever in Cali, Colombia Acta Tropica 164 169-176.
[6] LeSage J, Pace RK 2009 Introduction to Spatial Econometrics Taylor & Francis Group USA.
[7] Montgomery, Douglas C, Runger, George C 2011 Applied Statistics and Probability for Engineers John Wiley & Sons Inc. USA.
[8] Gujarati and Damodar N 1978. *Ekonometrika Dasar* Jakarta: Erlangga.
[9] Kleinbaum and David G 2014 *Applied Regression Analysis and Other Multivariable Methods, Fifth Edition* Boston: Cengage Learning.
[10] LeSage and James P 1999 *The Theory and Practice of Spatial Econometrics* Department of Economics University of Toledo.
[11] Anselin L 2003 *A Companion to Theoretical Econometrics*. UK: Blackwell Publishing Ltd.