



THE APPLICATION OF SINGULAR SPECTRUM ANALYSIS METHOD IN FORECASTING THE NUMBER OF FOREIGN TOURISTS VISIT TO SPECIAL CAPITAL REGION OF JAKARTA

M A SODIQIN¹, W SULANDARI^{2*}, AND RESPATIWULAN³

^{1,2,3}Study Program of Statistics, Sebelas Maret University, Indonesia

*winita@mipa.uns.ac.id

ABSTRAK

Tourism has an important role for a society and a state as one of the supporting sectors for national development. Besides that, it is also an important factor for increasing people's income. Information regarding the estimated number of tourists in the past, present, and future is needed to establish a strategy or policy related to development in the tourism sector. Meanwhile, information about the required number of foreign tourists can be obtained using forecasting methods, one of which is the Singular Spectrum Analysis (SSA). This study aims to discuss the application of singular spectrum analysis (SSA) in predicting the number of foreign tourist visits to the Special Capital Region of Jakarta. There are two types of SSA forecasting methods, recurrent methods and vector methods. In its implementation, the performance of forecasting accuracy is influenced by the window length parameter. In this case, we are comparing several window length values, 10, 20, 30, and 40. In this study, a monthly number of foreign tourists visit to the Special Region of Jakarta from January 2011 to December 2019 was used. The results showed that the recurrent method with a window length of 40 resulted in a 16.6% smaller MAPE than the vector method. So, it can be concluded that the SSA method can predict the number of foreign tourists visiting the Special Capital Region of Jakarta well.

Keywords: Recurrent method, Singular Spectrum Analysis (SSA), Vector method

1 Introduction

Tourism has an important role for a society and a state. Tourism is one of the supporting sectors for national development as well as an important factor in increasing people's income. Tourism has also been designated as the leading sector by the Indonesian government and the second largest contributor to foreign exchange after the palm oil industry. The Ministry of Tourism recorded that foreign exchange earned by the tourism sector reached US \$ 9 billion. The realization of foreign tourist visits continues to grow until the end of 2018 and is targeted to reach seventeen million people, Anisa [1].

The provincial government of Capital city, Jakarta has targeted the number of foreign tourists in 2020 to be 3.1 million visits. Appropriate strategic planning in the tourism sector is necessary to achieve this target. In an effort to improve the number of foreign tourist visits, it is necessary to have appropriate strategic planning in the tourism sector. Meanwhile, information on the required number of foreign tourists can be obtained using forecasting methods. One of the most frequently used forecasting methods is time series analysis, Wei [2]. The Singular Spectrum Analysis (SSA) method is a new method in time series analysis

introduced by Broomhead and King [3]. Basically, the SSA method can decompose time series data into time series components, namely trends, seasonality, and noise. This makes the SSA method independent from assumptions like other time series analysis. Several studies related to the SSA method are as follows. Khaeri et al [4] conducted a study on the application of the SSA method in forecasting the number of train passengers on the island of Java in 2017. The results of this study concluded that the SSA method is quite good for extracting time series data based on time series components, namely trends, seasonality, and noise. Sari et al [5] from Udayana University conducted research on the application of the SSA method in forecasting the number of foreign tourist visits to Bali. The data used shows an uptrend pattern. In addition, there is a recurring pattern in the data every year which indicates a seasonal element in the data. The results of this study concluded that the SSA method can predict the number of foreign tourist visits to Bali with a Mean Absolute Percentage Error (MAPE) value of 7.64% which indicates very good forecasting according to Zhang [6]. Safitri et al [7] conducted a study on forecasting Jabodetabek train passengers using a singular spectrum analysis and the Holt-Winters method. The results of this study conclude that the SSA method is better than the Holt-Winters method in predicting the number of Jabodetabek trains passenger. Subanar and Sulandari [8] conducted a study on the method of forecasting the comparison of trends and time series of Indonesian seasonal tourist arrivals. The results of this study concluded that SARIMA-FTS is the most appropriate model to capture seasonal trends and patterns of the series in terms of Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE).

SSA is a powerful tool of analysis and forecasting of time series, Golyandina and Korobeynikov [9]. There are two forecasting methods in the SSA method, namely the recurrent and vector methods. The recurrent method is the basic method that is often used because it is easier [10]. The vector method is a modification of the recurrent method. The difference between the two forecasting methods is that the recurrent method performs direct continuation with the help of the Linear Recurrent Formula (LRF), while the vector method is related to L-Continuation. This causes the approximate continuation usually gives different results, Golyandina et al [10]. For this reason, this study also compares the two methods.

2 Literature Review

2.1 Singular Spectrum Analysis (SSA)

SSA is an alternative and newest method of analyzing time series data. Basically, SSA consists of two stages, namely: decomposition and reconstruction. The basic SSA algorithm divides the initial time series data into new time series data consisting of trends, seasonality, and noise. The decomposition stage consists of embedding and SVD stages, while the reconstruction stage consists of grouping and diagonal averaging stages.

2.2 Embedding

In the embedding stage, the time series data is converted into a trajectory matrix (trajectory matrix), which transforms one-dimensional data (vector) into multidimensional data (matrix). Suppose that time series data of length N , without missing data is represented by $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, the data is transformed into a matrix of size $L \times K$, where L is the window length with $1 < L < N$. There is no specific method to determine the value of L with certainty, so to determine the value of L is done by trial and error and $K = N - L + 1$. The form of the matrix can be written as follows:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_K \\ \mathbf{x}_2 & \mathbf{x}_3 & \dots & \mathbf{x}_{K+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_L & \mathbf{x}_{L+1} & \dots & \mathbf{x}_N \end{bmatrix}$$

The X matrix is also known as the Hankel matrix, where all the anti-diagonal elements have the same value. So at this stage the output obtained is a Hankel matrix of size $L \times K$.

2.3 Singular Value Decomposition (SVD)

After the embedding process that produces the Hankel matrix, the next stage is the singular matrix decomposition process. The determination of the singular matrix in SSA can be explained, $S = XX^T$ for example $\lambda_1, \dots, \lambda_L$ is the eigenvalue of the S matrix with $\lambda_1 \geq \dots \geq \lambda_L \geq 0$ and $U_1 \dots U_L$ is the eigenvector of each value eigen. $V_i = X^T U_i / \sqrt{\lambda_i}$ for $i = 1, \dots, d$ with $d = \max \{i, \text{so that } \lambda_i > 0\}$ then the SVD of the path matrix X is obtained as follows:

$$X = X_1 + \dots + X_d, \text{ where } X_i = \sqrt{\lambda_i} U_i V_i^T$$

The X_i matrix is also called the eigentriple. So at this stage the output produced is an eigentriple, namely a singular matrix, eigen spectrum, and principal component matrix.

2.4 Grouping

The grouping stage in the reconstruction step is the grouping of the $X_{(L \times K)}$ matrix into subgroups based on the pattern of forming the time series data, namely trends, seasonality, and noise. The results of the grouping are then added together, in other words the matrix will partition the index set $\{1, \dots, d\}$ into m independent subsets, I_1, \dots, I_m .

Suppose $I = i_1, \dots, i_p$, then the resulting matrix X_I corresponds to group I defined as $X_I = X_{i_1} + \dots + X_{i_p}$. The SVD expansion is calculated with $I = I_1, \dots, I_m$ causing the decomposition to become $X = X_{I1} + \dots + X_{Im}$. The procedure for selecting I_1, \dots, I_m is called eigentriple clustering.

2.5 Diagonal Averaging

The purpose of the diagonal averaging stage is to get the singular value of the components that have been separated, which will then be used in forecasting. At this stage the X_{Ij} matrix obtained in the grouping stage is rearranged into a new data series with length N . Let Y be a matrix of size $L \times K$ with elements y_{ij} , with $1 \leq i \leq L$ and $1 \leq j \leq K$. Suppose we define $L^* = \min(L, K)$ and $K^* = \max(L, K)$, $N = L + K - 1$. Let $y_{ij}^* = y_{ij}$ if $L \leq K$ and $y_{ij}^* = y_{ji}$ if $L > K$. Using the diagonal averaging method, the Y matrix is transformed into series again, $y_1 \dots y_N$ using the following equation:

$$y_k = \begin{cases} \frac{1}{k} \sum_{m=1}^k y_{m, k-m+1}^* & \text{untuk } 1 \leq k < L^* \\ \frac{1}{L^*} \sum_{m=1}^{L^*} y_{m, k-m+1}^* & \text{untuk } L^* \leq k \leq K^* \\ \frac{1}{N-k+1} \sum_{m=k-K^*+1}^{N-K^*+1} y_{m, k-m+1}^* & \text{untuk } K^* \leq k \leq N \end{cases} \quad \text{Equation (1)}$$

Equation (1) is applied to the resultant matrix X_{Ik} resulting in a reconstructed series $\tilde{X}^{(k)} = (\tilde{x}_1^{(k)}, \dots, \tilde{x}_N^{(k)})$ therefore, the initial series x_1, \dots, x_N is decomposed into a number of m series which are reconstructed as follows:

$$x_n = \sum_{k=1}^m \tilde{x}_n^{(k)}, \text{ (with } n = 1, \dots, N)$$

The series of reconstructions produced by elementary grouping will be called the basic reconstruction series.

2.6 Forecasting

a. Recurrent

Recurrent is the basic method that is often used because it is relatively easier. Basically, forecasting with the recurrent method continues directly with the help of LRF. The time series $y_{N+M} = (y_1, \dots, y_{N+M})$ is obtained by the following formula:

$$y_i = \begin{cases} \hat{y}_i & \text{for } i = 1, \dots, N \\ \sum_{j=1}^{L-1} a_j y_{i-j} & \text{for } i = N + 1, \dots, N + M \end{cases}$$

The SSA forecasting method has two general stages, namely: diagonal averaging and continuation. The stages of the recurrent method, diagonal averaging is used to obtain reconstruction and continuation is performed by applying LRF.

b. Vector

Vector is a modified result of the recurrent method. Basically SSA forecasting assumes that the continuation sequence vector $\hat{\mathbf{X}}_1, \dots, \hat{\mathbf{X}}_K$ (which belongs to the \mathbf{L}_r subspace) for step M so the continuation vector \mathbf{Z}_m ($K < m \leq K + M$) is included in the subspace \mathbf{L}_r and matrix $\mathbf{X}_M = \hat{\mathbf{X}}_1 : \dots : \hat{\mathbf{X}}_K : \mathbf{Z}_{K+1} : \dots : \mathbf{Z}_{K+M}$ are approximately Hankel.

After obtaining the \mathbf{X}_M matrix, the \mathbf{Y}_{N+M} forecast series is obtained for the diagonal averaging of this matrix and the diagonal averaging series is $\mathbf{y}_1, \dots, \mathbf{y}_{N+M+L-1}$ with $\mathbf{y}_{N+1}, \dots, \mathbf{y}_{N+M}$ which is the M term of the forecast vector. Basically, vector method is related to L-Continuation. This causes the approximate continuation usually gives different results.

The forecasting stages of the vector method are almost the same as the recurrent method, while in the vector method the two stages are used in reverse order. First, vector forecasting is performed and then diagonal averaging provides forecast values to obtain M periods in the future. The vector method uses an $M+L-1$ step procedure, so the vector has an additional $L-1$ step.

2.7 Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error (MAPE) is a relative determination measure used to determine the percentage deviation of forecasting results and also indicates how much error in forecasting is compared to the actual value. Measurement of the level of forecasting error in this study using MAPE with the following formula:

$$\text{MAPE} = \frac{1}{n} \sum_{j=1}^n \left| \frac{y_j - \hat{y}_j}{y_j} \right| \times 100\%$$

where \hat{y}_j and y_j are the estimated and actual values and n is the number of data.

3 Research Methods

In this study, a monthly number of foreign tourists visit to the Special Region of Jakarta from January 2011 to December 2019 was used. The data is divided into two, namely in-sample and out-sample data. Data on the number of foreign tourists visit to Special Capital Region of Jakarta from January 2011 to December 2018 are in-sample data and data on the number of foreign tourist visits to Special Capital Region of Jakarta from January-December 2019 are out-sample data.

The steps for forecasting the number of foreign tourists to Special Capital Region of Jakarta using the SSA method are as follows:

1. Dividing the data used into in-sample and out-sample data with a ratio of 1: 8.
2. Determining the parameter value of the windows length to be used, parameter selection is carried out based on trial and error by considering the smallest MAPE value.
3. Performing decomposition, namely embedding and SVD.
4. Performing reconstruction, namely grouping and diagonal averaging.
5. Performing forecasting with the recurrent and vector methods.
6. Determining the best SSA model based on the MAPE value with the following formula:

$$\text{MAPE} = \frac{1}{n} \sum_{j=1}^n \left| \frac{y_j - \hat{y}_j}{y_j} \right| \times 100\%$$

where \hat{y}_j and y_j are the estimated and actual values and n is the amount of data.

7. Interpreting forecasting results and draw conclusions.

4 Results and Discussion

This section discusses the application of the SSA method to predict the number of foreign tourist visits to Special Capital Region of Jakarta. The first step is to divide the data into two, in-sample and out-sample data. Data on the number of foreign tourist visits to Special Capital Region of Jakarta from January 2011 to December 2018 are in-sample data and data on the number of foreign tourist visits to Special Capital Region of Jakarta from January-December 2019 are out-sample data.

In SSA, the window length (L) parameter selection is carried out based on trial and error by considering the smallest MAPE value. In this in-sample data, the amount of data is 96 so that the L value ranges from 2 to 48. To facilitate the search for optimum L , a grid method is used by trying the $L = 10, 20, 30$, and 40 values, then the L value with the smallest MAPE is selected. The author tries to use $L = 40$.

On resistance to decomposition with the embedding process, the X matrix (Hankel) can be structured as follows:

$$X = [X_1, X_2, \dots, X_K] = \begin{bmatrix} 145179 & 149645 & \dots & 217994 \\ 149645 & 166393 & \dots & 203444 \\ \vdots & \vdots & \ddots & \vdots \\ 191494 & 175391 & \dots & 193788 \end{bmatrix}$$

The next step in the decomposition stage is the singular value decomposition (SVD). From the obtained Hankel matrix, then SVD which produces 40 eigentriples (as much as L). Eigentriple consists of a singular value (λ_i), eigenvector (U_i) and the main component (V) which is presented in Table 1.

Table 1. The eigentriple component consists of a singular value (λ_i), eigenvector (U_i) and the main component (V)

No.	Singular Value (λ_i)	Eigenvector (U_i)			Main Component (V)		
1	8.88×10^{14}	-0.15	...	-0.28	-0.12	...	0.08
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
40	1.47×10^{10}	-0.16	...	0.13	-11.59	...	-0.00

The initial step in the reconstruction phase is the grouping eigentriples related to trends, seasonality, and noise. Grouping effect (r) is a the parameter used at the grouping stage to limit the number of eigentriples that will be used in the process of identifying trend and seasonal components. The grouping effect (r) parameter value is determined based on the number of

eigentriples that do not reflect noise in the singular value plot. In a singular value plot, a slow descending sequence of singular values reflects the noise component.

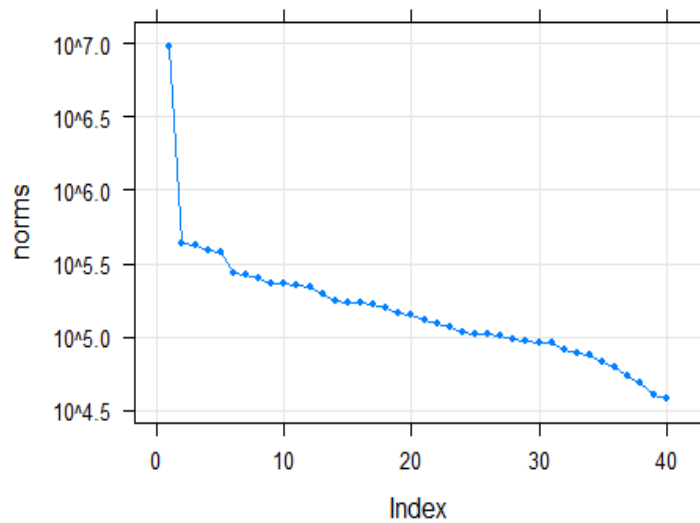


Figure 1. Plot plot of eigen value from biggest to smallest

The Figure 1 above shows the plot of the singular value starting to decline slowly from eigentriple 6. This results in eigentriple 6 to eigentriple 48 being identified as a noise component. Therefore, the eigentriple which is not included as a noise component is the grouping effect (r) parameter value set to 5. Thus, the number of eigentriples that will be used to identify the trend and seasonal components is 5 eigentriples. Plots of the reconstructed series can be used to identify eigentriples associated with trends and seasonality.

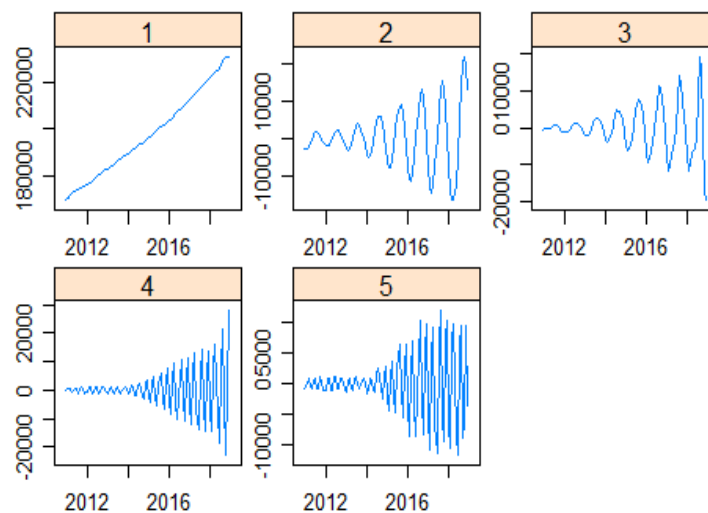


Figure 2. The reconstructed series plot from the first 5 eigen value with axis x is date and axis y is number of visits

The Figure 2 above shows that the series reconstructed by eigentriple 1 contains slow varying components. Therefore, eigentriple 1 is grouped into trend components. Furthermore, the grouping of eigentriple which is related to the seasonal component is carried out based on the similarity of the singular value of the consecutive eigentriples. In the reconstructed series plot, the similarity of the singular values results in the series reconstructed by an eigentriple having the same seasonal patterns and periods as the series reconstructed by other eigentriples.

So it can be seen that there are several consecutive pairs of eigentriples that have a similar pattern, namely eigentriple 2 and 3, eigentriple 4 and 5.

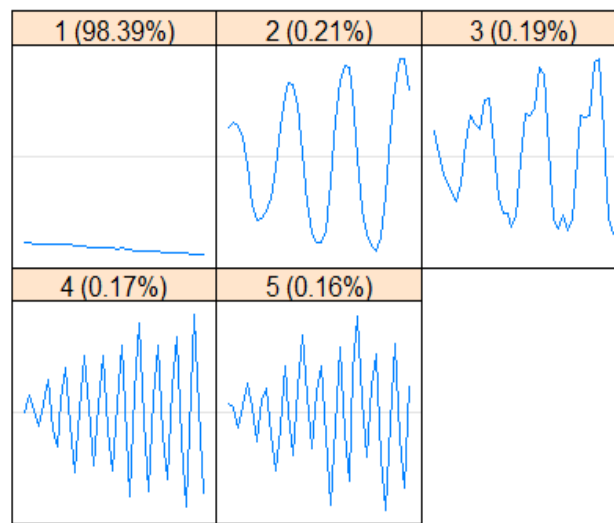


Figure 3. The eigenvector plot of the reconstructed series

The Figure 3 above shows that the series reconstructed by eigentriple 2,3,4 and 5 have different seasonal periods. In this case, we refer to them as oscillating components. The eigentriple identification process that reflects the trend and seasonal components can be seen in the W-correlation plot. The W-correlation plot is used to see the magnitude of the correlation between eigentriples. The darker the color, the higher the correlation. The lighter grayish color indicates that the eigentriple tends to be noise.

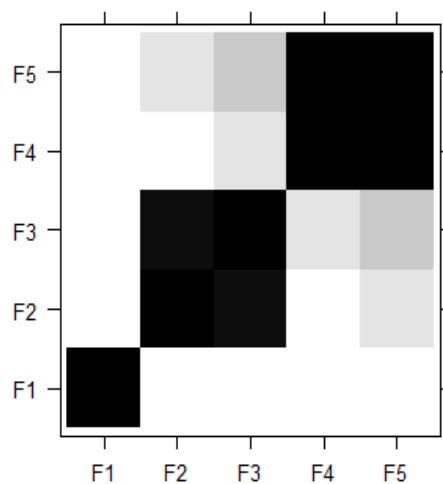


Figure 4. Plot W-correlation of the reconstructed series obtained from the first 5 eigen value

The Figure 4 above shows that between the eigentriple 2 and 3, and eigentriple 4 to 5 have a strong correlation, respectively. Moreover, there is also a correlation between those two groups of eigentriples. The correlation value is indicated by a color gradient from black (correlation = 1) to white (correlation = 0). From the above analysis, several possible combinations of groups with the same grouping using the two SSA methods are presented in Table 2.

Table 2. Possible group combinations

Method	Group
Recurrent method	Trend(1) and Oscillation (2,3,4,5)
Vector method	Trend(1) and Oscillation (2,3,4,5)

In the diagonal averaging step, each component can be reconstructed using their respective eigentriple. In this case, the trend component is reconstructed by eigentriple 1. The resulting plot of the reconstructed trend component is presented in Figure 5.

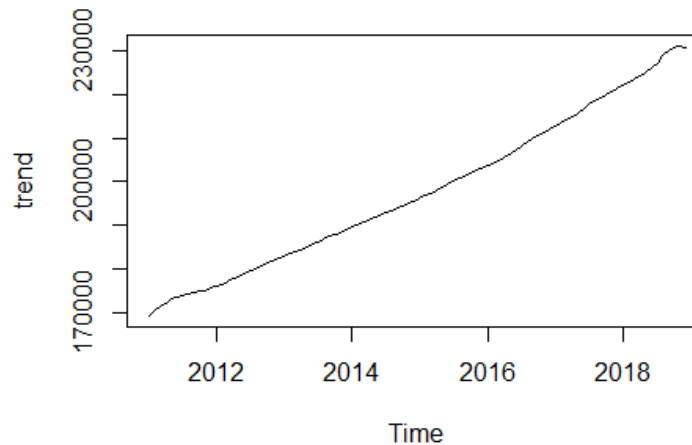


Figure 5. The plot of the results of the reconstructed trend components

Figure 5 above shows that the plot of the reconstructed trend component appears to be steadily increasing. Furthermore, the oscillation components are reconstructed by eigentriple 2,3, 4 and 5. The plot of the reconstructed seasonal components is presented in Figure 6.

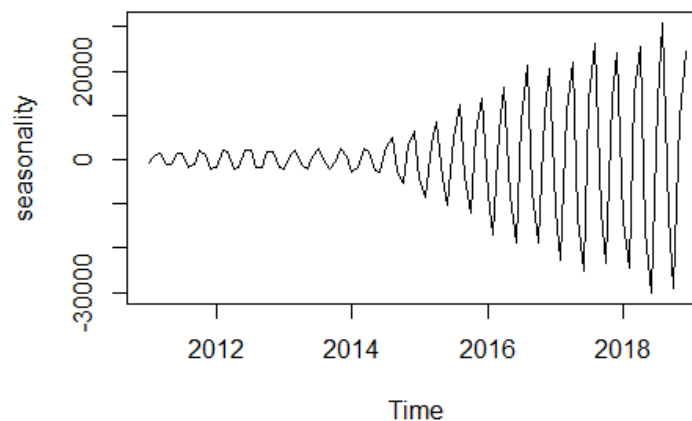


Figure 6. The plot of the results of the reconstructed seasonal components

Figure 6 above shows that the plot of the reconstruction of the seasonal components looks seasonal every year. Furthermore, the plot of the results of the reconstructed noise components is presented in Figure 7.

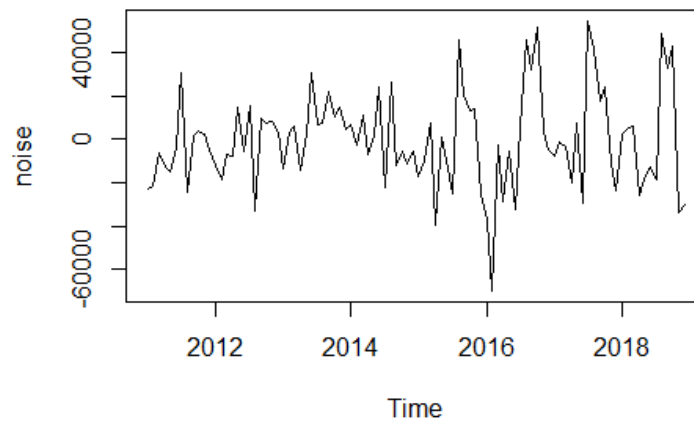


Figure 7. The plot of the results of the reconstructed noise components

Figure 7 above shows that the plot of the reconstruction of the noise component looks very fluctuating. Based on several group combinations that have been obtained, forecasting is carried out by trying the L values of 10, 20, 30, and 40 because in this case, the amount of data is 96 so that the L value ranges from 2 to 48. The respective MAPE values are obtained which are presented in Table 3.

Table 3: MAPE values from multiple trials

L	MAPE	
	Recurrent	Vector
10	19.8	20.7
20	18.6	20.8
30	18.3	19.8
40	16.6	18.1

From the Table 3, it is resolved that the L value with the smallest MAPE is 40, namely 16.6%. With this MAPE, it is expected that the prediction results obtained from the model will not differ greatly from the actual data values. Therefore, furthermore, this study uses an L value of 40. According to Zhang et al [6], MAPE value of less than 10% indicate that the forecast is very accurate. Meanwhile, the MAPE value between 10% and 20% indicates that the forecast is good. Thus, it can be concluded that the Recurrent method can predict the number of foreign tourist visits to Special Capital Region of Jakarta well and it is obtained that the eigentriple which is grouped into trend groups is eigentriple 1. Furthermore, the eigentriple which is grouped into oscillation groups is between the eigentriple 2 and 3, and eigentriple 4 to 5, the rest of the eigentriple which is not grouped into trend and seasonal groups is the noise group. After the SSA model used for forecasting has been formed, the next step is to forecast based on the best model that has been obtained. So that the forecast results for January-December 2019 are presented in Figure 8.

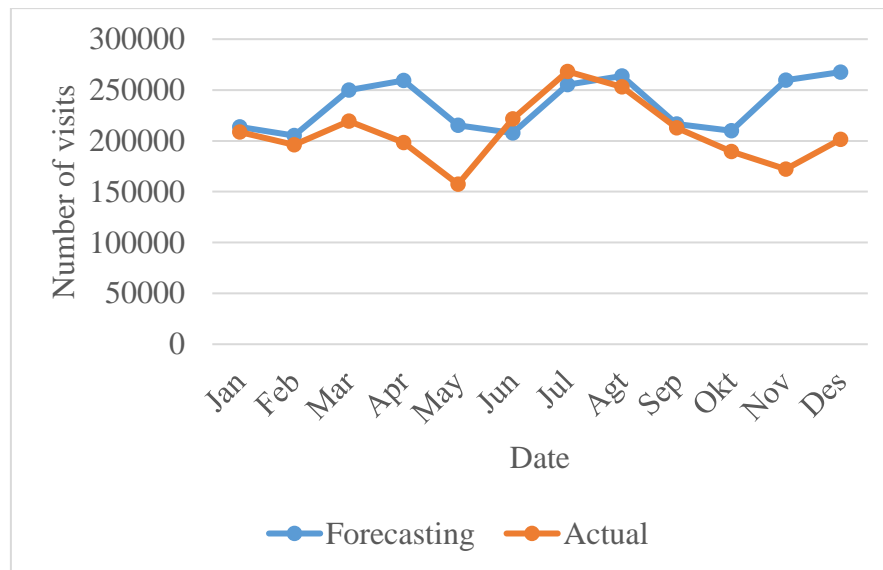


Figure 8. Actual value and forecast results for January - December 2019 with Recurrent method

With the same steps as above, using out-sample data, the MAPE value is 19.7%. So that the forecast results for January 2020-December 2021 are presented in Figure 9.

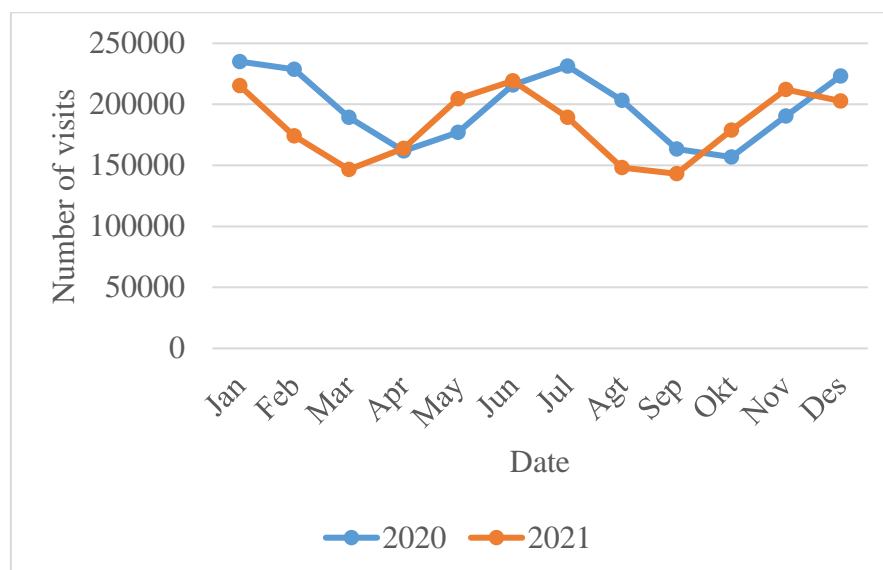


Figure 9. Forecast results for January 2020-December 2021 with out-sample data

Based on the Figure 9, it is known that the peak of foreign tourist visits occurs in January 2020. Through this information, it is hoped that the government can make new breakthroughs and the right policies in order to attract more foreign tourists to visit the Special Capital Region of Jakarta.

5 Conclusion

From the overall process above, it can be concluded that the SSA method can predict the number of foreign tourist visits to the Special Capital Region of Jakarta well and the best SSA model is a model with a value of $L = 40$. In the in-sample data, the mape value is 16.6% and the out-sample data is 19.7%. Future researchers are expected to be able to produce better forecasts and produce smaller MAPE values. Based on the MAPE value, it can be said that the recurrent method of SSA has better accuracy in predicting the number of foreign tourist visits

to the Special Capital Region of Jakarta compared to the vector method. Based on the forecast results, it is known that the peak of foreign tourist visits occurs in January 2020. Through this information, it is hoped that the government can make new breakthroughs and the right policies in order to attract more foreign tourists to visit the Special Capital Region of Jakarta.

References

- [1] D. F. Anisa, "Sektor Pariwisata Berpeluang Geser Sawit sebagai Penyumbang Devisa Terbesar," 2019. [Online]. Available: <https://www.beritasatu.com>.
- [2] W. W. S. Wei, "Time Series Analysis Univariate and Multivariate Methods; Second Edition," USA, Pearson Addison Wesley, 2006.
- [3] D. S. Broomhead dan G. P. King, "Extracting Qualitative Dynamics From Experimental Data," *Physica 20D*, pp. 217-236, 1986.
- [4] H. Khaeri, . E. Yulian dan G. Darmawan, "Penerapan Metode Singular Spectrum Analysis (SSA) pada Peramalan Jumlah Penumpang Kereta Api di Indonesia Tahun 2017," *Jurnal Euclid*, p. 8, 2018.
- [5] M. A. N. Sari, W. Sumarjaya dan M. Susilawati, "Peramalan Jumlah Kunjungan Wisatawan Mancanegara ke Bali Menggunakan Metode Singular Spectrum Analysis," *Jurnal Euclid*, pp. 303-308, 2019.
- [6] T. Zhang, K. Wang dan X. Zhang, "Modeling and Analyzing the Transmission Dynamics of HBV Epidemic in Xinjiang, China," *Journal Plos One*, pp. 1-14, 2005.
- [7] D. Safitri, Subanar, U. H dan S. W, "Forecasting of jabodetabek train passengers using singular spectrum analysis and holt-winters methods," *Journal of Physics: Conference Series*, 2020.
- [8] Subanar and W. Sulandari, "A comparison forecasting methods for trend and seasonal Indonesia tourist arrivals time series," *AIP Conference Proceedings*, vol. 2329, pp. 060012-1 - 060012-10, 2021.
- [9] G. Nina dan K. Anton, "Basic Singular Spectrum Analysis and forecasting with R," *Computational Statistics & Data Analysis*, vol. 71, pp. 934-954, 2014.
- [10] N. Golyandina, V. Nekrutkin dan A. A. Zigljavsky, *Analysis of Time Séries Structure: SSA and Related Thechniques*, London: Chapman & Hall, 2001.