Performance of W-AMOEBA and W-Contiguity matrices in Spatial Lag Model

Jajang\textsuperscript{1} and Pratikno, B.\textsuperscript{2}
\textsuperscript{1,2}Department of mathematics  
Faculty Mathematics and Natural Science  
Jenderal Soedirman University  
Purwokerto, Indonesia  
ruzajang@yahoo.com and bpratikto@gmail.com

Abstract - The paper discussed a parameter estimation methods and construction of spatial weighted matrix in the modeling of the spatial data. Many options can be used to construct a spatial weighted matrix, one of them is a matrix AMOEBA (W_AMOEBA). Here, we studied about the W_AMOEBA and Contiguity matrix spatial lag model (SLM) using two step least square (two-SLS). For simulation, we used human development index (HDI) data. The results showed that the relative W_AMOEBA is more accurate than W-contiguity.

2010 Mathematics Subject Classification: Primary 62H10 and Secondary 60E05

1. INTRODUCTION

1.1 Background

Many authors have studied about spatial models such as Folmer and Oud (2008), Liu, et.al (2011a, 2011b) and Aldstadt and Getis, 2006. They discussed W-structural equation model (W_SEM) and W-an multidirectional optimum ecotope base algorithm (W_AMOEBA), respectively.

Spatial data is a stochastic process which have more than two indexes. There were researchers have already studied modeling of spatial data in theoretical and application. The objective of spatial data analysis is to determine the pattern of the spatial data. The spatial data are usually dependent each others. It is called spatial autocorrelation. In the model of spatial data, the autocorrelation is represented by spatial weighted matrix.

Human development index (HDI) is one of the indicators for seeing development of regions (area). Here, we face problem on correlation among adjacent areas (neighbored area). To accommodate dependence among those areas, the spatial weighted matrix is used. Here, there are several ways to construct a spatial weighted matrix such as contiguity, inverse distance contiguity, and k-nearest neighbor (Anselin,1995, Stakhovych and Bijmolt, 2008), W-SEM (Folmer and Oud, 2008, Liu, et.al 2011a, 2011b) and W-AMOEBA (Aldstadt and Getis, 2006).

1.2. Objectives

The objective of this study is to evaluate of W-AMOEBA and W-contiguity on spatial lag model (SLM) in the case of the HDI.

2. RESEARCH METODOLOGY

2.1 Spatial Lag Model

In linear regression model, spatial dependence can be incorporated in two distinct way: as an additional regression in the form of a spatially lag dependence ($Wy$) or in the error structure. Spatial lag
model (SLM) (or spatial autoregressive, SAR) is appropriate when the focus of interest is the assessment of existence and strength of spatial interaction (Anselin, 1995). Generally, a spatial lag model or spatial autoregressive model is given as

\[ y = \rho Wy + X\beta + \varepsilon, \]  

where

- \( y \) = vector of dependent variable
- \( X \) = matrix of independent variable
- \( W \) = spatial weighted matrix
- \( \rho \) = spatial autoregressive coefficient
- \( \beta \) = vector of parameter
- \( \varepsilon \) = vector of error terms

Here, the spatial autoregressive coefficient (\( \rho \)) is assumed to stationary when \( |\rho| < 1 \).

2.2. W Contiguity

Spatial weighted matrix is an essential component of the spatial model. Generally, spatial weight matrix that used to spatial model based on contiguous area (geographically) and inverse distances. Due to of the geographical proximity, the spatial weighted matrix is created as follows. Let \( W = \{w_{ij}\} i, j = 1,2, ..., n \), is contiguities matrix with \( w_{ij} \) represents the value of spatial unit \( i \) and \( j \), and here we also present spatially located at some positions see Figure 1.

![Contiguity matrices](image)

Figure 1 Contiguity matrices rook (b), bishop (c) and queen (d) of spatial unit on (a) that close to \( F \).

Inverse distance matrix is another type of spatial weighted matrix that often used in spatial modeling. Generally, type of distance that use in constructing spatial weight matrix is the Euclidean distance. For two coordinate spatial units \( i \) and \( j \), \((x_i, y_i)\) and \((x_j, y_j)\), the inverse distance is expressed as

\[ W_{ij} = 1/d_{ij}, \text{ where } d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}. \]

2.3 W_AMOEBA

A multidirectional optimum ecotope-based algorithm (AMOEBA) is one of an illustration on the spatial weighted matrix. The AMOEBA is usually depended on the behavior data. Furthermore, W_AMOEBA are designed to clustering spatial units and construct spatial weighted matrix on empirical data (Alsdstadt and Getis, 2006). Later, the W_AMOEBA is defined as combination between the geographic and behavior data (Stakhovych and Bijnol 2008; Alsdstadt and Getis, 2006). Here, the W_AMOEBA has a special procedure that developed by Alsdstadt and Getis (2004). Note that in the W_AMOEBA, the local Getis statistic is used to divide into high and low spatial units.

Let an area is divided for \( n \) regions, \( i=1,2,...,n \), \( G_i = \frac{\sum_{j=1}^{n}w_{ij}x_j}{\sum_{j=1}^{n}x_j} \) is local Getis statistic, and \( G_i^* = \frac{G_i - E(G_i)}{\sqrt{\text{Var}(G_i)}} \) be standard local Getis statistic. Let \( G_i^*(k) \) is the value of statistics \( G_i^* \) for link \( k \). The AMOEBA algorithm (Alsdstadt and Getis, 2006) is given as follows.

At the outset of the AMOEBA procedure, we compute \( G_i^*(0) \) here the ecotope consist of just the \( i^{th} \) unit (\( k = 0 \)). The value of \( G_i^*(0) \) is greater than zero, it indicates that the value at location \( i \) is larger than mean of all unit, and otherwise. For \( k=1, G_i^*(1) \), this value shows that for each areas/region that contains units \( i \) and all combinations of its contiguous are neighbors. If \( G_i^*(0) \) is greater than the
Combination that maximizes $G_i^*(1)$, it be new high ecotope. If $G_i^*(0)$ is less than the combination that maximizes $G_i^*(1)$, it be new low ecotope. At each succeeding step, for the contiguous units in the ecotope, they are then not considered. Likewise, units included in the ecotope remain in the ecotope. Subsequent steps evaluate all combinations of contiguous neighbors and new members of the ecotope are then identified. This process continues for $k, k = 2, 3, ..., \text{max}$. The final ecotope ($k_{\text{max}}$) is identified when the addition of any set of contiguous units fail to increase the absolute value of the statistics $G_i^*$. The results of the AMOEBA procedure are then used to construct $W$ using several steps as follow.

(a) when $k_{\text{max}} > 1$,

$$W_{ij} = \begin{cases} 
\{P[z \leq G_i(k_{\text{max}})] - P[z \leq G_i(k_j)]\}, & 0 < k_j \\
0, & \text{otherwise}
\end{cases}$$

(b) when $k_{\text{max}} = 1$, $w_{ij} = 1$ for $k_j = 1$ and 0 otherwise.

(c) When $k_{\text{max}} = 0$, $w_{ij} = 0$ for all $k$

where $k_j$ is link that connecting between $i$ and $j$ in ecotope.

### 2.4. Instruments Variables Method

The simple method of estimation parameter in regression models is ordinary least square (OLS). One of the advantages of the OLS is robust to error distribution. Here, the error and exogenous variable must be independent. If the model containing endogenous variable and residual distribution is not known, we need other estimation methods, one of them is the method of instrumental variables of two step least square (two-SLS) (Verbeek, 2008). We see that SLM contains endogenous variable ($Wy$), so the OLS can’t be used to estimate model parameters. To overcome this problem, we must used other method. Instrument variable method or two-stage method is one of methods that can be used to solve endogenous problem in spatial model. The principle of the instrument variable method is to use new variable that correlated to the response variable, but uncorrelated with the residual.

### 3. RESULTS

#### 3.1 Violation OLS on SLM

Let $Z = (Wy \ X)$, and $\theta = (\beta \ \beta)$, so the SLM model can be written as

$$y = Z\theta + \varepsilon,$$

where the assumptions of the OLS method are $E(Z' \varepsilon) = 0$ and $\text{Cov}(\varepsilon, Z) = 0$. Due to $Z$ consists of $Wy$ and $X$ variables, so the OLS is fail to estimate parameter model on the SLM. This is due to the assumption is violate. The assumption of the OLS $E(Z' \varepsilon) = 0$, we then have $E(W' \varepsilon) = 0$ and $E(X' \varepsilon) = 0$. Here, $\text{Cov}(\varepsilon, Wy) = \text{Cov}(\varepsilon, W(\rho Wy + X\beta + \varepsilon))$ where $\text{Cov}(\varepsilon, Wy) \neq 0$ because $Wy = W(y + X\beta + \varepsilon) = Wy + WX\beta + W\varepsilon$, so $\text{Cov}(\varepsilon, Wy) = \text{Cov}(\varepsilon, Wy + WX\beta + W\varepsilon) \neq 0$. Therefore the moment, i.e $E(Z' \varepsilon) = 0$, is violate.
3.2 Method of Variable Instrument

The method of variable instrument (IV) is also called a two-stage OLS method. In this method, the endogenous variable is instrumented by new variables, but in SLM method, the endogenous variable, \( W_y \), is instrumented with new variable that correlated to response, but uncorrelated to the error terms.

Due to SLM model, let, \( Z = (W_y X) \) and \( \theta = \left( \frac{\beta}{\beta} \right) \), so (1) can be expressed as (2). To estimate parameter \( \theta \), we use **IV method** as follows:

1. determine instrument variable \( H \).
   
   \[
   E(\varepsilon | H) = 0, \quad \text{Cov}(Z, H) \neq 0
   \]
   
   \[
   H = [X WX W^2 X \ldots]\n   \]

2. estimate \( Z \) using variable \( H \).
   
   \[
   \hat{Z} = H(H' H)^{-1} H' Z
   \]

   \[
   \hat{\theta} = (\hat{Z}' Z)^{-1} \hat{Z}' y,
   \]

   \[
   \text{Var}(\hat{\theta}) = \hat{\sigma}^2 (\hat{Z}' Z)^{-1}
   \]

   \[
   \hat{\sigma}^2 = \hat{\varepsilon}' \hat{\varepsilon} / n,
   \]

   \[
   = (y - Z \hat{\theta})' (y - Z \hat{\theta}) / n
   \]

3.3 Performance Investigation of W-Contiguity and W AMOEBA with HDI Data

To evaluate the spatial weighted matrix W_AMOEBA (WG) and W_Contiguity (WC), we used the HDI data of the central of Java (Jateng). Here, we assumed that the HDI between districts have spatial relationships. HDI data is from SUSENAS 2013 BPS JATENG (2014) with single response and four predictors, namely (1) \( Y \) = human development index (HDI), (2) \( X_1 \) = the average length of school, (3) \( X_2 \) = overcrowding, (4) \( X_3 \) = the number of physicians per health center, and (5) \( X_4 \) = income per capita.

![Graph showing HDI vs. school years, population density, ratio of doctors per public health center, and income per capita](image)
Figure 2. Scatterplot of response variable versus predictors

From Figure 2, we see that relationship between response variable, predictor variables and duration of school factor have significant association to HDI, but not for others factor.

Table 1. ANOVA for SLM model

<table>
<thead>
<tr>
<th></th>
<th>Est.</th>
<th>Stdev</th>
<th>t.value</th>
<th>Est.</th>
<th>Stdev</th>
<th>t.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$</td>
<td>-0.0060</td>
<td>0.0037</td>
<td>-1.6413</td>
<td>0.1064</td>
<td>0.0977</td>
<td>1.0884</td>
</tr>
<tr>
<td>X1</td>
<td>1.9844</td>
<td>0.2501</td>
<td>7.9353</td>
<td>1.8399</td>
<td>0.2417</td>
<td>7.6126</td>
</tr>
<tr>
<td>X2</td>
<td>-0.0001</td>
<td>0.0001</td>
<td>-1.8362</td>
<td>-0.0002</td>
<td>0.0001</td>
<td>-1.9517</td>
</tr>
<tr>
<td>X3</td>
<td>-0.3844</td>
<td>0.1413</td>
<td>-2.7207</td>
<td>-0.3398</td>
<td>0.1415</td>
<td>-2.4014</td>
</tr>
<tr>
<td>X4</td>
<td>0.0932</td>
<td>0.0023</td>
<td>41.2132</td>
<td>0.0821</td>
<td>0.0112</td>
<td>7.3251</td>
</tr>
</tbody>
</table>

Table 1 showed that the coefficient $R^2$ on WG is greater than $R^2$ on WC. We see also that the root mean square error (RMSE) on WG is less than the RMSE on WC. Finally, we conclude that weighted spatial matrix WG is better than weighted spatial matrix WC.

Prediction of HDI in term of the model with WG and WC are presented Table 2.

Table 2. Predicting values of HDI_Y_hat(WG), HDI WC and actual data

<table>
<thead>
<tr>
<th>No</th>
<th>Y_actual</th>
<th>Y_hat(WG)</th>
<th>Y_hat(WC)</th>
<th>No</th>
<th>Y_actual</th>
<th>Y_hat(WG)</th>
<th>Y_hat(WC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>71.13</td>
<td>71.93</td>
<td>71.79</td>
<td>18</td>
<td>78.54</td>
<td>78.9</td>
<td>78.93</td>
</tr>
<tr>
<td>2</td>
<td>73.96</td>
<td>74.08</td>
<td>74.11</td>
<td>19</td>
<td>79.1</td>
<td>79.32</td>
<td>78.98</td>
</tr>
<tr>
<td>3</td>
<td>72.03</td>
<td>71.72</td>
<td>71.6</td>
<td>20</td>
<td>75.02</td>
<td>74.85</td>
<td>75.09</td>
</tr>
<tr>
<td>4</td>
<td>72.1</td>
<td>72.34</td>
<td>72.45</td>
<td>21</td>
<td>74.09</td>
<td>74.78</td>
<td>75.13</td>
</tr>
<tr>
<td>5</td>
<td>71.88</td>
<td>72.74</td>
<td>72.8</td>
<td>22</td>
<td>73.67</td>
<td>73.25</td>
<td>73.48</td>
</tr>
<tr>
<td>6</td>
<td>69.85</td>
<td>71.32</td>
<td>71.14</td>
<td>23</td>
<td>74.58</td>
<td>74</td>
<td>73.65</td>
</tr>
<tr>
<td>7</td>
<td>73.34</td>
<td>72.82</td>
<td>72.88</td>
<td>24</td>
<td>73.14</td>
<td>72.73</td>
<td>72.89</td>
</tr>
<tr>
<td>8</td>
<td>73.85</td>
<td>73.36</td>
<td>73.42</td>
<td>25</td>
<td>71.26</td>
<td>71.87</td>
<td>71.67</td>
</tr>
<tr>
<td>9</td>
<td>72.37</td>
<td>72</td>
<td>72.38</td>
<td>26</td>
<td>73.49</td>
<td>72.86</td>
<td>72.93</td>
</tr>
<tr>
<td>10</td>
<td>74.13</td>
<td>73.63</td>
<td>73.8</td>
<td>27</td>
<td>74.18</td>
<td>74.74</td>
<td>74.38</td>
</tr>
<tr>
<td>11</td>
<td>75.27</td>
<td>75.46</td>
<td>75</td>
<td>28</td>
<td>73.53</td>
<td>73.47</td>
<td>73.64</td>
</tr>
<tr>
<td>12</td>
<td>72.25</td>
<td>72.53</td>
<td>72.74</td>
<td>29</td>
<td>75.48</td>
<td>74.35</td>
<td>73.96</td>
</tr>
<tr>
<td>13</td>
<td>72.03</td>
<td>72.59</td>
<td>72.85</td>
<td>30</td>
<td>72.31</td>
<td>72.94</td>
<td>72.78</td>
</tr>
<tr>
<td>14</td>
<td>74.91</td>
<td>75.56</td>
<td>75.58</td>
<td>31</td>
<td>74.91</td>
<td>75.48</td>
<td>75.45</td>
</tr>
<tr>
<td>15</td>
<td>77.91</td>
<td>78.33</td>
<td>78.18</td>
<td>32</td>
<td>72.22</td>
<td>72.58</td>
<td>72.35</td>
</tr>
<tr>
<td>16</td>
<td>75.75</td>
<td>74.68</td>
<td>75.09</td>
<td>33</td>
<td>75</td>
<td>72.95</td>
<td>72.75</td>
</tr>
<tr>
<td>17</td>
<td>77.54</td>
<td>77.14</td>
<td>77.39</td>
<td>34</td>
<td>73.09</td>
<td>73.1</td>
<td>73.1</td>
</tr>
<tr>
<td></td>
<td>35</td>
<td>71.9</td>
<td>71.39</td>
<td>71.42</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. CONCLUSION
For spatial modeling, the response that has a spatial relationship need weighted spatial matrix for accommodating its relationship. The choice of the spatial weighted matrix should consider the characteristics and behavior of the data and adjacent geographical area. AMOeba weighted matrix is a significant matrix to improve the accuracy of predicting results.

5. REFERENCES


