Spatial Durbin Model for Identifying the Factors Affecting Diarrhea in East Java

Dyra Fitri Kesuma Dewi[1], Asep Saefuddin[2], Utami Dyah Syafitri[3]
{dyrafitrikd@gmail.com[1], asaefuddin@gmail.com[2], utamids@gmail.com[3]}

Department of Statistics, IPB University, Bogor, Indonesia[1,2,3]

Abstract. Diarrhea is an infectious disease that has been a major health problem in Indonesia. To reduce the number of diarrhea cases in Indonesia, the government must examine the factors causing diarrhea in order to make the right decision. An analysis that can be used forthis purpose is regression analysis. When regression analysis is conducted with spatial data, the residuals are usually spatially autocorrelated. There are some methods which can handlethis spatial effect, such as spatial autoregressive (SAR), spatial error (SEM), and spatial durbin(SDM) models. In this research, those models were used to model the number of diarrhea cases in East Java in 2017. SDM was the best model because it had the smallest AIC and the greatest pseudoR^2. There were two variables significantly affecting diarrhea cases in East Java, percentage of people with clean and healthy lifestyle, and spatial lag of the population's density.

Keywords: diarrhea, spatial autoregressive, spatial durbin, spatial error, spatial regression.

1 Introduction

Diarrhea is one of the major health problems in Indonesia. In 2016, there were about 3.17 million people who suffered from diarrhea in Indonesia. This number increased to 4.27 million people in 2017 [1]. Diarrhea itself is caused by an infection of the intestine which can lead to an abnormal condition of the intestinal function. The symptom of diarrhea is an increased number of watery stools that happens more than three times a day. Young children who have limited access to clean toilet, safe drinking water, and health center are the most likely to suffer from diarrhea[2].

The factors affecting the number of diarrhea casesmust be identified so that the government can implement the right decision to reduce the number of people suffered from diarrhea. The analysis which can be used to identify the factors affecting the number of diarrhea cases is regression analysis. Montgomery et al. stated that regression analysis is a statistical method used to analyze causal relationship between variables [3]. There are some assumptions that must be fulfilled to use linear regression analysis. The assumptions are linear relationship between dependent and independent variable(s), normality of errors, homoscedasticity, and independence of errors [4].

Adjacent regions usually have similar characteristics. Thus, the number of diarrhea cases in a region is possibly influenced by the number of diarrhea cases in the regions surrounding it. When regression analysis is used to model a data consisted of correlated regions as its observations, the residuals are usually autocorrelated. If this spatial effect is ignored while using linear regression analysis, the assumption that the errors are independent

ICSA 2019, August 02-03, Bogor, Indonesia

DOI 10.4108/eai.2-8-2019.2290532

is violated. There are some methods that have been developed to overcome this spatial effect, such as spatial autoregressive model (SAR) and spatial error model (SEM). Various studies on SAR and SEM have been done in many different fields. In poverty, Higazi et al. applied SEM to identify the factors affecting poverty ratios in Egypt [5]. Meanwhile, Andresen applied SAR for modeling the factors affecting crime rates in Vancouver [6].

Later, Anselin introduced a special case of SAR model which also considers spatial lag effect on explanatory variables, known as spatial durbin model (SDM) [7]. This model was developed because in reality spatial effect does not only occur in the dependent variable, but also in the independent variable. A study of Nugroho showed that SDM is better than SAR model to identify the factors affecting national examination results of Madrasah Aliyah Negeri on Java Island [8].

In this research, SAR, SEM, and SDM were used to examine the factors affecting the number of diarrhea cases. The case study was diarrhea cases in East Java in 2017 as East Java was the province with the second highest number of diarrhea cases in Indonesia in 2017. The results of the three models were compared based on the value of AIC and Pseudo R^2.

2 Materials

The data used in this research was a secondary data collected from the publication of Ministry of Health of East Java entitled "Profil Kesehatan Jawa Timur 2017". There were 38 districts and cities in East Java as the observations. The variables are listed at Table 1.

Table 1. List of variables.

Variables	Description	Unit of measurement
Y	The number of diarrhea cases	Cases
X_1	Population's density	Persons/ km^2
X_2	Percentage of people with healthy lifestyle	%
X_3	Percentage of healthy houses	%
X_4	Percentage of houses with access to safe drinking water	%
X_5	Percentage of safe drinking water suppliers	%
X_6	Percentage of houses with clean toilet	%

3Methods

Data analysis is conducted using R 3.4.3 software. The procedures are as follows:

1. Mapping the distribution of diarrhea cases in East Java.
2. Forming a spatial weight matrix using queen contiguity method. The form of the spatial weight matrix is shown at equation (1).

$$\mathbf{W}=\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{21} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{bmatrix} \#(1)$$

with n is the number of regions or observations. LeSage illustrate the procedures to form a spatial weight matrix by queen contiguitymethod as follows [9]:

a. Define w_{ij} = 1 for entities that share a common side or vertex with theregion of interest.
b. Standardize each row of the weight matrix by dividing every element ofthe matrix with the number of elements in row as shown in equation (2).

$$w_{ij}^{*}=\frac{w_{ij}}{\sum_{j=1}^{n} w_{ij}};w_{ij}^{*}\in\mathbf{W}^{*} \quad \#(2)$$

3. Checking the existence of spatial autocorrelation on the number of peoplesuffered from diarrhea and its risk factors by using Moran's Index. The hypothesis is as follows:
   H0 : There's no spatial autocorrelation
   H1 : There's spatial autocorrelation
   The test statistics of Moran I is written at equation (3).

$$z=\frac{\text{I-E(I)}}{\text{V(I)}} \#(3)$$

   where

$$\text{I}=\frac{(\mathbf{x}-\bar{\mathbf{x}})'\mathbf{W}^{*}(\mathbf{x}-\bar{\mathbf{x}})}{(\mathbf{x}-\bar{\mathbf{x}})'(\mathbf{x}-\bar{\mathbf{x}})}$$

   with**x** is the vector of the variable of interest, $\bar{\mathbf{x}}$ is the vector containing the meanof the variable of interest, and **W*** is the spatial weight matrix that has been row-standardized.
4. Estimating multiple linear regression parameters using ordinary least square estimation and checking the assumptions of regression based on the residuals.
5. Choosing spatial dependence model by doing Lagrange Multiplier test.
   a. Spatial autoregressive model
   H0: There is no spatial dependence on the number of people suffered from diarrhea.
   H1: There is spatial dependence on the number of people suffered from diarrhea.
   The test statistics is written at equation (4).

$$\text{LM}_{\rho}=\left(\frac{\boldsymbol{\varepsilon}'\mathbf{W}^{*}\mathbf{y}}{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}n^{-1}}\right)^{2}\frac{1}{\text{D}} \#(4)$$

   where

$$\text{D}=\left\{\left(\mathbf{W}^{*}\mathbf{X}\boldsymbol{\beta}'\right)\left[\mathbf{I}-\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\left(\mathbf{W}^{*}\mathbf{X}\boldsymbol{\beta}\right)\sigma^{-2}\right\}+\text{tr}(\mathbf{W}^{*'}\mathbf{W}^{*}+\mathbf{W}^{*}\mathbf{W}^{*})$$

   The null hypothesis is rejected if $LM_{\rho}>\chi^{2}_{(1)}$
   b. Spatial error model
   H0: There is no spatial dependence on error.

H1: There is spatial dependence on error.
The test statistics is shown at equation (5).

$$LM_e=\left(\frac{\boldsymbol{\varepsilon}'\mathbf{W}\boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}\mathrm{n}^{-1}}\right)^2\frac{1}{\mathrm{tr}(\mathbf{W}'\mathbf{W}+\mathbf{WW})}\ \#(5)$$

where

$$\mathrm{D}=\left\{(\mathbf{WX}\boldsymbol{\beta}')\left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right](\mathbf{WX}\boldsymbol{\beta})\sigma^{-2}\right\}+\mathrm{tr}(\mathbf{W}'\mathbf{W}+\mathbf{WW})$$

The null hypothesis is rejected if $LM_e>\chi^2_{(1)}$.

6. Estimating SAR parameters using maximum likelihoood estimation and checking the assumptions based on the residuals.Anselinexpressed the model of SAR as written at equation (6).

$$\boldsymbol{y}=\rho\mathbf{W}^*\boldsymbol{y}+\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\#(6)$$

In this spesification, ρ is the coefficient of spatially lagged dependent variable, and the other notations are as explained before.By using maximum likelihood estimation, the estimator of $\boldsymbol{\beta}$ in the spatial autoregressive model is shown at equation (7).

$$\hat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}-(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\rho}\mathbf{W}^*\boldsymbol{y}\#(7)$$

The estimation of ρ can't be obtained by maximizing equation (7) as it does not have a close-form solution. Thus, the estimate of ρ is obtained by doing numerical iteration to get an estimate that maximizes the log-likelihood function.

7. Estimating SEM parameters using maximum likelihoood estimation and checking the assumptions based on the residuals. The model of SEM is expressed as in equation (8).

$$\boldsymbol{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{u}$$
$$\boldsymbol{u}=\lambda\mathbf{W}^*\boldsymbol{u}+\boldsymbol{\varepsilon}\ \#(8)$$

where u is the vector of residuals having spatial autocorrelation and λ is the coefficient of spatial error autocorrelation. By using maximum likelihood estimation, the estimator of $\boldsymbol{\beta}$ in the spatial error model is shown at equation (9).

$$\hat{\boldsymbol{\beta}}=\left[(\mathbf{X}-\lambda\mathbf{W}^*\mathbf{X})'(\mathbf{X}-\lambda\mathbf{W}^*\mathbf{X})\right]^{-1}(\mathbf{X}-\lambda\mathbf{W}^*\mathbf{X})(\boldsymbol{y}-\lambda\mathbf{W}^*\boldsymbol{y})\#(9)$$

The estimation of λ can't be done by maximizing equation (9) as it does not have a close-form solution. Thus, the estimate of λ is obtained by doing numerical iteration to get an estimate that maximizes the log-likelihood function.

8. Estimating SDM parameters using maximum likelihoood estimation and checking the assumptions based on the residuals.Bekti et al. [10] showed that SDM model can be expressed in matrix notation as in equation (10).

$$\boldsymbol{y}=\rho\mathbf{W}^{*}\boldsymbol{y}+\mathbf{Z}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\ \#(10)$$

where

$$\mathbf{Z}=[\mathbf{I}\mathbf{X}\mathbf{W}^{*}\mathbf{X}]$$

where u is the vector of residuals having spatial autocorrelation and λ is the coefficient of spatial error autocorrelation. By using maximum likelihood estimation, the estimator of $\boldsymbol{\beta}$ in the spatial durbin model is shown at equation (11).

$$\hat{\boldsymbol{\beta}}=\left(\mathbf{Z}^{'}\mathbf{Z}\right)^{-1}\mathbf{Z}^{'}\boldsymbol{y}-\left(\mathbf{Z}^{'}\mathbf{Z}\right)^{-1}\mathbf{Z}^{'}\hat{\rho}\mathbf{W}^{*}\boldsymbol{y}\#(11)$$

The estimation of ρ can't be obtained by maximizing equation (11) as it does not have a close-form solution. Thus, the estimate of ρ is obtained by doing numerical iteration to get an estimate that maximizes the log-likelihood function.

9. Choosing the best model by using model selection criteria. Model selection criteria used in this study are pseudo-R^2 and Akaike's Information Criterion (AIC). Pseudopseudo-R^2 is the value of R^2 obtained from regressing response variable y and $\hat{y}$. The model is better if the pseudo-R^2 is greater. The formula of AIC as a model selection's criteria is shown at equation (12).

$$AIC=n+n\log(2\pi)+n\log\left(\frac{SSE}{n}\right)+2p\#(12)$$

SSE is the sum of squared error ($\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$) and p is the number of parameters. The model is better if the AIC is smaller.

10. Interpret the results.

4Results and Discussion

East Java was one of the provinces with the highest number of diarrhea cases in Indonesia in 2017. The total number of diarrhea cases registered in health facilities of East Java in 2017 was 841 874 incidents. The distribution pattern of diarrhea cases of East Java in 2017 can be seen in Figure 1. Regions with darker color indicated a higher number of diarrhea cases while regions with lighter color indicated a lower number of diarrhea incidents.

The regions having the highest number of diarrhea cases (60 – 70 thousands cases) in 2017 were Sidoarjo and Mojokerto. There were 65 543 diarrhea cases in Sidoarjo and 64 468 diarrhea cases in Mojokerto. The regions surrounding Sidoarjo and Mojokerto also had a quite high number of diarrhea cases although it was not as high as the number of diarrhea cases in Sidoarjo and Mojokerto. It indicated that the diarrhea-causing bacterias in Sidoarjo and Mojokerto spread to other regions surrounding them. Based on the map in Figure 1, it could be visually seen that there was a spatial effect of diarrhea cases between one region and other regions as nearby regions had a relatively similar number of diarrhea cases.

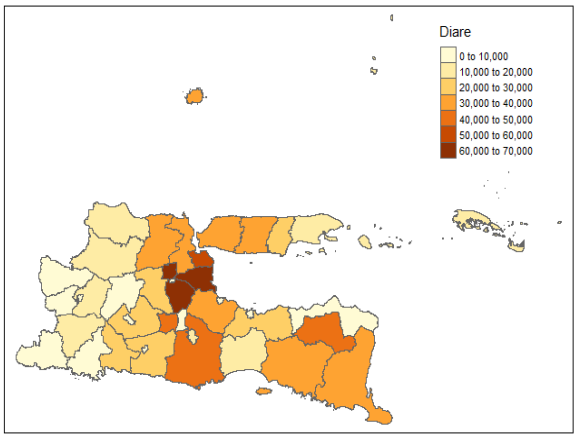

Fig. 1. Diarrhea cases distribution in East Java in 2017

4.1 Moran I Test

The results of spatial autocorrelation test by using Moran I test is shown at Table 2. Based on Table 2, it could be concluded that there were one variable which had a significant spatial autocorrelation at α = 5% and three variables which had a significant spatial autocorrelation at α = 10%. The variables which had a significant spatial autocorrelation at α = 5% was Y (the number of diarrhea cases), while the variables which had a significant spatial autocorrelation at α = 10% were X_2 (percentage of people with healthy lifestyle), X_3 (percentage of healthy houses), and X_4 (percentage of houses with access to safe drinking water) .

Table 2. Moran index of each variable.

Variables	Moran I	P-value
Y	0.2097	0.0338**
X_1	0.0647	0.2351
X_2	0.1727	0.0598*
X_3	0.1670	0.0662*
X_4	0.1656	0.0641*
X_5	-0.1313	0.7860
X_6	0.1076	0.1444

**significant at α = 5%; *significant at α = 10%

4.2Multiple Linear Regression

The number of diarrhea cases is a count or discrete variable. A count variable tends to follow Poisson distribution. As the number of diarrhea has a high enough mean (λ in Poisson distribution), it can be approximated as a normal distribution. The factors that influence the number of diarrhea incidents can be identified by using linear regression analysis assuming the data is normally distributed. Before doing a regression analysis, multicollinearity checking of the explanatory variables must be carried out. Multicollinearity is a condition where there is a

strong relationship between explanatory variables. If there is multicollinearity between explanatory variables of a regression model, the parameter estimates will produce a large standard error. Hence, the regression model might have high coefficient of determination but the parameter estimator are not likely to be significant. In this study, multicollinearity checking would be carried out by using variance inflation factor (VIF). The VIF value of each variable in the data is listed on Table 3. Multicollinearity exists if the VIF value of a variable is more than 10. Based on Table 3, there was no variable which had a VIF value more than 10. It could be concluded that there was no multicollinearity problem in the regression model.

Table 3. Initial regression model.

Coefficients	Estimate	T value	Pr(>|t|)	VIF
Intercept	9 701.016	0.371	0.713	
X_1	-1.122	-0.654	0.518	1.556
X_2	195.662	0.897	0.377	1.249
X_3	-97.349	-0.422	0.676	1.629
X_4	49.898	0.205	0.839	2.482
X_5	62.107	0.326	0.747	1.065
X_6	25.486	0.086	0.932	3.007

**significant at α = 5%; *significant at α = 10%

The normality of residuals can be formally tested by using Shapiro-Wilk normality test. The null hypothesis (H0) of this test is that the residuals are normally distributed while the alternative hypothesis (H1) of this test is that the residuals are not normally ditributed. The Shapiro-Wilk value of the test was 0.915 and the p-value was 0.007. The p-value of the normality test was less than α = 5%, so the null hypothesis was rejected. It could be concluded that the residuals were not normally distributed.

The homogeneity of variance can be formally tested by using Breusch Pagan (BP) test. The null hypothesis of the BP test is that the variance of the residuals is homogeneous while the alternative hypothesis is that the variance of the residuals is not homogeneous. The BP value of the test was 5.709 and the p-value was 0.4565, more than α = 5%. Therefore, the null hypothesis was not rejected. It meant that the variance of the residuals was homogeneous.

The assumption of error independency can be formally tested by using runs test. In the runs test, the null hypothesis is that there is no autocorrelation in the residuals while the alternative hypothesis is that there is autocorrelation in the residuals. The p-value of this test is 0.003, smaller than α = 5%, so the null hypothesis was rejected. It could be concluded hat there was a significant autocorrelation in the residuals. The violation of this assumption might be caused by the presence of spatial autocorrelation in the dependent or in the independent variables as shown at the result of Moran index test in the previous explanation.

4.3Lagrange Multiplier Test

The value of LM statistics for SAR model was 3.942 with a p-value of 0.047. The p-value was smaller than 5% so there was a significant spatial dependency on the lag of dependent variable at α = 5%. Meanwhile, the value of LM statistics for SEM model was 4.511 with a p-value of 0.034. The p-value was smaller than 5% so there was also a significant spatial dependency of the residuals at α = 5%. Based on the result obtained from the lagrange

multiplier test of the regression model, both SAR and SEM were suitable to model the number of diarrhea cases in East Java.

4.4Spatial Autoregressive Model

Parameter estimates of the spatial autoregresive model can be seen in Table 4. From Table 4, it can be seen that ρ coefficient had a p-value smaller than α = 5%, so the null hypothesis was rejected. It means that spatial dependency on the lag of dependent variable significantly influenced the number of diarrhea incidents in East Java. There was only one explanatory variable which significantly affected the number of diarrhea incidents at α = 10%. The variable was X_1, the population's density.

Table 4. Spatial autoregressive model.

Coefficients	Estimate	Z value	Pr(>|t|)
Intercept	-61.060	-0.074	0.940
X_1	-2.201	-1.661	0.096*
X_2	207.464	1.244	0.213
X_3	-121.673	-0.698	0.485
X_4	52.792	0.307	0.758
X_5	62.144	0.517	0.604
X_6	52.043	0.225	0.821
ρ	0.14093		0.040**

**significant at α = 5%; *significant at α = 10%

The Shapiro-Wilk value of the residuals was 0.914 and the p-value was 0.006, which was smaller than α = 5%. It can be concluded that the null hypothesis of normality test was rejected. In other words, the residuals were not normally distributed. The BP value of the residuals was 5.222 and the p-value was 0.515. The p-value of the BP test was more than α = 5%, so the null hypothesis of the homogeneity of variance assumption was not rejected. It means that the variance of the residuals was homogeneous. Lastly, the result of runs test had a p-value of 0.003. The p-value is smaller than α = 5%, so the null hypothesis of the error independency assumption was rejected. It meant that there was autocorrelation in the residuals.

4.5Spatial Error Model

Parameter estimates of the spatial error model can be seen in Table 5. From Table 5, it can be seen that λ coefficient had a p-value smaller than α = 5%, so the null hypothesis was rejected. It means that spatial dependency on the residuals of the regression model significantly influenced the number of diarrhea incidents in East Java. There was only one explanatory variable which significantly affected the number of diarrhea incidents at α = 5% in this model. The variable was X_1, the population's density.

Table 5. Spatial error model.

Coefficients	Estimate	Z value	Pr(>|t|)
Intercept	-9 619.480	0.429	0.667
X_1	-2.925	-2.128	0.033**
X_2	205.449	1.168	0.242
X_3	-174.327	-0.935	0.349
X_4	110.195	0.519	0.603
X_5	34.679	0.240	0.809
X_6	60.150	0.247	0.804
Λ	0.456		0.016**

**significant at α = 5%; *significant at α = 10%

The Shapiro-Wilk value of the residuals was 0.902 and the p-value was 0.003, which was smaller than α = 5%. It could be concluded that the null hypothesis of normality test was rejected. In other words, the residuals were not normally distributed. The BP value of the residuals was 3.622 and the p-value was 0.727. The p-value of the BP test was more than α = 5%, so the null hypothesis of the homogeneity of variance assumption was not rejected. It means that the variance of the residuals was homogeneous. Lastly, the result of runs test had a p-value of 0.323. The p-value was greater than α = 5%, so the null hypothesis of the error independency assumption was not rejected. It means that there was no autocorrelation in the residuals.

4.6Spatial Durbin Model

Parameter estimates of the spatial durbin model can be seen in Table 6. From Table 6, it can be seen that $\hat{\rho}$ had a p-value greater than α = 5%, so the null hypothesis was accepted. It means that spatial dependency on the lag of dependent variable did not significantly influence the number of diarrhea incidents in East Java. There were two variables that significantly influenced the number of diarrhea incidentsat α = 5%. The variables were X_2 (percentage of people with healthy lifestyle) and lag of X_1 (lag of population's density).

Table 6. Spatial durbin model.

Coefficients	Estimate	z value	Pr(>|t|)
Intercept	28 883.959	0.749	0.453
X_1	-1.113	-0.890	0.373
X_2	427.530	2.416	0.015**
X_3	-88.276	-0.565	0.571
X_4	89.272	0.513	0.607
X_5	142.647	1.101	0.270
X_6	-223.185	-1.013	0.310
Lag of X_1	17.954	4.499	0.000**
Lag of X_2	-38.804	-0.120	0.904
Lag of X_3	78.495	0.419	0.674
Lag of X_4	-483.577	-1.355	0.175
Lag of X_5	200.848	0.786	0.431
Lag of X_6	-355.939	-0.908	0.363
ρ	0.134		0.447

**significant at α = 5%

The Shapiro-Wilk value of the residuals was 0.960 and the p-value was 0.2017, which was greater than $\alpha = 5\%$. It can be concluded that the null hypothesis of normality test was not rejected. In other words, the residuals were normally distributed. The BP value of the residuals was 11.485 and the p-value was 0.487. The p-value of the BP test was more than $\alpha = 5\%$, so the null hypothesis of the homogeneity of variance assumption was not rejected. It means that the variance of the residuals was homogeneous. Lastly, the result of runs test had a p-value of 0.7422. The p-value was more than $\alpha = 5\%$, so the null hypothesis of the error independency assumption was not rejected.It means that there was no autocorrelation in the residuals.After considering spatial effect on the independent variables by applying spatial durbin model, the assumptions of error independency and normality were no longer violated. In this model, all assumptions of regression analysis had been fulfilled.

4.7The Best Model

AIC and pseudo-R^2 values of each model are listed at Table 7. Based on Table 7, spatial durbin model was the best model because it had the smallest AIC value (850.310) and the greatest pseudo-R^2 values (50.667%). Furthermore, spatial durbin model was the only model where all the assumptions of regression are fulfilled. The summary of the assumptions checking of each model is shown at Table 8. In spatial durbin model, the residuals were normally distributed, had a homogeneous variance, and were not autocorrelated. It could be concluded that spatial durbin model was the best model compared to SEM and SAR in modeling the number of diarrhea cases in East Java.

Table 7.AIC and Pseudo-R^2

Values	SEM	SAR	SDM
AIC	857.080	858.690	850.310
Pseudo-R^2	24.660%	20.000%	50.670%

Table 8. Assumptions checking of each model.

Assumptions	Normality	Homogeneity	Autocorrelation
SEM	No	Yes	Yes
SAR	No	Yes	No
SDM	Yes	Yes	Yes

According to parameter estimates of spatial durbin model shown at Table 6, the variables that significantly influenced the number of diarrhea casesat $\alpha = 5\%$ are x_2 (percentage of people with healthy lifestyle) and lag of x_1 (lag of population's density). The positive coefficient of x_2 showed that an increase in the percentage of people with healthy lifestyle would increase the number of diarrhea incidents in East Java. It was not in accordance with existing theories and studies. Based on previous studies, household with a clean and healthy lifestyle tend to be prevented from diarrheal disease as diarrhea-causing bacterias were usually spread in dirty environment. This discrepancy might occur because of unsuitable data selection. The data of the number of diarrhea cases was taken from health facilities, while the data of households with clean and healthy lifestyle was taken from the head of each neighborhood. As both data was taken separately at different times, the data

obtained might not be suitable for the purpose of this study. For example, supposed that there was someone who suffered from diarrhea at the beginning of the year. This person's data had been recorded in the data of diarrhea cases in the health facilities.When that person came to health facilities, the health worker who handled this person suggested them to carry out a clean and healthy lifestyle to prevent them from suffering from diarrhea. When the data of households with a clean and healthy lifestyle was collected by the head of the neighborhood at a later time, this person would be considered to have undergone a clean and healthy lifestyle as that person followed the advice of the health worker. Therefore, further research regarding the relationship between the number of diarrhea cases and healthy lifestyle is needed so that there is no misleading conclusion.

Spatial lag of population's density also had a positive regression coefficient. It means that a region tend to have a high number of diarrhea cases if its surrounding regions have a high population's density. In densely populated regions, diarrhea-causing bacterias can spread more easily as there are more contact between humans both intentionally and unintentionally. To limit the spread of diarrheal bacteria in a region surrounded by densely populated regions, the government should re-promote family planning programs to suppress the population's density.

5Conclusion and Suggestion

There was a significant spatial effect on the numberof diarrhea cases in East Java in 2017 based onMoran index and LM test. SDM was the best model to handle thisspatial effect because it had the smallest AIC value and greatest Pseudo-R^2 value compared to OLS, SAR, and SEM. Furthermore, by using SDM, the violation of assumptions in the regression caused by spatial effect in the data could be overcame. In the SDM model, the factors that significantly affectedthe number of diarrhea cases in East Java were thepercentage of people with healthy lifestyle and thepopulation's density of nearby regions.

In this research, the spatial weight matrix was formed based on the queen contiguity approach. For the next research, using another type of spatial weight matrix to model the number of diarrhea cases is suggested. Distance-based approach can be used as another method to form the spatial weight matrix.

References

[1]Ministry of Health Indonesia. 2018. Profil Kesehatan Indonesia Tahun 2017. Jakarta (ID): Kementerian Kesehatan RI

[2]Grimwood K, Forbes DA. 2009. Accute and persistent diarrhea. Pediatr Clin N Am. 56: 1343-1361

[3]Montgomery DC, Peck EA, Vining GG. 2012. Introduction to Linear Regression Analysis – 5th Edition. New Jersey(US): John Wiley and Sons

[4]Ernst AF, Albers CJ. 2017. Regression assumptions in clinical psychology research practice—a systematic review of common misconceptions. PeerJ. 5: 1-16

[5]Higazi SMF, Abdelhady D, Al-Oulfi SA. 2013. Application of spatial regression models to income poverty ratios in middle delta contiguous counties in Egypt. Pakistan Journal of Statistics and Operation Research. 9(1): 93-110

[6]Andresen MA. 2006. A spatial analysis of crime in Vancouver, British Columbia: a synthesis of social disorganization and routine activity theory. Le Géographe canadien. 50(1): 487-502

[7]Anselin L. 1988. Spatial Econometrics: Methods and Models. Netherlands (NLD): Kluwer Academic Publishers
[8]Nugroho S. 2016. Kajian spatial durbin model (SDM) untuk menelaah faktor-faktor yang mempengaruhi nilai Ujian Nasional (UN) di Madrasah Aliyah Negeri (MAN) [tesis]. Bogor (ID): Institut Pertanian Bogor
[9]Lesage J, Kelly P. 2009. Introduction to Spatial Econometrics. Boca Raton (FL): CRC Press
[10]Bekti RD, Nurhadiyanti G, Irwansyah E. 2014. Spatial pattern of diarrhea based on regional economic and environment by spatial autoregressive model. AIP Conference Proceeding. 1621(1): 454-461