Performance of Bootstrap Method in Singular Spectrum Analysis on Forecasting Rainfall Data

Gumgum Darmawan[1], Dedi Rosadi[2] and Budi Nurani R[3]
{gumgum@unpad.ac.id[1]}

[1,3]Padjadjaran University, Jatinangor, Sumedang, Indonesia
[1,2]Department of Mathematics, Gadjah Mada University, Jogjakarta, Indonesia

Abstract: SSA forecasting based on bootstrap is used to forecast rainfall data. The execution of the strategy is inspected as far as an assortment grouping and Window length. Application to original data are directed to research the precision of this method in contrast with other prediction methods i.e linear recurrent (LRF) and vector forecasting method. After running computation by software R, The bootstrap forecasting method released the best result then any other methods i.e LRF and vector method.

Keywords: bootstrap forecasting, rainfall, singular Spectrum Analysis (SSA)

1. Introduction

Singular spectrum analysis is a relatively new tool in time series analysis that flexible to use but accurate enough in forecasting. There are several books devoted to SSA(Elsner and Tsonis 1996; N. Golyandina and Osipov 2007; Nina Golyandina et al. 2013; Hassani Hossein; Sanei 2016) as well as many papers related to theory of SSA and especially to various applications of SSA. The scope of application SSA is sufficient wide, from biomedics to astronomy and from structural change to filtering.

SSA has been utilized effectively in a few regions like hydrology, geophysics, climatology (Vautard and Ghil 1989); economy ((de Carvalho, Rodrigues, and Rua 2012);Medical Enginering (Sanei, Ghodsi, and Hassani 2011), Reliability analysis (Rocco S 2013), trading (Sasmita and Darmawan 2017) among others. One of the principal favoured point of view of SSA stood out from parametric techniques.

According to Singular Spectrum Analysis literature, numerous analysists look for to discover determining strategies in SSA can create more noteworthy forecast precision. See example, (Nina Golyandina and Stepanov 2005) considered MSSA (multivariate Singular Spectrum Analysis) forecast, their experiment showed that accuracy of forecast globally corresponds to accuracy of reconstruction. (Hassani, Soofi, and Zhigljavsky 2010) had bootstrapped the original series data to reduce noise and improve forecasting accuracy. Furthermore, (Rodrigues and de Carvalho 2013) had proved that recurrent forecast method could be applied for time series data with missing value. Moreover, (Gao, Qu, and Zhang 2016), proposed two combination techniques in SSA forecasting, neural network and firefly algorithm. Furthermore, (Hassani, Kalantari, and Yarmohammadi 2017) proposed algorithm of filtered LRF (Linear Recurrent Formula) prediction.

We interested to SSA forecasting based on Bootstrap method because of accuracy. In (Nina Golyandina et al. 2013) , bootstrap forecast based SSA was described for exploring the accuracy of forecast interval (FI). Furthermore, in (Nina Golyandina et al. 2013) bootstrap FI was connected for some time series like GDP (Gross Domestic Product) and Manufacturing Product. In this paper, the precision of bootstrap prediction SSA will be examined using time

ICASI 2019, July 18-19, Banda Aceh, Indonesia

DOI 10.4108/eai.18-7-2019.2288596

series. Moreover, forecasting results are compared to linear recurrent and vector forecasting methods.

The remainder of the paper is composed as take after: Section 2 contains the general description of the Basic SSA from embedding to diagonal averaging step. Section 3 reports the empirical results of forecasting comparison among three methods: Recurrent, Vector and Bootstrap forecasting and the paper and with some conclusion in section 4.

2. Singular Spectrum Analysis

SSA used in this paper is basic SSA from Embedding to Diagonal averaging. Detail of information about SSA is available in (Nina Golyandina et al. 2013).

2.1 Decomposition

An important characteristic of the SSA decomposition is the way that the reconstructed time series always satisfies some linear recurrent formula (LRF) which means that the series can be expressed as a linear combination of results of polynomials, sinusoids and exponential functions.

2.1.1 Embedding

Embedding is the first stage in SSA, this step is to change univariate time series data to multivariate data. Multivariate data is formulated in a matrix, called trajectory matrix. The dimension of matrix is determined by Window length (L). This window length is an integer value and the only one SSA parameter in the first step. Moreover, researcher usually use value of L Between 2 to N/2. N is length of time series data Dimension of trajectory matrix is LxK, where K=N-L+1. Element of this matrix can be formulated in equation (1),

$$X_i = \left(x_i, ..., x_{i+L-1}\right)^T \quad \left(1 \le i \le K\right) \qquad (1)$$

That vectors from i to k, called lagged vectors. Then, the trajectory matrix of that vector can be seen in equation (2). The best value of *length window* can be tracked from 2 to N/2.

$$X = \left[X_1 : X_2 : ... : X_K\right] = \left(x_{ij}\right)_{i,j=1}^{L,K} = \begin{pmatrix} x_1 & x_2 & x_3 & ... & x_K \\ x_2 & x_3 & x_4 & ... & x_{K+1} \\ x_3 & x_4 & x_5 & ... & x_{K+2} \\ . & . & . & ... & . \\ x_L & x_{L+1} & x_{L+2} & ... & x_N \end{pmatrix} \quad (2)$$

2.1.2 Singular Value Decomposition

Singular Value Decomposition is the second step in SSA. In this step, trajectory matrix is multiplied by transpose of trajectory matrix $S = XX^T$. S has dimension LxL.

From matrix S, we determine the value of eigenvalue $\lambda_1, \lambda_2, ..., \lambda_d$,eigen vector V_i and the corresponding eigenvectors (U_i). The three values are called eigentriple $\left(\sqrt{\lambda_i}, U_i, V_i\right)$. Eigenvalues used in the analysis is non-zero values, where d is the highest value of the eigenvalue. If $i=1,2,,...d$ then a sum of matrices $X = X_1 + X_2 + ... + X_d$.

2.2 Reconstruction

The point of the SSA reconstruction step is to get an estimate for the signal components.

2.2.1 Grouping

Grouping is a step in SSA that can be explained by figure. In grouping, the same pattern of eigenvectors is grouped as trend, seasonal with many periods and noise.

$$X = X_{I_1} + ... + X_{I_m} \qquad (3)$$

The procedure of choosing the sets $I_1,...,I_m$ is called the eigentriple grouping. For given group I the contribution of the component X_I into the expansion **X** is measured by the share of the corresponding eigenvalues: $\sum_{i \in I} \lambda_i \Big/ \sum_{i=1}^{d} \lambda_i$.

2.2.2 Diagonal Averaging

Diagonal Averaging is the last step in SSA. In this step, multivariate data in grouping step back to univariate data. This process usually called *hankelization*. The equation used in diagonal averaging is equation (4).

Let Y be a an LxK matrix with elements y_{ij} , $1 \le i \le L, 1 \le j \le K$. Set $L^* = \min(L,K), K^* = \min(L,K)$ and $N=L+K-1$. Let $y_{ij}^* = y_{ij}$ if $L<K$ and $y_{ij}^* = y_{ij}$ otherwise. By making the *diagonal averaging* we transfer the matrix **Y** into the series $y_1, y_2, ..., y_N$ using the formula;

$$y_k = \begin{cases} \dfrac{1}{k}\sum_{m=1}^{k} y_{m,k-m+1}^* & 1 \le k < L \\ \dfrac{1}{L^*}\sum_{m=1}^{L^*} y_{m,k-m+1}^* & L^* \le k \le K^* \\ \dfrac{1}{N-k+1}\sum_{m=1}^{L^*} y_{m,k-m+1}^* & K^* < k \le N \end{cases} \qquad (4)$$

equation (4) is used to an addition matrix X_{I_k} produces a reconstructed series $\tilde{X}^{(k)} = \left(\tilde{x}_1^{(k)}, ..., \tilde{x}_N^{(k)}\right)$. Therefore, the initial series $x_1,...,x_N$ is decomposed into sum of m reconstructed series:

$$x_n = \sum_{k=1}^{m} \tilde{x}_n^{(k)} \qquad (5)$$

The reconstructed series produced by the elementary grouping will be called *elementary reconstructed series*.

2.3 Forecasting

Actually, there are three methods in forecasting, LRF, Vector and Bootstrap method. Here we focus on the last method as new forecasting method.

2.3.1 The SSA forecasting algorithm

Recurrent forecasting algorithm (briefly, R-forecasting) can be formulated as follows. Algorithm (R-forecasting):

1. The time series $Y_{N+M} = \left(y_1, y_2, ..., y_{N+M} \right)$ is characterized by

$$y_i = \begin{cases} \tilde{x}_i & i = 1, ..., N, \\ \sum_{j=1}^{L-1} a_j y_{i-j} & i = N+1, ..., N+M. \end{cases}$$

2. The numbers $y_{N+1}, ..., y_{N+M}$ form the M terms of the recurrent forecast.
   Thus, the R-forecasting is performed by the direct use of the LRF with coefficients $\left\{ a_j, j = 1, 2, ..., L-1 \right\}$

2.3.2 Bootstrap SSA

Bootstrap forecasting SSA proposed by (Rahmani 2014), In the process of SSA, there is a representation $\hat{Y} = \hat{S}_N + \hat{E}_N$. Here a stochastic model is assumed for the residual series $\hat{E}_N$. By means of bootstrap technique B independent "copies" $E_{N,i} \left(i = 1, 2, ...B \right)$ of $\hat{E}_N$ are generated from the presumed model. Step of SSA-bootstrap algorithm is as follow :

Step 1: for instance time series as $\hat{Y} = \hat{S}_N + \hat{E}_N$.

Step 2: Sample with replacement form $\hat{E}_N$ and denote it as $\hat{E}_{N,i}$ and calculated $Y^* = \hat{S}_N + \hat{E}_{N,i}$.

Step 3: Use SSA algorithm on the generated time series as Y^*.

Step 4: (Bootstrapping) Repeat the steps 2 and 3 for B times.

Bootstrap LRF coefficient is considered via following algorithm:

Step 1: Repeat step 1 to 4 of bootstrap SSA.

Step 2: Either compute the mean of the coefficients

$\bar{A} = \frac{\sum_{i=1}^{n} A_i}{n}$ or compute the 5% and 95% percentile of $A_{i..}$

2.4 Evaluation of Forecast Accuracy

The forecasting precision is assessed with MAPE criteria. The differences between the observed and predicted values are measured by utilizing forms more sensitive to significant over or underprediction such as MAPE value.

$$MAPE = \left[\frac{1}{N_v} \sum_{i=1}^{N_v} \left| x_i - \hat{x}_i \right| / x_i \right] x100\% \quad (6)$$

Where N_v is the validation data set size, x_i is the i-th observed value. $\hat{x}_i$ is the i-th forecasted value.

2.5 Flowchart

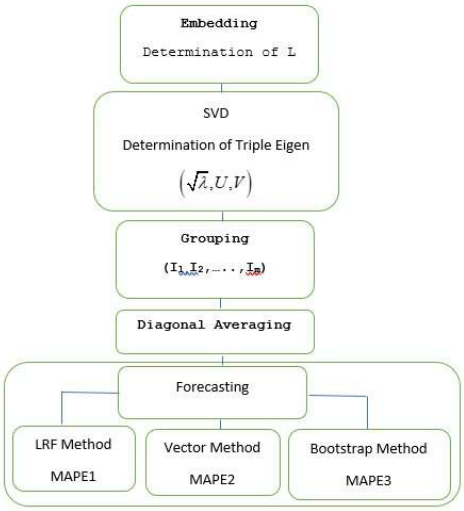

Fig. 1. Flowchart of Analysis.

According to that figure, the process starts from Embedding to forecast step. In forecast stage, we used three methods LRF, Vector and Bootstrap. All results are evaluated by MAPE values.

3. Results And Discussion

3.1 Data

The atmosphere data utilized as a part of this investigation is monthly precipitation time series data. These time series data are from Aceh, Indonesia from March 2000 to December 2017, portrayed in Fig 2.

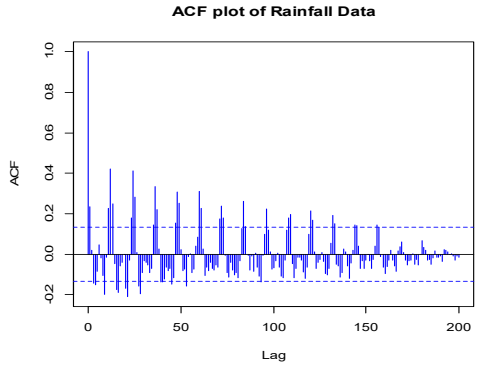

Fig. 2. Autocorrelation Function Plot of Rainfall Data.

The graph shows seasonal pattern from beginning to end of lag. So, monthly time series data from Aceh have seasonal pattern with period s. The value of period s can be identified by clicking every peak of line, rainfall time series data usually have period 12. To compute the rainfall data, we use software R with many packages. Packages R for this data are Rssa, tseries, svd and forecast. The main package is Rssa, this package operates R program from embedding to forecasting step. Complete R program is available in 2.5.

3.2 Forecasting Result

In order to assessed the performance of bootstrap forecasting method in comparison with LRF and vector technique, data were separated into two groups. The primary group is training data set and the second group is testing data set (three months). From testing data, we computed MAPE values of three forecasting methods.

Not all of the tracking L are showed here, the best window length is L=101 and r=4, trend, season 1, season 2 and season 3. The best values of forecasting result are printed red. All of The best tenth values are showed in table 1.

Table 1. Tracking Result SSA forecasting.

No	Grouping				Mape		
	Trend	Season 1	Season 2	Season 3	Boostrap	Reccurent	Vector
1	1	2,3,5,6	7,8,9	10,11,13	12.96	17.46	21.18
2	1	2,3,5,6	7,8	10,11,13	16.35	18.04	23.07
3	1	2,3,5,6	7,8	9, 10,11,13	14.32	19.01	23.16
4	1	2,3,5	6,7,8,9	10,11,13	12.66	15.59	13.15
5	1	2,3	5,6,7,8,9	10,11,13	9.89	15.38	13.10
6	1	2,3,4	5,6,7,8,9	10,11,13	13.21	16.29	12.83
7	1	2,3	5,6,7,8,9	10,11,12,13	12.75	16.86	15.08
8	1	2,3	4, 5,6,7,8,9	10,11,13	12.32	15.49	12.62
9	1	2	3, 5,6,7,8,9	10,11,13	12.98	13.85	12.76
10	1	2,3	5,6,7,8	9, 10,11,13	11.93	16.83	15.72

According to table 1, Recurrent method has the smallest MAPE 13.85% in grouping No 9 but at that grouping both bootstrap and vector methods have smaller value than recurrent MAPE. Vector forecasting method has the smallest value also in grouping no.9 with MAPE 12.76% . Finally, the smallest value of all tracking is MAPE 9.89%, this value come from bootstrap forecasting method with grouping no.5. Despite the successful forecasting of bootstrap method compared to LRF and vector method, three real data ahead will be forecasted by bootstrap forecasting method.

Table 2. Three Months Forecast Ahead.

Month	Forecasts	Lower Bound	Upper Bound
January	261.19	132.52	400.66
February	248.50	248.50	384.55
March	230.49	230.49	361.28

Forecast values in table 2 are forecasting results for real time series data. Lower and Upper bound of forecast are included in this table to see the maximum and minimum values with alpha =5%.

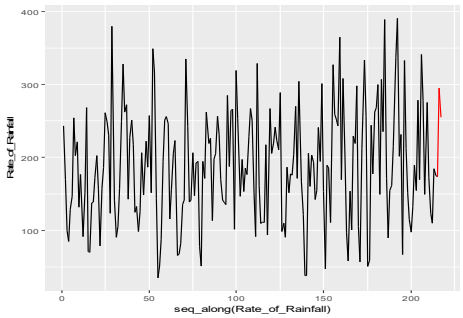

Fig. 3. Plot of Data and 3 Months Forecast Ahead.

According to figure 3, the values of the forecast results had been acceptable because it appeared among other real time series data.

4. Conclusion

Our work has led us to conclude that the bootstrap forecasting method can be successfully utilized as a prediction precipitation data. According to this result (especially in rainfall time series data), Bootstrap forecasting method is the best then any other methods (recurrent and vector method).

References

[1] de Carvalho, Miguel, Paulo C. Rodrigues, and António Rua. 2012. "Tracking the US Business Cycle with a Singular Spectrum Analysis." *Economics Letters* 114(1): 32–35. http://dx.doi.org/10.1016/j.econlet.2011.09.007.

[2] Elsner, J.B., and A.A. Tsonis. 1996. *Singular Spectrum Analysis: A New Tool in Time Series Analysis*. https://books.google.com.br/books/about/Singular_Spectrum_Analysis.html?id=pHsGF9WIBxkC&pgis=1.

[3] Gao, Yuyang, Chao Qu, and Kequan Zhang. 2016. "A Hybrid Method Based on Singular Spectrum Analysis, Firefly Algorithm, and BP Neural Network for Short-Term Wind Speed Forecasting." *Energies* 9(10).

[4] Golyandina, N., and E. Osipov. 2007. "The 'Caterpillar'-SSA Method for Analysis of Time Series with Missing Values." *Journal of Statistical Planning and Inference* 137(8): 2642–53.

[5] Golyandina, Nina, and D. Stepanov. 2005. "SSA-Based Approaches to Analysis and Forecast of Multidimensional Time Series." *Proceedings of the 5th St.Petersburg Workshop on Simulation* (September): 293–98.

[6] Golyandina, Nina, Anatoly Zhigljavsky, Golyandina N., and Zhigljavsky A. 2013. *Singular Spectrum Analysis for Time Series*. http://www.springerlink.com/index/10.1007/978-3-642-34913-3.

[7] Hassani, Hossein, Mahdi Kalantari, and Masoud Yarmohammadi. 2017. "Un Algorithme de Prévision SSA Amélioré Reposant Sur Des Séries Filtrées." *Comptes Rendus Mathematique* 355(9): 1026–36. http://dx.doi.org/10.1016/j.crma.2017.09.004.

[8] Hassani, Hossein, Abdol S. Soofi, and Anatoly A. Zhigljavsky. 2010. “Predicting Daily Exchange Rate with Singular Spectrum Analysis.” *Nonlinear Analysis: Real World Applications* 11(3): 2023–34. http://dx.doi.org/10.1016/j.nonrwa.2009.05.008.

[9] Hassani Hossein; Sanei, Saeid. 2016. *Singular Spectrum Analysis of Biomedical Signals*. http://gen.lib.rus.ec/book/index.php?md5=4836ce2af7f156bc7987ea7da4c98fcc.

[10] Rahmani, Donya. 2014. “A Forecasting Algorithm for Singular Spectrum Analysis Based on Bootstrap Linear Recurrent Formula Coefficients.” *International Journal of Energy and Statistics* 2(4): 287–99.

[11] Rocco S, Claudio M. 2013. “Singular Spectrum Analysis and Forecasting of Failure Time Series.” *Reliability Engineering and System Safety* 114(1): 126–36. http://dx.doi.org/10.1016/j.ress.2013.01.007.

[12] Rodrigues, Paulo C., and Miguel de Carvalho. 2013. “Spectral Modeling of Time Series with Missing Data.” *Applied Mathematical Modelling* 37(7): 4676–84. http://dx.doi.org/10.1016/j.apm.2012.09.040.

[13] Sanei, Saeid, Mansoureh Ghodsi, and Hossein Hassani. 2011. “An Adaptive Singular Spectrum Analysis Approach to Murmur Detection from Heart Sounds.” *Medical Engineering and Physics* 33(3): 362–67. http://dx.doi.org/10.1016/j.medengphy.2010.11.004.

[14] Sasmita, Y., and G. Darmawan. 2017. “Accuracy Evaluation of Fourier Series Analysis and Singular Spectrum Analysis for Predicting the Volume of Motorcycle Sales in Indonesia.” In *AIP Conference Proceedings*,.

[15] Vautard, R, and Michael Ghil. 1989. “Singular Spectrum Analysis in Nonlinear Dynamics, with Applications to Paleoclimatic Time Series.” *Physica D: Nonlinear Phenomena* 35(3): 395–424. http://linkinghub.elsevier.com/retrieve/pii/0167278989900778.