Simulation study of singular spectrum analysis from time series data with outlier

G Darmawan¹,²*, D Rosadi² and B N Ruchjana¹

¹Universitas Padjadjaran, Bandung-Sumedang km 21 Jatinangor, West Java, Indonesia
²Universitas Gadjah Mada, Jln.Bulak Sumur,Yogjakarta, Indonesia

*gumgum@unpad.ac.id.

Abstract. In this simulation study, Singular Spectrum Analysis (SSA) will be used to analyze time series data with outlier. Time series data with outliers will be generated by open source software R (OSSR) for many replications. The goal of this research is to evaluate robustness of the SSA model. SSA analyzes both types of time series data, with outlier and without outlier. Performance of SSA will be evaluated by changing of SSA parameter (Window length) and other parameter is determined by automatic grouping. Moreover, the MAPE (mean absolute percentage error) of SSA is measured for both types of data. The result of Simulation study shows the larger the data the smaller the effect of the outliers. However, the larger the data the greater the shift in the value of L. The accuracy of SSA will decrease by 0.245 (MAPE) if there is available outlier in Moving Average (MA) data. For Autoregressive data, the accuracy of SSA will decrease by 0.264 (MAPE) if there is available single outlier in time series data.

1. Introduction
SSA is relatively new tool in time series modeling. At the beginning, SSA is used to decompose time series data, however most of SSA is used to forecast time series data. Nowadays, SSA is applied for many purposes for example Imputation time series data, linearity test, filtering and structural change detection.

SSA also has been applied for many disciplines such as Astronomy, Physics, Climatology, hydrology, economic and finance. The SSA technique has been used in a variety of fields such as, nonlinear dynamics [1], climate [2], medicine [3], mathematical statistics [4] and trading [5]. One of the interesting topic in SSA is robustness. In SSA forecasting, vector method is more robust then recurrent method [6]. A new robust approach for SSA procedure had been suggested by [7]. Moreover, the robustness of the SSA modeling comparison to the other forecasting techniques evaluated in [8]. A new algorithm Factorized Hankel Optimal Singular Spectrum Analysis (FHSSA) also had been proposed by [9], the result of this method is comparatively faster and robust to the existing classical SSA method.

In this research, we evaluate robustness of SSA by adding outlier in time series data. Robustness of SSA is measured by alternating MAPE and L (window length) values in clean data and data with outlier. Single outlier is put in the middle of time series generated data with values 3 to types of data (Autoregressive and Moving Average).
2. Methodology

In this session, we describe step of simulation methodology in SSA with outlier time series. The step of methodology can be seen in the figure 1.

- **Step 1**: Time series Data are generated by R with two models ARIMA (1,1,0) and ARIMA(0,1,1), these models are nonstationary models. We chose values of autoregressive and moving average 0.4 with $N = 100, 125, 150, 175, 200, 225$ and $250$ with 1000 replication. The detail of models can be seen in Simulation study below.
- **Step 2**: Generated data are split into two series, training and testing. Training series data are used for SSA modelling and testing data are used for evaluating the performance of SSA. The value of testing series called validation is 12.
- **Step 3**: Add single outlier to the middle of data, the value of outlier here 10000. If the number of generated data 100, we change data in position 50 by 10000.
- **Step 4**: Use SSA as a model for both clean and data with outlier. In all generated data, we use the same SSA procedure from Embedding, Singular value decomposition, auto grouping and forecasting.
- **Step 5**: Compute MAPE, SD and L, three statistic values of this simulation study are used to evaluated the robustness of SSA. MAPE value is measurement of model accuracy (SSA) to data. SD values is standard deviation of MAPE values of 1000 replication. The last, L is window length of SSA model of generated data.
- **Step 6**: Compare the three values (MAPE, SD and L) between clean data and data with outlier. From 1000 generated clean data and data with outlier, we have 1000 MAPE, SD and L values respectively. After getting these values (MAPE) we calculated the mean of MAPE of both types of data(clean data and data with outlier). The last Mean MAPE value of clean data are subtracted by MAPE value of data with outlier. The process of determining L is also calculated from both types of data. The L values of clean data were subtracted by L values of data with outlier.

![Figure 1. Flowchart of methodology.](image)
3. Simulation study

In this session, we generate two models, ARIMA (1,1,0) and ARIMA (0,1,1). The mathematics models are as follows;

\[ z_t = 1.4z_{t-1} - 0.4z_{t-2} + a_t \]  \hspace{1cm} (1)

\[ z_t = z_{t-1} - 0.4a_{t-1} + a_t \]  \hspace{1cm} (2)

Data were generated for various N = (100,125,150,175,200,225,250) with 1,000 replication. We used Autoregressive and moving autoregressive parameter 0.4. We used R.3.4.3 for running this simulation with package Rssa.

The simulation of time series data by SSA modeling had been done by many researchers [10], [11],[12],[13] and [14]. Many of these studies were done in multidimensional Singular Spectrum Analysis (MSSA). However, they use spectral equation instead of ARIMA models.

Output for these simulations are Mean, Standard deviation and L. Mean is average of MAPE values between data without outlier and data with outlier of 1,000 data generated. Same as Mean, Sd is standard deviation of MAPE values between data without outlier and data with outlier of 1,000 data generated. L is window length value of subtraction between data without outlier and data with outlier of 1,000 data generated.

4. Result and discussion

Performance of SSA is measured by MAPE, this criteria has been used by many researchers [15],[16] and [17] in simulation results. Result of simulation study is shown in the table below;

| N       | AR Model | MA Model |
|---------|----------|----------|
|         | Mean     | Sd       | L   | Mean     | Sd       | L   |
| 100     | 0.264    | 0.375    | 11  | 0.245    | 0.323    | 11  |
| 125     | 0.168    | 0.206    | 14  | 0.179    | 0.219    | 14  |
| 150     | 0.135    | 0.168    | 17  | 0.133    | 0.167    | 18  |
| 175     | 0.112    | 0.140    | 19  | 0.113    | 0.138    | 20  |
| 200     | 0.098    | 0.123    | 25  | 0.102    | 0.134    | 24  |
| 225     | 0.089    | 0.122    | 30  | 0.093    | 0.117    | 31  |
| 250     | 0.077    | 0.097    | 34  | 0.080    | 0.097    | 34  |

In the first column, the number of sample that had been generated, the values are from 100 to 250. The three next columns are result of AR simulated model. The difference between MAPE value of clean data and data with outlier for N=100 is 0.264. It means that if the data have single outlier in the middle of data (N=100), the accuracy of SSA model (MAPE) will change with value 0.264 and standard deviation 0.375. The larger the data the smaller the effect of the outliers. Moreover, the difference between L value of clean data and data with outlier for N=100 is 11. It means that if the data have single outlier in the middle of data (N=100), the value of L will change with value 11. The larger the data the greater the shift in the value of L.

The three last columns are result of MA simulated model. The difference between MAPE value of clean data and data with outlier for N=100 is 0.245 (the first value). It means that if the data have single outlier in the middle of data (N=100), the accuracy of SSA model (MAPE) will change with value 0.245 and standard deviation 0.323. The larger the data the smaller the effect of the outliers. Moreover, the difference between L value of clean data and data with outlier for N=100 is 11. It means that if the data have single outlier in the middle of data (N=100), the value of L will change with value 11. The larger the data the greater the shift in the value of L.
5. Conclusion
According to session before, the larger the data the smaller the effect of the outliers. That effect was happened not only for Autoregressive model but also for Moving average Model. However, effect of outlier to window length (L) value, the larger the data the greater the shift in the value of L both for autoregressive and for moving average models.

References
[1] Vautard R and Ghil M 1989 Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series Phys. D Nonlinear Phenom. 35 395–424
[2] Raynaud S, Yiou P, Kleeman R and Speich S 2015 Using MSSA to determine explicitly the oscillatory dynamics of weakly nonlinear climate systems Nonlinear Process. Geophys 12 807–815
[3] Hassani Hossein and Sanei S 2016 Singular spectrum analysis of biomedical signals (CRC Press)
[4] Moskvina V and Zhigljavsky A 2003 An Algorithm Based on Singular Spectrum Analysis for Change Point Detection Commun. Stat. Simul. Comput. 32 319–352
[5] Sitohang Y O and Darmawan G 2017 The accuracy comparison between ARFIMA and singular spectrum analysis for forecasting the sales volume of motorcycle in Indonesia in AIP Conference Proceedings 1868
[6] Golyandina N A Z 2003 Singular spectrum analysis
[7] Sanei S, Ghodsi M and Hassani H 2011 An adaptive singular spectrum analysis approach to murmur detection from heart sounds Med. Eng. Phys. 33 362–367
[8] Silva E S and Hassani H 2015 On the use of singular spectrum analysis for forecasting U.S. trade before, during and after the 2008 recession Int. Econ. 141 34–49
[9] Tiwari R K and Rajesh R 2014 Factorized Hankel optimal singular spectral approach for erratic and noisy seismic signal denoising J. Appl. Geophys. 111 95–101
[10] Hassani H 2010 Singular spectrum analysis based on the minimum variance estimator Nonlinear Anal. Real World Appl. 11 2065–2077
[11] Hassani H, Xu Z and Zhigljavsky A 2011 Singular spectrum analysis based on the perturbation theory Nonlinear Anal. Real World Appl. 12 2752–2766
[12] Alexandrov T and Golyandina N 2005 Automatic extraction and forecast of time series cyclic components within the framework of SSA Proc. St. Petersbg. Work. Simul. 45–50
[13] Golyandina N, Korobeynikov A, Shlemov A and Usevich K 2013 Multivariate and 2D Extensions of Singular Spectrum Analysis with the Rssa Package arXiv preprint arXiv 67
[14] Shlemov A and Golyandina N E 2014 Shaped extensions of singular spectrum analysis arXiv Prepr. arXiv1401.4980 1813–1820
[15] Gao Y, Qu C and Zhang K 2016 A hybrid method based on singular spectrum analysis, firefly algorithm, and BP neural network for short-term wind speed forecasting Energies 9
[16] Ghodsi Z, Silva E S and Hassani H 2015 Bicoid Signal Extraction with a Selection of Parametric and Nonparametric Signal Processing Techniques Genomics Proteomics Bioinforma 13 183–191
[17] Hassani H and Ghodsi Z 2014 Pattern Recognition of Gene Expression with Singular Spectrum Analysis Med. Sci. 2 127–139