Spatial Regression Model with Optimum Spatial Weighting Matrix on GRDP Data of Sulawesi Island

N Paramita, M Masjkur*, and Indahwati
Department of Statistics, IPB University, Bogor, Indonesia

*E-mail: masjkur@apps.ipb.ac.id

Abstract. The two main spatial regression models are the spatial autoregressive model (SAR) and spatial error model (SEM). The extension of the SAR model is a spatial Durbin model (SDM), which considers the spatial dependence of response and explanatory variables. However, the determination of the spatial weight matrix is critical for the best estimation results. We consider two distance-based spatial weight matrices, i.e., the k-Nearest Neighbour (k-NN) and Inverse Distance Weighting (IDW). The objective of this study was to compare the performance of the Ordinary Least Squares (OLS) regression, SAR, SEM, and SDM models with k-NN and IDW on the estimation of Growth Regional Domestic Product (GRDP) and identify the critical factors that influence the value of GRDP of Sulawesi island. The study used the GRDP data of 81 districts/cities in Sulawesi island in 2018 with six explanatory variables. The results show that the 4-NN weighted SAR model outperforms the OLS, the 4-NN SEM, SDM models, IDW SAR, SEM, and SDM models. The factors that influence the value of GRDP in Sulawesi island are HDI (Human Development Index), population size, open unemployment, small/micro and medium industries, and the spatial lag autoregressive coefficient.

1. Introduction
Spatial data is cross-section data with additional spatial effects on the data. Spatial data is oriented to geographic structure and has a coordinate system (latitude and longitude) presented in map form. Analysis of spatial data is different from ordinary data analysis since there are the effect of the environment and characteristics of a region. Regions that tend to be close together will have a closer relationship when compared to regions that are far apart [1]. This relationship is called the spatial effect, which means a spatial autocorrelation between the regional unit. Adding an autoregressive component to the response variable, residual/error, or both accounts for spatial autocorrelation [2, 3].

Two main spatial regression models are the spatial autoregressive model (SAR) and spatial error model (SEM). The SAR model is a model with a spatial dependency on the response variable, while the SEM model is a model with a spatial dependency on the residual errors. If the spatial dependence occurs not only on the response variable but also on the explanatory variables, we can use the spatial Durbin model (SDM) [3, 4].

In the spatial model, [5] stated that the spatial weight matrix is an essential component because the analysis results are susceptible to the specifications of the weight matrix, such as spatial contiguity, inverse distance, or k nearest neighbors [2, 6]. Monte-Carlo's simulation study [7] demonstrates that applying a weights matrix selection procedure based on information criteria increases the probability of identifying its accurate specification.
Gross regional domestic product (GRDP) is the added value generated from all economic activities in a region. Generally, GRDP is used to evaluate the economic growth of a region. Based on Indonesia's islands’ economic growth rate in 2018, Sulawesi island is the second-highest economic growth with 4.91% from the previous year [8]. The first position is Java island, with a growth rate of 5.78%. Meanwhile, Bali and Nusa Tenggara island are the lowest economic growth with 2.33%. The increasing value of GRDP generated by each district/city is the main factor for the high economic growth in Sulawesi island.

The study by [9] observes that the factors that influence GRDP in Java Island are HDI (Human Development Index), education, the average length of schooling, and the percentage of poor people. Meanwhile, the factors that influenced the GRDP in Sumatra Island in 2000-2006 were export value, domestic investment, and labor [10].

The objective of this study was to compare the performance of four models, i.e., the OLS (Ordinary Least Squares) regression, SAR, SEM, and SDM models with a spatial weight of k-Nearest Neighbour (k-NN) and Inverse Distance Weighting (IDW) on the estimation of GRDP and identify the critical factors that influence the value of GRDP of Sulawesi island.

2. Methodology

2.1. Data

The study used the district/city's GRDP data for each province in Sulawesi Island [8]. The response variable GRDP is at constant prices from 81 districts/city in Sulawesi Island in billion rupiahs. The explanatory variables are selected based on three main aspects that affect economic growth. These are capital accumulation, population growth, and technological advances, the related explanatory variables presented in table 1.

Table 1. The explanatory variables in this study

| Code | Variable                              | Unit       |
|------|---------------------------------------|------------|
| 𝑋₁   | Human development index (HDI)         | -          |
| 𝑋₂   | Population size                       | thousand   |
| 𝑋₃   | Percentage of the poor population     | percent    |
| 𝑋₄   | Unemployment rate                     | percent    |
| 𝑋₅   | The number of small/micro and medium  | unit       |
|      | industries                            |            |
| 𝑋₆   | The number of medium and large        | unit       |
|      | manufacturing industries              |            |

2.2. Data Analysis Procedure

The procedure of data analysis are as follows:

- We are exploring the data to determine the distribution of GRDP in Sulawesi Island and then test the correlation between variables.
- Conducting regression modeling using the ordinary least-squares (OLS) estimation, as follows [11]:

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

where:
- \( y \) – the response variable
- \( X \) – matrix of explanatory variables
- \( \beta \) – vector of coefficient (model parameters)
- \( \varepsilon \sim N(0, \sigma^2I) \) is the vector of model errors

- Checking the regression model's assumptions, such as normality of the residual distribution, homogeneity of the residuals, and independence of the residuals. Also, checking the multicollinearity assumption between the explanatory variables by looking at the VIF (Variance Inflation Factor) value, which is formulated by:
\[ VIF_p = \frac{1}{1 - R_p^2} \]

where \( R_p^2 \) is the coefficient of determination value obtained by regressing the \( p \)th explanatory variable on the remaining ones. A strong correlation between explanatory variables in the model indicates multicollinearity. The tolerance limit for there is no multicollinearity if the VIF value is less than 10 [11].

- Forming a spatial weight matrix \( W \) using the \( k \)-nearest neighbor (\( W_{mn} \)), the elements of which are \( w_{ij} = 1 \) when \( j \) is among the \( k \)-nearest neighbors of \( i \) with \( k = 1, 2, ..., 20 \) and the inverse distance weight (\( W_{id} \)) the elements of which are \( w_{ij} = 1/d_{ij} \), where \( d_{ij} \) is the distance between units \( i \) and \( j \). All weights matrices are row-standardized [7, 12].

- Checking the existence of spatial autocorrelation using Moran's test, the hypothesis is:
  \[ H_0: I = 0 \text{ (there is no spatial autocorrelation)} \]
  \[ H_1: I \neq 0 \text{ (there is spatial autocorrelation)} \]

The formula for calculating the Moran's index is as follows:
\[ I = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \]
where \( y_i \) the response variable for \( i = 1, 2, ..., n \), \( \bar{y} \) the response average and \( w_{ij} \) is the spatial weighting matrix that consists of \( i \)th row and \( j \)th column [2, 13].

- Conducting the Lagrange Multiplier test to select the spatial dependence model, with the hypothesis and the test statistics for each model as follows [2, 6]:
  a) Spatial autoregressive model (SAR):
  \[ y = \rho W y + X \beta + \varepsilon \]
  where:
  \( W \) – the spatial weight matrix
  \( \rho \) – the spatial autocorrelation coefficient
  \( H_0: \rho = 0 \) (there is no spatial dependence on the response variables)
  \( H_1: \rho \neq 0 \) (there is spatial dependence on the response variable)
  The test statistics is:
  \[ LM_{lag} = \frac{1}{D} \left[ \varepsilon' W y \right]^2 \]
  with
  \[ D = \frac{1}{\sigma^2} \left( W X \hat{\beta} \right)' \left( I - X (X'X)^{-1} X' \right) (W X \hat{\beta}) + tr(W^2 + W'W) \]
  We reject the null hypothesis if \( LM_{lag} > \chi^2(1) \)
  b) Spatial error model (SEM):
  \[ y = X \beta + u \]
  \[ u = \lambda W u + \varepsilon \]
  where:
  \( W u \) – the spatial lag error
  \( \lambda \) – the spatial autocorrelation coefficient, \( \varepsilon \) – the independent model error
  \( H_0: \lambda = 0 \) (there is no spatial dependence on the residuals)
  \( H_1: \lambda \neq 0 \) (there is spatial dependence on the residuals)
  The test statistics is:
  \[ LM_{error} = \frac{\left[ \varepsilon' W \varepsilon \right]^2}{\left[ \sigma^2 \right]} \]
  \[ \left[ \sigma^2 \right] = tr(W^2 + W'W) \]
We reject the null hypothesis if $LM_{error} > \chi^2_{(1)}$

- We are estimating parameters of the selected model obtained in the previous step using maximum likelihood estimation.
- Estimating the spatial Durbin model parameters:
  $$y = \rho W y + X\beta + WX\theta + \varepsilon$$
  where:
  $WX$ – the spatial lag explanatory variables
  $\theta$ – the coefficient
  $\varepsilon$ – the independent model error
- Evaluating the goodness of fit for the model based on Akaike's information criterion (AIC) and Bayesian information criterion (BIC) values. The formula is as follows [12, 14]:
  $$AIC = -2\log(\hat{L}) + 2k$$
  $$BIC = -2\log(\hat{L}) + k \log(n)$$
  where $k$ is the number of parameters estimated by the model, $\hat{L}$ is the maximum value of the likelihood function of the model, and $n$ the number of observations. The smaller AIC and BIC values, the model is better.
- We are checking the assumptions of the best regression model based on the residuals analysis.
- Interpreting the results and make conclusions.

3. Result and Discussion

3.1. Data Exploration
The average GRDP value of districts/cities on Sulawesi Island in 2018 is 7999 billion rupiahs, while the median is 5472 billion rupiahs. The highest GRDP in Sulawesi Island is 112568 billion rupiahs for Makassar city, while the lowest GRDP is for Konawe Kepulauan at 981 billion rupiahs. The difference between the average and the median values indicates that the distribution of GRDP value was not normally distributed but rather the positively skewed distribution.

We divide the GRDP of districts/cities in Sulawesi Island into three categories: low, medium, and high. Districts/cities with the lighter color on the map of figure 1 indicated a lower GRDP, the orange color indicated a medium GRDP, and the darker color indicated a higher GRDP. The distribution of GRDP shows that most districts/cities in Sulawesi Island have a lower GRDP, that are 41 districts/cities. Meanwhile, there are 20 districts/cities that have a medium GRDP and then a higher GRDP. A district/city island has a lower GRDP.

![GRDP distribution of each district/city in Sulawesi Island.](image)

Figure 1. GRDP distribution of each district/city in Sulawesi Island.
Based on the correlation plot or correlogram presented in figure 2 indicate that the percentage of the poor population \( (X_3) \) with a small and lighter circle has a low and negative correlation with the GRDP value. It means that increasing the poor population will decrease the amount of GRDP value. The explanatory variables population size \( (X_2) \) and the number of large and medium manufacturing industries \( (X_6) \) have a high and positive correlation with GRDP with a correlation value is about 0.8. It means that increasing the population size and the number of industries also increase the amount of GRDP value.

Figure 2. Correlation plot of variables.

3.2. OLS Multiple Linear Regression

Parameter estimates of the multiple linear regression using the ordinary least-squares method (OLS) showed three variables that have significant effects on GRDP \( (p\text{-value} < 0.05) \), that are the population size \( (X_2) \), the number of small/micro and medium industries \( (X_5) \), and the number of medium and large manufacturing industries \( (X_6) \) (table 2).

The residuals are not normally distributed based on the Kolmogorov-Smirnov normality test; the statistic value is 0.1086 with a \( p\)-value of 0.0194. We use the Glejser test for the assumption of variance homogeneity of the residuals. It has a \( p\)-value of 0.0000 indicate that the variance of the residuals was not homogenous. Then, we check the residual’s independence assumption using the Runs-test. It has a \( p\)-value of 0.1464 which means that the residuals are independent of each other. Several assumptions have not been fulfilled based on the residual analysis. We then transformed the response variable is into a natural logarithmic form.

Table 2. Parameter estimation of OLS multiple regression model

| Parameter | Estimated value | Transformation variable | VIF |
|-----------|----------------|-------------------------|-----|
| \( \beta_0 \) | -20967.4 | 2.7183 ** | |
| \( \beta_1 \) | 214.16 | 0.0790 ** | 2.3594 |
| \( \beta_2 \) | 4673.03 ** | 0.2415 ** | 2.8178 |
| \( \beta_3 \) | 248.10 | -0.0006 | 1.7049 |
| \( \beta_4 \) | 243.32 | -0.0816 ** | 2.1269 |
| \( \beta_5 \) | -633.70 ** | 0.0395 ** | 1.4702 |
| \( \beta_6 \) | 248.25 ** | 0.0054 | 3.0837 |

** Significant at \( \alpha = 5\% \)
Parameter estimation of the multiple regression model using the transformation variable shows that the human development index \( (X_1) \), population size \( (X_2) \), the unemployment rate \( (X_4) \) and the number of small/micro and medium industries \( (X_5) \) are significant at \( \alpha = 5\% \). We observe that the residuals are not normally distributed based on the transformed variables. However, the residuals are homogeneous and independent (table 3).

**Table 3. Assumption of the multiple linear regression model**

| Assumption                      | P-value  |
|---------------------------------|----------|
| Normal distribution             | 0.0469   |
| Homogeneity of the residuals    | 0.2026 **|
| Independence of the residuals   | 0.2192 **|

**3.3. Spatial Weighted Matrix**

The optimum number of nearest neighbors is \( k = 4 \) because this is the highest Moran index value (figure 3). The spatial dependence on the response variable for \( k = 3 \) through 20 is significant at \( \alpha = 5\% \). However, on \( k = 2 \), the spatial dependence was significant at \( \alpha = 10\% \), while \( k = 1 \), the spatial dependence was not significant. Therefore, we proceed with 4-NN spatial weighting and inverse distance weighting matrices.

**Figure 3.** Plot the k values and Moran's index.

**3.4. Spatial Effects**

Based on table 4, testing for global spatial autocorrelation of residuals produces Moran’s index value of 0.0891 for the 4-NN and 0.02 for the IDW weighting, which is not significant at \( \alpha = 5\% \). However, the spatial autocorrelation for the response and the explanatory variables significant at \( \alpha = 5\% \), except for the human development index variable \( (X_1) \) for the 4-NN and the number of medium and large manufacturing industries \( (X_6) \) for IDW weighting, which is significant at \( \alpha=10\% \). The human development index \( (X_1) \) was not significant for IDW weighting.

**3.5. Selection of SAR and SEM**

Based on Lagrange Multiplier testing, we observed that there is a significant spatial dependency on the response variable for both 4-NN and IDW spatial weighting (p-value < 0.05) (table 5). On the other hand, there is no spatial dependency on the residual of the model for both 4-NN and IDW weighting (p-value > 0.05). Therefore, we selected the SAR model for the GRDP data.
Table 4. Spatial autocorrelation testing

| Variable | Moran's index |
|----------|---------------|
|          | 4-NN | IDW  |
| Residuals| 0.0891| 0.0200|
| 𝑌        | 0.2343**| 0.0681**|
| 𝑥₁       | 0.1045*  | 0.0228 |
| 𝑥₂       | 0.2689** | 0.0685**|
| 𝑥₃       | 0.3254** | 0.1035**|
| 𝑥₄       | 0.3384** | 0.1385**|
| 𝑥₅       | 0.3763** | 0.1645**|
| 𝑥₆       | 0.1286** | 0.0249* |

*) Significant at 𝛼 = 10%; **) Significant at 𝛼 = 5%

Table 5. Lagrange Multiplier test

| Model | LM statistics |
|-------|---------------|
|       | 4-NN | IDW  |
| SAR   | 4.7316** | 4.0693**|
| SEM   | 1.4343  | 0.3841 |

**) Significant at 𝛼 = 5%

3.6. Spatial Autoregressive Model (SAR)

Parameter estimation of SAR model with 4-NN and IDW spatial weighting presented in table 6. We found that the human development index (𝑋₁), the population size (𝑋₂), the unemployment rate (𝑋₄), and the number of small/micro and medium industries (𝑋₅) have a significant effect on GRDP at 𝛼 = 5%, while the number of medium and large manufacturing industries (𝑋₆) was not significant. Also, the spatial lag on the response variable (ρ) has a significant effect on GRDP at 𝛼 = 5% for 4-NN and 𝛼 = 10% for IDW weighting. The spatial lag coefficient (ρ) shows the average effect of the neighborhood on the GRDP produced by a region.

Table 6. Parameter estimation of SAR model

| Parameter | Estimated value |
|-----------|----------------|
|            | 4-NN | IDW  |
| 𝛽₀        | 0.9325| -0.8708|
| 𝛽₁        | 0.0819** | 0.0818**|
| 𝛽₂        | 0.2286** | 0.2321**|
| 𝛽₄        | -0.0868** | -0.0811**|
| 𝛽₅        | 0.0358** | 0.0741**|
| 𝛽₆        | 0.0052 | 0.0052|
| ρ         | 0.1907** | 0.3962* |

*) Significant at 𝛼 = 10%, **) Significant at 𝛼 = 5%

3.7. Spatial Durbin Model (SDM)

Table 7 shows the parameter estimation of the SDM model with 4-NN and IDW spatial weighting. We observed that as well as the SAR model, the human development index (𝑋₁), the population size (𝑋₂), the unemployment rate (𝑋₄), and the number of small/micro and medium industries (𝑋₅) have a significant effect on GRDP at 𝛼 = 5%. Additionally, on 4-NN weighting, the lag of the population size and the lag of the number of medium and large manufacturing industries has a significant effect at a
significant level of 5%. However, for IDW weighting, the lag of the explanatory variable has no significant effect. Moreover, the spatial lag on the response variable (\( \rho \)) has no significant effect on GRDP for both 4-NN and IDW weighting.

### Table 7. Parameter estimation of SDM model

| Parameter | Estimated value | 4-NN | IDW |
|-----------|----------------|------|-----|
| \( \beta_0 \) | -1.0662 | -3.0362 |
| \( \beta_1 \) | 0.0793 ** | 0.0842 ** |
| \( \beta_2 \) | 0.2358 ** | 0.2289 ** |
| \( \beta_4 \) | -0.0771 ** | -0.0813 ** |
| \( \beta_5 \) | 0.0476 ** | 0.0345 * |
| \( \beta_6 \) | 0.0043 | 0.0057 |
| Lag \( \beta_1 \) | 0.0408 | 0.0965 |
| Lag \( \beta_2 \) | 0.1506 ** | 0.0483 |
| Lag \( \beta_4 \) | -0.0193 | -0.2026 |
| Lag \( \beta_5 \) | -0.0381 | 0.0989 |
| Lag \( \beta_6 \) | -0.0254 ** | 0.0333 |
| \( \rho \) | 0.1117 | -0.1320 |

*Significant at \( \alpha = 10\% \), **Significant at \( \alpha = 5\% \)

### 3.8. The Best Model Selection

Based on the AIC and BIC values, we found that the SAR model has smaller AIC and BIC values than the SDM and OLS models for both 4-NN and IDW weighting (table 8). However, the 4-NN spatial weighting for SAR and SDM models has a smaller AIC and BIC value than the IDW weighting. Therefore, the SAR model with 4-NN spatial weighting is the best model for GRDP data of Sulawesi Island. The model has fulfilled the assumption. The residuals are independent, normally distributed, and have constant variance (p-value > 0.05) (table 9).

### Table 8. Evaluating the best fit of the model

| Model | Spatial weight | AIC  | BIC  |
|-------|----------------|------|------|
| MKT   | -              | 101.44 | 118.21 |
| SAR   | 4-NN           | 98.91 | 118.07 |
|       | IDW            | 100.14 | 119.30 |
| SDM   | 4-NN           | 99.40 | 130.53 |
|       | IDW            | 107.69 | 138.81 |

### Table 9. Checking the assumptions of the SAR model

| Assumption                  | P-value |
|-----------------------------|---------|
| Normal distribution         | 0.1762  |
| Homogeneity of the residuals| 0.2776  |
| Independence of the residuals| 0.9099 |

### 3.9. Marginal Effects

The interpretation of the coefficients in the spatial regression model cannot be explained by the estimated value of the parameters because changes in unit value from the explanatory variable \( X_p \) in a
A particular region will affect the region itself (direct impact) and potentially affect all other regions indirectly (indirect effect) [3]. Table 10 shows the magnitude of the direct, indirect, and total effects generated by changes in unit value from each explanatory variable on GRDP in a region.

The direct and indirect effects of the population size \(X_2\) have the most significant positive effect on GRDP. One hundred thousand inhabitants increase in the population size in a given region \(i\) will give the average increasing effect on GRDP in the same region of 1.2591 billion rupiahs, and give the average effect on GRDP in other regions \(j\) is 1.0535 billion rupiahs. Cumulatively, increasing one hundred thousand inhabitants of population size in a given region gives the average effect on GRDP in all regions is 1.3264 billion rupiahs.

The unemployment rate \(X_4\) decreases the GRDP value. One percent increase in the unemployment rate in a given region will give the average effect on GRDP in the same region of 0.9162 while increasing one percent of the unemployment rate in other regions will give the average effect on GRDP of 0.9804. Cumulatively, changes in one percent of the unemployment rate in a given region give the average effect on GRDP in all regions of 0.8983.

An increase in one unit of the human development index \(X_1\) in a given region gives the average effect on GRDP in all regions of 1.1064. An increase in one thousand unit number of small/micro and medium industries \(X_5\) in a given region gives the average effect on GRDP in all regions of 1.0453. Meanwhile, the number of medium and large manufacturing industries \(X_6\) has a minor effect among the other variables because it has no significant effect on GRDP.

| Variable | Direct effect | Indirect effect | Total effect |
|----------|---------------|----------------|-------------|
| \(X_1\)  | 0.0825        | 0.0187         | 0.1012      |
| \(X_2\)  | 0.2304        | 0.0521         | 0.2825      |
| \(X_4\)  | -0.0875       | -0.0198        | -0.1073     |
| \(X_5\)  | 0.0361        | 0.0082         | 0.0443      |
| \(X_6\)  | 0.0052        | 0.0012         | 0.0064      |

4. Conclusion
The best model for modeling the district/city's GRDP data on Sulawesi Island in 2018 is the 4-NN weighted SAR model. The significant factors that influence the amount of districts/city's GRDP are the human development index \(X_1\), the population size \(X_2\), the unemployment rate \(X_4\), the number of small/micro and medium industries \(X_5\), and the spatial lag of the response variable \(\rho\).

References
[1] Tobler W R 1970 A Computer Movie Simulating Urban Growth in the Detroit Region Economic Geography 46 234-240
[2] Anselin L 1988 Spatial Econometrics: Methods and Models Dordrecht (NL): Kluwer Academic
[3] LeSage J and Pace R K 2009 Introduction to Spatial Econometrics New York (US): CRC Press
[4] Elhorst J P 2010 Applied Spatial Econometrics: Raising the Bar Spat. Econ. Anal. 5 9–28.
[5] Getis A and Aldstadt J 2004 Constructing the spatial weights matrix using local statistics Geographical Analysis 36(2):90-104
[6] Anselin L and Bera A 1998 Spatial dependence in linear regression models with an introduction to spatial econometrics In: Ullah A, Giles DE (eds) Handbook of applied economic statistics Marcel Dekker, New York
[7] Stakhovych S and Bijmolt T H A 2009 Specification of spatial models: a simulation study on weights matrices Papers in Regional Science 88 389-408
[8] BPS 2019 PDB Indonesia Triwulan 2015-2019 Jakarta (ID): Badan Pusat Statistik
[9] Yulita I, Djuraidah A and Wigena A H 2016 Pemodelan regresi komponen utama dan lasso terboboti geografis (global dan lokal) (Studi kasus: Data Produk Domestik Regional Bruto (PDRB) pada 113 kabupaten/kota di Pulau Jawa) [tesis] Bogor (ID): Institut Pertanian Bogor

[10] Arisuaka K and Purmiyati A 2010 Faktor-faktor yang mempengaruhi produk domestik regional bruto di Pulau Sumatera tahun 2000-2006 (Studi kasus bencana tsunami) [tesis] Surabaya (ID): Universitas Airlangga

[11] Draper NR and Smith H 1998 Applied Regression Analysis 3rd ed New York (US): John Wiley & Sons

[12] Cellmer R 2013 Use of spatial autocorrelation to build regression models of transaction prices Real Estate Management and Valuation 21(4) 65-74

[13] Chen X, Xuan C, and Qiur R 2021 Understanding spatial spillover effects of airports on economic development: New evidence from China’s hub airports Transportation Research Part A 143 48–60

[14] Burnham K P and Anderson D R 2004 Multimodel inference: Understanding AIC and BIC in model selection Sociological Methods Research 33(2) 261-304