SCIREA Journal of Mathematics
http://www.scirea.org/journal/Mathematics
May 4, 2023
Volume 8, Issue 2, April 2023
https://doi.org/10.54647/mathematics110393

# Forecasting of Air Passengers using Singular Spectrum Analysis

**Sisti Nadia Amalia and Zul Amry**

Department of Mathematics, State University of Medan, Indonesia

Email: sistinadia@unimed.ac.id and zul.amry@gmail.com

Corresponding author: sistinadia@unimed.ac.id (Sisti Nadia Amalia)

## Abstract

Air transportation is the most appropriate option for extremely vast distances, such as those between cities, provinces, and countries. While unpredictability, high volatility, and seasonality sometimes result in complex behavior in air passenger time series, this research applies the Singular Spectrum Analysis technique for air passengers data and uses the linear recurrent type for forecasting. Trends, seasonality, cyclists, and noise can all be found and extracted using Singular Spectrum Analysis. Singular Spectrum Analysis has the potential to be a highly effective forecasting method.

**Keywords**: Singular Spectrum Analysis, Linear Reccurent Forecasting, Air Passenger

## 1. Introduction

Kualanamu International Airport, the fourth largest commercial airport in Indonesia with economic potential and a strategic geographical position, has the main capital as the second international and domestic hub in Indonesia besides Soekarno Hatta International Airport.

Kualanamu Airport is one of the busiest airports providing local and international routes due to its strategic location in Medan City, which serves as the entry point to western Indonesia.

According to estimates, there are over 10 million passengers a year, with almost 90% of people going on domestic flights. Data on the number of passengers on flights is a time series in which the data are systematically forecasted according to previous and present information that is known to be capable of minimizing the impact of fluctuating and tending to increase air passengers.

Prediction is a form of forecasting based on theoretical assumptions. Estimating future values using past data requires high sensitivity and the ability to understand the characteristics of the data so that future possibilities can be estimated. Forecasting can also be used as a reference in planning and establishing policies so as to provide the best alternative courses of action to choose from among the various possibilities available.

The methods that can be used for prediction are very diverse, and one of the methods that can be used to predict is using Singular Spectrum Analysis. Singular Spectrum Analysis is a non-parametric method that does not require initial assumptions about the data, so it can collect all patterns and is flexible to the data. Thus, it is expected to get powerful forecasting results.

## 2. Materials and Methods

### 2.1 Singular Spectrum Analysis

The concepts of SSA have been explained in detail in [5]. In brief, the basic SSA algorithm has two stages: decomposition and reconstruction.The initial stage of SSA is decomposition by embedding and singular value decomposition (SVD). Embedding is the process of forming the original series into the path matrix; SVD decomposes the track matrix and breaks the data into trend, seasonality, monthly components, and noise according to their single values. Then the next stage is reconstruction, which involves clustering to create groups from the decomposition of the path matrix and diagonal averages to reconstruct a new time series from subgroups.

**Step 1. Decomposition**

The basic stage of decomposition is called the embedding step. The concept of embedding is the path matrix, which transfers the one-dimensional time series of length $N$, into a sequence of L-dimensional vectors ($i = 1, K = N - L + 1$). The window length $L$, ($1 < L < N$)

is the only parameter in this delay procedure. The columns of the $(LxK)$ trajectory matrix will be formed of the $K$ vectors $Xi$. The trajectory matrix $X$ is a Hankel matrix with equal elements on the diagonals( $i + j = const$).

$$\mathrm{X} = \begin{bmatrix} x_1 & x_2 & x_3 & . & x_K \\ x_2 & x_3 & x_4 & . & x_{K+1} \\ . & . & . & . & . \\ x_L & x_{L+1} & x_{L+2} & . & x_N \end{bmatrix} \quad (1)$$

The next stage in the decomposition is singular value decomposition (SVD). This stage is the most important stage in SSA : converting the trajectory matrix into a sum of rank-one biorthogonal elementary matrices. From the $S = XXT$ matrix, the eigenvalues $(\lambda_k)$ of $S$ taken in decreasing order of magnitude $\lambda_k$ $(1 \le k \le L)$ and eigenvector $(U_k)$ can be calculated and then sorted by the ortho-normal system of the eigenvectors of the matrix $S$ corresponding to these eigenvalues. Then calculate the $i$-th element of the $k$ principal components that will be formed $V_i = {\mathrm{X}^T U_i}/{\sqrt{\lambda_i}}$ . Then the SVD of the path matrix can be calculated by $\mathrm{X}_i = \sqrt{\lambda_i} U_i V_i^T$. In this notation, the SVD of trajectory matrix X can be written as

$$X = X_1 + X_2 + ... + X_d \quad (2)$$

**Step 2. Reconstruction**

The initial stage of reconstruction in SSA is grouping. The concept of grouping involves grouping the results of the path matrix decomposition and uniting those that are considered similar to form several groups. In the next reconstruction stage called diagonal averaging, each elementary matrix of the grouped decomposition is transformed into a new time series by applying a linear transformation known as diagonal averaging or Hankelization. The diagonal averaging algorithm transforms Y into the reconstructed time series $y_1, ..., y_N$ using the formula:

$$y_k = \begin{cases} \frac{1}{k}\sum_{m=1}^{k} a^*_{m,k-m+1}, 1 \le k < O \\ \frac{1}{O-1}\sum_{m=1}^{O-1} a^*_{m,k-m+1}, O \le k < P+1 \\ \frac{1}{N-k+1}\sum_{m=k-P+1}^{N-P+1} a^*_{m,k-m+1}, P+1 \le k < N \end{cases} \quad (3)$$

By applying the Hankelization procedure to all matrix components, we obtain another expansion:

$$\widetilde{Y}^{(k)} = \left(\tilde{y}_1^{(k)}, ..., \tilde{y}_N^{(k)}\right) \quad (4)$$

.

## 2.2 SSA Reccurent Forecasting

In forecasting using the SSA Linear Recurrent Formula (LRF), the concept is to create a model

$$y_{i+d} = \sum_{k=1}^{d} r_k \, y_{i+d-k} \, , 1 \leq i \leq N - d \quad (5)$$

The LRF coefficient $(r_1, ..., r_d)$ obtained from the singular value decomposition using the formula

$$(r_{L-1}, ..., r_1)^T = \frac{1}{1-v^2} \sum_{i=1}^{L-1} \pi_i \, \mathrm{P}_i^{\nabla}, \; v^2 = \sum_{i=1}^{L-1} \pi_i^2 \quad (6)$$

where $\mathrm{P} = (\mathrm{p}_1, \mathrm{p}_2, ..., \mathrm{p}_{\mathrm{L}-1}, \mathrm{p}_{\mathrm{L}})$ and $\mathrm{P}^{\nabla} = (\mathrm{p}_1, \mathrm{p}_2, ..., \mathrm{p}_{\mathrm{L}-1})$

In forecasting using SSA recurrent forecasting, the series used is from the reconstruction results obtained from the diagonal average results. Then $M$ new data points will be determined to be forecasted. The series of forecasting results that will be formed can be written $G_{N+M} = (g_1, g_2, ..., g_{N+M})$ as follows :

$$g_i \begin{cases} \tilde{y}_i \; , & i = 0,1, ..., N \\ \sum_{j=1}^{L-1} r_j g_{i-j} \; , & i = N+1, ..., N+M \end{cases} \quad (7)$$

## 2.3 Forecasting Accuracy

### 2.3.1 *Mean Absolute Percentage Error (MAPE)*

The accuracy of the forecast used MAPE as a reference for the suitability of the method with the data. MAPE is used to measure the accuracy of prediction results with actual data in the form of an average absolute error proportion, as follows:

$$MAPE = \frac{1}{n} \sum_{t=1}^{n} \frac{|Y_t - \widehat{Y_t}|}{Y_t} \times 100\% \quad (8)$$

The MAPE criteria in [4] are as follows:

| | |
|---|---|
| < 10% | Highly accurate forecasting |
| 10 – 20% | Good forecasting |
| 20 – 50% | Reasonable forecasting |
| > 50% | Weak and inaccurate forecasting |

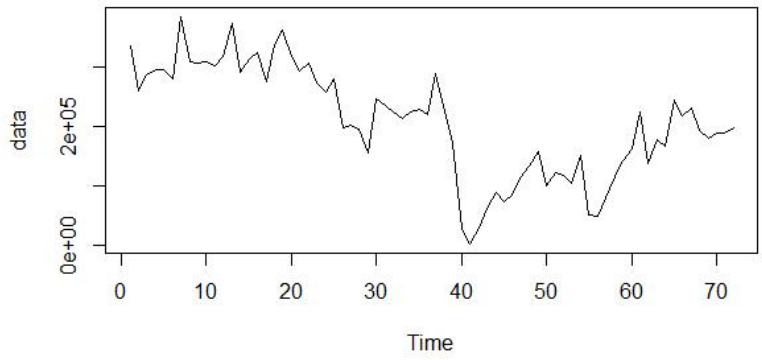
### 2.3.2 *Tracking Signal*

The tracking signal is a measure of tolerance that can be used to determine the possibility of using the forecasting results, which estimate if the basic pattern changes. In [2], if the tracking signal value is outside the acceptable limit (+/- 5), then the forecasting model must be reviewed. as follows:

$$Tracking\ Signal = \frac{\sum_1^n e_n}{\sum_1^n \frac{|e_n|}{n}} \tag{9}$$

## 3 Results

The data used in this research is passenger data for domestic routes from Kualanamu Airport for the observation period of January 2018–December 2022 from [15].




**Pict 1. Plot of Passenger Data for Kualanamu Airport Domestic Route**

From the plot above, at a glance, we can recognize the influence of trends and seasonality over several periods.

**Step 1. Decomposition**

In the decomposition process, the first step is embedding by determining the window length. In this case, the number of data points is 60. In determining the optimum $L$, it is done by trying 10, 20, and 30. Then the MAPE with the minimum value is selected. The results are as follows:

| **L** | **10** | **20** | **30** |
|---|---|---|---|
| **MAPE** | 54,26 | 63,68 | 31,96 |

Obtained with a minimum MAPE of 30. In the same way that tracking around for 30 is carried out to get the most appropriate, the result is:

| L | 27 | 28 | 29 |
|---|---|---|---|
| **MAPE** | 18,59 | 14,47 | 17,41 |

Obtained with a minimum MAPE, $L$ = 28 of 14.47 percent. It is expected that the prediction results obtained from the model do not differ much from the actual data values. With $L$ = 28 and $K$ = 33, the Hankel matrix can be arranged as follows:

$$\mathrm{X} = \left(x_{ij}\right)_{i,j=1}^{28;33} = \begin{bmatrix} 334556 & 259457 & 287522 & . & 214364 \\ 259457 & 287522 & 294602 & . & 225484 \\ . & . & . & . & . \\ 195395 & 155731 & 245820 & . & 161457 \end{bmatrix} \quad (10)$$

Based on the Hankel matrix, the SVD stage produces 28 eigentriples. Eigentriple consists of singular value $(\boldsymbol{\lambda_i})$,eigenvector $(U_i)$, and principal component $(V_i)$ as follows:

Singular value $(\boldsymbol{\lambda_i})$

| NO | $\boldsymbol{\lambda_i}$ | $\sqrt{\boldsymbol{\lambda_i}}$ |
|---|---|---|
| 1 | 4.741555e+13 | 6885894.94 |
| 2 | 6.654375e+11 | 815743.50 |
| ⋮ | ⋮ | ⋮ |
| 27 | 2.606720e+09 | 51056.04 |
| 28 | 2.204908e+09 | 46956.44 |

Eigenvector $(U_i)$

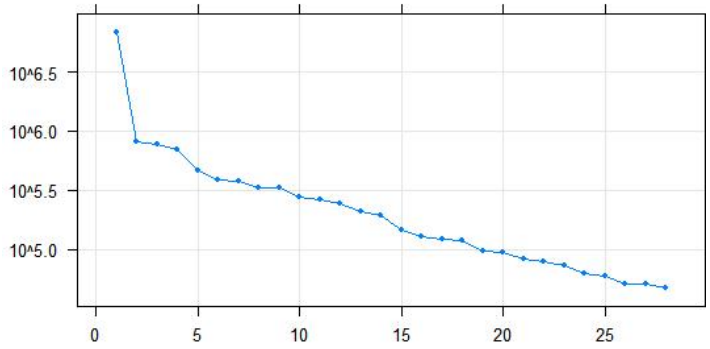
| $\boldsymbol{U_1}$ | $\boldsymbol{U_2}$ | ... | $\boldsymbol{U_{28}}$ |
|---|---|---|---|
| -0.236 | 0.144 | ... | -0.279 |
| -0.234 | 0.177 | ... | -0.263 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| -0.123 | -0.264 | ... | 0.129 |
| -0.121 | -0.209 | ... | -0.233 |

Principal Component $(V_i)$

| NO | $\boldsymbol{V_1}$ | $\boldsymbol{V_2}$ | ... | $\boldsymbol{V_{28}}$ |
|---|---|---|---|---|
| 1 | -0.226 | -0.160 | ... | -0.061 |
| 2 | -1.874 | -0.117 | ... | -0.203 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 32 | -1.024 | 0.668 | ... | -0.113 |
| 33 | -1.484 | 0.551 | ... | 0.052 |

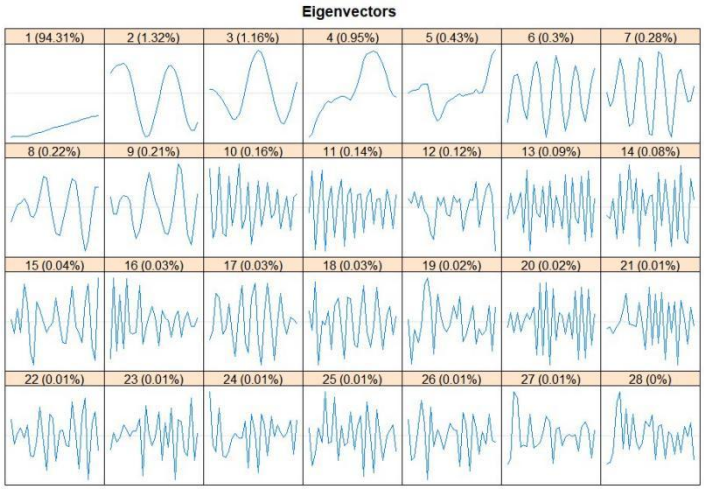
**Step 2. Reconstruction**

In the grouping process, the estimated number of groups formed can be seen from the scree plot, as follows:




**Pict 2. Scree Plot**

From the picture 2, it can be seen that there will be around 4 or 5 groups formed. To determine the members of the group, we can use the plot of the eigenvector.




**Pict 3. Eigenvector Plot**

Grouping can be created on a subjective basis through plots; it can also be created by calculating the period values of each eigenvector. The adjacent periodicities are indicated to be in the same group. By using the R program obtained:

| **Eigenvector** | **Period** |
|---|---|
| 1,2,3,4,5 | Trend |
| 6 | 5.6 |
| 7 | 5.6 |
| ⋮ | ⋮ |
| 27 | 4.7 |
| 28 | 2 |

From the table above, several possible groups can be formed. Here are some possible group combinations along with their respective MAPE values:

| **No** | **Model** | **MAPE** |
|---|---|---|
| **1** | Trend (1,5)<br>Seasonal 1 (6,9)<br>Seasonal 2 (10,13) | 14,99 |
| **2** | Trend (1,5) | 14,61 |

| | | |
|---|---|---|
| | Seasonal 1 (6,9)<br>Seasonal 2 (10,13)<br>Seasonal 3 (15,17) | |
| **3** | Trend (1,5)<br>Seasonal 1 (6,9)<br>Seasonal 2 (10,13)<br>Seasonal 3 (15,17)<br>Seasonal 4 (21,28) | 14,47 |

In the diagonal averaging stage, the groups that have been formed are rearranged into a new series. The results are obtained:

| **No** | **Actual** | **Reconstruction** | | | | | **Diagonal Averaging** | **Residual** |
|---|---|---|---|---|---|---|---|---|
| | | **Trend** | **Seasonal 1** | **Seasonal 2** | **Seasonal 3** | **Seasonal 4** | | |
| 1 | 334556 | 367328.3 | 19664.15 | 28259.5 | -864.27 | 1290.48 | 415678.14 | -81122.14 |
| 2 | 259457 | 362377.9 | 3556.45 | -21060.56 | -377.18 | 1828.45 | 346325.01 | -86868.0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 59 | 143606 | 197106.6 | -2078.20 | -7074.18 | -16511.41 | 1130.5 | 172573.34 | -28967.3 |
| 60 | 161457 | 207478.2 | 9338.58 | 374.17 | 19016.38 | -682.33 | 235525.01 | -74068 |

**SSA Reccurent Forecasting**

In this research, using the SSA recurrent forecasting type, the LRF coefficient was obtained as L-1 = 27, as follows:

| No | LRF Coefficient |
|---|---|
| 1 | -2,7108 |
| 2 | 0,7851 |
| ⋮ | ⋮ |
| 27 | -3,4222 |

The LRF coefficient is used to create the forecasting model as follows:

$$g_i = \sum_{j=1}^{L-1} r_j g_{i-j} = -2{,}7108(g_{i-1}) + 0{,}7851(g_{i-2}) + \ldots + (-3{,}4222)(g_{i-27})$$

**Forecasting Accuracy**

Forecasting accuracy using MAPE and the tracking signal is based on out-of-sample data and the forecast, which we obtain as follows:

| No | Outsample $D_t$ | Forecast $F_t$ | Error | $\sum(D_t - F_t)$ | MAD | Tracking Signal | MAPE |
|---|---|---|---|---|---|---|---|
| 61 | 225304 | 202061.47 | 23242.53 | 23242.53 | 23242.53 | 1.00 | 14,47% |
| 62 | 137536 | 177627.30 | -40091.30 | -16848.77 | 43288.18 | -0.39 | |
| 63 | 177348 | 169834.02 | 7513.982 | -9334.79 | 45792.84 | -0.20 | |
| 64 | 166129 | 164388.92 | 1740.081 | -7594.71 | 46227.86 | -0.16 | |
| 65 | 245423 | 171692.88 | 73730.12 | 66135.42 | 60973.88 | 1.08 | |
| 66 | 217171 | 178129.89 | 39041.11 | 105176.53 | 67480.73 | 1.56 | |
| 67 | 230623 | 176567.50 | 54055.50 | 159232.03 | 75202.95 | 2.12 | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 68 | 193350 | 173908.35 | 19441.65 | 178673.67 | 77633.15 | 2.30 | |
| 69 | 179483 | 167684.67 | 11798.33 | 190472.01 | 78944.08 | 2.41 | |
| 70 | 189447 | 165912.17 | 23534.83 | 214006.83 | 81297.56 | 2.63 | |
| 71 | 188598 | 164156.32 | 24441.69 | 238448.52 | 83519.53 | 2.86 | |
| 72 | 198739 | 168029.09 | 30709.91 | 269158.43 | 86078.69 | 3.13 | |

It is known that the tracking signal values for the 12 forecasted time periods are within acceptable tolerance limits. This shows that the forecasting model can be used to forecast $M$ future time periods. Based on the MAPE criteria, the forecasting model created is good. The results of the air passenger forecast for January through July 2023 are as follows:

| Januari | Februari | Maret | April | Mei | Juni | Juli |
|---|---|---|---|---|---|---|
| 155468 | 123775 | 89814 | 53539 | 23821 | 4828 | 245 |

## 4 Conclusion

From the whole process, it can be concluded that the best SSA model is the model with L = 28 and groups = 5. The SSA method is quite good for extracting time series data based on its constituent components. The prediction of the number of airplane passengers from Kualanamu Airport for seven months in 2023 with SSA using R-forecasting shows that the number of passengers in 2023 will relatively not experience a significant increase. With forecasting models $g_i = \sum_{j=1}^{L-1} r_j g_{i-j} = -2{,}7108(g_{i-1}) + 0{,}7851(g_{i-2}) + \ldots + (-3{,}4222)(g_{i-27})$.

## References


[1] A.Shlemov, N. Golyandina, D. Holloway and A.Spirov Shaped 3D Singular Spectrum Analysis for Quantifying Gene Expression, with Application to the Early Zebrafish Embryo,Biomed Research International,v.2015,p.1-18,(2015).

[2] Abraham B, Ledolter J, et al. Statistical Methods for Forecasting: John Willey and Sons, New York (1983).

[3] Bougas C. Forecasting air passenger traffic flows in Canada:An evaluation of time series

models and combination methods. Quebec : Laval University, 2013.

[4] C. D. Lewis, Industrial and business forecasting methods: A practical guide to exponential smoothing and curve fitting. Butterworth-Heinemann (1982).

[5] Golyandina N and A.Zhigljavsky, Singular Spectrum Analysis for Time Series. Springer,Heidelberg, (2013).

[6] Golyandina N, V.Nekritkuin and A.Zhigljavsky, Analysis of Time Series Structure,SSA and Related Techniques, Chapman & HALL /CRC (2001).

[7] Gumgum Darmawan, et al, Forecasting of Internet Usage by Singular Spectrum Analysis with Trend Extraction Method. Cite as: AIP Conference Proceedings 2192, 090002; https://doi.org/10.1063/1.5139172. Published Online: 19 December 2019

[8] Hassani H, Heravi S, Brown G, et al. Forecasting before, during, and after recession with singular spectrum analysis. Journal of Applied Statistics, 2013, 40(10):2290-2302

[9] Hassani H and Mahmoudvand R. Singular Spectrum Analysis Using R Macmillan Publishers Ltd. London (2018).

[10] Liang Xiaozhen, et al. An integrated forecasting model for air passenger traffic in China based on singular spectrum analysis. Xitong Gongcheng Lilun yu Shijian/System Engineering Theory and Practice, 2017, 37(6):1479-1488

[11] Sun, Y.; Zhang, G.; Yin, H. Passenger flow prediction of subway transfer stations based on nonparametric regression model. Discret. Dyn. Nat. Soc. 2014, 1–8.

[12] Tsui W, Balli H, Gilbey A, et al. Forecasting of Hong Kong airport's passenger throughput. Tourism Management, 2014, 42(6):62-76.

[13] Wei Zhou,et al, Passenger Flow Forecasting in Metro Transfer Station Based on the Combination of Singular Spectrum Analysis and AdaBoost-Weighted Extreme Learning Machine. Sensors, 2020, 20 : 3555

[14] Xiao Y, Liu J J, Hu Y, et al. A neuro-fuzzy combination model based on singular spectrum analysis for air transport demand forecasting[J]. Journal of Air Transport Management, 2014, 39(7):1-11.

[15] Badan Pusat Statistik, www.bps.go.id