

SPATIAL MODELING IN DATA PANELS WITH LEAST SQUARE DUMMY VARIABLE TO IDENTIFY FACTORS AFFECTING UNEMPLOYMENT IN INDONESIA

Sri Indriani Amil¹, Sri Astuti Thamrin^{2*}, Siswanto³

^{1,2,3}Department of Statistics, Faculty of Mathematics and Natural Sciences, Hasanuddin University
Perintis Kemerdekaan Street Km.10, Makassar, 90245, Indonesia.

Corresponding author's e-mail: * tuti@unhas.ac.id

ABSTRACT

Article History:

Received: 23rd January 2023

Revised: 13th July 2023

Accepted: 23rd July 2023

Keywords:

Dummy variable;

Least square;

Open unemployment rate;

Panel data;

Spatial autoregressive

model;

Spatial error model.

Unemployment is a serious issue that must be addressed. Unemployment has a negative impact on the national economy, making economic growth unpredictable. In 2015, Indonesia was ranked third with the highest unemployment rate in ASEAN. It is estimated that the unemployment rate in each province of Indonesia is influenced by the surrounding provinces. Therefore, spatial modelling on panel data with Least Square Dummy Variable (LSDV) is needed to identify factors that influence unemployment in Indonesia. The data used is Open Unemployment Rate (OUR) data, and influencing factors are population, average length of schooling, Gross Regional Domestic Product (GRDP) rate, and Human Development Index (HDI) in 34 provinces of Indonesia from 2015 to 2020. Spatial model on panel data with appropriate LSDV for OUR data are the Spatial Autoregressive Model (SAR) and Spatial Error Model (SEM). The SAR model with fixed effects has an R^2 value of 90.289%, which is greater than the SEM model with fixed effects (82.708%) and LSDV model (87.864%). The root mean square error value for SAR model with fixed effects is 0.58951, less than SEM model with fixed effects (0.78669) and LSDV model (0.65903). The best model is the SAR model with fixed effects. Based on this model, the factors that influence OUR in Indonesia from 2015-2020 are obtained, namely the rate of GRDP and HDI.



This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 International License.

How to cite this article:

S. I. Amil, S. A. Thamrin and Siswanto., "SPATIAL MODELING IN DATA PANELS WITH LEAST SQUARE DUMMY VARIABLE TO IDENTIFY FACTORS AFFECTING UNEMPLOYMENT IN INDONESIA," *BAREKENG: J. Math. & App.*, vol. 17, iss. 3, pp. 1381-1392, September, 2023.

Copyright © 2023 Author(s)

Journal homepage: <https://ojs3.unpatti.ac.id/index.php/barekeng/>

Journal e-mail: barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id

Research Article • Open Access

1. INTRODUCTION

Unemployment is still a socio-economic problem that must be solved by the state apart from poverty. An unemployed is not working but is looking for work. The Open Unemployment Rate (OUR) is one indicator to see the unemployment rate [1]. Therefore, to formulate an effective public policy to reduce unemployment, it is necessary to know the factors that influence OUR in Indonesia [2].

The unemployment rate is influenced by several factors, including the population, education in this case the average length of schooling, Gross Regional Domestic Product (GRDP), and the Human Development Index (HDI). Previous research on OUR by Arjun et al (2019) found that the average length of schooling, labor force participation rate, percentage of the poor, the rate of economic growth, and the number of large industries had an effect on OUR in Kalimantan [3]. According to the results of research by Palindangan and Bakar (2021) HDI has a significant effect on OUR in Mimika Regency [4]. Handayani's research (2019) shows that the population, education, minimum wage, and GRDP together have a significant influence on the open unemployment rate in Central Java Province [5]. When observing a problem, it is not enough to only observe one time at the same time, but observe at several periods. Panel data regression analysis is one approach that includes the effect of time on the model [6].

Approach methods for estimating panel data regression models include pooled Ordinary Least Square [7], and Least Square Dummy Variable (LSDV). Pooled Ordinary Least Square (OLS) is the simplest way to estimate model parameters by combining time series data and cross-section data, pooled OLS estimates do not differentiate between time series and between cross-sections [8]. LSDV regression model can be used to explain individual heterogeneity through dummy variables [9]. In addition to using panel data, spatial analysis is needed to overcome heterogeneous errors caused by inter-regional linkages.

Spatial analysis is an analysis that incorporates the influence of the region into the model. To incorporate the influence or regional effects into the model, a spatial weighting matrix is required. The weighting matrix is obtained from neighboring or distance information. Indonesia consists of 34 provinces in the form of islands, so the appropriate weighting matrix is the inverse distance matrix, which is the weight of the distance between an area and the surrounding area by the distance between the two regions [10]. Spatial regression has various methods including the Spatial Autoregressive Model (SAR) and the Spatial Error Model (SEM). SAR has a dependence on the response variable and SEM has a dependence on errors between locations. The parameter estimation used to estimate the SAR and SEM parameters is Maximum Likelihood Estimation (MLE) and both models require a weighting matrix [6]. These two methods are often used for OUR problems.

A previous study on spatial modeling of panel data to determine factors that affect health in Papua Province by Ira Rosianal Hikmah and Yulial Hikmah (2020), found that the SAR panel model with fixed effects proved to be better than the regular panel model [11], contrast to this research focuses on spatial panel data with LSDV to determine the factors that influence OUR in Indonesia. Research by Reydita (2020) on spatial analysis of panel data on the human development index in West Java Province and Banten Province. Reydita (2020) found that the best spatial model for panel data is a SAR fixed effect model using a queen contiguity spatial weighting matrix without transformation [12]. In this study, spatial modeling of panel data with LSDV will be carried out. So that through the spatial model on panel data with LSDV obtained it will be known what factors have a significant effect on OUR in Indonesia from 2015 to 2020. This is expected to help the government to overcome the problem of unemployment in Indonesia.

2. RESEARCH METHODS

The data used in this study is secondary data, namely data sourced from the website of the Central Statistics Agency (<https://bps.go.id/>) from 2015-2020. Research variables consist of response variables and predictor variables. The response variable in this study is the Open Unemployment Rate (Y). The predictor variables in this study were the population (X_1), the average length of schooling (X_2), the rate of Gross Regional Domestic Product (X_3), and the Human Development Index (X_4).

2.1 Research Procedure

The procedures performed in the spatial panel data analysis are as follows:

1. Describe the characteristics of the open unemployment rate variable using a thematic map.

2. Calculating the value of the correlation coefficient of the response variable to the predictor variable.
3. Modeling panel data. The following are the stages in conducting panel data modeling:
 - a. Estimating OLS pooled parameters.

OLS pooled model **Equation (1)**:

$$Y_{it} = \alpha + \sum_{k=1}^p \beta_k X_{ikt} + \varepsilon_{it} \quad (1)$$

with $i = 1, \dots, n$; $t = 1, \dots, T$; and $k = 1, \dots, p$

n : number of local units

T : number of periods

p : number of predictor variables

Y_{it} : the response variable of local unit i in period t

α_{it} : the intercept of local unit i regression model in period t

β_k : the slope of the predictor variable k of local unit i in period t

X_{ikt} : the k -th predictor variable from local unit i in period t

ε_{it} : error for local i in period t

Pooled OLS are estimated using the OLS method, in **Equation (1)** can be written in matrix form as follows:

$$Y = X\beta + \varepsilon$$

Estimation of β :

$$S = \varepsilon' \varepsilon$$

$$S = (Y - X\beta)'(Y - X\beta)$$

$$S = Y'Y - 2\beta'X'Y + \beta'X'X\beta$$

$$\frac{dS}{d\beta} = \frac{d(Y'Y - 2\beta'X'Y + \beta'X'X\beta)}{d\beta}$$

$$\frac{dS}{d\beta} = -2X'Y + 2X'X\beta$$

$$X'X\hat{\beta} = X'Y$$

$$\hat{\beta} = (X'X)^{-1}X'Y$$

$$\hat{\beta} = \begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

- b. Estimating LSDV parameters.

LSDV model **Equation (2)**:

$$Y_{it} = \alpha_i \sum_{j=2}^n D_{ij} + \sum_{k=1}^p \beta_k X_{ikt} + \varepsilon_{it} \quad (2)$$

with

α_i : regression model intercept

D_{ij} : dummy variable from j location. D_{ij} is 1 if $j = i$ and 0 if $j \neq i$

Example of specifying a dummy variable:

Suppose there are 3 locations

So that

$i = 1$ and $j = 1$ dummy variable value is 1 ($D_{11} = 1$)

$i = 1$ and $j = 2$ dummy variable value is 0 ($D_{12} = 0$)

$i = 1$ and $j = 3$ dummy variable value is 0 ($D_{13} = 0$)

$i = 2$ and $j = 1$ dummy variable value is 0 ($D_{21} = 0$)

$i = 2$ and $j = 2$ dummy variable value is 1 ($D_{22} = 1$)

$i = 2$ and $j = 3$ dummy variable value is 0 ($D_{23} = 0$)

$i = 3$ and $j = 1$ dummy variable value is 0 ($D_{31} = 0$)

$i = 3$ and $j = 2$ dummy variable value is 0 ($D_{32} = 0$)

$i = 3$ and $j = 3$ dummy variable value is 1 ($D_{33} = 1$)

LSDV parameter estimation, in **Equation (2)** can be written in matrix form as follows:

$$\begin{aligned} Y &= D\alpha + X\beta + \varepsilon \\ &= [D \ X][\alpha \ \beta] + \varepsilon \end{aligned}$$

Supposen $M = [D \ X]$ dan $\theta = [\alpha \ \beta]$, it can be written as:

$$Y = M\theta + \varepsilon$$

Next look for the estimated parameter θ . Estimation of θ uses the least squares method by minimizing the total square error function.

$$S = \varepsilon' \varepsilon$$

$$S = (Y - M\theta)'(Y - M\theta)$$

$$\frac{dS}{d\theta} = \frac{d((Y - M\theta)'(Y - M\theta))}{d\theta}$$

$$\frac{dS}{d\theta} = \frac{d((Y' - \theta' M')(Y - M\theta))}{d\theta}$$

$$2 M' M \hat{\theta} = M' Y$$

So obtained:

$$[D' \ X'] [D \ X] [\alpha \ \beta] = [D' \ X'] Y$$

$$D' D \hat{\alpha} + D' X \hat{\beta} = D' Y \quad (3)$$

$$X' D \hat{\alpha} + X' X \hat{\beta} = X' Y \quad (4)$$

Forming the parameter estimation $\hat{\alpha}$ using **Equation (3)** is as follows:

$$D' D \hat{\alpha} + D' X \hat{\beta} = D' Y$$

$$\hat{\alpha} = (D' D)^{-1} D' Y - (D' D)^{-1} D' X \hat{\beta} \quad (5)$$

The parameter estimation from $\hat{\beta}$ is obtained by substituting **Equation (5)** into **Equation (4)**.

$$X' D \hat{\alpha} + X' X \hat{\beta} = X' Y$$

$$X' D [(D' D)^{-1} D' Y - (D' D)^{-1} D' X \hat{\beta}] + X' X \hat{\beta} = X' Y$$

$$X' D (D' D)^{-1} D' Y + X' [I - D (D' D)^{-1} D'] X \hat{\beta} = X' Y$$

Suppose $D (D' D)^{-1} D' = P$, then we get:

$$X' P Y + X' (I - P) X \hat{\beta} = X' Y$$

$$\hat{\beta} = [X' (I - P) X]^{-1} X' (I - P) Y$$

c. Testing the significance to determine the combined model or fixed effect model using the Chow test.

4. Spatial modeling on panel data with LSDV:

a. Create a spatial weighting matrix using a distance inverse weighting matrix.

b. Doing spatial autocorrelation testing using the Moran Index method.

Index Moran **Equation (6)**:

$$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n W_{ij} (x_i - \bar{x})(x_j - \bar{x})}{S_0 \sum_{i=1}^n (x_i - \bar{x})} \quad (6)$$

with

$$S_0 = \sum_{i=1}^n \sum_{j=1}^n W_{ij}^*$$

- W_{ij} : elements on a weighting between regions i and j
 I : Moran Index
 x_i : value on location i
 x_j : value on location j
 \bar{x} : the average of the number of variables or values

- c. Identify spatial patterns using the Lagrange Multiplier lag and Lagrange Multiplier error.
- d. Perform SAR modeling with the following steps:
 - i. Finding the value of estimating the parameters of the SAR model using the MLE method.
 - ii. Modeling SAR.

SAR model **Equation (7)**:

$$Y_{it} = \rho \sum_{j=1}^n W_{ij} Y_{jt} + \sum_{k=1}^p \beta_k X_{ikt} + \alpha_i + \varepsilon_{it} \quad (7)$$

with

- ρ : spatial autoregressive coefficient
 W_{ij} : elements on a weighting between regions i and j , using a distance inverse weighting matrix
 Y_{it} : individual response variable i at the time t
 Y_{jt} : response variable j neighbor location at the time t
 α_i : intercept LSDV location i

- e. Perform SEM modeling with the following steps:
 - i. Find the parameter estimator value of the SEM model with the MLE method.
 - ii. Modeling SEM.

SEM model **Equations (8)**:

$$Y_{it} = \sum_{k=1}^p \beta_k X_{ikt} + \alpha_i + \phi_{it} \quad (8)$$

with

- $\phi_{it} = \lambda \sum_{j=1}^N W_{ij} \phi_{jt} + \varepsilon_{it}$
 ϕ_{it} : the error from individual i spatial autocorrelation and the period t
 λ : spatial autocorrelation coefficient

- f. Selection of the best model by:
 - i. Seeing the largest coefficient of determination (R^2) [6].

$$R^2 = 1 - \frac{SSE}{SST}$$

dengan

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

y_i : the results of the data observation i

\hat{y}_i : data prediction results from i

\bar{y} : the average result of data observation

- ii. Look at the smallest Root Mean Squared Error (RMSE) value [18].

$$RMSE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}}$$

3. RESULTS AND DISCUSSION

3.1 Description of Unemployment Variable Characteristics with Thematic

Classification of data distribution is divided into 5 categories, namely very high, high, medium, low, and very low. The higher the OUR value in an area is indicated by a darker color gradation.

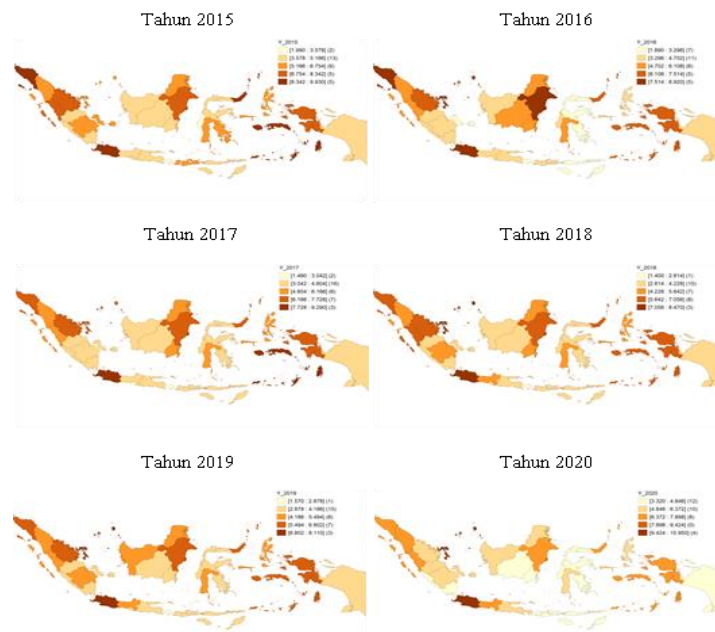


Figure 1. Thematic map of the open unemployment rate in Indonesia in the period of 2015 to 2020

Figure 1 shows that there is an annual change in OUR in most provinces of Indonesia. OUR classified as very high, marked with a maroon color, was found in the provinces of Banten, West Java, DKI Jakarta, Kep. Riau and Maluku. Other provinces marked with light orange colors remain consistent from year to year, namely Jambi, East Java, and North Sumatra. This shows that there is a relationship between one region and another. According to Tobler's first law, everything is related to everything else, but something closer is more influential than something that is far away.

3.2 Panel Data Analysis

In this analysis, the Chow test will be carried out with a significance level of 5%. Before doing the test, the first thing to do is look for the estimation results of the pooled OLS and LSDV models. The following is the estimation result of the pooled OLS model.

Table 1. Pooled OLS Model Parameter Estimation Results

Variable	Coefficient	P-value
Intercept	9.76379	0.00000
X_1	0.00008	0.00000
X_2	1.95211	0.00000
X_3	-0.10238	0.00018
X_4	-0.29893	0.00000
R^2		0.43210

Table 1 shows that the overall estimation results of the pooled OLS model are significant at the 5% level. There is an inverse relationship between the rate of GRDP (X_3), and HDI (X_4) to the open unemployment rate (Y). However, the open unemployment rate has a positive relationship with the population (X_1), and the average length of schooling (X_2). R^2 is worth 43.210% explaining that all predictor variables in the model can explain the diversity of data by 43.210% to the response variable. While the remaining 56.790% is explained by other variables not tested in this study. Furthermore, the parameter estimation of the LSDV model is carried out.

Table 2. LSDV Model Parameter Estimation Results

Variable	Coefficient	P-value
X_1	0.00074	0.01306
X_2	2.99659	0.00114
X_3	-0.10593	0.00000
X_4	-1.11906	0.00000
R^2		0.87864

Table 2 shows that the overall estimation results for the LSDV model are significant at the 5% level. R^2 is 87.864% which explains that all predictor variables in the model can explain 87.864% of the data diversity for the response variable. The remaining 12.136% is explained by other variables not tested in this study. Next, perform the Chow test to determine the best panel data model. The following hypotheses in the Chow test:

Hypotheses:

$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_{n-1} = 0$ (there is no effect of location on the model or the model follows the pooled OLS model)

$H_1 : \text{there is at least one } \alpha_i \neq 0$ (there are one or more location effects on the model or the model follows the LSDV model).

The test statistics used are:

$$F_0 = \frac{(RRSS - URSS)/(n - 1)}{URSS/(nT - n - p)}$$

Test criteria:

Decision reject H_0 if $F_0 > F_{n-1, n(T-1)-p}$ or if $p\text{-value} < \alpha$ [19].

Table 3. Chow Test Results

F	Df_1	Df_2	P-value	F_{table}
18.51000	33	166	0.00000	1.50674

Table 3 present the results of the chow test. The results of $F_0 > F_{table}$ are $18.51000 > 1.50674$, meaning that H_0 is rejected at the 5% significance level, or in other words the model has one or more individual effects (the model follows the LSDV model). So that the panel data regression model produces the LSDV model is the best model. Here is the LSDV model:

$$y_{it} = \hat{\alpha}_i + 0.00074x_{1it} + 2.99659x_{2it} - 0.10593x_{3it} - 1.11906x_{4it}$$

In the LSDV model, the intercept value for individual units is different and the slope is constant so that every 1% increase in population will increase OUR by 0.00074% and a 1% increase in the average length of school can also increase OUR by 2.99659%. On the other hand, an increase in GRDP will reduce OUR cases by 0.10593% and an increase in HDI by 1% will reduce OUR by 1.11906%.

3.3 Spatial Autocorrelation Test

A spatial autocorrelation test is used to determine whether there is a correlation between regions. Before performing the autocorrelation test, first make a weighting matrix inverse distance. Spatial weighting is an important step in measuring and modeling spatial dependence in area data [20]. Neighborhood relationships are formally expressed in an $n \times n$ matrix with elements (W_{ij}) representing the size of the relationship between locations i and j . The specification of the spatial weighting matrix begins by identifying the neighboring structures of each cell [21]. Mathematically, the calculation of the distance inverse spatial weighting matrix is [22]:

$$W = [W_{ij}] = \begin{cases} 0, & i = j \\ \frac{1}{d_{ij}}, & i \neq j \\ \frac{1}{\sum_{j=1}^n \frac{1}{d_{ij}}}, & i \neq j \end{cases}$$

with

$$d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2}$$

d_{ij} : location distance i and j

u_i : longitude location i

u_j : longitude location j

v_i : latitude location i

v_j : latitude location j

n : number of individual units/locations

The following is an example of calculating the spatial distance function for Aceh Province.

$$d_{11} = \sqrt{(96.94310 - 96.94310)^2 + (4.28439 - 4.28439)^2} = 0$$

$$d_{12} = \sqrt{(96.94310 - 115.11643)^2 + (4.28439 - (-8.35546))^2} = 22.13675$$

⋮

$$d_{134} = \sqrt{(96.94310 - 99.15446)^2 + (4.28439 - 2.30447)^2} = 2.96820$$

After calculating the spatial distance function, then calculate the inverse spatial distance weights as follows.

$$W_{ij} = \begin{cases} 0, & i = j \\ \frac{1}{d_{ij}}, & i \neq j \\ \frac{1}{\sum_{j=1}^n \frac{1}{d_{ij}}}, & i \neq j \end{cases}$$

$$W_{11} = 0$$

$$W_{12} = \frac{1}{d_{12}} = \frac{1}{22.13675} = 0.02914$$

⋮

$$W_{134} = \frac{1}{d_{134}} = \frac{1}{2.96820} = 0.14278$$

The following is the form of the W weighting matrix:

$$W = \begin{bmatrix} 0 & W_{12} & \dots & W_{134} \\ W_{21} & 0 & \dots & W_{234} \\ \vdots & \vdots & \ddots & \vdots \\ W_{341} & W_{342} & \dots & 0 \end{bmatrix}$$

After obtaining the spatial weighting matrix, then the spatial autocorrelation test was carried out. The following are the results of the spatial autocorrelation test.

The hypothesis for the Moran Index is as follows:

H_0 : $I = 0$ (There is no spatial autocorrelation)

H_1 : $I > 0$ (There is a positive spatial autocorrelation) or

H_1 : $I < 0$ (There is a negative spatial autocorrelation)

Test Statistics:

$$Z(I) = \frac{I - E(I)}{\sqrt{Var(I)}}$$

with

$$E(I) = -\frac{1}{n-1}$$

$$Var(I) = \frac{n^2 S_1 - n S_2 + 3 S_0}{(n^2 - 1) S_0^2}$$

$$S_0 = \sum_{i=1}^n \sum_{j=1}^n W_{ij}^*$$

$$S_1 = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (W_{ij}^* + W_{ji}^*)^2$$

$$S_2 = \sum_{i=1}^n (\sum_{j=1}^n W_{ij}^* + \sum_{j=1}^n W_{ji}^*)^2$$

Test Criteria:

Reject H_0 at the significance level $\alpha = 5\%$ if $Z(I) > Z_{1-\alpha}$ or $Z(I) < -Z_{1-\alpha}$ or p-value $< \alpha$ [23].

Table 4. Spatial Autocorrelation Test Results

	2015	2016	2017	2018	2019	2020
<i>Y</i>	2.12970	1.91630	2.54940	2.63720	3.16030	4.29610
<i>X</i> ₁	3.00970	3.01410	3.01860	3.02320	3.02780	3.03250
<i>X</i> ₂	0.25803	0.19625	0.23268	-0.03977	-0.18913	-0.10720
<i>X</i> ₃	1.57770	2.24290	1.71200	1.17280	2.03250	2.46770
<i>X</i> ₄	2.79500	2.84260	2.76060	2.72760	2.60100	2.63220

Table 4 shows that the number in bold is a variable with a value greater than $Z_{tabel} = 1.96$ meaning that it rejects H_0 , i.e., there is a spatial autocorrelation. Variable *Y* has a spatial autocorrelation for 2015, 2017-2020 and no spatial autocorrelation for 2016. The test statistic value is $1.91630 < 1.96$, meaning that it accepts H_0 . Variables *X*₁ and *X*₄ for the years 2015-2020 have spatial autocorrelation. Variable *X*₂ has no spatial autocorrelation for each year. Variable *X*₃ obtains different $Z(I)$, in 2016, 2019 and 2020 $Z(I)$ is more than 1.96 which means there is a spatial autocorrelation and in 2015, 2017 and 2018 $Z(I)$ is less than 1.96 meaning there is no autocorrelation spatial. The statistical results of the Moran Index test show that the variables *Y*, *X*₁, *X*₃ and *X*₄ have positive autocorrelation.

3.4 Panel Data Spatial Analysis

Before doing the spatial modeling of the panel data, first perform the Lagrange Multiplier (LM) test to determine the appropriate spatial model of the panel data.

The hypothesis used in the LM lag is:

$H_0 : \rho = 0$ (No lag spatial dependencies)

$H_1 : \rho \neq 0$ (There are spatial dependencies lag)

The test statistics used in LM lag are:

$$LM_{\rho} = \frac{\left[\frac{e'_{1 \times nT} (I_{T \times T} \otimes W_{n \times n}) Y_{nT \times 1}}{\hat{\sigma}^2} \right]^2}{J}$$

with

$$J = \frac{1}{\hat{\sigma}^2} \left[(a'_1 a_2 a_1) + TT_w \hat{\sigma}^2 \right]$$

$$a_1 = (I_{T \times T} \otimes W_{n \times n}) X_{nT \times k} \hat{\beta}_{k \times 1}$$

$$a_2 = (I_{nT \times nT} - X_{nT \times k} (X'_{k \times nT} X_{nT \times k})^{-1} X'_{k \times nT})$$

$$T_w = tr((W'_{n \times n} + W_{n \times n}) W_{n \times n})$$

Test Criteria:

Decision making on LM lag will reject H_0 if the value of $LM_{\rho} > \chi^2_{(\alpha,1)}$. While the LM Error test uses the following hypothesis test.

LM hypothesis error:

$H_0 : \delta = 0$ (No error spatial dependencies)

$H_1 : \delta \neq 0$ (There are spatial dependencies errors)

The test statistic used in the LM error is:

$$LM_{\delta} = \frac{\left[\frac{e'_{1 \times nT} (I_{T \times T} \otimes W_{n \times n}) e_{nT \times 1}}{\hat{\sigma}^2} \right]^2}{T \times T_w}$$

Test Criteria:

Decision making on LM error will be rejected H_0 if the value of $LM_{\delta} > \chi^2_{(\alpha,1)}$ [24].

Table 5. Lagrange Multiplier Test Results

LM Test	Coefficient	P-value	$\chi^2_{(0.05,1)}$
Lag	66.78300	0.00000	3.84150
Error	41.46700	0.00000	

Table 5 shows that the coefficient of LM lag and LM error is greater than the value of Chi-square ($\chi^2_{(0.05,1)}$) which is $66.78300 > 3.84150$ and $41.467000 > 3.84150$ which means rejecting H_0 or in other words, there is a spatial dependence of lag and error so that it can perform SAR and SEM modeling on panel data.

The SAR model with fixed effect has a dependence on the response variable between locations. The following are the results of the SAR model parameter estimation with fixed effects, because the results of the chow test in table 3 show the model has one or more individual effects (the model follows the LSDV model).

Table 6. Estimation Results of SAR Model Parameters with Fixed Effects

Variable	Coefficient	$ t_{count} $	P-value
ρ	0.64246	7.94440	0.00000
X_1	0.00047	1.96130	0.04984
X_2	0.64207	0.84340	0.39903
X_3	-0.05534	3.63400	0.00028
X_4	-0.38645	2.05580	0.03980
R^2			0.90426

Table 6 shows that the p-value for the variable X_1 is significant at the 5% level, but the value of $t_{count} < t_{table}$ ($1.96130 < 1.97196$) means that it is not significant at the 5% level. The results of the p-value and t_{count} for the variable X_2 are not significant at the 5% significance level. Therefore, the modeling was carried out again without involving the X_1 and X_2 variables.

Table 7. Estimation Results of SAR Model Parameters (Without X_1 and X_2) with Fixed Effect

Variable	Coefficient	P-value
ρ	0.68813	0.00000
X_3	-0.05707	0.00001
X_4	-0.16931	0.00262
R^2		0.90289

Table 7 shows that the results of the p-value of the variables X_3 and X_4 are significant at the 5% significance level. So that the variables X_3 and X_4 can be included in the SAR model with a fixed effect. Next, we estimate the parameters of the SEM model with fixed effects. The following are the results of the estimation of the parameters of the SEM model with fixed effects.

Table 8. The Result of Parameter Estimation of SEM Model with Fixed Effect

Variable	Coefficient	P-value
λ	0.79141	0.00000
X_1	0.00027	0.27630
X_2	0.04465	0.96047
X_3	-0.05301	0.00192
X_4	-0.55371	0.02350
R^2		0.83569

Table 8 shows that the p-value of the variables X_1 and X_2 is not significant at the 5% significance level. The modeling is done again without involving X_1 and X_2 variables. The following is the result of parameter estimation without variables X_1 and X_2 .

Table 9. Estimation Results of SEM Model Parameters (Without X_1 and X_2) with Fixed Effect

Variable	Coefficient	P-value
λ	0.05536	0.00000
X_3	-0.05237	0.00219
X_4	-0.53508	0.00196
R^2		0.82708

The p-value results from **Table 9** shows that the variables X_3 and X_4 are significant at the 5% significance level. Therefore, these variables can be included in the model.

3.5 Evaluation Model

Model evaluation was carried out to select the best panel data spatial model based on the value of R^2 and Root Mean Square Error (RMSE). The results of the evaluation of the best models are presented in **Table 10**.

Table 10. Best Evaluation Model

Model	R^2	RMSE
LSDV	0.87864	0.65903
SAR with fixed effect	0.90289	0.58951
SEM with fixed effect	0.82708	0.78669

Table 10 shows the results of R^2 SAR with fixed effects that are greater than the LSDV and SEM models with fixed effects, namely $0.90289 > 0.87864$ and $0.90289 > 0.82708$ and the RMSE value of the SAR model with fixed effects are less than the LSDV and SEM models with fixed effects, namely $0.58951 < 0.65903$ and $0.58951 < 0.78669$. So that the spatial model of panel data with the best LSDV is the SAR model with fixed effects.

SAR model with fixed effects:

$$\hat{Y}_{it} = 0.68813 \sum_{j=1}^{34} W_{ij} Y_{jt} - 0.05707 X_{3it} - 0.16931 X_{4it} + \hat{\alpha}_i$$

The SAR model with fixed effects assumes that other variables are constant so that a 1% increase in the rate of GRDP reducing the number of OUR cases by 0.05707% and an increase of 1% in HDI can also reduce the number of OUR cases by 0.16931% with the influence of each province that surrounds it by the coefficient value is 0.68813 multiplied by the average from the province that has OUR cases around it.

4. CONCLUSIONS

The best spatial model on panel data with LSDV to identify the factors of the Open Unemployment Rate (OUR) in 34 Indonesian provinces from 2015 to 2020 is the SAR panel model with fixed effects. The modeling results are as follows:

$$\hat{Y}_{it} = 0.68813 \sum_{j=1}^{34} W_{ij} Y_{jt} - 0.05707 X_{3it} - 0.16931 X_{4it} + \hat{\alpha}_i$$

Based on this model, the factors that have a significant effect on OUR in 34 provinces of Indonesia are the rate of GRDP (X_3) and HDI (X_4). GRDP and HDI factors have a negative influence on the open unemployment rate in Indonesia. Suggestions for the government to raise Indonesia's GRDP and HDI in order to minimize the percentage of OUR.

REFERENCES

- [1] E. Amalia and L. K. Sari, "Analisis Spasial untuk Mengidentifikasi Tingkat Pengangguran Terbuka Berdasarkan Kabupaten/Kota di Pulau Jawa Tahun 2017," *Indones. J. Stat. Its Appl.*, vol. 3, no. 3, pp. 202–215, 2019.
- [2] R. Ilahi, M. Syamsuddin, and Y. Suparman, "Model Spasial Durbin dengan Efek Tetap untuk Tingkat Pengangguran Terbuka di Provinsi Kepulauan Bangka Belitung," pp. 1–12, 2014.
- [3] D. A. Arjun, Sifriyani, and Syaripuddin, "Analisis Faktor-Faktor Yang Mempengaruhi Menggunakan Regresi Nonparametrik Spline," *Pros. Semin. Nas. Mat. Stat. dan Apl.*, vol. 1, pp. 115–121, 2019.
- [4] J. Palindangan and A. Bakar, "Analisis Pengaruh Tingkat Pertumbuhan Ekonomi Dan Indeks Pembangunan Manusia (IPM) Terhadap Tingkat Pengangguran Di Kabupaten Mimika," *J. Krit. (Kebijakan, Riset, dan Inovasi)*, vol. 5, no. 1, pp. 65–80, 2021.
- [5] D. P. H. R. Handayani, "Analisis Pengaruh Jumlah Penduduk, Pendidikan, Upah Minimum, dan PDRB terhadap Tingkat Pengangguran Terbuka Di Provinsi Jawa Tengah," *Diponegoro J. Econ.*, vol. 1, no. 1, pp. 159–169, 2019.
- [6] J. P. Elhorst, *Spatial Panel Data models*. New York: Springer, 2010.
- [7] Ö. B. Soyulu, İ. Çakmak, and F. Okur, "Economic growth and unemployment issue: Panel data analysis in Eastern European Countries," *J. Int. Stud.*, vol. 11, no. 1, pp. 93–107, 2018, doi: 10.14254/2071-8330.2018/11-1/7.
- [8] J. Brüderl and V. Ludwig, "Fixed-effects Panel Regression," *Regres. Anal. Casual Inference*, no. 1, pp. 327–357, 2015.
- [9] G. K. Appiah, K. Prempeh, S. Benn, and E. Oduro, "Panel Data Analysis of Employment and Unemployment in Sub-Saharan Africa and Its Impact on Annual Percentage Change in GDP Growth," *Open Sci. J.*, vol. 5, no. 2, pp. 1–26, 2020.
- [10] R. Dubin, A. S. Fotheringham, and P. A. Rogerson, *Spatial Weights*. London: Sage Publications, 2009.
- [11] I. R. Hikmah and Y. Hikmah, "Pemodelan Spasial Data Panel Untuk Menentukan Faktor-Faktor Yang Mempengaruhi

- Kesehatan Di Provinsi Papua,” *Statmat J. Stat. Dan Mat.*, vol. 2, no. 2, pp. 176–188, 2020.
- [12] T. Reydita, “Analisis Spasial Data Panel Pada Indeks Pembangunan Manusia Provinsi Jawa Barat dan Provinsi Banten,” 2020.
- [13] D. N. Gujarati and D. C. Porter, *Basic Econometrics*, Fifth Edit. New York, 2013.
- [14] C. Hsiao, *Analysis of panel data, second edition*. The Press Syndicate of the University of Cambridge, 2014.
- [15] L. Anselin, *Spatial Econometrics: Methods and Models*. London, 1988.
- [16] L. Anselin, A. S. Fotheringham, and P. A. Rogerson, *Spatial Regression*. London: Sage Publications, 2009.
- [17] J. P. Elhorst, “Specification and estimation of spatial panel data models,” *Int. Reg. Sci. Rev.*, vol. 26, no. 3, pp. 244–268, 2003.
- [18] T. Chai and R. R. Draxler, “Root Mean Square (RMSE or Mean Absolute Error (MAE) Arguments against Avoiding RMSE in the Literature,” *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, 2014.
- [19] B. H. Baltagi, *Econometric Analysis of Panel Data*, vol. 5, no. 7. 2005.
- [20] R. S. Bivand, E. J. Pebesma, and V. G. Rubio, *Applied Spatial Data Analysis with R*. New York: Springer, 2013.
- [21] C. F. Dormann *et al.*, “Methods to account for spatial autocorrelation in the analysis of species distributional data : a review,” *Ecography (Cop.)*, pp. 609–628, 2007.
- [22] O. Babak and C. V. Deutsch, “Statistical approach to inverse distance interpolation,” *Stoch. Environ. Res. Risk Assess.*, vol. 23, no. 5, pp. 543–553, 2009.
- [23] M. Mathur, “Spatial autocorrelation analysis in plant population: An overview,” *J. Appl. Nat. Sci.*, vol. 7, no. 1, pp. 501–513, 2015.
- [24] J. P. Elhorst, “Dynamic Spatial Panels: Models, Methods and Inferences,” *Journal of geographical systems.*, vol. 14, no. 1, pp. 5-28, 2014.