Spatial autoregressive with a spatial autoregressive error term model and its parameter estimation with two-stage generalized spatial least square procedure

DRS Saputro¹, RY Muhsinin², PWidyaningisih³ and Sulistyaningsih⁴
¹,²,³,⁴Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Sebelas Maret, Jl. Ir. Soetami 36 A, Kentingan, Surakarta 57126, Indonesia
E-mail: dewiretnoss@staff.uns.ac.id

Abstract. Spatial relationship models often use dependence relationships into covariance structures through the autoregressive model. The autoregressive process is shown through the dependence relationship between a set of observations or a location from now on called the dependent spatial model. The Spatial dependent model is divided into two categories: spatial lag and spatial error. The spatial lag regression model is a model that considers dependent variables on an area with other areas associated with it, and the spatial error regression model is a model that takes into account the dependency of error values of an area with errors in other areas associated with it. Models with both dependencies are expressed as spatial autoregressive models with a spatial autoregressive error term (SAR-SAR). These dependencies resulted in the estimation of parameters by the ordinary least square method (OLS) resulting in inconsistent estimators. Therefore a special estimation method is required which results in a consistent estimate of the generalized spatial two-stage least square (GS2SLS). In this paper, we review the parameter estimation of SAR-SAR model with GS2SLS. To complete this paper, we also apply a SAR-SAR model in dengue hemorrhagic fever (DHF) case in Surakarta, Central Java.

1. Introduction
The regression model is used to find out factors which exert influence towards a response [1,2,3]. Such regression is known as Ordinary Least Square (OLS). It assumes an independent identical normally distributed residual. In case that an assumption is not fulfilled, it can be said that there is a spatial effect [4,5]. Meanwhile, Tobler’s first law of geography states that everything is related to everything else, but near things are more related than distant things [6]. It is a fundamental concept of a problem with spatial effects/position.

Spatial regression belongs to a regression model which includes spatial effects/position. The model follows an autoregressive process, which is indicated by the presence of dependence relationship among a set of observations called spatial dependence [7,8]. One of these is spatial regression models, such as simultaneous autoregressive models [9,10], which augment the standard linear regression model with an additional term that incorporates the spatial autocorrelation structure of a given data set. The spatial dependence is categorized into the spatial lag model and spatial error model [2]. The former takes dependence in the dependent variable of a spatial unit and the corresponding neighboring units into account, while the latter considers spatial dependence in the error term of a spatial unit and the corresponding neighboring units. A model having both forms of spatial dependence is termed a spatial autoregressive model with a spatial autoregressive error term (SAR-
SAR). Such forms of dependence cause parameter estimation with ordinary least square (OLS) method to yield inconsistent estimators. Therefore, a parameter estimation method which results in consistent estimators, generalized spatial two-stage least square (GS2SLS), is required [11].

An incidence associated with SAR-SAR model is dengue hemorrhagic fever. DHF is a vector-borne disease and is the leading cause of public health problems in Indonesia. According to Ministry of Health of the Republic of Indonesia [12], since 1968 until 2009, World Health Organization (WHO) had noted Indonesia as a country with the highest number of DHF cases in Asia. The excessive number of sufferers and broad transmission areas are increasing along with an increase in mobility and population density. DHF is caused by dengue viruses belonging to the genus flavivirus, the family Flaviviridae and is transmitted to humans through the bites of Dengue virus-infected Aedes mosquitoes. Surakarta is included as a Dengue-endemic area. According to [13], in 2008 out of 51 existing villages, 40 villages were Dengue-endemic areas with a total of 826 sufferers; twelve of which were dead. In 2009, there were 46 Dengue-endemic villages with a total of 684 sufferers; seven of which were dead. In 2010, 49 villages were claimed Dengue-endemic areas with a total of 533 sufferers; nine of which were dead.

About inconsistent parameter estimation in SAR-SAR model with OLS and several studies by its implementation, in this paper, we review the conceptual SAR-SAR model and its parameter estimation with GS2SLS. Moreover, we illustrated how the SAR-SAR model could be applied in Data’s DHF.

2. SAR-SAR Model
The SAR-SAR regression model follows an autoregressive process, which is indicated by the presence of dependence relationship among a set of observations or spatial units. It is written as

\[ y = \rho Wy + X\beta + u \]
\[ u = \lambda Mu + \epsilon \] (1)

where \( \epsilon \sim N(0; \sigma^2I) \), \( I \) is identity matrix, \( y \) is a vector of observations on the dependent variable \((nx1)\), \( X \) is a matrix of observations \((k)\) on the independent variables \(nxk\), \( \rho \) is spatial autoregressive coefficient and \( |\rho| < 1 \), \( \lambda \) is autoregressive coefficient and \( |\lambda| < 1 \), \( u \) is a vector of error terms \((n \times 1)\) assumed to have autocorrelation, \( \beta \) is a vector of parameters (regression coefficients) \((k \times 1)\), and \( W, M \) are standardized spatial weights matrix \((n \times n)(\sum_j w_{ij} = 1, \sum_j m_{ij} = 1)\) where \( n \) is the number of observations.

Below are the two types of the spatial regression model.
1. The spatial lag regression model
   The model is suitably used if the \( y \) value in spatial unit \( i \) is directly influenced by \( y \) value in its corresponding neighboring units. The model emerges from dependence in observation value of dependent variable in a spatial unit and the corresponding neighboring units. The spatial unit \( i \), for example, corresponds to the spatial unit \( j \); the observation value in the spatial unit \( i \) is therefore the function of that in the spatial unit \( j \), where \( i \neq j \) and finally the function \( y_i = f(y_j) \) is expressed. The spatial lag model is obtained if \( \rho \) in (1) is non-zero and \( \lambda \) in (1) is zero, that is \( y = \rho Wy + X\beta + \epsilon \), where \( \epsilon \sim N(0; \sigma^2I) \).
2. The spatial error regression model
   The model emerges from the presence of spatial dependence in the error term of a spatial unit and the corresponding neighboring units. It occurs in the case that some variables influencing dependent variable value but excluding in the model correlate among spatial units. The spatial error model is obtained if \( \lambda \) in the model (1) is non-zero and \( \rho \) in model (1) is zero, that is \( y = X\beta + u, u = \lambda Mu + \epsilon \), where \( \epsilon \sim N(0; \sigma^2I) \).

3. Results and Discussion
3.1 Two-Stage Least Squares (2SLS) and Generalized Spatial Two-Stage Least Square (GS2SLS) Procedure
Assumptions required in the model (1) include:
   a. matrix \((I - \lambda W)\) and \((I - \rho M)\) are nonsingular matrix,
   b. matrix \( X \) has full column rank,
c. a vector of innovation $\epsilon$ consists of random components $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ where $E(\epsilon_i^2) = \sigma^2$ and $E(\epsilon_i) = 0$ where $0 < \sigma^2 < b$ and $b < \sim$, and

d. $X$ does not correlate with $u$.

If variable $X$ correlates with $u$, the assumption stating that $E(u|X) = 0$ cannot be fulfilled. For that reason, $\hat{\beta}$ is not a consistent estimator for $\beta$. In order to obtain a consistent estimator, a procedure called instrumental variables is needed. Such procedure requires instrumental variables which should meet some criteria. $H$, for example, is matrix $n \times p$ which belongs to instrumental variable matrix having $p$ instrumental variable. $H$ should meet the following criteria:

a. $H$ does not correlate with $u$ or $E(u|H) = 0$,

b. $H$ should have at least $k$ variables or $p \geq k$ in such a way that $H$ correlates with $X$.

The more specific procedure is Two Stage Least Squares (2SLS). It employs OLS carried out in two following steps:

1. Doing regression for $X$ on instrumental variable matrix, and consequently, it results in the following equation.

$$\bar{X} = H(H'H)^{-1}H'X.$$  \hspace{1cm} (2)

The equation (2) denotes that the regressor cannot be directly used, but instead it should be a value resulted from regression on the instrumental variables. $H$ is matrix $n \times p$ which belongs to instrumental variable matrix having $p$ instrumental variable.

2. Doing regression with OLS for $Y$ on $\bar{X}$ is to obtain parameter estimation $\hat{\beta} = [\bar{X}'\bar{X}]^{-1}\bar{X}'Y$.

with 2SLS estimation, $\hat{\beta}$ is proved to be a consistent estimator for $\beta$.

Afterwards, model (1) is written as

$$y = Z\delta + u, u = \lambda Mu + \epsilon$$ \hspace{1cm} (3)

with $Z = (X, Wy)$ is $nx(k+1)$ sized matrix and $\delta = (\beta^T, \lambda)^T$ is a vector of parameters $(k+1) \times 1$. To eliminate autocorrelation of the error terms on model (1), Cochrane-Orcutt transformation is performed. If model (3) is multiplied by $\rho M$ and is subtracted by model (1), the following function is obtained:

$$y' = Z'\delta + \epsilon$$ \hspace{1cm} (4)

with $y' = y - \rho My; Z' = (I - \rho M)Z$ and $\epsilon = u - \rho Mu$. Model (4) does not have autocorrelation of the error terms, but $y'$ and $Z'$ are functions of $\rho$ and parameter $\hat{\alpha}$ is unknown so that parameter $\delta$ estimation cannot be conducted. For that reason, three stages of procedure known as generalized spatial two stage least square (GS2SLS) [14, 15,16] are required. The stages include parameter estimation of spatial autoregressive model, parameter $\lambda$ estimation, and final model estimation, explained as follows.

1. The Parameter estimation of the spatial autoregressive model. In general spatial model, a spatial autoregressive element, vector $Wy$, correlates to residual $u$. Consequently, $\delta$ cannot be consistently estimated with OLS since $E[(Wy)u^T] \neq 0$. Therefore, $\delta$ is estimated with 2SLS and $\hat{\delta} = (Z'Z)^{-1}Z'y$ is resulted, where $Z' = P_HZ = (x, Wy), Wy = P_HWy$ and $P_H = H(H'H)^{-1}H^T$.

Estimation with 2SLS method requires instrumental variable $H$. The instrumental variable which is used involves a combination of $X$ and $WX$ or $H = (X, WX)$. The stage results in estimation model of spatial autoregressive which is written below.

$$\bar{y} = \hat{\beta}Wy + X\hat{\beta}.$$  \hspace{1cm} (5)
Model (5) can be used to obtain parameter estimation of spatial error on the second stage.

2. Parameter Estimation. Parameter \( \lambda \) is estimated with generalized moment method (GMM). In reference to the model resulted from stage 1, a residual value denoted by \( \hat{e} \) is obtained. The value serves as a vector of observations for random variable \( u \) on spatial error model. Based on model (1) \( \varepsilon = u - \lambda W u \) is obtained and consider \( \bar{u} = W u \) so that

\[
\varepsilon = u - \lambda \bar{u}
\]  

(6)

If the two sides of equation (6) are multiplied by \( W \), the following equation is obtained

\[
W \varepsilon = W u - W \lambda \bar{u} = \bar{u} - \lambda \bar{u}
\]  

(7)

with \( \bar{e} = W \varepsilon \) and \( \bar{u} = W u \). The following three equation system is obtained by squaring equation (6) and (7), multiplying, and dividing each of the equation by \( n \).

\[
2 \lambda n^{-1} u^T \bar{u} - \lambda^2 n^{-1} \bar{u}^T u + n^{-1} \bar{e}^T \bar{e} = n^{-1} u^T u
\]

\[
2 \lambda n^{-1} u^T \bar{u} - \lambda^2 n^{-1} \bar{u}^T u + n^{-1} \bar{e}^T \bar{e} = n^{-1} \bar{u}^T \bar{u}
\]

\[
\lambda n^{-1} (u^T \bar{u} + \bar{u}^T u) - \lambda^2 n^{-1} \bar{u}^T u + n^{-1} \bar{e}^T \bar{e} = n^{-1} u^T \bar{u}
\]

(8)

The equation system (8) can be written

\[
2 \lambda n^{-1} E(u^T \bar{u}) - \lambda^2 n^{-1} E(\bar{u}^T u) + n^{-1} E(\bar{e}^T \bar{e}) = 0
\]

\[
2 \lambda n^{-1} E(u^T \bar{u}) - \lambda^2 n^{-1} E(\bar{u}^T u) + n^{-1} E(\bar{e}^T \bar{e}) = 0
\]

\[
\lambda n^{-1} E(u^T \bar{u} + \bar{u}^T u) - \lambda^2 n^{-1} E(\bar{u}^T u) + n^{-1} E(\bar{e}^T \bar{e}) = 0
\]

(9)

The equation system (10) expressed in the matrix is

\[
\begin{bmatrix}
2 \lambda n^{-1} E(u^T \bar{u}) \\
2 \lambda n^{-1} E(u^T \bar{u}) \\
n^{-1} E(u^T \bar{u} + \bar{u}^T u)
\end{bmatrix}
- \lambda^2 n^{-1} E(\bar{u}^T u) - \lambda^2 n^{-1} E(\bar{u}^T u) - n^{-1} E(\bar{e}^T \bar{e}) = 0
\]

where \( \alpha \) represents a vector having spatial error parameter, that is \( \alpha = [\lambda \quad \lambda^2 \quad \sigma^2]^T \). The value of parameter \( \alpha \) can be estimated with \( \hat{\alpha} = [G^T G]^{-1} G^T g \).

The estimator value of spatial error parameter \( \hat{e} \) is used to perform Cochran-Orcutt transformation on the third stage. The transformation is conducted to find out \( X_i^* y_i^* \) to estimate the final model. The Cochran-Orcutt transformation which is used is

\[
X_i^* = X_i - \hat{\lambda} \sum_{j=1, j \neq i} W_{ij} X_i \text{ and } y_i^* = y_i - \hat{\lambda} \sum_{j=1, j \neq i} W_{ij} y_i
\]

3. The Final model estimation. The final stage of GS2SLS is a final model estimation. The estimation is performed with 2SLS method. At this stage, the variable used is the transformation result of Cochran Orcutt. \( \hat{\delta} \) is denoted as generalized spatial two stage least square (GS2SLS) estimator obtained from

\[
\hat{\delta} = [Z^T \hat{Z}]^{-1} \hat{Z}^T y^*\]

where \( \hat{Z}^* = P_H Z^*\); \( P_H = H(H^T H)^{-1} H^T \); \( y^* = y - \lambda W y\); \( Z^* = Z - \lambda W Z\); \( X^* = X - \lambda W X\); \( W y^* = W y - \lambda W^2 y\). \( H \) is instrumental variable \( (X, WX) \).

The estimation results in the estimation model

\[
\hat{y}^* = \beta W y^* + X^* \hat{\beta}
\]
For further discussion on estimation with GS2SLS, see [14] and [15]. For articles of Kelejian and Prucha [11], [16], and [17] and complete explanation on the model, see [15] and [18].

3.2 Application

In DBF case in Surakarta City in 2009, the factors affecting the number of DBF patients in Surakarta city were population number ($X_1$) and car number ($X_2$). The linear regression model with both factors is written as follows

$$\hat{y}_i = 0.892 + 0.001x_{1i} + 0.002x_{6i}$$

With $R$-squared is 0.8362. It means that 83.62% of DBF patient number in Surakarta in 2009 can be accounted for by the population number and the car number. In overall linear regression model testing, alpha of 10% was used, thereby obtaining $F$ value = 122.510 > $F_{(0.1;2;48)} = 2.417$ so that it could be stated that at least one regression parameter affected the number of DBF population significantly. The parameter estimation value and $t$-statistic value of model can be shown in Table 1.

**Table 1.** The parameter estimation and $t$-statistic values of the best linear regression model

| Independent variable | Parameter estimation value | $t_{statistic}$ |
|----------------------|----------------------------|-----------------|
| Constant             | 0.892                      | 0.686           |
| $X_1$                | 0.001                      | 9.767           |
| $X_2$                | 0.002                      | 3.829           |

Considering the Table 1, the constant did not affect significantly because $|t_{statistic}| = 0.686 < t_{(0.05;48)} = 1.677$. The assumption of normality and non-multicollinearity in the linear regression model was met, but that of homoscedasticity was not met. The unmet homoscedasticity assumption made the number of DBF patients in Surakarta City could not be presented using a linear regression model with *ordinary least square* (OLS). Heteroscedasticity occurring was assumed due to the presence of spatial autocorrelation effect.

The indication of spatial autocorrelation effect could be found through the scattered chart of Moran index depicted in Figure 1. The Moran index value was 0.166 meaning that there was a positive spatial autocorrelation in the model. For that reason, DBF case in Surakarta city could be presented with a spatial regression model.

![Figure 1](image-url). Moran index scattered chart for the error regression model with the mean error regression among interrelated areas

In addition to using the scattered chart, the indication of spatial autocorrelation effect can also be made by determining Moran index value. Previously, the spatial weighing matrix should be determined for the standardized queen intersection. The value yielded $|Z(I)| = 2.110 > z_{0.005} = 1.645$ meaning that $H_0$ was not supported meaning that there was a spatial autocorrelation.

Moran index can only be used to find out whether or not there is a spatial autocorrelation in an area but not spatial autocorrelation in the model. For that reason, there should be Lagrange multiplying test to detect spatial dependency more specifically, lag and error dependency. The lag and error Lagrange multiplier has statistic value $9.685 > X_{(0,1;2)}^2 = 4.61$ so that $H_0$ is rejected. It means there is an
interrelationship between regions or there are dependence between spatial lag and error. The presence of lag and error spatial dependency made the number of DBF patients could be presented with SAR-SAR model. Based on the process of determining the model, the SAR-SAR model can be written as

$$\hat{y}_i = 3.392 - 0.355W y_i + 0.001x_{1i} + 0.002x_{6i} + u_i, u_i = 0.587W u_i$$

With pseudo-R-squared is 83.71%. The parameter estimation value of SAR-SAR model and z-statistic with queen contiguity spatial weighing matrix is shown in Table 2.

| Independent variable | Parameter estimation value | t-statistic |
|----------------------|---------------------------|-------------|
| Constant             | 3.392                     | 1.058       |
| $X_1$                | 0.001                     | 12.554      |
| $X_6$                | 0.002                     | 4.898       |
| $W_y$                | -0.355                    | -1.858      |
| $\lambda$            | 0.587                     | 2.430       |

Considering the Table 2, the constant did not affect significantly because $z_{stat} = 1.058 < 1.645$. The increased numbers of population and car would improve the number of DBF patients. If the population number in a village increases by 1000 populations, the number of DBF patients will increase 1 patient and if the car number in a village increases by 1000 populations, the number of DBF patients will increase 2 patients with lag and error spatial correlation between villages. The presence of spatial autocorrelation is shown by the value of the lag spatial parameter, which is 0.355 and value of spatial error parameter, which is 0.587.

The chart of predicted DBF patient number with SAR-SAR model and its comparison with actual data are shown in Figure 2.

![Figure 2](image)

Figure 2. The number of DBF patients with predicted DBF patient number in 51 villages in 2009 using SAR-SAR model.

Based on Figure 2, it can be seen that SAR-SAR model has a good ability to replicate the pattern of DBF patients. Nevertheless, there is a discrepancy between the predicted and the actual data. It indicates that the model has a limitation in predicting because there is another factor that cannot be taken into account in SAR-SAR model.

The data of 2010 was used as model control with validation value of 7.674 according to RMSE. The predicted number of DBF patients in 2011 is shown in Figure 3. The result of prediction represents the number of DBF patients in every village of Surakarta City.
Based on Figure 3, it can be seen that the district with the high predicted DBF patient number were Mojosongo (adjacent to Karanganyar Regency), and Semanggi (adjacent to Sukoharjo Regency). It was assumed that the areas bordering on other areas are an endemic area of DBF.

4. Conclusion

Based on the results and discussion, can be concluded that using model (1), (2) and (3) we get model (4) and model (5), so we obtain parameter estimation for SAR-SAR model with GS2SLS is \( \hat{\beta} = \hat{\beta}^* + X' \hat{\beta} \) with \( \hat{\beta} \) is estimator of generalized spatial two stage least square or estimator of GS2SLS obtained by \( \hat{\delta} = \left[ \hat{Z}' \hat{Z} \right]^{-1} \hat{Z}' \hat{y} = \hat{P}_H \hat{Z} \); \( \hat{P}_H = H(H' H)^{-1} H' \); \( y^* = y - \lambda Wy' ; \ Z^* = Z - \lambda WZ ; \ X^* = X - \lambda WX \); (Wy)\(^*\) = Wy - \lambda W^2 y. H is an instrument variable where (X,WX). In addition, the estimation of parameter of SAR-SAR model with GS2SLS is consistent estimator. Based on the incidence of dengue hemorrhagic fever (DHF) in Surakarta, it could be concluded that the factors affecting the number of DBF patient in Surakarta city were population and car numbers. The SAR-SAR model of the number of DBF cases in Surakarta City is written as \( \hat{y}_i = 3.392 - 0.355 Wy_i + 0.001 x_{1i} + 0.002 x_{6i} + u_i, u_i = 0.587 Wu_i. \)

It shows that the increase in population and the number of cars will increase the number of DBF case. The presence of spatial autocorrelation is shown by the value of the spatial lag parameter, which is -0.355 and value of spatial error parameter is 0.587. The prediction of the number of DHF cases with this model has a value of RMSE, which is 58.89.5

References

[1] Gujarati D 2011 Basic Econometrics, five editions (USA : McGraw-Hill)
[2] Gujarati D 2012 Econometrics by Example, first edition (USA : Palgrave Macmillan)
[3] Rawlings J O, Pantula S Gand Dickey D A 1998. *Applied Regression Analysis: Research Tool*, Second edition (New York: Springer Verlag Inc.)

[4] Fotheringham A S and Rogerson P A 2009. *The Sage Handbook of Spatial Analysis* (London: Sage)

[5] Anselin L 1988. *Spatial Econometrics: Methods and Models* (London: Kluwer Academic Press)

[6] Tobler W R 1970 A Computer Movie Simulating Urban Growth in The Detroit Region, *Economic Geography*, 46: 234–40

[7] Kelejian H H, Harry H and Robinson D P 1995 “Spatial Correlation: A Suggested Alternative to the Autoregressive Model,” in Luc Anselin and Raymond J. G. M. Florax, eds., *New Directions in Spatial Econometrics*, Berlin: Springer, pp. 75–95.

[8] Harry H. Kelejian and Prucha I R 2010 *J Econom* 157 53.

[9] Cressie NAC 1993 *Statistics for spatial data* (New York :Wiley Series in Probability and Mathematical Statistics. Wiley)

[10] Haining R 2003 *Spatial data analysis: theory and practice*. (Cambridge :Cambridge University Press)

[11] Kelejian H H, Prucha I R 1998 A Generalized Spatial Two Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbance. *Journal of Real Estate Finance and Economics* 17 99.

[12] Kementerian Kesehatan RI 2010. *Buletin Jendela Epidemiologi*, 2010

[13] Dinas Kesehatan Kota Surakarta 2011. *Profil Kesehatan Kota Surakarta Tahun 2010*, Surakarta

[14] Arraiz I, Drukker D M, Kelejian H H, and Prucha I R 2010 *Journal of Regional Science* 50 59.

[15] Drukker D M, Egger P, and Prucha I R 2013 *Econometric Reviews* 32 686.

[16] Kelejian H H, Prucha I R 1999 *International Economic Review* 40 509.

[17] Kelejian H H, Prucha I R 2010 *Journal of Econometrics* 157 53.

[18] Drukker D M, Prucha I R and Raciborski R 2013 *The Stata Journal* 13 221.