Available online at: http://journal.uny.ac.id/index.php/pythagoras
PYTHAGORAS: Jurnal Matematika dan Pendidikan Matematika, 19 (2), 2024, 165-178

Modeling Human Development Index of East Java Using Spatial Autoregressive and Spatial Error Ensemble

Nadia Aulia Jelita, Sri Sulistijowati Handajani*, Irwan Susanto
Study Program of Statistics, Universitas Sebelas Maret, Surakarta, Indonesia
* Corresponding Author. E-mail: rr_ssh@staff.uns.ac.id

ARTICLE INFO

Article History:
Received: 27-Oct. 2024
Revised: 04-Dec. 2024
Accepted: 17-Dec. 2024

Keywords:
HDI, spatial regression, ensemble method, homogeneity

ABSTRACT

The human development index (HDI) is an indicator used to monitor the government's success in developing the quality of human life. East Java Province's HDI is the lowest compared to other provinces on Java Island. Therefore, it is necessary to improve human development in this province. Attention must be paid to all aspects of human development, including the relationship between neighboring regions. The spatial regression method is an analysis method that considers the spatial dependency of the data. Ensemble spatial regression combines several spatial models by adding noise to the response variable, which is expected to reduce the diversity in the data. This research aims to use ensemble spatial regression to examine the East Java HDI. East Java HDI has spatial lag and spatial error dependence, modeled with SAR and SEM. Queen contiguity is used as a spatial weight. The SEM model does not fulfill the homogeneity assumption, so it is continued with the ensemble method. The ensemble method is proven to reduce diversity, so SEM Ensemble fulfills the assumption of homoscedasticity. After analysis using SAR and SEM Ensemble, the SAR model was chosen as the best model with the largest R^2 and lowest AIC value. Significant variables on East Java HDI are life expectancy, expected years of schooling, average years of schooling, and expenditure per capita.

How to Cite:
Jelita, N. A., Handajani, S. S., & Susanto, I. (2024). Modeling human development index of East Java using spatial autoregressive and spatial error ensemble. *Pythagoras: Jurnal Matematika dan Pendidikan Matematika, 19*(2), 165-178. https://doi.org/10.21831/pythagoras.v19i2.78621

https://doi.org/10.21831/pythagoras.v19i2.78621

INTRODUCTION

Development is important for improving the progress and welfare of society. It refers to positive change in various aspects of human life, involving improved quality of life, economic growth, and social improvement. The term "human development" was first used in 1990 by the United Nations Development Program (UNDP). It is important and needs attention because, in reality, high economic growth does not always solve welfare problems, such as poverty and the standard of living of the community at large (Si'lang et al., 2019).

The Central Bureau of Statistics has mapped out three basic dimensions of the HDI: longevity and healthy living, knowledge, and a decent standard of living. These three dimensions must be given equal attention because they are equally important. This is expected to lead to good human development. In the last decade, Indonesia's HDI has grown quite well, even though it has slowed down due to the COVID-19 pandemic. Since 2016, the country's human development status has improved from "medium" to "high". In the last 12 years, the HDI of Indonesia has increased by an average of 0.77 percent per year (Badan Pusat Statistik, 2022). Based on data from the BPS, Indonesia's HDI in 2022 reached 72.91 percent.

After West Java Province, East Java Province has the second-highest population in Indonesia. The development of the East Java HDI shows growth every year. The HDI of this province is also relatively high because it reached 72.75 percent in 2022 (Badan Pusat Statistik Provinsi Jawa Timur, 2022). Unfortunately, this province has been unable to keep up with the HDI growth of other provinces in Indonesia, especially the provinces on Java Island. This can be seen from BPS data, where in 2022, East Java Province was ranked 14th out of 34 provinces in Indonesia and 6th out of 6 provinces on Java Island.

To achieve community welfare, various efforts to improve human development in East Java Province must continuously be realized. Attention must be given to all aspects of human development. The relationship between neighboring regions is no exception. Everything is interconnected with one another, but something close will have more influence than something further away (Anselin, 1988). A set of observations with spatial dependence indicates that observation at location i depends on another observation at location j where locations i and j are close together (Santoso et al., 2022). If an observation is proven to have spatial dependence, it can be analyzed using the spatial regression method.

Classical linear regression is developed into spatial regression. This method will include the effect of location in the analyzed data. Spatial regression is generally divided into two categories, which are point spatial regression and area spatial regression (Hidayah & Indrasetianingsih, 2019). Point spatial regression focuses on distance information as its weight. Meanwhile, area spatial regression uses the intersection between neighboring locations. In spatial area regression, there are several commonly used modeling models, namely spatial autoregressive (SAR) and spatial error model (SEM). The SAR model is a model that contains spatial dependence on the response variable. SEM is a model that contains spatial dependence in the error model. If the model has spatial dependence on response variables and errors, it can be analyzed with the SARMA model. The SARMA model is not widely used because there is no supporting theory regarding this model where this model uses the same weights, so there may be identification problems (Viton, 2010).

The regression model must fulfill the assumption test to result in the correct parameter estimation, likewise for spatial regression models. In the spatial regression model, if the homogeneity assumption is not fulfilled, it will result in incorrect parameter estimation (Savita et al., 2017). One way to solve these problems is to apply the ensemble method.

In the ensemble technique, results from multiple regression models will be combined to improve prediction performance. Applying ensemble techniques to regression analysis can provide more powerful, accurate, and reliable results when compared to using a single regression model. An ensemble has two approaches: hybrid ensemble techniques and non-hybrid ensemble techniques (De Bock et al., 2010). The hybrid ensemble technique combines two methods of prediction results into one final prediction model. Meanwhile, the non-hybrid ensemble technique uses only one method that is modeled repeatedly so that many predictions are generated. The prediction results are then combined into one final model.

Previous research on East Java HDI with the spatial regression method has been conducted by Santoso et al. (2022). Researchers used three methods, the ordinary least square (OLS), SAR, and SEM models, to analyze the 2020 East Java HDI data. Of the three models, the SEM model was chosen as the best model with the largest R^2 value, 81.3772%, and the smallest AIC, 183.772. Novitasari and Khikmah (2019) analyzed the application of spatial regression models on Central Java HDI in 2017. Researchers used two spatial area regression methods: the SAR and SEM models. The SAR model was chosen as the best model of the two models, with an AIC of 143.49 (Novitasari & Khikmah, 2019).

Sazaen et al. (2020) analyzed Central Java HDI data using the non-hybrid ensemble spatial regression method. This research proved that ensemble SAR gave the best results compared to classical regression and SAR (Sazaen et al., 2020). In their research, Handajani et al. (2018) explained that ensemble spatial regression reduced diversity in poverty data in Central Java. This means that the ensemble technique can overcome violations of the assumption of homogeneity in the data (Handajani et al., 2018). Based on the description above, this research aims to apply spatial regression and non-hybrid ensemble spatial regression models to the HDI data of East Java Province.

METHODS

This research used spatial statistics and ensemble methods to model the Human Development Index in East Java. The data used in this research are data on the human development index of 38 districts and cities in East Java

Province in 2022. This data comes from the Central Bureau of Statistics of East Java Province, which consists of one response variable and six predictor variables. The research variables used are in Table 1.

Table 1. Research variable

Variable	Description
Y	East Java human development index 2022
X_1	Life expectancy
X_2	Expected years of schooling
X_3	Average years of schooling
X_4	Adjusted expenditure per capita
X_5	Number of poor people
X_6	Open unemployment rate

Using the above variables, quantitative analysis is carried out using statistical approach techniques to obtain appropriate conclusions. The steps in the data analysis stage in this study are in Figure 1.

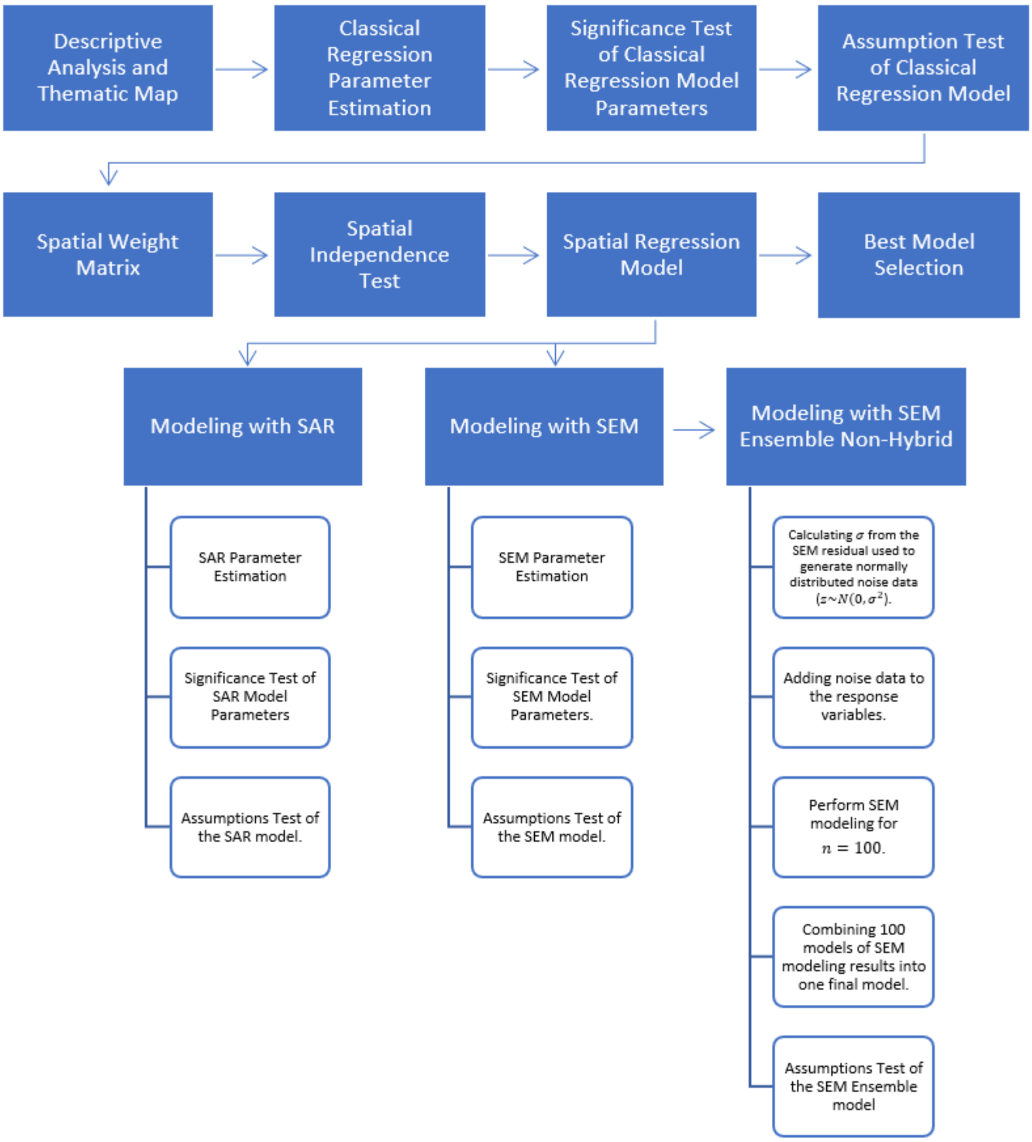

Figure 1. Analysis steps

Ordinary Least Square Regression

Regression analysis is a technique for data analysis that allows one to infer important information about how one variable depends on another (Draper & Smith, 1992). The general equation of multiple linear regression is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i \tag{1}$$

where Y_i is the response variable at the i-th observation, X_{ik} is the k-the predictor variable at the i-th observation, β_0 is the regression coefficient, $\beta_1 \dots \beta_k$ is the $1 \dots k$-the regression parameters, ε_i is the i-th data residual, n is the number of observations, and k is the number of predictor variables.

Classical Assumption Tests

1. *Normality Test*

   The residuals of a good model will be normally distributed. The test used is the Jarque-Bera test, with H_0: the residuals are normally distributed and H_1: the residuals are not. Gujarati (2004), in his book, states the Jarque Bera Test as:

$$JB = n\left(\frac{S^2}{6} + \frac{(K-3)^2}{24}\right) \tag{2}$$

$$S = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}} \quad \text{and} \quad K = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} \tag{3}$$

   where JB is the value of the Jarque-Bera test statistic, n is the sample size, S is skewness, and K is kurtosis. This test decides to reject H_0 if $JB > \chi^2_{\alpha;2}$.

2. *Multicollinearity Test*

   The multicollinearity test determines whether predictor variables are correlated in a regression model (Purba et al., 2021). Detecting multicollinearity in the data can be done by looking at the tolerance value and VIF (variance inflation factor) value. Gujarati (2004) in his book explains that the inverse of VIF is tolerance so it can be written as:

$$TOL_j = \frac{1}{VIF_j} = (1 - R_j^2), \qquad j = 1,2,..,k \tag{4}$$

   where R_j^2 is the coefficient of determination between X_j and other predictor variables. If $VIF < 10$, the data does not have multicollinearity.

3. *Homogeneity Test*

   The homogeneity test aims to determine whether the residuals of one observation differ from those of another (Purba et al., 2021). A good model is a homogeneous one (Gujarati, 2004). The statistical test used is the Breusch-Pagan test, with H_0: the residuals model is homogeneous and H_1: the residuals model is heterogeneous. Anselin (1988), in his book, states this test as:

$$BP = \frac{1}{2}[f'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'f]; \qquad f = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_n \end{pmatrix} \tag{5}$$

$$f_i = \left(\frac{{\varepsilon_i}^2}{\sigma^2}\right) - 1\ ; \ \ i = 1,2,\dots,n \tag{6}$$

   where BP is the value of the Breusch-Pagan test statistic, $\boldsymbol{Z}$ is a matrix of predictor variables of size $n \times (k+1)$, and ε_i is the residual at the i-th observation. This test decides to reject H_0 if $BP > X^2_{(a;k)}$

4. *Non-Autocorrelation Test*

A non-autocorrelation test assesses whether the residuals from a regression model exhibit a correlation pattern over time or space. The test used is the Durbin-Watson, with H_0: no autocorrelation in the model and H_1: there is autocorrelation. This test can be written as:

$$d = \frac{\sum_{i=2}^{i=n}(\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1})^2}{\sum_{i=1}^{i=n}\hat{\varepsilon}_i^2} \tag{7}$$

where d is the Durbin-Watson test statistic value, εi is the error in the i-th observation, and $\varepsilon i{-}1$ is the error in the $(i-1)$-th observation. Decision-making is done by comparing the value of the d statistic with the upper limit value (d_U) and the lower limit (d_L) (Gujarati, 2004).

a. If $0 < d < d_L$ or $4 - d_U < d < 4$, then there is autocorrelation between residuals.
b. If $d_L \leq d \leq d_U$ or $4 - d_U \leq d \leq 4 - d_L$, then the test is inconclusive so that it cannot be concluded that there is autocorrelation between residuals.
c. If $d_U < d < 4 - d_U$, there is no autocorrelation between the residuals.

Spatial Weight Matrix

The spatial weight matrix is a spatial dependency (contiguity) matrix with the notation $\boldsymbol{W}$. This matrix describes the connection between regions and is obtained based on distance or neighboring information. The dimension of this matrix is $n \times n$, where n is the number of observations or units across individuals. Three common types of spatial dependence or contiguity matrices are as follows (Dubin, 2009).

1. *Rook Contiguity.* This intersection concept assigns a value 1 to areas adjacent to the north, south, west, and east, called the common side, and 0 to the others.
2. *Bishop Contiguity.* This intersection concept defines a value of 1 for the common vertex of the region being observed and 0 otherwise.
3. *Queen Contiguity.* This intersection concept defines a value of 1 for regions whose sides and corners intersect with the region being observed and 0 for others.

After determining the spatial weight matrix, row standardization is performed on the spatial weight matrix. It means that the matrix is standardized so that the sum of each row of the matrix becomes equal to one (Dubin, 2009).

Moran's I Test

A non-autocorrelation test is used to determine whether or not autocorrelation occurs between the residuals of one observation and another. A spatial non-autocorrelation is conducted to see whether observations in a region affect other adjacent regions. A commonly used statistical test is Moran's I test. Moran's I is a measure of the correlation between observations in a region and other adjacent regions (Ward & Gleditsch, 2008), with H_0: there is no spatial autocorrelation in the residuals data and H_1: there is spatial autocorrelation in the residuals data. The Moran's I index equation is:

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}(y_i - \bar{y})(y_j - \bar{y})}{S_0\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{8}$$

$$S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij} \tag{9}$$

where I is Moran's I index, n is the number of observations, x_i is the value at the i-the location, x_j is the value at the j-the location, $\bar{x}$ is the average of x across n observations, and W_{ij} is the spatial weight matrix element. The expected value of Moran's I is (Novitasari & Khikmah, 2019):

$$E(I) = I_0 = -\frac{1}{n-1} \tag{10}$$

The statistical test used is:

$$Z(I) = \frac{I - E(I)}{\sqrt{Var(I)}} \tag{11}$$

The value of the Moran's I index is between 1 and -1. If $I > I_0$, the data has a positive autocorrelation, meaning that neighboring locations have similar values. If $I < I_0$, then the data has a negative autocorrelation, which means that neighboring locations have values that are not similar to each other (Anselin, 1988; Ward & Gleditsch, 2008). The critical region of this test $Z > Z_{\alpha/2}$.

Spatial Dependence Test

The spatial dependency test is used to detect whether there is a spatial effect between observations (Novitasari & Khikmah, 2019). The lagrange multiplier (LM) test detects spatial effects on data. This LM test is divided into two parts, as follows.

1. *Test for spatial dependence on the response variable or spatial lag dependence.*
   The hypotheses of this test are $H_0: \rho = 0$ (there is no spatial dependence on lag) and $H_1: \rho \neq 0$ (there is spatial dependence on lag). This test can be expressed as:
   $$LM_{lag} = \frac{\left(\frac{\hat{\boldsymbol{\varepsilon}}' \boldsymbol{WY}}{\hat{\sigma}^2}\right)^2}{\frac{1}{\hat{\sigma}^2}\left[(\boldsymbol{WX\beta})' \boldsymbol{M} (\boldsymbol{WX\beta}) + T\hat{\sigma}^2\right]} \tag{12}$$
   $$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}, \;\; \hat{\sigma}^2 = \frac{1}{n}(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}), \;\; T = tr(\boldsymbol{WW} + \boldsymbol{W}'\boldsymbol{W}), \;\; M = [\boldsymbol{I} + \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'] \tag{13}$$
   where tr is the $trace$ matrix, which is the sum of the main diagonals of a square matrix. If $LM_{lag} > \chi_{(1)}$, then H_0 is rejected, so there is spatial dependence on the lag. Therefore, the data can be analyzed using the SAR method.
2. *Test for the spatial dependence of errors.*
   The hypotheses of this test are $H_0: \lambda = 0$ (there is no spatial dependence on the error) and $H_1: \lambda \neq 0$ (there is spatial dependence on the error). This test can be expressed as:
   $$LM_{error} = \frac{\left(\frac{\hat{\boldsymbol{\varepsilon}}' \boldsymbol{W}\hat{\boldsymbol{\varepsilon}}}{\hat{\sigma}^2}\right)^2}{tr[(\boldsymbol{W}' + \boldsymbol{W})\boldsymbol{W}]} \tag{14}$$
   $$\hat{\sigma}^2 = \frac{1}{n}(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}) \tag{15}$$
   If $LM_{error} > \chi_{(1)}$, then H_0 is rejected, so there is spatial dependence on the error. Therefore, the data can be analyzed using the SEM method.

Spatial Regression

Spatial regression is a method for determining the connection between response and predictor variables based on location/spatial correlations. In general, this model can be written as:

$$\boldsymbol{y} = \rho \boldsymbol{W_1 y} + \boldsymbol{X\beta} + \boldsymbol{u}, \qquad \boldsymbol{u} = \lambda \boldsymbol{W_2 u} + \boldsymbol{\varepsilon} \tag{16}$$

where $\boldsymbol{y}$ is a $n \times 1$ vector of response variables, $\boldsymbol{X}$ is a $n \times k$ matrix of predictor variables, $\boldsymbol{\beta}$ is a $k \times 1$ vector of regression parameters, ρ is a spatial lag coefficient parameter, λ is a spatial error coefficient parameter, $\boldsymbol{u}$ is a $n \times 1$ vector of errors containing autocorrelation, $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of errors approximating the $N(0, \sigma^2 I)$ distribution, $\boldsymbol{W}$ is a $n \times n$ spatial weight matrix with zero diagonal elements, and $\boldsymbol{I}$ is a $n \times n$ identity matrix.

1. *Spatial Autoregressive (SAR)*
   The spatial autoregressive model (SAR) combines a linear regression model with a spatial lag on the response variable and cross-section data. Spatial lag occurs when the observation value of the response variable at one site is associated with the observation value of the response variable at its neighboring location; in other

words, there is a spatial correlation between response variables. If in equation (16) the value of $\rho \neq 0$ and $\lambda = 0$, then the SAR model can be expressed as:

$$\boldsymbol{y} = \boldsymbol{\rho W Y} + \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{17}$$

2. *Spatial Error Model (SEM)*
   The spatial error model (SEM) arises when the error value at a location is correlated with the error value at the surrounding location. If in equation (16) the value of $\rho = 0$ and $\lambda \neq 0$, then the SEM model can be expressed as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}, \qquad \boldsymbol{u} = \boldsymbol{\lambda W u} + \boldsymbol{\varepsilon} \tag{18}$$

Ensemble Regression

An ensemble method is a method that can be used to improve the accuracy and reduce the diversity of a model. In some previous studies, this ensemble technique reduced the diversity contained in the prediction model. The ensemble regression method combines the parameter estimation results of the two resulting models into one more accurate final estimate. There are two techniques in the ensemble: the hybrid ensemble and the non-hybrid ensemble (De Bock et al., 2010).

The hybrid ensemble technique uses two different models, and the predictions generated from the two models are combined into one model type. Meanwhile, the non-hybrid ensemble technique only uses one model. However, the model is used repeatedly to obtain many different predictions. The results of the different model predictions are then combined into one model.

The ensemble spatial regression model is performed by adding noise (z) normally distributed $(z \sim N(0, \sigma^2))$ to the response variable. Noise is an irregular disturbance in the data (Wu & Huang, 2009). The additive noise equation is written as:

$$\boldsymbol{m} = \boldsymbol{y} + \boldsymbol{z} \tag{19}$$

where $\boldsymbol{m}$ is the vector of response variables after adding noise, $\boldsymbol{y}$ is the vector of response variables, and $\boldsymbol{z}$ is the data generation with $z \sim N(0, \sigma^2)$.

RESULTS AND DISCUSSIONS

Based on Figure 2, the green color indicates that the district/city has a high HDI. Meanwhile, the red indicates that the area has a low HDI that requires special attention. By looking at the thematic map, it can be seen that the geographical location of each district/city tends to be close together. Regions with high HDI (colored green) are located in the northeastern part of Java Island (Surabaya City, Sidoarjo Regency, Gresik Regency, Lamongan Regency, and Mojokerto Regency). Meanwhile, areas on Madura Island (Kab. Bangkalan, Kab. Sampang, Kab. Pamekasan, and Kab. Sumenep) tend to have a fairly low HDI. This is possible due to the geographical location of these regions, which are not directly neighboring with other districts/cities in Java. Looking at the thematic map above, we can see a spatial dependency between neighboring regions. To confirm this, we analyzed and tested the HDI of districts/cities in East Java Province.

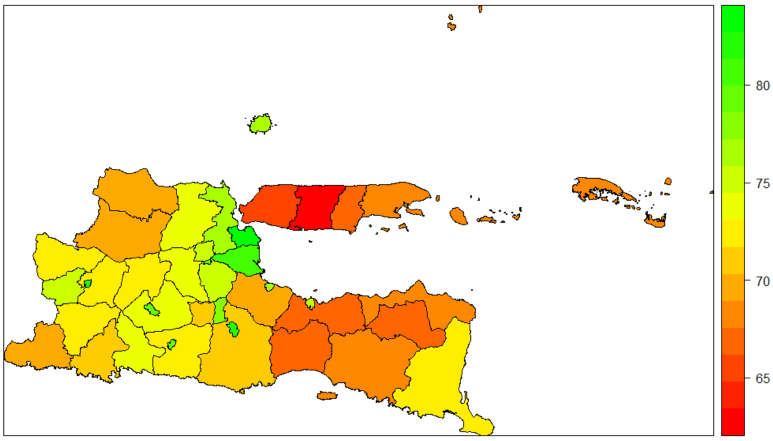

Figure 2. Thematic map of East Java HDI

Modeling East Java HDI Using Classical Regression

The first step of classical regression modeling is estimating the model's parameter, which includes all variables, to identify the significant effect of the independent variables on the dependent variable. The parameter estimation results obtained are presented in Table 2.

Table 2. Parameter estimation of classical regression model

Variable	Coefficient	Std. Error	t-Stat.	$p\ value$
Constant	6.6310	2.2930	2.89	0.007
X_1	0.4605	0.0329	13.99	0.000
X_2	1.0003	0.0752	13.29	0.000
X_3	1.3342	0.0893	14.94	0.000
X_4	0.000728	0.000041	17.83	0.000
X_5	-0.000473	0.000903	-0.52	0.604
X_6	0.0127	0.0288	0.44	0.663

In Table 2 above, several variables do not have a significant effect on the model (have a $p\ value > 0.05$), which are X_5 (number of poor people) and X_6 (open unemployment rate) so the model is still not good enough. Therefore, the best model parameters will be selected using the backward elimination method, and the results are in Table 3.

Table 3. Parameter estimation with backward test

Variable	Coefficient	Std. Error	t-Stat.	$p\ value$	Decision
Constant	6.7420	2.1760	3.10	0.004	significant
X_1	0.4584	0.0301	15.23	0.000	significant
X_2	0.9872	0.0697	14.16	0.000	significant
X_3	1.3657	0.0667	20.47	0.000	significant
X_4	0.000726	0.000037	19.70	0.000	significant

The regression equation model formed is as follows

$$\hat{y} = 6.74 + 0.4584X_1 + 0.9872X_2 + 1.36572X_3 + 0.000726X_4$$

1. *Parameter Significance Test*.
   Parameter significance tests are carried out simultaneously and partially. A simultaneous significance test determines whether all regression parameters significantly affect the linear regression model. The test used is the F test and obtained $F_{test} = 4363.74$ with $p\ value = 0.00$ so that the model is significantly affected by at least one predictor variable. Partially, whether each regression parameter significantly affects the linear

regression model will be examined. The t-test was used to conduct this test, and the results showed that the four variables tested were significant to the model.

2. *Regression Assumption Test*
If the regression model parameters are significant but do not meet the assumptions of identical, independent, and normal distribution, the resulting model is not considered good enough to use. The results of the model regression assumption test are presented in Table 4.

Table 4. Regression assumption results

Regression assumptions	Test	Test result	Critical area	Decision
Normality	Jarque-Bera	$JB = 4.5770$	$JB > \chi^2_{0.05;2} = 5.99$	Normal
Homogeneity	Breusch-Pagan	$BP = 5.0349$	$BP > \chi^2_{0.05;4} = 9.488$	Homogeneous
Multicollinearity	*VIF*	$VIF < 10$ for all variables	$VIF > 10$	There is no multicollinearity
Non-Autocorrelation	Durbin-Watson	$d = 1.4509$	$d_L = 1.2614$ and $d_U = 1.7223$	No decision

In the Durbin-Watson test, $d_L < d < d_{U,}$ so that no conclusion can be drawn. This means that either there is autocorrelation or there is no autocorrelation. The test continued with the spatial non-autocorrelation test.

Morans'I Test

Moran's I test is conducted to detect whether there is a spatial relationship (spatial autocorrelation) of observations close to each other. By using Equation (8), (9), (10), (11), and queen contiguity, the following results were obtained in Table 5.

Table 5. Moran's I test

I Index	$E(I)$	Z-Stat.	$p\ value$
0.4510	-0.027027	3.6566	0.0001

Table 5. shows that $Z_{stat} = 3.6566 > Z_{0.025} = 1.960$ and $p\ value = 0.0001 < \alpha = 0.05$, so H_0 is rejected which means that there is spatial autocorrelation in the classical regression model. This data is proven to have a spatial relationship in adjacent observations, so it does not meet the assumption of non-spatial autocorrelation. Therefore, this data will be analyzed using spatial area regression.

Spatial Dependency Test

Spatial dependency is a condition when there is a correlation between a region and another region. The Lagrange Multiplier (LM) test can do the spatial dependency test. The results of this test are shown in Table 6 below.

Table 6. Spatial dependence test

Test	Result test	$p\ value$
LM (Lag)	$LM_{lag} = 4.8597$	0.02749
LM (Error)	$LM_{error} = 6.0124$	0.01421

From Table 6 above, it is known that the statistical values of LM_{lag} and $LM_{error} > \chi^2_{0.05;1} = 3.841$ and both have $p\ value < \alpha = 0.05$, so there is spatial dependence on the response variables and errors. Therefore, this data will be modeled with SAR and SEM.

Modeling East Java HDI with Spatial Autoregressive (SAR)

Spatial autoregressive model (SAR) is a linear regression model that contains spatial dependence on lag. The SAR model estimation results for all significant predictor variables are shown in Table 7 below.

Table 7. Parameter estimation of the SAR model

Variable	Coefficient	Z-Stat.	$p\ value$
ρ	0.03480	2.28492	0.0223
Constant	6.69613	3.52840	0.0004
X_1	0.42391	13.9289	0.0000
X_2	1.00899	16.4145	0.0000
X_3	1.36702	23.4902	0.0000
X_4	0.0007029	20.8909	0.0000

The SAR equation formed is as follows.

$$\hat{y}_i = 0.0348 \sum_{i=1}^{n} W_{ij} y_j + 6.69613 + 0.42391\, X_1 + 1.00899\, X_2 + 1.36702\, X_3 + 0.0007029\, X_4$$

where i refers to the location for which the prediction $(\hat{y}_i)$ is being calculated, and j refers to the locations interacting with location i, such as its neighbor.

The model above has a R^2 value of 0.99835. This means that the regression model can explain 99.835% of the 2022 East Java Province HDI cases, while other factors outside the model explain the rest. From the parameter estimates formed, it shows that $\rho = 0.0348058$ with $p\ value < \alpha = 0.05$ and all variables tested also have a $p\ value < \alpha = 0.05$ which means that the spatial lag coefficient value and all these variables have a significant effect on the model (life expectancy, expected years of schooling, average years of schooling, and per capita expenditure).

After the model is formed, the assumptions of normality and homogeneity are tested on the residuals generated from the model, and the results are shown in Table 8 below.

Table 8. SAR assumption test results

Regression assumptions	Test	Test result	Critical area	Decision
Normality	Jarque-Bera	$JB = 2.9089$	$JB > \chi^2_{0.05;2} = 5.99$	Normal
Homogeneity	Breusch-Pagan	$BP = 5.6515$	$BP > \chi^2_{0.05;4} = 9.488$	Homogeneous

The SAR model can fulfill both assumptions. Thus, the HDI of East Java Province can be presented with the SAR model.

Modeling East Java HDI with Spatial Error Model (SEM)

Spatial error model (SEM) is a linear regression model containing spatial dependency on errors. The SEM model estimation results for all predictor variables are shown in Table 9 below.

Table 9. SEM model parameter estimation

Variable	Coefficient	Z-Stat.	$p\ value$
Constant	5.99487	3.24849	0.0012
X_1	0.46686	18.2743	0.0000
X_2	1.04628	19.3835	0.0000
X_3	1.30117	29.0194	0.0000
X_4	0.0007118	28.2912	0.0000
λ	0.62514	5.13651	0.0000

The SEM equation formed is as follows.

$$\hat{y}_i = 5.99487 + 0.46686\, X_1 + 1.04628\, X_2 + 1.30117\, X_3 + 0.0007118\, X_4 + 0.62514 \sum_{i=1}^{n} W_{ij} u_j$$

where i refers to the location for which the prediction $(\hat{y}_i)$ is being calculated, and j refers to the locations interacting with location i, such as its neighbor.

The model above has a R^2 value of 0.99875. This means that the regression model can explain 99.875% of the 2022 East Java Province HDI cases, while other factors outside the model explain the rest. From the parameter estimates formed, it shows that $\lambda = 0.625142$ with $p\ value < \alpha = 0.05$ and all variables tested also have a $p\ value < \alpha = 0.05$ which means that the spatial error coefficient value and all these variables have a significant effect on the model (life expectancy, expected years of schooling, average years of schooling, and per capita expenditure).

After the model is formed, the assumptions of normality and homogeneity are tested on the residuals generated from the model, and the results are presented in Table 10 below.

Table 10. SEM assumption test results

Regression assumptions	Test	Test result	Critical area	Decision
Normality	Jarque-Bera	$JB = 1.8862$	$JB > \chi^2_{0.05;2} = 5.99$	Normal
Homogeneity	Breusch-Pagan	$BP = 11.384$	$BP > \chi^2_{0.05;4} = 9.488$	Heterogeneous

Table 10 above shows that the SEM model does not fulfill the assumption of homogeneity, so further handling is needed to reduce the diversity in this model.

Modeling East Java HDI with Ensemble Spatial Regression

An ensemble spatial regression model reduces the diversity in spatial regression models. Non-hybrid ensemble spatial regression will be applied to the SEM model because this model does not fulfill the homogeneity test. The first step in predicting the SEM ensemble model is to add noise (s) derived from the generation $s \sim N(0, \sigma^2)$ data on the percentage of HDI of districts and cities in East Java. The value of σ used is 0.179. The addition of noise is done 100 times so that the SEM model is obtained as follows.

$$1: \hat{y}_i = 8.0797 + 0.4381\, X_1 + 0.9771\, X_2 + 1.3875\, X_3 + 0.000730\, X_4 + 0.33889 \sum_{i=1}^{n} W_{ij} u_j$$

$$2: \hat{y}_i = 6.0166 + 0.4668\, X_1 + 1.0020\, X_2 + 1.3503\, X_3 + 0.00073\, X_4 + 0.14545 \sum_{i=1}^{n} W_{ij} u_j$$

$$\vdots$$

$$100: \hat{y}_i = 8.0491 + 0.4391\, X_1 + 0.9768\, X_2 + 1.3933\, X_3 + 0.00072\, X_4 + 0.39143 \sum_{i=1}^{n} W_{ij} u_j$$

The final model of the SEM ensemble is obtained by calculating the average parameter estimation results of 100 models, so there is only one final model. The model is expressed as follows.

$$\hat{y}_i = 6.8254 + 0.4568\, X_1 + 0.9894\, X_2 + 1.3620\, X_3 + 0.0007277\, X_4 + 0.31226 \sum_{i=1}^{n} W_{ij} u_j$$

where i refers to the location for which the prediction ($\hat{y}_i$) is being calculated, and j refers to the locations interacting with location i, such as its neighbor.

The model has a R^2 value of 0.9981. This shows that the regression model can explain 99.81% of the 2022 East Java Province HDI cases, while other factors outside the model explain the rest. Furthermore, normality and homogeneity tests will determine whether the above model meets the regression assumptions. The results of the tests are presented in Table 11 below.

Table 11. SEM ensemble assumption test results

Regression assumptions	Test	Test result	Critical area	Decision
Normality	Jarque-Bera	$JB = 4.7313$	$JB > \chi^2_{0.05;2} = 5.99$	Normal
Homogeneity	Breusch-Pagan	$BP = 8.7397$	$BP > \chi^2_{0.05;4} = 9.488$	Homogeneous

The SEM ensemble model can fulfill both assumptions. Thus, the HDI of East Java Province can be presented with an SEM ensemble model. In this case, the ensemble method proved to reduce the diversity in the data.

Best Model Selection

After analyzing several spatial regression models on the HDI of East Java Province, one of the best models will be selected from the SAR and SEM ensemble models. The best model selection will be made by looking at the most significant coefficient of determination (R^2) and the lowest AIC value. Table 12 below summarizes the research results using the spatial regression model.

Table 12. Summary of model analysis results

Model	R^2 value	AIC value
SAR	0. 99835	0.9055
SEM Ensemble	0.99812	4.0017

Table 12 above shows that the SAR model has a larger R^2 value compared to the ensemble SEM model. Also, the SAR model has the lowest AIC value compared to the ensemble SEM model. Therefore, the SAR model is chosen as the best-recommended model. The SAR equation formed is as follows.

$$\hat{y}_i = 0.0348 \sum_{i=1}^{n} W_{ij} y_j + 6.69613 + 0.42391\, X_1 + 1.00899\, X_2 + 1.36702\, X_3 + 0.0007029\, X_4$$

The model can be interpreted as follows.

1. The HDI estimation of an area surrounded by other areas (with sides and corners) will be affected by 0.0348 times the average HDI of the surrounding areas.
2. Each one-year increase in life expectancy will increase the human development index by 0.42391 percent.
3. Each one-year increase in expected years of schooling will increase the human development index by 1.00899 percent.
4. Each one-year increase in average years of schooling will increase the human development index by 1.36702 percent.
5. Each one thousand rupiah increase in per capita expenditure value will increase the human development index by 0.7029 percent.

Figure 3. Thematic map of East Java HDI

The SAR equation formed can be depicted in a region, for example, when looking at Pacitan Regency. In Figure 3, it can be seen that this regency is neighboring the Ponorogo Regency and Trenggalek Regency, so the equation model formed is:

$$\hat{y}_{Pacitan} = 0.0348(0.5y_{Ponorogo} + 0.5y_{Trenggalek}) + 6.69613 + 0.42391\, X_1 + 1.00899\, X_2 + 1.36702\, X_3 + 0.0007029\, X_4$$

This means that the HDI of Ponorogo Regency contributes 50% to the spatial influence. The HDI of Trenggalek Regency also contributes 50%, and the spatial lag coefficient parameter (0.0348) reflects the overall strength of spatial dependence.

CONCLUSION

The ensemble method has proven to be able to reduce the diversity that exists in the SEM model. However, in this case, the ensemble method has not been able to increase the model's goodness, so the best model chosen is the SAR model. This is because the SAR model has a greater R^2 value (0.99835) and a smaller AIC value (0.9055) than the ensemble SEM model. This means that the regression model can explain 99.835% of the 2022 East Java Province HDI cases, while other factors outside the model explain the rest. Based on the SAR model, the HDI of East Java Province is significantly influenced by life expectancy, expected years of schooling, average years of schooling, and per capita expenditure.

ACKNOWLEDGEMENTS

The authors would like to thank Universitas Sebelas Maret and all those who have helped with this research.

REFERENCES

Anselin, L. (1988). *Spatial Econometrics: Methods and Models* (Vol. 4). Springer Netherlands. https://doi.org/10.1007/978-94-015-7799-1

Badan Pusat Statistik. (2022). *Berita Resmi Statistik Indeks Pembangunan Manusia (IPM) Tahun 2022*. Badan Pusat Statistik. https://www.bps.go.id/id/pressrelease/2022/11/15/1931/indeks-pembangunan-manusia--ipm--indonesia-tahun-2022-mencapai-72-91--meningkat-0-62-poin--0-86-persen--dibandingkan-tahun-sebelumnya--72-29-.html

Badan Pusat Statistik Provinsi Jawa Timur. (2022). *Berita Resmi Statsitsik Indeks Pembangunan Manusia (IPM) Jawa Timur Tahun 2022*. Badan Pusat Statistik Provinsi Jawa Timur. https://jatim.bps.go.id/pressrelease/2022/11/15/1309/indeks--pembangunan-manusia--ipm--jawa-timur-pada-tahun-2022--mencapai-72-75.html

De Bock, K. W., Coussement, K., & Van den Poel, D. (2010). Ensemble classification based on generalized additive models. *Computational Statistics & Data Analysis*, *54*(6), 1535–1546. https://doi.org/10.1016/j.csda.2009.12.013

Draper, N., & Smith, H. (1992). *Analisis Regresi Terapan : Edisi Kedua* (Edisi Kedua). Gramedia Pustaka Utama.

Dubin, R. (2009). Spatial Weights. In *The SAGE Handbook of Spatial Analysis* (pp. 125–157). SAGE Publications, Ltd. https://doi.org/10.4135/9780857020130.n8

Gujarati, D. N. (2004). *Basic Econometrics* (4th ed.). The McGraw-Hill Companies. http://13.235.221.237:8080/jspui/bitstream/123456789/443/1/Basic-Econometrics---Gujaratipdf.pdf

Handajani, S. S., Savita, C. A., Pratiwi, H., & Susanti, Y. (2018). Best weighted selection in handling error heterogeneity problem on spatial regression model. *Proceedings of the International Conference on Mathematics and Islam*, 293–299. https://doi.org/10.5220/0008521002930299

Hidayah, N. R., & Indrasetianingsih, A. (2019). Analisis regresi spatial durbin model (SDM) untuk pemodelan kemiskinan Provinsi Jawa Timur tahun 2017. *J Statistika: Jurnal Ilmiah Teori Dan Aplikasi Statistika*, *12*(1), 40–46. https://doi.org/10.36456/jstat.vol12.no1.a1994

Novitasari, D., & Khikmah, L. (2019). Penerapan model regresi spasial pada indeks pembangunan manusia (IPM) di Jawa Tengah. *Jurnal Statistika*, *19*(2), 123–134. https://doi.org/10.29313/jstat.v19i2.5068

Purba, D. S., Tarigan, W. J., Sinaga, M., & Tarigan, V. (2021). Pelatihan penggunaan software spss dalam pengolahan regressi linear berganda untuk mahasiswa fakultas ekonomi Universitas Simalungun di masa pandemi covid 19. *Karya Abadi*, *5*(2), 202–208. https://doi.org/10.22437/jkam.v5i2.15257

Santoso, E., Jumiati, A., Hadi Priyono, T., & Putomo Somaji, R. (2022). Determinan indeks pembangunan manusia di Provinsi Jawa Timur: model crossectional spasial. *JAE (JURNAL AKUNTANSI DAN EKONOMI)*, *7*(1), 103–112. https://doi.org/10.29407/jae.v7i1.17884

Savita, C. A., Handajani, S. S., & Winarno, B. (2017). Penerapan model regresi ensemble non-hybrid pada data kemiskinan di Provinsi Jawa Tengah. *The 6th University Research Colloquium 2017*, 127–134. https://journal.unimma.ac.id/index.php/urecol/article/view/1237

Sazaen, E. A., Wasono, R., & Nur, I. M. (2020). Non-hybrid ensemble spatial regression on human development index (IPM) in Central Java. *Jurnal Litbang Edusaintech*, *1*(1), 23–34. https://doi.org/10.51402/jle.v1i1.4

Si'lang, I. L. S., Hasid, Z., & Priyagus. (2019). Analisis faktor-faktor yang berpengaruh terhadap indeks pembangunan manusia. *Jurnal Manajemen*, *11*(2), 159–169. http://journal.feb.unmul.ac.id/index.php/JURNALMANAJEMEN

Viton, P. A. (2010). *Notes on Spatial Econometric Models*. The Ohio State University. https://www.yumpu.com/en/document/read/3858779/notes-on-spatial-econometric-models-the-ohio-state-university

Ward, M., & Gleditsch, K. (2008). *Spatial Regression Models*. SAGE Publications, Inc. https://doi.org/10.4135/9781412985888

Wu, Z., & Huang, N. E. (2009). Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Advances in Adaptive Data Analysis*, *01*(01), 1–41. https://doi.org/10.1142/S1793536909000047