Forecasting of Rainfall in Sumatera Barat: Singular Spectrum Analysis (SSA) Application

F Fitri1*, R Rahmat2, A D Pengestuti3

1Mathematics Department, Universitas Negeri Padang, Padang, Indonesia
2Climate Early Warning Sub Division, Indonesia Agency for Meteorology, Climatology and Geophysics (BMKG), Jalan Angkasa 1 Jakarta, 10610, Indonesia
3Chemistry Department, Universitas Negeri Padang, Padang, Indonesia

*corresponding author: fadhilahfitri@fmipa.unp.ac.id

Abstract. West Sumatra has 2 peak of rainy seasons, October-November and March to May. The high intensity of rainfall causes West Sumatra has potential to floods. Precise forecasting is needed to be a reference for the government. One of the method can be used is SSA. This method is flexible because it does not require specific form of time series data, as well as parametric assumptions. Thus, accurate forecasting results is expected to be provided from the SSA method. From the forecasting process that has been done, obtained MAPE of 17% with the tracking signal value within the tolerance range.

1. Introduction
Rain is related to the water cycle on earth. Based on rainfall classification zone [1], Indonesia can be devided into 3 zone: A, B and C. Zone A includes the south and central of Indonesia: the south part of Sumatera to Timor Island, some part of Kalimantan, some part of Sulawesi and some part of Irian Jaya. Zone B located in northwest of Indonesia and includes the north part of Sumatera and northwest of Kalimantan. While Zone C region covers Maluku and some part of Sulawesi, as can be seen on Figure 1.

Figure 1. Rainfall Classification Zone, Source:[2]
One of the provinces in zone B is Sumatera Barat. Zone B is the only region which has 2 peak rainy seasons, October-November and March to May. The highest rainfall intensity for Zone B is 320 mm/month [1]. The high intensity of rainfall caused West Sumatra to be a disaster-prone area. Therefore, it is necessary to do a forecast to find out how the rainfall prediction, thus the government can determine the right policy.

Forecasting is the process of predicting conditions of observation in the future using its historical value. In forecasting methods, data patterns are very important because there are several methods that are only good for certain patterns. There are 4 data patterns which are stationary / horizontal, trend, seasonal and cycle. In their applications, time series data patterns often contain a combination of all four patterns [3]. Hence, an approach that no longer requires the specification of time series data is developed, called the non-parametric approach.

One method of non-parametric approach is Singular Spectrum Analysis (SSA). SSA is a flexible method because in its application, this method does not require model specifications from time series data, as well as parametric assumptions, hence it is expected that the SSA method can provide accurate forecasting results. The advantage of this method lies in the stage where the SSA decomposes the time series data series into special components, then reshapes the data set without involving noise components from the data. In addition, the SSA can work well even though the data has a small sample size [4]. Therefore, SSA will be used in this study. Thus, the formulation of the problem in this study is "How is the application of the SSA method in predicting rainfall intensity in West Sumatra and how accurate are the estimated results?".

2. Singular Spectrum Analysis (SSA)
SSA is a method which involves the analysis of classical time series data, multivariate statistics, multivariate geometry, dynamic systems, and signal processing [5]. This method is used to solve various problems, such as finding trends from different resolutions, smoothing, extracting various time series components, finding the structure of short-term time series, etc [6]. SSA method requires no assumptions in model formation. Beside that, it also does not require stationary data and can be used to analyze short-term periods data [5]. Furthermore, it does not require model specifications on time series data, meaning that no hypothesis testing is done in this method. However, the data used must have a seasonal effect [7].

2.1. Decomposition
2.1.1. Embedding
Embedding is the stage where the initial time series data is converted into trajectory matrix, which means changing the initial data in the form of one-dimensional data into multidimensional data. The trajectory matrix has $L \times K$ dimensions, and L is a window length to be a matrix line, while $K = N - L + 1$ becomes a matrix column. The selection range for the value of L is $2 < L < \frac{N}{2}$, and the assumption that time series data throughout the N period do not contain missing data, and $X = \{x_i; i = 1, 2, \ldots, N\}$ [5].

Trajectory matrix ($T_x$) formed is the Hankel matrix. This matrix is a matrix where all elements along the diagonal $i + j$ are constant. It can be written as follows:

$$ T_x = (T_{i,j})_{L \times K} = \begin{bmatrix} x_1 & x_2 & \ldots & x_K \\ x_2 & x_3 & \ldots & x_{K+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & \ldots & x_N \end{bmatrix} $$ (1)
2.1.2. Singular Value Decomposition
Singular Value Decomposition aims to obtain the separation of components in decomposition from time series data. Singular Value Decomposition in its application has similarities with Principal Component Analysis, which is to reduce the component of the initial data and reduce dimensions. Singular Value Decomposition starts by determining eigenvalue \((\lambda_1, \lambda_2, \ldots, \lambda_L)\) from S matrix \(S = T_x T_x^T\) where \(\lambda_i \geq \lambda_j > 0\), and eigenvector \((U_1, U_2, \ldots, U_L)\) from these S matrix. Singular Value Decomposition on Trajectory Matrix \(T_x\) will produce:

\[
T_u = U_i D V_i
\]  

where: \(U_i\) is a \(K \times L\) orthonormal matrix
\(D\) is a \(L\)-order diagonal matrix
\(V_i\) is a \(L\) square orthonormal matrix

so that it produces:

\[
T_u = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^T
\]  

with \(i = 1, 2, \ldots, d\) and \(d = \max \{i\} ; \lambda_i > 0\)

The three components in the matrix \(T_u\): singular value \((\sqrt{\lambda_i})\), eigenvector \((U_i)\), and principal component \((V_i)\) called i-eigentriple of Singular Value Decomposition. Next, it can be written as:

\[
T_u = T_{u1} + T_{u2} + \ldots + T_{ud}
\]

2.2. Reconstruction

2.2.1. Grouping
Grouping is the stage of separating additive component such as trend, seasonal and noise contained in time series data. The grouping process is done by grouping index \(1, 2, \ldots, d\) sets into \(m\) subset that can be symbolized by \(I = I_1, I_2, \ldots, I_m\). Then forming a matrix based on the Singular Value Decomposition for the trajectory matrix \(T_u\) as follows:

\[
T_{i_k} = T_{i1} + T_{i2} + \ldots + T_{im}
\]

The step for selecting sets \(I = I_1, I_2, \ldots, I_m\) is called eigentriple grouping.

2.2.2. Diagonal Averaging
Diagonal averaging is the last step in SSA. At this stage, the reconstruction of each matrix in \(T_{i_k}\) matrix becomes a new time series data with length \(N\). Let matrix \(Y = L \times K\); \(y_{ij}, 1 \leq i \leq L, 1 \leq j \leq K\). The Y matrix is converted to a time series \(g_0, \ldots, g_{N-1}\) through a diagonal averaging which is

\[
g_k = \begin{cases} 
\frac{1}{k+1} \sum_{m=1}^{k+1} y_{m,k-m+2} & ; 0 \leq k < L' - 1 \\
\frac{L'}{L} \sum_{m=1}^{L'} y_{m,k-m+2} & ; L' - 1 \leq k < K' \\
\frac{1}{N-k-K'+2} \sum_{m=k-K'+2}^{N-k} y_{m,k-m+2} & ; K' \leq k < N
\end{cases}
\]
With: \( L' = \min(L, K) \); \( K^* = \max(L, K) \)
\[
N = L + K - 1
\]
\[
y^*_j = \begin{cases} 
 y_j & : L < K \\
 y_{ji} & : \text{other}
\end{cases}
\]
g\(_k\) is the average of the matrix elements along the diagonal \( i + j = k + 2 \). \( k = 0 \) will produce \( g_0 = y_{11} \), \( k = 1 \) and \( g_1 = \frac{(y_{11} + y_{22})}{2} \) and so on.

2.3. Forecasting
R-Forecasting is related to Linier Recurrent Formula (LRF) estimation: \( a_1, \ldots, a_d \) which is the eigenvectors obtained from SVD step. Let \( \mathbf{U} = (u_1, u_2, \ldots, u_{L-1}, u_L)^T \), \( \mathbf{U}^Y = (u_1, u_2, \ldots, u_{L-1})^T \) and \( \pi_i \) is the last component of eigenvectors \( \mathbf{U} \) or can be written as \( \pi_i = u_L \) the the LRF coefficient can be calculated using equations:
\[
9R = \mathbf{U}^T \mathbf{a} = ((a_{L-1}, u_{L-2}, \ldots, a_2, u_1)^T = \frac{1}{1-v^2} \sum_{i=1}^{r} \pi_i U_i^Y
\]
with \( v^2 = \sum \pi_i^2 \)

The time series data used in R-forecasting is the result of reconstruction obtained from diagonal averaging. The next step is to determine the M-new points to be predicted. So that the series of forecast results can be shown by:
\[
G_{N+M} = (g_1, \ldots, g_{N+M})
\]
Forecasting results obtained based on:
\[
g_k = \begin{cases} 
 y_i & : i = 0,1, \ldots, N \\
 \sum_{j=1}^{k-1} a_i g_{i-j} & : i = N + 1, \ldots, N + M
\end{cases}
\]

2.4. Accuracy of Forecasting
2.4.1. Mean Absolute Percentage Error (MAPE)
MAPE will be used to measure the accuracy of the forecasting result in this research. The formulation is as follows:
\[
\text{MAPE} = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{X_i - F_i}{X_i} \right| \cdot 100\%
\]
with:
\( X_i \): actual time series data
\( F_i \): forecast data

2.4.2. Tracking Signal
Tracking signal is a measure of tolerance that can be used to determine the possibility of whether the results of forecasting can be used. The tolerance limit of the acceptable signal tracking value is ± 5 [8].
3. Methodology

3.1. Data
Data on “Rainfall in West Sumatra from January 2001 to December 2018” used in this study were obtained from "Stasiun Meteorologi Minangkabau", BMKG.

3.2. The Calculation Steps and Result

3.2.1. Determine the Window Length (L)
Windows Length (L) is a dimension of the trajectory matrix of data. L determination is still subjective because there are no rules regarding the optimum L [9]. Trial and error method used to select L value, which is choosing L that produces the smallest MAPE.

3.2.2. Perform the decomposition process: embedding stages
Embedding is the step where the initial one-dimensional data change into multidimensional by forming trajectory matrix. The chosen L is 50, N is 216, thus \( K = N - L + 1 = 216 - 50 + 1 = 167 \). Then the dimension of the trajectory matrix is \( 50 \times 167 \), the matrix as follows:

\[
T_i = (T_{ij})_{50 \times 167} = \begin{bmatrix}
271.8 & 309.7 & \ldots & 300.3 \\
427.8 & 271.8 & \ldots & 442.6 \\
\vdots & \vdots & \ddots & \vdots \\
238.2 & 537.5 & \ldots & 642.3
\end{bmatrix}
\]

3.2.3. Continued with singular value decomposition stage
Eigentriple will be obtained at this stage using the trajectory matrix from the previous step. The steps to find the eigentriple:

a. Singular value (\( \sqrt{\lambda_i} \))
The eigenvalue and singular value are obtained as follows:

| No | Eigenvalues | Singular Values |
|----|-------------|-----------------|
| 1  | 1015549017  | 31867.67982     |
| 2  | 16701444.96 | 4086.740138     |
| \vdots | \vdots | \vdots |
| 50 | 1351154.823 | 1162.391854     |

b. Eigenvector (U)_i
From the calculation results obtained eigenvector as can be seen on Table 2:
Table 2. Eigenvectors

| No | \( U_1 \) | \( U_2 \) | ... | \( U_{50} \) |
|----|------------|------------|------|------------|
| 1  | -0.14191  | 0.189053   | ... | -0.08763   |
| 2  | -0.14193  | 0.220971   | ... | 0.140985   |
| ...| ...        | ...        | ... | ...        |
| 50 | -0.1404   | 0.149554   | ... | -0.13319   |

Then, each existing eigenvector value will be formed as a graph as shown in Figure 2 bellow:

Figure 2. Eigenvector Graph

Figure 2 shows that each eigenvector forms certain patterns. This shows that rainfall data in West Sumatra Province from 2001 - 2018 involved several additive components. Therefore, to facilitate the identification of cyclical and seasonal patterns that are formed, two-dimensional graphs of paired eigenvectors should be formed.

Figure 3. Pairs of Eigenvector graph
Based on the two-dimensional paired eigenvector graph in Figure 3 above, the first eigenvector has a horizontal pattern and is made as the first group, but for the next group it has an irregular pattern so the next grouping is done by checking the period for each eigenvector.

c. Principal Component (\(V_i\))
The principal component values obtained will be representative of the whole process of decomposition that has been done. This is based on the value of each component so that it can be used at the grouping stage.

| Table 3. Principal Component Value |
|-----------------------------------|
| No  | \(V_1\)  | \(V_2\)  | ... | \(V_{50}\) |
|-----|---------|---------|-----|---------|
| 1   | -2455.27| -80.88  | ... | 110.5406|
| 2   | -2466.38| -313.661| ... | -112.69 |
| ⋮   | ⋮      | ⋮     | ⋮ | ⋮ |
| 167 | -2419.09| 142.4343| ... | -50.2556|

3.2.4. Conducting a reconstruction process: grouping stages
Data is grouped into five groups based on existing patterns, \(m = 5\).

3.2.5. Continue the reconstruction process: diagonal averaging stages
At this stage, series data will be formed with one dimension and with length \(N\) based on five groups. Diagonal averaging can be calculated by adding up the results of reconstruction for each group.

| Table 4. Reconstruction Results |
|--------------------------------|
| Actual Data | Reconstruction Results | Diagonal Averaging |
| Pattern 1 | Pattern 2 | Pattern 3 | Pattern 4 | Pattern 5 | Pattern 5 |
| 238.2 | 348.4382 | 73.80348 | -17.4195 | -101.363 | -9.30687 | 294.1520003 |
| 537.5 | 349.2424 | 10.1112 | 3.422686 | 49.61304 | 41.49608 | 453.8854356 |
| 244.8 | 345.9599 | -63.8366 | 13.64615 | 7.298127 | -40.5649 | 262.5027475 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 300.3 | 339.6319 | 117.0428 | -50.0413 | -12.6773 | 62.88401 | 456.840117 |

Figure 4 below shows the plot between the actual data and the diagonal average.
3.2.6. Perform the R-forecastings stage

The data from reconstruction at the diagonal average step will be used in the R-forecasting stage. Table 6 below contains the LRF coefficient

| Table 6. Linear Recurrent Formula (LRF) Coefficients |
|----------------------------------|
| Coefficient |        |
| \( \alpha_1 \) | 0.114338 |
| \( \alpha_2 \) | 0.08575  |
| \( \vdots \) | \( \vdots \) |
| \( \alpha_{49} \) | 0.147666 |

The next step is to predict rainfall in West Sumatra for January 2019 to June 2019. The results of the forecast can be seen in Table 7.

| Table 7. Forecast Results for Outsample Data |
|---------------------------------------------|
| Period  | Actual Data | Forecast |
| Jan-19  | 335.2       | 365.3548 |
| Feb-19  | 333.3       | 400.0669 |
| Mar-19  | 487.2       | 397.8222 |
| Apr-19  | 235.8       | 334.4716 |
| May-19  | 208.9       | 221.6775 |
| Jun-19  | 223         | 207.8432 |

3.2.7. Calculate the accuracy of forecast

The forecasting result is good to use if the MAPE in between 10%-20% \([10]\). The MAPE obtained in this study is 17%. After that, calculated the tracking signal. The tolerance limit for the tracking signal value is \( \pm 5 \) \([8]\). Table 7 below shows the tracking signal for each period.

| Table 8. Tracking Signal Result |
|--------------------------------|
| Period  | Actual Data | Forecast | Tracking Signal |
| Jan-19  | 335.2       | 365.3548 | -1              |
| Feb-19  | 333.3       | 400.0669 | -2              |
| Mar-19  | 487.2       | 397.8222 | -0.12148        |
| Apr-19  | 235.8       | 334.4716 | -1.4909         |
| May-19  | 208.9       | 221.6775 | -1.99821        |
| Jun-19  | 223         | 207.8432 | -1.99107        |

From the MAPE and the tracking signal as can be seen in Table 8, we can conclude that the forecasting method by using SSA with a window length of 50 is good to forecast the rainfall in Sumatera Barat.

3.2.8. Forecasting result

The results of rainfall forecasting in Sumatera Barat from July 2019 to November 2019 using the SSA method are as seen on Table 9.
Table 9. Forecast Results for Outsample Data

| Period | Forecast |
|--------|----------|
| Jul-19 | 335.2993 |
| Aug-19 | 394.8990 |
| Sep-19 | 320.2594 |
| Oct-19 | 294.1171 |
| Nov-19 | 306.5552 |

4. Conclusion

The forecasting process for rainfall in West Sumatra has been done using the SSA method. Accuracy of forecast results is calculated using MAPE and Tracking Signal. The MAPE of the forecast for the SSA method is 17% with the tracking signal value within the tolerance range. Based on the MAPE and Tracking Signal values, it can be said that the SSA method is suitable for future forecasting. The rainfall forecasting result for July were 335.2993 mm, August 394.8990 mm, September 320.2594 mm, October 294.1171 mm and November 306.5552 mm.

References

[1] Aldrian E and Susanto RD 2003 Identification of three dominant rainfall regions within Indonesia and their relationship to sea surface temperature *International Journal of Climatology* **23** 1435-53

[2] Aldrian E, Karimin M, and Budiman 2011 Adaptasi dan Mitigasi Perubahan Iklim di Indonesia Available from: https://www.researchgate.net/publication/309721670_Adaptasi_dan_Mitigasi_Perubahan_Iklim_di_Indonesia [accessed Mar 05 2019].

[3] Makridakis S, Wheelright SC 1999 *Forecasting: Methods and Applications Second Edition* (New York: John Wiley & Sons)

[4] Hassani H 2007 Singular Spectrum Analysis: Methodology and Comparison Cardiff University and Central Bank of Islamic Republic of Iran

[5] Golyandina N, Nekrutkin V, and Zhigljavsky A 2001 *Analysis of Time Series Structure: SSA and Related Technique* (London: Chapman & Hall)

[6] Zhigljavsky A 2010 Singular spectrum analysis for time series: introduction to this special issue *Statistics and Its Interface* **3**

[7] Fitri F, Gamayanti NF and Darmawan G 2017 Metode SSA pada data produksi perikanan tangkap di provinsi jawa barat *Jurnal Ilmiah Matematika dan Pendidikan Matematika* **9** 95-110

[8] Bovas and Ledolter 1983 *Statistical Method for Forecasting* (New York: John Wiley and Son Inc)

[9] Amdayani NJ 2016 Penentuan Window Length (L) Terbaik pada Metode Singular Spectrum Analysis(SSA) *Skripsi* Program Studi Statistika Universitas Padjadjaran

[10] Lewis C D 1982 *International and Business Forecasting Methods* (London: Butterworths)