Media Statistika

p-ISSN 1979 – 3693 e-ISSN 2477 – 0647
MEDIA STATISTIKA 15(1) 2022: 72-82
http://ejournal.undip.ac.id/index.php/media_statistika

LIFE EXPECTANCY MODELING USING MODIFIED SPATIAL AUTOREGRESSIVE MODEL

Hasbi Yasin[1], Budi Warsito[1], Arief Rachman Hakim[1], Rahmasari Nur Azizah[2]
[1] Department of Statistics, Diponegoro University, Indonesia
[2] Data Science Institute, I-Biostat, Hasselt University Belgium, Belgium

e-mail: hasbiyasin@live.undip.ac.id

DOI: 10.14710/medstat.15.1.72-82

Article Info:
Received: 17 October 2021
Accepted: 5 July 2022
Available Online: 27 July 2022

Keywords:
Life Expectancy; MSOM; Outlier detection; SAR
.

Abstract: The presence of outliers will affect the parameter estimation results and model accuracy. It also occurs in the spatial regression model, especially the Spatial Autoregressive (SAR) model. Spatial Autoregressive (SAR) is a regression model where spatial effects are attached to the dependent variable. Removing outliers in the analysis will eliminate the necessary information. Therefore, the solution offered is to modify the SAR model, especially by giving special treatment to observations that have potentially become outliers. This study develops to modeling the life expectancy data in Central Java Province using a modified spatial autoregressive model with the Mean-Shift Outlier Model (MSOM) approach. Outliers are detected using the MSOM method. Then the result is used as the basis for modifying the SAR model. This modification, in principle, will reduce or increase the average of the observed data indicated as outliers. The results show that the modified model can improve the model accuracy compared to the original SAR model. It can be proved by the increased coefficient of determination and decreasing the Akaike Information Criterion (AIC) value of the modified model. In addition, the modified model can improve the skewness and kurtosis values of the residuals getting closer to the Normal distribution.

1. INTRODUCTION

Measurement of Life Expectancy (LE) is one way for the government to evaluate its performance to improve the welfare and health of the population. Based on data from the Central Statistics Agency, LE in Indonesia shows an increase every year (BPS-Statistics of Jawa Tengah Province, 2019). However, Indonesia's life expectancy is still below Malaysia and Singapore. Therefore, research is needed to find factors that significantly affect the increase in life expectancy. This study discusses the modeling of life expectancy in Central Java Province, Indonesia. Central Java province occupies the second position nationally based on the life expectancy of its population. In the last three years, Life Expectancy in Central Java Province increased from 72.28 to 72.51 for the male population and increased from 76.10 to 76.30 for the female population. While the factors studied were education, health, and economic factors. One of the methods in analyzing the factors that influence the Life Expectancy is to use a spatial regression model approach (Hakim et al., 2020;

Musyarofah et al., 2020; Yasin et al., 2020). The existence of spatial dependence between locations causes the regression model to include a spatial component in the model. If the spatial dependence only occurs on the response variable, then the Spatial Autoregressive (SAR) model can be used.

In many cases, the presence of outliers in data modeling will make parameter estimates biased and ineffective. The presence of outliers will also cause the accuracy of the model to be very doubtful. According to Draper & Smith (1998), simply rejecting the presence of outliers is not a wise move. Sometimes outliers provide information that other data points cannot. Therefore, the outliers that arise are not the result of recording errors or errors when preparing the equipment, so the outliers cannot be erased. The in-depth investigation must be done carefully. Especially in spatial analysis, eliminating an outlier can result in changes in the composition of spatial effects on the data. Outliers will also affect statistical inference. Likewise, methodologies for the detection and accommodation of outliers have always been an important topic in data analysis (Beckman & Cook, 1983).

Statistically, spatial data observations that are spatially correlated are different from independent data observations. For example, the value of one observation might be depending on their neighbors. Based on this case, several authors suggest several methods or algorithms that detect outliers in spatial data. Powerful estimation methods for dealing with outliers in the data have also been proposed (Genton, 1998, 2001; Militin et al., 2003). However, the outlier detection in spatial data is certainly different when compared to the usual statistical data. Then, in some cases, the detected outliers are discarded without taking any new information that the outliers might provide into consideration. This results in lost information from the data set and can lead to misleading conclusions. Therefore, an appropriate method is needed to detect outliers and accommodate them in the spatial regression model to improve the model performance of the spatial regression model in general. Improvement of the spatial regression model contaminated by outliers has been developed, including the robust spatial regression method (Hakim et al., 2019; Mukrom et al., 2021; Musyarofah et al., 2020; Yasin et al., 2020). However, one of the weaknesses of this method is that it does not improve the model based on observations that potentially become outliers specifically. Therefore, in this study, the Mean-Shift Outlier Model (MSOM) method is used to detect observations that have the potential as outliers in the SAR model (Dai et al., 2016). Then based on the detection results, the SAR model was modified by shifting the average value only on the data indicated as outliers. This model is called Modified SAR.

2. LITERATURE REVIEW

2.1. Spatial Autoregressive Model

The Spatial Autoregressive Model (SAR) is also known as the Spatial Lag Model (SLM). This model is a spatial model with an area approach by taking into account the effect of spatial lag only on the dependent variable. This model is also called Mixed Regressive-Autoregressive because it combines the usual regression model with the spatial lag regression model on the dependent variable (Anselin, 1988, 1992). In Mathematics notation, the SAR model can be written as in equation (1), or in matrix form, it can be written as in equation (2).

$$y_i = \rho \sum_{j=1}^{n} w_{ij} y_j + \beta_0 + \sum_{k=1}^{p} \beta_k x_{ki} + \varepsilon_i \quad (1)$$

$$\boldsymbol{y} = \rho \boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \text{and } \boldsymbol{\varepsilon} \sim \mathbf{N}(0, \sigma_{\varepsilon}^2 \boldsymbol{I}_n) \tag{2}$$

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

where $\boldsymbol{y}$ is a vector of the response variable, ρ is a spatial lag coefficient of the response variable, $\boldsymbol{W}$ is a spatial contiguity weights matrix of size $n \times n$, such as queen or rook contiguity (LeSage, 1999). $\boldsymbol{X}$ is a matrix of predictor variables, $\boldsymbol{\beta}$ is a vector of regression coefficients, and $\boldsymbol{\varepsilon}$ is a vector of error model. The parameters of this models are estimated using the Maximum Likelihood Estimation (MLE) method (LeSage & Pace, 2009). The equation describes the variation in y as a linear combination of adjacent units with no independent variables. Parameter estimation in the Spatial Autoregressive model uses the Maximum Likelihood Estimation (MLE) method. The parameter estimation of the SAR model using the OLS method produces biased and inconsistent parameters.

2.2. Mean-Shift Outlier Model (MSOM) in SAR Model

Based on equation (2), the dependent variable will be univariate normally distributed as shown in equation (3).

$$\boldsymbol{y} \sim \boldsymbol{N}(\boldsymbol{A^{-1}X\beta}, \boldsymbol{\sigma^2 I}) \tag{3}$$

where $\boldsymbol{A} = (\boldsymbol{I} - \rho \boldsymbol{W})$. Then, the likelihood function of $\boldsymbol{y}$ is as follows:

$$L_{\boldsymbol{y}} = -\frac{n}{2} ln\, \sigma^2 + ln|\boldsymbol{A}| - \frac{1}{2\sigma^2} \boldsymbol{e}^T \boldsymbol{e} \tag{4}$$

where $\boldsymbol{e} = \boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$. By partially deriving the function $L_{\boldsymbol{y}}$ for each parameter ρ, σ, and β, then equating to zero, the following equation will be obtained:

$$\sigma^2 = \frac{1}{n} \boldsymbol{e}^T \boldsymbol{e}, \qquad \sigma^2 tr(\boldsymbol{A^{-1}W}) = \boldsymbol{e^T W y}$$

$$\boldsymbol{\beta} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T A y}$$

Estimator of parameter ρ, σ and β are obtained iteratively as in (LeSage, 1999).

Equation (2) can also be written as follows:

$$\boldsymbol{A}\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{5}$$

It's means that $\boldsymbol{e} = \boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \sim N(0, \sigma^2 \boldsymbol{I})$, and the estimator for the parameters in equation (2) is:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T} \widehat{\boldsymbol{A}} \boldsymbol{y} \tag{6}$$

$$\hat{\boldsymbol{e}} = \widehat{\boldsymbol{A}} \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \big(\boldsymbol{I} - \widehat{\boldsymbol{P}}\big)\widehat{\boldsymbol{A}} \boldsymbol{y} = \widehat{\boldsymbol{Q}}\widehat{\boldsymbol{A}} \boldsymbol{y} \tag{7}$$

where $\widehat{\boldsymbol{Q}} = (\boldsymbol{I} - \widehat{\boldsymbol{P}})$, $\widehat{\boldsymbol{P}} = \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T$, and $\widehat{\boldsymbol{A}} = (\boldsymbol{I} - \hat{\rho} \boldsymbol{W})$.

Therefore, the studentized residual in spatial autoregressive model can be defined as (Shi & Chen, 2009):

$$t_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - \hat{p}_{ii}}} \tag{8}$$

where e_i is the i-th element of $\hat{\boldsymbol{e}}$, and $\hat{p}_{ii}$ is the i-th diagonal element of matrix $\widehat{\boldsymbol{P}}$.

Suppose a represents the index of observations that are considered as outliers in the spatial regression model. To test whether some of these observations are outliers or not, MSOM in the SAR model is defined as follows (Dai et al., 2016):

$$\boldsymbol{y} = \rho\boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}_a\boldsymbol{\delta} + \boldsymbol{\varepsilon} \tag{9}$$

where $\boldsymbol{\varepsilon}\sim N(0,\sigma^2\boldsymbol{I})$, $\boldsymbol{\delta}$ is an $m\times 1$ vector parameter, $a=\{i_1,\cdots,i_m\}(m<n)$, $\boldsymbol{D}_a=\left\{\boldsymbol{d}_{i_1},\cdots,\boldsymbol{d}_{i_m}\right\}$ is an $n\times m$ indicator matrix indexed by a matrix, and $\boldsymbol{d}_{i_k}$ is an $n\times 1$vector with i_k-th element equal to 1, and the rest equal to 0, $k=1,2,\cdots,m$. This shows that to detect some outliers in equation (9) is equivalent to testing the hypothesis on the significance of the parameter $\boldsymbol{\delta}$.

Hypothesis:

H_0: $\boldsymbol{\delta}=\boldsymbol{0}$

H_1: $\boldsymbol{\delta}\neq\boldsymbol{0}$

Test Statistics:

$$SC_a = \hat{\boldsymbol{e}}_a^T\left(\widehat{\boldsymbol{Q}}_{aa} - \hat{\boldsymbol{\kappa}}^{-1}\widehat{\boldsymbol{b}}_a\widehat{\boldsymbol{b}}_a^T\right)^{-1}\frac{\hat{\boldsymbol{e}}_a}{\hat{\sigma}^2} \tag{10}$$

where $\hat{\boldsymbol{e}}_a$ is the sub-vector of $\hat{\boldsymbol{e}}$ indexed by a, $\widehat{\boldsymbol{b}}_a$ is the sub-vector of $\widehat{\boldsymbol{b}}$ indexed by a, $\widehat{\boldsymbol{b}}=\widehat{\boldsymbol{Q}}\hat{\boldsymbol{\eta}}$, $\hat{\boldsymbol{\eta}}=\boldsymbol{W}\widehat{\boldsymbol{A}}^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{Q}}_{aa}$is the diagonal block matrix of $\widehat{\boldsymbol{Q}}$ indexed by a, and $\widehat{\boldsymbol{Q}}=(\boldsymbol{I}-\widehat{\boldsymbol{P}})$, $\hat{\boldsymbol{\kappa}}=\hat{\sigma}^2\hat{c}_{11}+\hat{\boldsymbol{\eta}}^T\widehat{\boldsymbol{Q}}\hat{\boldsymbol{\eta}}-\hat{\sigma}^2\frac{\left(n\hat{c}_{12}^2-4\hat{c}_1\hat{c}_2\hat{c}_{12}+2\hat{c}_1^2\hat{c}_{22}\right)}{\left(n\hat{c}_{22}-2\hat{c}_2^2\right)}$, $\hat{c}_i=tr(\widehat{\boldsymbol{C}}_i)$, $\hat{c}_{ij}=tr\left(\widehat{\boldsymbol{C}}_i^T\widehat{\boldsymbol{C}}_j+\widehat{\boldsymbol{C}}_i\widehat{\boldsymbol{C}}_j\right)$, $i,j=1,2$, and $\widehat{\boldsymbol{C}}_1=\boldsymbol{W}\widehat{\boldsymbol{A}}^{-1}$, $\widehat{\boldsymbol{C}}_2=\boldsymbol{I}$.

Decision:

Reject H_0 if $SC_a>\chi^2_{m;(1-\alpha)}$, where $\chi^2_{m;(1-\alpha)}$ is the $(1-\alpha)$-th quantile of the Chi-square distribution with degrees of freedom m.

If outlier detection is used a single observation, or $a=\{i\}(i=1,2,\cdots,n)$, then the MSOM model in equation (9) can be used to detect a single outlier, and the equation becomes:

$$\boldsymbol{y} = \rho\boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{d}_i\delta + \boldsymbol{\varepsilon} \tag{11}$$

where $\boldsymbol{d}_i$ is an $n\times 1$ vector with i -th element equal to 1, and the rest equal to 0, $(i=1,2,\cdots,n)$, and δ is a scalar. Therefore, the hypothesis test is as follows:

Hypothesis:

H_0: $\delta=0$

H_1: $\delta\neq 0$

Test Statistics:

$$SC_i = \frac{\hat{e}_i^2}{\hat{\sigma}^2\left(\hat{q}_{ii}-\frac{\hat{b}_i^2}{\hat{\kappa}}\right)} \tag{12}$$

where $\hat{q}_{ii}$ is the i-th diagonal element of matrix $\widehat{\boldsymbol{Q}}$, $\hat{e}_i$ and $\hat{b}_i$is the i-th element of vector $\boldsymbol{e}$ and $\boldsymbol{b}$, and $\hat{\kappa}$ is a scalar.

Decision:

Reject H_0 if $SC_i > \chi^2_{1;(1-\alpha)}$, where$\chi^2_{1;(1-\alpha)}$ is the $(1-\alpha)$-th quantile of the Chi-square distribution with degrees of freedom 1.

3. MATERIAL AND METHOD

3.1. Research Variables

This study uses life expectancy data in Central Java Province, Indonesia. This data is secondary data obtained from the Statistics of Central Java Province in 2018 issued by BPS-Statistics ((BPS-Statistics of Jawa Tengah Province), 2019), and the Health Profile of Central Java in 2018 issued by Dinas Kesehatan Provinsi Jawa Tengah ((BPS-Statistics of Jawa Tengah Province), 2018). This data consists of 35 regencies and cities in Central Java Province. Several aspects have a significant effect on Life Expectancy (LE) (Years), including health, education, and economic factors. The "Percentage of Households with Clean and Healthy Living Behavior (PCHLB) (%)" and the "Number of Integrated Health Post (IHP) (Unit)" are explain the health factors. Educational factor is explained using the "Average Length of School (ALS) (Year)" variable. The "Percentage of Poor Population (PP) (%)" and "Adjusted Per Capita Expenditure (APCE) (Thousand IDR)" are describe an economic factor. Table 1 shows the description of each variable in the data set. Each region has different characteristics and form some groups (see Fig. 1). Therefore, a spatial regression is needed in modeling the life expectancy in Central Java Province.

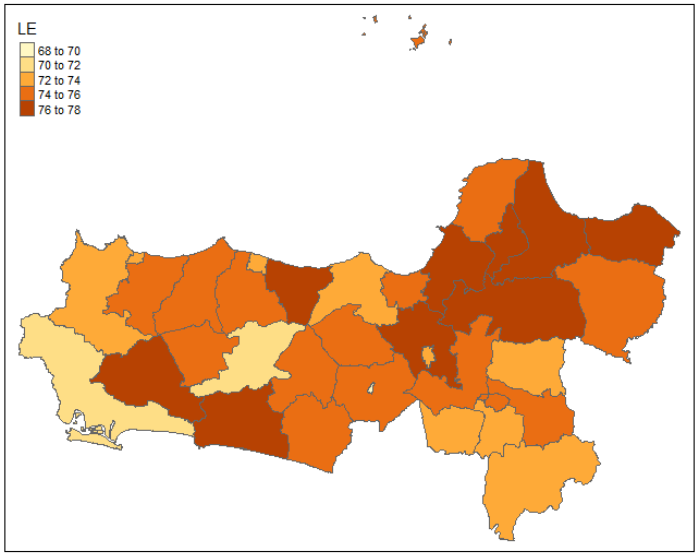


Figure 1. Spatial Distribution of Life Expectancy data in Central Java

Table 1. Description of Research Variable in Central Java Province

Statistics	LE	ALS	PCHLB	IHP	PP	APCE
Min	68.61	6.18	59.69	164.00	4.62	7,785
1st Qu	73.42	6.71	71.60	600.00	9.33	9,238
2nd Qu (Median)	74.46	7.29	78.30	901.00	12.42	9,813
Mean	74.63	7.58	79.22	925.60	12.49	10,414
3rd Qu	75.90	8.26	88.06	1,160.50	14.09	11,379
Max.	77.49	10.50	97.25	2,195.00	20.32	14,921

3.2. Modified Spatial Autoregressive Model

Based on equation (9), we can rewrite the Modified SAR model as follows:

$$\begin{aligned} \boldsymbol{y} &= \rho \boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}_a\boldsymbol{\delta} + \boldsymbol{\varepsilon} \\ &= \rho \boldsymbol{W}\boldsymbol{y} + [\boldsymbol{X} \quad \boldsymbol{D}_a]\begin{bmatrix}\boldsymbol{\beta} \\ \boldsymbol{\delta}\end{bmatrix} + \boldsymbol{\varepsilon} \\ &= \rho \boldsymbol{W}\boldsymbol{y} + \boldsymbol{Z}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \end{aligned} \tag{13}$$

where $\boldsymbol{Z} = [\boldsymbol{X} \quad \boldsymbol{D}_a]$, $\boldsymbol{\theta} = [\boldsymbol{\beta} \quad \boldsymbol{\delta}]^T$, and $\boldsymbol{\varepsilon} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\sigma^2 I})$. Thus, parameter estimation in the modified model can be solved by the MLE method, in the same way as in the original SAR model. Therefore, the procedure for modification of the SAR model can be carried out with the following steps:

1. Outlier detection based on MSOM to get m observations that have the potential to become outliers
2. Input $a = \{i_1, \cdots, i_m\}$ as an index of observational data that is considered to be an outlier
3. Create an $n \times m$ matrix $\boldsymbol{D}_a$
4. Create an $n \times ((p+1)+m)$join matrix of independent variable, $\boldsymbol{Z} = [\boldsymbol{X} \quad \boldsymbol{D}_a]$
5. Estimate the $\boldsymbol{\theta} = [\boldsymbol{\beta} \quad \boldsymbol{\delta}]^T$using MLE method.

Computation and data analysis in this study using R software with the "spatialreg" package as the main package (Bivand, 2022). We modified it into a web application using the "shiny" R package (R Team, 2021).

4. RESULTS AND DISCUSSION

4.1. Life Expectancy Model using Spatial Autoregressive

In this study, we modeled the Life Expectancy using the SAR model approach. Then, we detected outliers in the model using the MSOM method. Furthermore, based on the results of the outlier detection, we modified the SAR model to obtain a better model. The SAR model was used because based on the LM-Test (Anselin, 1988), it was found that there was significant spatial lag dependence (see Table 2). Then, Table 3 shows that parameter of SAR model based on the weight matrix, Queen and Rook contiguity. Based on the Table 3, we can conclude that the Spatial Autoregressive using Rook matrix contiguity is better than the Queen weighting matrix.

Table 2. Lagrange Multiplier Diagnostics for Spatial Dependence

LM Test	Queen		Rook	
	Statistics	P-Value	Statistics	P-Value
Spatial Error	1.67056	0.19618	1.73284	0.18805
Spatial lag	4.08648	0.04323*	4.15521	0.04151*

*) Significant in α = 0.05

Table 3. Spatial Autoregressive Coefficient of Life Expectancy Model

Weight Matrix	Variable	*Parameter*	*Coeff*	p-value	AIC	MSE
Queen	W_LE	ρ	-0.5430	0.0105*	125.61	1.2593
	Intercept	β_0	109.7200	0.0000*		
	ALS	β_1	0.6666	0.0369*		
	PCHLB	β_2	0.0353	0.1525		
	IHP	β_3	0.0004	0.3691		
	PP	β_4	-0.1941	0.0021*		
	APCE	β_5	0.0000	0.9644		
Rook	W_LE	ρ	-0.5453	0.0094*	125.43	1.2492
	Intercept	β_0	109.8600	0.0000*		
	ALS	β_1	0.6723	0.0345*		
	PCHLB	β_2	0.0342	0.1637		
	IHP	β_3	0.0004	0.3840		
	PP	β_4	-0.1924	0.0022*		
	APCE	β_5	0.0000	0.9840		

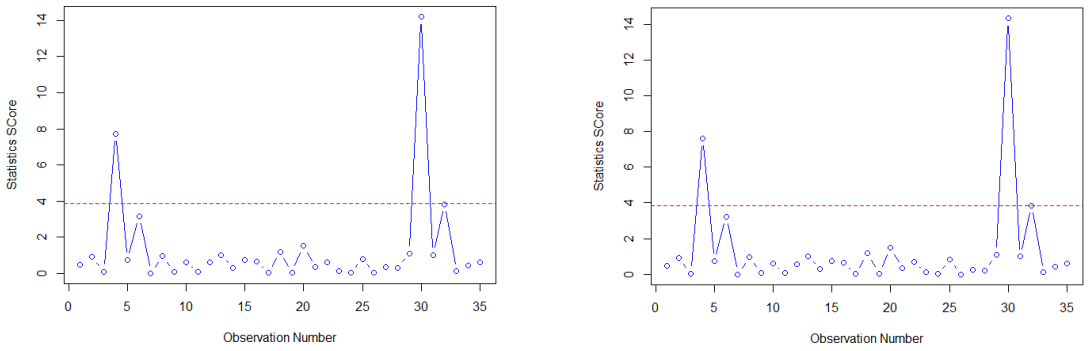
*) Significant in $\alpha = 0.05$

4.2. Detecting Outlier in SAR Model using MSOM

Modeling life expectancy with a spatial approach has been done by Hakim et al. (2019), Mukrom et al. (2021), and Yasin et al. (2020). This study used the Spatial Durbin Model (SDM) as the basic model and did not detect outliers in the model. Handling outliers in this study using a robust regression approach. In this study, after obtaining the SAR model, the next step is to detect outliers in the model. Outlier detection was carried out on each observation using the MSOM method. The results of outlier detection are shown in Table 4. If we use the queen weighting matrix, it can be concluded that two observations have the potential to become outliers in the model, namely the 4th (Temanggung Regency) and 30th (Cilacap Regency) observations. Meanwhile, if we use the Rook Weighting Matrix, the number of outliers detected becomes 3 locations with the addition of a 32nd (Boyolali Regency) location.

Table 4. Outlier Detection in SAR Model

Model	Number of Outliers	Observation Number
SAR-Queen	2	4 and 30
SAR-Rook	3	4, 30, and 32



(a) (b)

Figure 2. Plot of SC_i Value of SAR Model

The plot of SC_i values of the SAR-Queen model is shown in Fig. 2a. It shows that several observations are unusual, in which the 4th and 30th points have more large leverages than others. The plot of SC_i values of the SAR-Model is shown in Fig. 2b. This figure indicates that points 4, 30, and 32 have large values. So these are three outliers in the SAR-Rook Model.

4.3. Modified SAR Model to Accommodate the Outliers

In the MSOM method, to modify the SAR model, it is done by shifting the average of the observations indicated as outliers. Referring to Table 4, to accommodate outliers in the SAR-Queen model, we can modify by shifting the mean of the 4^{th} and 30^{th} observations. While in the SAR-Rook model, we can solve this by shift the means of 4^{th}, 30^{th}, and 32^{nd} observations. Table 5 shows the modified SAR model based on MSE, AIC, and BIC values. Based on this table, we can conclude that the modified model can improve the performance of the original SAR model because the MSE, AIC, and BIC values of the modified model are all smaller than the original model. It can also be seen from the distribution of the residuals wherein the skewness and kurtosis values of the modified model are close to the skewness and kurtosis values from the normal distribution, namely S=0 and K=3.

Table 5. Model Comparison

Model	Skewness	Kurtosis	MSE	AIC	BIC
SAR-Queen	-1.2914463	5.21488	1.2592624	125.61210	138.0549
Modified SAR-Queen	-0.3141667	2.29985	0.5368186	99.94502	115.4985
SAR-Rook	-1.2888626	5.26154	1.2491448	125.42521	137.8680
Modified SAR-Rook*	-0.4355527	2.61593	0.4486395	96.96485	114.0737

*) The Best Model

To determine whether there is a significant improvement or not, we can look at the Likelihood Ratio Test (LR Test) output. We can see that the p-value of the test is very small (p-value << 0.05), so we can conclude that the modified model can provide significant improvements (see Table 6). In detail, we can see the changes in parameters and other measures in Table 7.

Table 6. Likelihood Ratio Test (Modified Vs Original)

Model	**Likelihood Ratio**	**df**	**p-value**
Modified SAR-Queen **vs** SAR-Queen	29.667	2	0.00000
Modified SAR-Rook **vs** SAR-Rook	34.46	3	0.00000

Look at Table 7, some predictor variables in the model that are not significant to the response. We can solve this problem by excluding some insignificant variables. We can eliminate these variables starting from the variable with significance the weakest gradually. Elimination repeatedly until all variables are significant. We can see the final results in equation 14.

Table 7. Parameter of Modified SAR-Rook Model

Variable	Variable	*Parameter*	*Coeff*	p-value	AIC	MSE
Predictor	W_LE	ρ	-0.70104	0.0000*	96.667	0.4962
	Intercept	β_0	122.47654	0.0000*		
	ALS	β_1	0.56818	0.0005*		
	PCHLB	β_2	0.03533	0.0269*		
	PP	β_4	-0.17425	0.0000*		
Shifted Observations	d_4	δ_4	-2.75846	0.0004*		
	d_{30}	δ_{30}	-4.27363	0.0000*		
	d_{32}	δ_{32}	1.41939	0.0587*		

$$\hat{\boldsymbol{y}} = -0.70104\boldsymbol{W}\boldsymbol{y} + [\mathbf{1} \quad ALS \quad PCHLB \quad PP]\begin{bmatrix}122.476536\\0.568184\\0.035325\\-0.274252\end{bmatrix} + [\boldsymbol{d}_4 \quad \boldsymbol{d}_{30} \quad \boldsymbol{d}_{32}]\begin{bmatrix}-2.758458\\-4.273630\\1.419392\end{bmatrix} \tag{14}$$

Based on the equation 14, it can be explained that:

a. $\rho = -0.70104$, means that the life expectancy of each district/city has an effect of -0.70104 times the average life expectancy of each neighboring district/city.
b. An increase of 1 year in the Average Length of Schooling (ALS) will increase the Life Expectancy by 0.568184 years, where other variables are considered constant.
c. An increase of 1% in the Percentage of Households with Clean and Healthy Living Behavior (PHBS) will increase the Life Expectancy Rate by 0.035325 years, where other variables are considered constant.
d. An increase of 1% in the Percentage of Poor Population (PP) will decrease the Life Expectancy Rate of 0.174252 years, where other variables are considered constant.
e. The mean of Life Expectancy in 4th, 30th, and 32nd locations will be shifted by -2.758458, -4.273630, and 1.419392 year to reduce the outliers impact and improve the SAR Model.

5. CONCLUSION

Outliers will affect the parameter estimation results and model accuracy, so this study develops a modeling of life expectancy in Central Java Province using a modified SAR model using the Mean-Shift Outlier Model (MSOM) approach. Outliers are detected using the MSOM method. Then the result is used as the basis for modifying the SAR model. This modification, in principle, will reduce or increase the average of the observed data indicated as outliers. The modified SAR model can improve the model accuracy compared to the original SAR model. It can be proved by the increased coefficient of determination and decreasing the AIC value of the modified model. In addition, the modified SAR model can improve the skewness and kurtosis values of the residuals by getting closer to the Normal distribution. Further research needs to improve other spatial models using a different approach.

ACKNOWLEDGMENT

This research was fully funded by "DRPM-BRIN Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia" through the PDUPT scheme with contract number 225-92/UN7.6.1/PP/2021. The authors would like to thank the Directorate General of Higher Education for all their support.

REFERENCES

(BPS-Statistics of Jawa Tengah Province). (2018). *Profil Kesehatan Provinsi Jawa Tengah 2017*. BPS-Statistics of Jawa Tengah Province. https://jateng.bps.go.id/publication/2018/08/03/0392a381b71c2bc8f708f794/profil-kesehatan-provinsi-jawa-tengah-2017.html

(BPS-Statistics of Jawa Tengah Province). (2019). *Jawa Tengah Province in Figures 2019*. BPS-Statistics of Jawa Tengah Province.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers.

Anselin, L. (1992). Spatial Data Analysis with GIS: An Introduction to Application in the Social Sciences, *Technical Report 92-10*. *August*.

Beckman, R. J., & Cook, R. D. (1983). Outlier... ... . s. *Technometrics*, *25*(2), 119–149. https://doi.org/10.1080/00401706.1983.10487840

Bivand, R. et al. (2022). *spatialreg: Spatial Regression Analysis Version 1.2-3*.

Dai, X., Jin, L., Shi, A., & Shi, L. (2016). Outlier Detection and Accommodation in General Spatial Models. *Statistical Methods and Applications*, *25*(3), 453–475. https://doi.org/10.1007/s10260-015-0348-1

Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis Third Edition* (3rd ed.). John Wiley & Sons, Inc.

Genton, M. G. (1998). Spatial Breakdown Point of Variogram Estimators. *Math Geol*, *30*, 853–871.

Genton, M. G. (2001). *Robustness Problems in the Analysis of Spatial Data*, *Spatial Statistics: Methodological Aspects and Applications* (M. Moore (ed.); 159th ed.). Springer Lecture Notes in Statistics.

Hakim, A. R., Warsito, B., & Yasin, H. (2020). Live Expectancy Modelling using Spatial Durbin Robust Model. *Journal of Physics: Conference Series*, *1655*(1). https://doi.org/10.1088/1742-6596/1655/1/012098

Hakim, A. R., Yasin, H., & Rusgiyono, A. (2019). Modeling Life Expectancy in Central Java Using Spatial Durbin Model. *Media Statistika*, *12*(2): 152-163. https://doi.org/10.14710/medstat.12.2.152–163

LeSage, J. P. (1999). *The Theory and Practice of Spatial Econometrics*. University of Toledo.

LeSage, J. P., & Pace, R. K. (2009). *Introduction to Spatial Econometrics*. Taylor & Francis Group.

Militin, A. F., Palacios, M. B., & Ugarte, M. D. (2003). Robust Trend Parameters in a Multivariate Spatial Linear Model. *Test*, *12*(2), 445–457.

Mukrom, M. H., Yasin, H., & Hakim, A. R. (2021). Pemodelan Angka Harapan Hidup Provinsi Jawa Tengah Menggunakan Robust Spatial Durbin Model. *Jurnal Gaussian*, *10*(1), 44–54. https://doi.org/10.14710/j.gauss.v10i1.30935

Musyarofah, H., Yasin, H., & Tarno, T. (2020). Robust Spatial Autoregressive untuk Pemodelan Angka Harapan Hidup Provinsi Jawa Timur. *Jurnal Gaussian*, *9*(1), 26–40. https://doi.org/https://doi.org/10.14710/j.gauss.v9i1.27521

R Team. (2021). *shiny: Web Application Framework for R*.

Shi, L., & Chen, G. (2009). Influence Measures for General Linear Models with Correlated Errors. *Am Stat*, *63*(1), 40–42.

Yasin, H., Warsito, B., & Hakim, A. R. (2020). Development Life Expectancy Model in Central Java Using Robust Spatial Regression With M-Estimators. *Communications in Mathematical Biology and Neuroscience*, *2020*(69), 1–16. https://doi.org/10.28919/cmbn/4984