Forecasting of jabodetabek train passengers using singular spectrum analysis and holt-winters methods

D Safitri¹,², Subanari¹, H Utami¹, W Sulandari¹

¹Department of Mathematics, Universitas Gadjah Mada, Indonesia
²Department of Statistics, Universitas Diponegoro, Indonesia
³Study Program of Statistics, Universitas Sebelas Maret, Indonesia

Corresponding author: diharma@live.undip.ac.id, diharma@live.ugm.ac.id

Abstract. Train is one of the convenient transportation, it also reduce traffic jam in Jabodetabek. In order to maintain the convenience of passengers using train services, it is necessary to predict the number of train passengers as a consideration in the planning of Jabodetabek train services. In this study the Singular Spectrum Analysis (SSA) and Holt-Winters methods are used to predict the number of Jabodetabek train passengers. SSA is a powerful technique for nonparametric time series analysis and forecasting, which decomposes the original time series into the sum of a small number of independent and interpretable components such as a slowly varying trend, oscillatory components and noise. Holt-Winters method are suitable methods for data that contains trends and seasonality. The Holt-Winters method is based on three smoothing equations, one for level, one for trend, and one for seasonality. SSA method in forecasting the number of Jabodetabek train passengers give lower Mean Absolute Percentage Error (MAPE) value than Holt-Winters method. This means SSA method is better than Holt-Winters method in forecasting the number of Jabodetabek train passengers.

1. Introduction
Jabodetabek (Jakarta-Bogor-Dekop-Tangerang-Bekasi) is commonly abbreviated as Jabodetabek. It is the most populous metropolitan area in Indonesia. Many people live and work in Jabodetabek.

Train has a great demand by the public especially in the Jabodetabek as daily transportation to go to work or just visiting a place in Jabodetabek. Train is the convenient transportation and it can reduce the traffic jam in Jabodetabek. To maintain the convenience of passengers, it is necessary to predict the number of train passengers as a consideration in planning the Jabodetabek train service.

The monthly number of Jabodetabek train passengers exhibits trend and seasonal patterns. Efendi [1] has implemented Seasonal Autoregressive Integrated Moving Average (SARIMA) method to model the data. Other literature, such as Hanke [2] and Sulandari et. al. [3] discussed that Holt Winter exponential smoothing and Singular Spectrum Analysis (SSA) successfully modeled this kind of series, respectively. Gamberini, et. al. [4] did forecasting of sporadic demand patterns with seasonality and trend components and make an empirical comparison between Holt-Winters and SARIMA methods.
The Singular Spectrum Analysis (SSA) is a powerful technique for nonparametric time series analysis and forecasting. SSA decomposes the original time series into the sum of a small number of independent and interpretable components such as a slowly varying trend, oscillatory components and noise [5]. In Hassani [6], SSA is very useful tool which can be used to find trends of different resolution, smoothing, extraction of seasonality components, simultaneous extraction of cycles with small and large periods, extraction of periodicities with varying amplitudes, simultaneous extraction of complex trends and periodicities, finds structure in short time series, and changes point detection.

In this paper, we discuss the implementation of SSA and Holt Winter method to the monthly number of train passenger forecasting. We then compare the accuracy forecasting performance between the two methods, based on the Mean Absolute Percentage Error (MAPE).

2. Literature Review

Hassani [6] analyzed methodology of singular spectrum analysis in monthly accidental deaths in the usa and compared with another methodology. Hassani, et. al. [7] wrote about forecasting European industrial production with singular spectrum analysis. Khan and Poskitt [8] wrote about window length selection and signal-noise separation and reconstruction in singular spectrum analysis. Hassani, et. al. [9] analyzed separability and window length in the singular spectrum analysis, Hassani, et. al. [10] analyzed separability between signal and noise in singular spectrum analysis. Khan and Poskitt [8] wrote about theory and practice of singular spectrum analysis forecasting.

In SSA there are two steps that should be done, which are decomposition and reconstruction. In the decomposition step there are two stages, embedding and Singular Value Decomposition (SVD), in the reconstruction step also has two stages which are grouping and diagonal average [11].

a. Embedding

In Golyandina et. al. [11] and Golyandina & Zhigljavsky [12], time series \( F = (f_0, ..., f_{N-1}) \) has a length of \( N \), with \( N > 2 \). The embedding procedure maps the time series to multidimensional lag vectors. L notation for window length is an integer, with \( L \leq \frac{N}{2} \). Embedding forms lag vectors \( K = N - L + 1 \).

\[
X_i = (f_{i-1}, ..., f_{i+L-2})^T, 1 \leq i \leq K
\]

Which has L dimensions, trajectory matrix of F series:

\[
X = [X_1 : \ldots : X_K]
\]

Which has a lag vector in each column or it can be written:

\[
X = \begin{bmatrix}
  f_0 & f_1 & f_2 & \cdots & f_{K-1} \\
  f_1 & f_2 & f_3 & \cdots & f_K \\
  f_2 & f_3 & f_4 & \cdots & f_{K+1} \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  f_{L-1} & f_L & f_{L+1} & \cdots & f_{N-1}
\end{bmatrix}_{L \times K}
\]

b. Singular Value Decomposition (SVD)

The Singular Value Decomposition (SVD) of the X trajectory matrix can be written as:

\[
X = X_1 + X_2 + \ldots + X_d
= U_1 \sqrt{\lambda_1} V_1^T + U_2 \sqrt{\lambda_2} V_2^T + \ldots + U_d \sqrt{\lambda_d} V_d^T
\]

The eigenvector \( U_1 \), singular value \( \sqrt{\lambda_1} \), and principal component \( V_1^T \) are called eigentriple from SVD.

The reconstruction step also has two stages which are grouping and diagonal averaging [11]:

a. Grouping
After Equation (1) is obtained, the grouping procedure partitions the set of indices \{1, \ldots, d\} into m disjoint subsets \(I_i, \ldots, I_m\). Let \(I = \{i_1, \ldots, i_p\}\), then the resultant matrix \(X_i\) corresponding to the group \(I\) is defined as:

\[X_i = X_{i_1} + \ldots + X_{i_p}\]

These matrices are computed for \(I=I_i, \ldots, I_m\) and Equation (1) leading to decomposition:

\[X = X_{i_1} + \ldots + X_{i_m}\] (2)

The procedure of choosing the sets \(I_i, \ldots, I_m\) is called the eigentriple grouping.

b. Diagonal Averaging

The last step in SSA transforms each matrix of the group decomposition (2) into a new series of length \(N\). Let \(Y\) be a matrix of \(L \times K\) size with elements \(y_{ij}\), \(1 \leq i \leq L, 1 \leq j \leq K\). With \(L^* = \min(L, K)\), \(K^* = \max(L, K)\) and \(N = L \times K - 1\), \(y_{ij}^* = y_{ij}\) if \(L < K\) and \(y_{ij}^* = y_{ij}\) otherwise.

Diagonal averaging transfers the \(Y\) matrix to the series \(g_0, \ldots, g_{N-1}\) with the formula:

\[
g_k = \begin{cases} 
\frac{1}{k+1} \sum_{m=1}^{k+1} y_{mk+m+2}^* & \text{for } 0 \leq k < L^* - 1 \\
\frac{1}{L^*} \sum_{m=1}^{L^*} y_{mk+m+2}^* & \text{for } L^* - 1 \leq k < K^* \\
\frac{1}{N-K^*+1} \sum_{m=K^*}^{N-K^*+1} y_{mk+m+2}^* & \text{for } K^* \leq k < N 
\end{cases}\] (3)

Diagonal averaging (3) applied to the resultant matrix \(X_{i_k}\), produces a series \(\tilde{F}^{(k)} = (\tilde{f}_0^{(k)}, \ldots, \tilde{f}_{N-1}^{(k)})\) and therefore the initial series \(f_0, \ldots, f_{M-1}\) is decomposed into the sum of \(m\) series

\[f_n = \sum_{k=1}^{m} \tilde{f}_n^{(k)}\]

In Golyandina and Zhigljavsky (2013), there are two forecasting methods in SSA which are the recurrent and the vector forecasting.

a. Algorithm of recurrent forecasting:

1. Time series \(G_{N+M} = (g_0, \ldots, g_{N+M-1})\) defined as:

\[g_i = \begin{cases} 
\tilde{f}_j & \text{for } i = 0, \ldots, N-1 \\
\sum_{j=1}^{L-1} a_{ij} g_{i-j} & \text{for } i = N, \ldots, N+M-1 
\end{cases}\]

2. The numbers \(g_N, \ldots, g_{N+M-1}\) form the \(M\) terms of the recurrent forecast.

b. Algorithm of vector forecasting:

The algorithm of vector method SSA forecasting:

1. Vector \(Z\) defined as:

\[Z_i = \tilde{X}_i \text{ for } i = 1, \ldots, K\]

\[P^{(i)}Z_{i-1} \text{ for } i = K+1, \ldots, K+M+L-1\]

2. By constructing the matrix \(Z = \left[Z_i; \ldots; Z_{K+M+L-1}\right]\) and making its diagonal averaging we obtain the series \(y_0, \ldots, y_{N+M+L-1}\).

3. The numbers \(g_0, \ldots, g_{N+M-1}\) form the \(M\) terms of the vector forecast.

Holt Winters method is based on three smoothing equations, one for the level, one for trend, and one for seasonality. There are two different Holt Winters methods, depending on whether seasonality is modeled in additive or multiplicative way [13].
The basic equations for Holt Winters additive method are as follows [13]:

Level: \[ L_t = \alpha(Y_t - S_{t-s}) + (1-\alpha)(L_{t-1} + b_{t-1}) \]

Trend: \[ b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1} \]

Seasonal: \[ S_t = \gamma(Y_t - L_t) + (1-\gamma)S_{t-s} \]

Forecast: \[ F_{t+h} = L_t + b_t h + S_{t+h} \]

Mean Absolute Percentage Error (MAPE) is the average of the absolute percentage errors of forecasts. Error is defined as actual or observed value minus the forecasted value. This measure is easy to understand because it provides the error in terms of percentages. MAPE has managerial appeal and is a measure commonly used in forecasting. The lower the MAPE the better the forecast [14].

3. Methods

3.1. Data source and research variables
The data used is secondary data which is monthly data on the number of Jabodetabek train passengers (in thousand people). This research uses the data from January 2012 to December 2016 as insample data and data from January 2017 to December 2017 as outsample data. The data was taken from the Badan Pusat Statistik website https://www.bps.go.id . Data processing in this paper using R software.

3.2. Data analysis method
The steps of data analysis:
1. Using SSA method to forecast the number of Jabodetabek train passengers
   a. Embedding.
   b. Singular Value Decomposition (SVD).
   c. Grouping.
   d. Doing the reconstruction using the diagonal averaging method.
   e. Forecasting using SSA method.
   f. Calculate MAPE value.
2. Using Holt-Winters method to forecast the number of Jabodetabek train passengers.
   a. Calculate the level, trend, and seasonal equations.
   b. Forecasting using Holt-Winters Method.
   c. Calculate MAPE value.
3. Compare the accuracy performance between SSA and Holt-Winters methods in forecasting the number of Jabodetabek train passengers based on MAPE value.

4. Results and discussions
The data used is monthly data on the number of Jabodetabek train passengers (in thousand people). The length of the data is 72, then the data is divide into two parts, from January 2012 to December 2016 as insample data and from January 2017 to December 2017 as outsample data. The length of insample data is 60 and the length of outsample data is 12. The time series plot of the data can be seen in Figure 1 which shows that the data contains trend and seasonal components.
Figure 1. Plot Time Series for the number of Jabodetabek train passengers (in thousand people) data

At the embedding step, a trajectory matrix with $X$ notation $L \times K$ window length denoted $L$ is an important parameter in the SSA with the term $L \leq \frac{N}{2}$, in this research $N=60$ and obtained $L \leq 30$.

Determination of $L$ is done by trial and error, tried for some value and then selected $L$ value with the lowest value of Mean Absolute Percentage Error (MAPE) because in Encyclopedia of Production and Manufacturing Management (2000), the lower the MAPE the better the forecast. Window length used is $L = 6$ because it has the lowest MAPE which is 3.92% with $N =$ number of data insample =60 and $K=N-L+1=60$.

W-correlation measure can be calculated for insample data to evaluate appropriateness of gruping [15]. The correlation between one component with another component can be seen in the W-correlation matrix, the stronger the correlation between components, the darker the color. W-correlation matrix for the number of Jabodetabek train passengers with $L=6$ in Figure 2, can be seen that the first component is signal extractor and the remaining components are noise.

Figure 2. W-correlation matrix for the number of Jabodetabek train passengers with $L=6$
In the results of forecasting outsample data using SSA method are listed in Table 1 and obtained a MAPE value of 3.92% which can be interpreted to have high forecasting accuracy. The results of the reconstruction can be seen in Figure 3, it consists of original series, trend, and noise components.

![Reconstructed Series](image)

**Figure 3.** Reconstructed original series (top), trend (middle), and noise (bottom).

The time series plot in Figure 1 shows that the data contains trends and additive seasonal components therefore the method used to make forecasting is the Holt-Winters Additive and obtained:

| Parameter | Value         |
|-----------|---------------|
| Alpha     | 0.4513347     |
| Beta      | 0.01818096    |
| Gamma     | 1             |

The models obtained are as follows:

Level: \( L_t = 0.4513347Y_t - S_{t, s} + (1 - 0.4513347)(L_{t-1} + b_{t-1}) \)

Trend: \( b_t = 0.01818096(Y_t - L_{t-1}) + (1 - 0.01818096)b_{t-1} \)

Seasonal: \( S_t = (Y_t - L_t) \)

Forecast: \( F_{t+h} = L_t + b_t h + S_{t+h} \)

Forecasting using Holt Winters method from January to December 2017 are listed in Table 1 and the actual and forecasting graph for the number of Jabodetabek train passengers can be seen in Figure 4 and it obtained a MAPE value of 4.02% which can be interpreted to have high forecasting accuracy.

| Outsample data (actual value) | Forecasting using SSA method | Forecasting using Holt- Winters method |
|-------------------------------|------------------------------|--------------------------------------|
| 1 24185                       | 25062.36                     | 23820.81                             |
| 2 21743                       | 25347.87                     | 22623.92                             |
Table 1. Data and MAPE value (2009-2016)

|   | Actual | Forecasting Using SSA Method | Forecasting Using Holt-Winters Method |
|---|--------|-------------------------------|-------------------------------------|
| 3 | 25775  | 25648.82                      | 25024.18                           |
| 4 | 25411  | 25965.45                      | 24931.72                           |
| 5 | 27385  | 26312.25                      | 25747.63                           |
| 6 | 24432  | 26679.29                      | 25139.21                           |
| 7 | 27016  | 27015.90                      | 23682.5                            |
| 8 | 27679  | 27362.89                      | 25968.44                           |
| 9 | 26158  | 27719.25                      | 25888.19                           |
| 10| 28765  | 28083.71                      | 27031.13                           |
| 11| 28246  | 28451.89                      | 26616.83                           |
| 12| 29059  | 28820.49                      | 27319.51                           |
|   | MAPE value | 3.92%                      | 4.02%                             |

**Figure 4.** Actual and forecasting graph for the number of Jabodetabek train passengers using SSA and Holt-Winters methods.

Figure 4 is actual data and forecasting graph for the number of Jabodetabek train passengers using SSA and Holt-Winters methods where the black plot for actual data, red plot for forecasting using SSA, and blue plot for forecasting using Holt-Winters method. Figure 4 shows that SSA method is better than Holt-Winters method in forecasting the number of Jabodetabek train passengers.

5. Conclusion
Singular Spectrum Analysis method in forecasting the number of Jabodetabek train passengers give lower MAPE value than Holt-Winters method. This means for forecasting the number of Jabodetabek train passengers, SSA method is better than Holt-Winters method. In the future research, we will compare the accuracy between SSA method and ARCH/GARCH method in forecasting the data with volatility.

**References**

[1] Efendi S R 2017 Analisis Peramalan Jumlah Penumpang Kereta Api dengan Metode SARIMA (Studi Kasus Jumlah Penumpang Kereta Api P.T. KAI Commuter Jabodetabek dan P.T. Kereta Api Indonesia Periode Januari 2009-Desember 2016)

[2] Hanke J E and Wichern D 2014 *Business Forecasting* Pearson New International
[3] Sulandari W, Subanan, Suhartono, Utami H 2017 proc of 3rd International Conference on Science in Information Technology (ICSITech) IEEE

[4] Gamberini R, Lolli F, Rimini B, Sgarbossa F 2010 Hindawi Publishing Corporation Mathematical Problems in Engineering Volume 2010

[5] Hassani H 2007 Journal of Data Science 5 239-257

[6] Hassani H, Heravi S, Zhigljavsky A 2009 International Journal of Forecasting 25 103-118

[7] Hassani H, Mahmoudvand R, Zokaei M 2011 Comptes Rendus Mathematiques 349 987-990

[8] Khan M A R and Poskitt D S 2011 Window Length Selection and Signal-Noise Separation and Reconstruction in Singular Spectrum Analysis, Working Paper 23/11, Department of Econometrics and Business Statistics Monash University Australia.

[9] Hassani H, Mahmoudvand R, Zokaei M, Ghodsi M 2012 Fluctuation and Noise Letters 11(2) 1-11

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

[11] Golyandina N, Nekrutkin V, Zhigljavsky A A 2001 Analysis of Time Series Structure : SSA and Related Techniques, Chapman & Hall/CRC

[12] Golyandina N, and Zhigljavsky A 2013 Singular Spectrum Analysis for Time Series Springer

[13] Makridakis S G and Wheelwright S C, Hyndman R J 1997 Forecasting Methods and Applications Wiley

[14] Encyclopedia of Production and Manufacturing Management 2000 Mean Absolute Percentage Error (MAPE) Springer

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

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