Modified singular spectrum analysis in identifying rainfall trend over peninsular malaysia

S.M. Shaharudin[1], N. Ahmad[2], N.H. Zainuddin[3]
[1,3]Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Malaysia
[2]Department of Mathematics, Faculty of Science, Universiti Teknologi Malaysia, Malaysia

Article Info

Article history:

Received Dec 17, 2018
Revised Feb 18, 2019
Accepted Maret 1, 2019

Keywords:

Iterative oblique singular spectrum analysis
Robust sparse K-means
Singular spectrum analysis
Trend

ABSTRACT

Identifying the local time scale of the torrential rainfall pattern through Singular Spectrum Analysis (SSA) is useful to separate the trend and noise components. However, SSA poses two main issues which are torrential rainfall time series data have coinciding singular values and the leading components from eigenvector obtained from the decomposing time series matrix are usually assesed by graphical inference lacking in a specific statistical measure. In consequences to both issues, the extracted trend from SSA tended to flatten out and did not show any distinct pattern. This problem was approached in two ways. First, an Iterative Oblique SSA (Iterative O-SSA) was presented to make adjustment to the singular values data. Second, a measure was introduced to group the decomposed eigenvector based on Robust Sparse K-means (RSK-Means). As the results, the extracted trend using modification of SSA appeared to fit the original time series and looked more flexible compared to SSA.

Corresponding Author:

S.M. Shaharudin,
Department of Mathematics,
Faculty of Science and Mathematics,
Universiti Pendidikan Sultan Idris, 35900 Tanjung Malim, Perak, Malaysia.
Email: shazlyn@fsmt.upsi.edu.my

1. INTRODUCTION

Understanding of rainfall patterns is essential for water resources planning and management especially with evidence of hydrological extreme events and variability in recent years. It represents one of the main potential causes of natural hazards such as floods in certain areas in which it causes economic losses, damage and sometimes, human losses. Thus, identifying local time scale to determine when torrential rainfall events could possibly occur at a particular location and concurrently to detect trend at that time period are critically important. In this aspect, local time scale can be defined as time or length of a process, observation or model in a particular region or area. The term trend based on local time scale can be defined as smooth additive component containing information that requires a predetermined time scale. Trend usually characterizes the shape of time series data.

There have been many research work done on identifying rainfall trend based on time scale [1-3]. Most of these studies were proposed using Kendall correlation coefficient for identify trend in the time series data. However, this method did not perform well since it is not possible to establish which kind of trend that could be extracted by its means [4]. In addition, it did not consider any noise that could potentially compromise the method used. In this study, we aim to identify potential trend that can specify the range of local time scale to determine whether or not there have been any high extreme rainfall values in order to detect the abnormally heavy rainfall that can cause torrential rainfall events. One of the methods in identifying the range of local time scale according to the trend is based on singular spectrum analysis (SSA).

In general, SSA provides a representation of a univariate time series which is transformed in terms of the eigenvalues and eigenvectors of a trajectory matrix. In this study, eigenvalues and eigenvector of this type of time series are referred as eigen time series. SSA is useful to separate the time series data into trend and noise by decomposing its time series eigen and reconstruct them into selection of group.

In torrential rainfall time series data, the daily amount of rainfall is approximately similar over a period of time. This situation leads to a problem when using SSA where there were coinciding singular values. It made disjoint sets of singular values and different series components actually mixed with each other. Another issue that cropped up when using SSA was that leading component from the eigenvector obtained from the decomposing the time series matrix was usually assessed subjectively by graphical inference, lacking a specific statistical measure. Thus, this had potentially affected the separability strength in SSA. In consequence of the issues stated above, the extracted trend from SSA tended to flatten out and did not show any distinct pattern. Therefore, it faced difficulties in determining the range of the local time scale in estimating when the torrential rainfall event occurs.

In order to effectively overcome the above shortcomings, two approaches were introduced in SSA to mitigate these problems. Firstly, an iterative oblique SSA (Iterative O-SSA) was presented which helped to improve the weak separability. This method performed a new decomposition of a part in the SSA which corrected the eigenvalues in order to avoid their possible mixture or disjoint sets of singular values in the group corresponding to different components. Iterative O-SSA found that the separating inner products by iterations were converged to a stationary point. Secondly, a measure was introduced to guide an effective grouping of decomposed eigenvector based on Robust Sparse K- means clustering method (RSK-means). Typically, basic clustering method like k-means method performed rather poorly in the presence of noise. Thus, it was unable to retrieve a very good partition for eigenvector in order to separate the trend and noise components accordingly. Moreover, the sparcity in this method was obtained by assigning weights to components where the optimal weights are positive which were used to determine the cluster.

In Section 2, the rainfall data that used in this study is described. Then, in Section 3, the methodologies related to the SSA and modification of SSA is presented. Next, in Section 4, the results and discussion to identify temporal cluster patterns using SSA and modified SSA are provided. Conclusion is given in the final section.

2. DEVELOPMENT OF MODIFIED SSA APPROACH

Over the years, developments in improving SSA gradually grow. Various studies look at issues such as weak separability between components, high dimensions of the trajectory matrices, change point detection and low-rank matrix approximations. Separability condition is one of the important aspects when using SSA as it determines proper decompositions and extraction of components to be achieved. The main objective of SSA is to determine the separability which entails how well the components of the time series can be separated from each other to allow further analysis to be meaningfully done. [5-7] proved that separability had a great influence in SSA process. [8] proposed two methods which are Iterative Oblique SSA and Nested SSA with derivatives (DerivSSA) in improving the separability conditions. Their methods focus on the inner products corresponding to oblique coordinate systems instead of the conventional Euclidean inner product. Both methods can be considered as refining of the decomposition obtained by SSA approach where they can considerably improve the separability and hence the reconstruction accuracy.

Improvement or replacement of the singular value decomposition (SVD) procedure in decomposition stage is one of the alternatives to extend SSA. There are two reasons SVD method is inappropriate to use in analysis. The first is when the dimension of the trajectory matrix of the data set is so large that it becomes too expensive to compute the data using SVD method. The second is, according to perturbation theory, that SVD seriously degrades the performance of the method. In classical SSA method, standard SVD is based on the least square estimate which has highest possible residual noise level. [9] introduced SSA based on the minimum variance estimation where the estimator gives the minimum total residual power. The difference between these two methods is that by adapting the weights of different singular components, it could separate the signal from noise components. Thus, a small adjustment of the eigenvalues and eigenvectors can counter the problems of standard SVD in SSA approach.

The combination method of SSA and classification is another statistical approach in improving the SSA. The improvement indicates that the procedure successfully eliminates most of the noise present in the signal in an efficient manner. Clustering techniques are widely used and often combined with SSA as shown by [10] in order to replace the traditional digital filtering and spline-based methods. The procedure according to this approach consists of obtain leading components in the frequency domain and remove noise present in the signal in a simple and intuitive way. A latter approach described by [11] used the same approach which combined SSA with neural network. This proposed method has the same goal which is to remove most of the noise in signal components.

3. MATERIAL AND METHOD

In this section, data and method are explained.

3.1. Data

The daily rainfall data from 75 stations over Peninsular Malaysia were obtained from Jabatan Pengairan dan Saliran (JPS). In this study, the focus is on the occurrence of extreme rainfall event described as torrential rainfall. Days which exhibited torrential characterized were strictly selected based on criteria described. It was therefore necessary to choose some criteria that would lead to the establishment of a threshold, in order to allow for a clear distinction between what constitutes a day of torrential rainfall in the Peninsular Malaysia region and what does not. Area with a tropical climate with 60 mm/day is the most common threshold applied for this purpose [12]. The period is based on observed series of daily maximum rainfall starting from 1 November 1975 and ends on 31 December 2007, covering a sequence of 61 consecutive days.

3.2. Singular Spectrum Analysis

Singular spectrum analysis (SSA) is a time series analysis method whush is constructed from several elements of classical time series, multivariate statistics, multivariate geometry, dynamic system and signal processing. The function of using SSA method is to decompose the original components to trend, seasonal and noise components [13]. SSA method consists of two complementary stages which are decomposition and reconstruction stage. In decomposition stage, it has two vital steps which are embedding and singular value decomposition while in reconstruction stage, it has another two main procedures which are grouping and diagonal averaging. Below is a brief discussion on the methodology of SSA method. The steps involved in

Stage 1: Decomposition

There are two steps in the decomposition stage which are embedding and singular value decomposition (SVD). In general, this stage aims to decompose the series to obtain eigen time series data.

Step 1: Embedding. The first step in basic SSA algorithm is to construct the trajectory matrix from a one dimensional series to a multidimensional series whose dimension is called the window length, L. Let a one dimensional time series, $\mathbb{Y}_T = \{y_1, y_2, \dots, y_T\}$ be transformed into a multi-dimensional series $X_1, \dots, X_K$ with vectors $X_i = (y_i, \dots, y_{i+L-1})^T$, where $2 < L < {}^{T}\!/_{2}$ is the window length and $K = T - L + 1$. The trajectory matrix is denoted by $\mathbf{X} = (X_1, \dots, X_K) = (x_{ij})_{i,j=1}^{L,K}$, such that

$$\mathbf{X} = [X_1 : \dots : X_K] \ = (x_{ij})_{i,j=1}^{L,K} = \begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_K \\ x_2 & x_3 & x_4 & \cdots & x_{K+1} \\ x_3 & x_4 & x_5 & \dots & x_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & x_{L+2} & \cdots & x_T \end{pmatrix} \tag{1}$$

The lagged vectors X_i are the columns of the trajectory matrix **X**. The rows and columns of **X** are subseries of the original time series data.

Step 2: Singular Value Decompostion (SVD). SVD transforms the trajectory matrix in step I into a decomposed trajectory matrix which will turn into trend and noise based on their singular values. In this step, SVD of the trajectory matric **X** is performed. The SVD can be represented as

$$\mathbf{X} = U^T \Sigma V \tag{2}$$

where $U = (u_1, \dots, u_L)$ is an $L \times L$ orthogonal matrix, $V = (v_1, \dots, v_k)$ is a $K \times K$ orthogonal matrix and Σ is an $L \times K$ diagonal matrix with nonnegative real diagonal entries $\Sigma_{ii} = \sigma_i$ for $i = 1, \dots, L$. The vectors u_i are called left singular vectors, the v_i are the right singular vectors and the σ_i are the singular values. Let the singular values be arranged in descending order such that ($\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_L$). Then, the SVD of the trajectory matrix X can be written as

$$\mathbf{X} = \mathbf{X}_1 + \dots + \mathbf{X}_{\mathrm{L}} \tag{3}$$

Where $\mathbf{X}_i = \sigma_i \mu_i v_i^T$. Note that, the matrices of $\mathbf{X}_i$ are called elementary matrices if $\mathbf{X}_\mathbf{i}$ has rank one. The collection (σ_i, u_i, v_i) is called the ith eigentriple of the SVD.

Stage 2: Reconstruction

There are two steps in the reconstruction stage which are grouping and diagonal averaging. In general, this stage aims to reconstruct the original series and use the reconstructed series for further analysis such as forecasting.

Step 1: Grouping. In grouping step, the trajectory matrix is split into two groups based on the trend and noise components. The indices set $\{1, \ldots, L\}$ is partitioned into m disjoint subsets $I_1, \ldots, I_m$, corresponding to spliting the elementary matrices into m groups. Set $I = \{i_1, \ldots, i_p\}$, and the resultant matrix $\mathbf{X}_I$ is defined as

$$\mathbf{X}_I = \mathbf{X}_{i1} + \cdots + \mathbf{X}_{ip} \tag{4}$$

The resultant matrices are computed for $I = I_1, \ldots, I_m$ and substituted in (5.3). The expansion is defined as:

$$\mathbf{X} = \mathbf{X}_{I1} + \cdots + \mathbf{X}_{Im} \tag{5}$$

where the trajectory matrix is represented as a sum of m resultant matrices. The choice of the sets $I = I_1, \ldots, I_m$ is known as eigentriple grouping.

Step 2: Diagonal averaging. Diagonal averaging transforms each resultant matrix obtained from (5) into a new one dimensional series of length T.

Let Z be an $L \times K$ matrix with elements $z_{ij}, 1 \le i \le L,\ 1 \le j \le K$. Set $L^* = \min(L, K), K^* = \max(L, K)$ and $N = L + K - 1$. Let $z_{ij}^* = z_{ij}$ if $L < K$ and $z_{ij}^* = z_{ji}$ otherwise. By using (6), matrix $\mathbf{Z}$ will transfer into the series $z_1, \ldots, z_T$

Once the diagonal averaging step in reconstruction stage is completed, it becomes a reconstruction data series $\widetilde{\mathbb{Y}}_T^{(k)}$ with entries $\tilde{y}_T^{(k)}$.

$$z_k \begin{cases} \frac{1}{k}\sum_{m=1}^{k} z_{m,k-m+1}^* & 1 \le k < L^* \\ \frac{1}{L^*}\sum_{m=1}^{L^*} z_{m,k-m+1}^* & L^* \le k \le K^* \\ \frac{1}{T-k+1}\sum_{m=k-k^*+1}^{T-K^*+1} z_{m,k-m+1}^* & K^* < k \le N \end{cases} \tag{6}$$

3.3. Modification of SSA

Large Modification of SSA in this study is the combination of two main strategies which are (i) Iterative Oblique SSA (Iterative O-SSA) based on restricted SVD to make adjustment to singular values obtained from the decomposing time series matrix and (ii) Robust Sparse K-means (RSK-means) which is to identify relevant cluster for eigenvector obtained from decomposing time series matrix in order to separate the trend and noise components in the time series torrential rainfall data appropriately.

3.4. Iterative Oblique SSA (Iterative O-SSA)

Number Iterative O-SSA was initially proposed by [8] to improve the separability of the series components by SSA. In general, the method consists of a basic algorithm of SSA and several variations to improve the separability between components in the original series data. Separability is closely related to the SVD which is the essential part of many statistical and signal processing methods. However, it has a drawback if the singular values obtained from SVD have coinciding singular values. Coinciding singular values mean that torrential rainfall time series data

have equal singular value which are generated from the decomposing time series matrix. This situation causes disjoint sets of singular values and different series components to mix with each other. Iterative O-SSA method could overcome the problem by modifying the basic SSA algorithm where the SVD step is changed to the (**L,R**)-SVD known as Restricted SVD (RSVD).

The algorithm of iterative O-SSA starts with an oblique SSA. Oblique SSA comes from two oblique bases referred to as **L**-orthonormal and **R**-orthonormal that corresponding to the row and column spaces based on **Definition 1**. Flow chart of SSA approach as shown in Figure 1 and Flow chart of Modified SSA approach as shown in Figure 2.

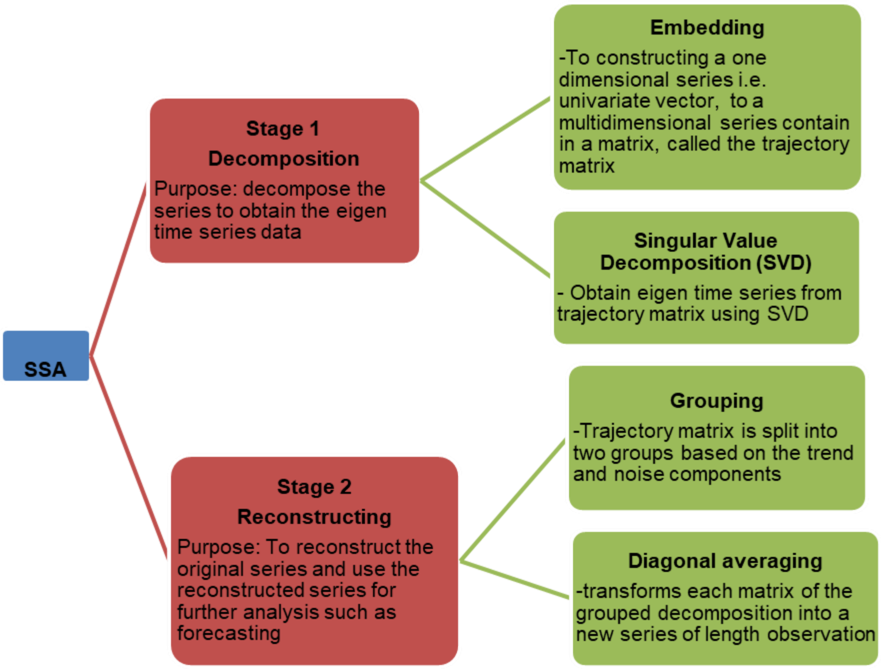

SSA
Stage 1
Decomposition
Purpose: decompose the series to obtain the eigen time series data
Embedding
-To constructing a one dimensional series i.e. univariate vector, to a multidimensional series contain in a matrix, called the trajectory matrix
Singular Value Decomposition (SVD)
- Obtain eigen time series from trajectory matrix using SVD
Stage 2
Reconstructing
Purpose: To reconstruct the original series and use the reconstructed series for further analysis such as forecasting
Grouping
-Trajectory matrix is split into two groups based on the trend and noise components
Diagonal averaging
-transforms each matrix of the grouped decomposition into a new series of length observation

Figure 1. Flow chart of SSA approach

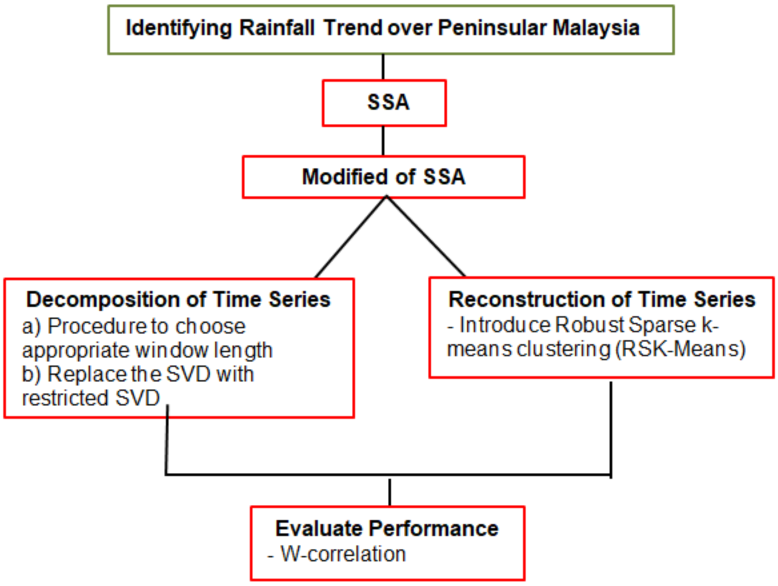

Identifying Rainfall Trend over Peninsular Malaysia
SSA
Modified of SSA
Decomposition of Time Series
a) Procedure to choose appropriate window length
b) Replace the SVD with restricted SVD
Reconstruction of Time Series
- Introduce Robust Sparse k-means clustering (RSK-Means)
Evaluate Performance
- W-correlation

Figure 2. Flow chart of Modified SSA approach

Definition 1: Consider a minimal decomposition of $\mathbf{Y}\epsilon\mathcal{M}_{L,K}$ of rank r in the form

$$\mathbf{Y} = \sum_{i=1}^{r} \sigma_i P_i Q_i^T \tag{7}$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, $\{P_i\}_{i=1}^{r}$ and $\{Q_i\}_{i=1}^{r}$are linearly independent. $\{P_i\}_{i=1}^{r}$ is a basis of the column space of **Y** of an **L**-orthonormal system while $\{Q_i\}_{i=1}^{r}$ is a basis of the row space of **Y** of an **R**-orthonormal system. (7) can be expressed in the matrix form,

$$\mathrm{Y} = \mathrm{P}\Sigma\mathrm{Q}^{\mathrm{T}} \tag{8}$$

where $\mathbf{P} = [P_1 : \dots : P_r]$, $\mathbf{Q} = [Q_1 : \dots : Q_r]$ and $\mathbf{\Sigma} = diag(\sigma_1, \dots, \sigma_r)$. The decomposition is $(\mathbf{L}, \mathbf{R})$-biorthogonal known as Restricted SVD (RSVD) given by the triple $(\mathbf{Y}, \mathbf{L}, \mathbf{R})$.
Let $\mathbf{O_L}$ be an orthonormalizing matrix of $\{P_i\}_{i=1}^{r}$ and $\mathbf{O_R}$ be an orthonormalizing matrix of $\{Q_i\}_{i=1}^{r}$. Then, (7), transformed to

$$\mathbf{O_L Y O_R^T} = \sum_{i=1}^{r} \sigma_i (\mathbf{O_L} P_i)(\mathbf{O_R} Q_i)^T \tag{9}$$

is an SVD of $\mathbf{O_L Y O_R^T}$ of $\epsilon\mathcal{M}_{r,r}$ with left singular vectors $\mathbf{O_L} P_i \epsilon R^r$ and right singular vectors $\mathbf{O_R} Q_i \epsilon R^r$. The result follows the fact that any bi-orthogonal decompositions is an SVD.
(9) is an SVD if and only if (7) is an ($\mathbf{L}$, $\mathbf{R}$)-SVD, where $\mathbf{L} = \mathbf{O_L^T O_L}$ and $\mathbf{R} = \mathbf{O_R^T O_R}$, $O_L$and $\mathbf{O_R}$ are orthonormalizing.

Oblique SSA (O-SSA) is the modification from the SSA where the SVD step is changed to ($\mathbf{L}$, $\mathbf{R}$)-SVD. The algorithm for O-SSA is as follows:
Algorithm 1: Oblique SSA (O-SSA)
Input: $\mathrm{Y}, (\mathrm{L}, \mathrm{R})$ consistent with Y
Output: $(\mathrm{L}, \mathrm{R}) - SVD$ in the form (7)
[1] Calculate $\mathbf{O_L}$ and $\mathbf{O_R}$using Cholesky decomposition
[2] Calculate $\mathbf{O_L Y O_R^T}$
[3] Find the ordinary SVD decomposition as follow:

$$\mathbf{O_L Y O_R^T} = \sum_{i=1}^{r} \sqrt{\lambda_i} U_i V_i^T \tag{10}$$

be ordinary SVD of $\boldsymbol{O_L Y O_R^T}$. Then, compare with the decomposition of (7) with $\sigma_i = \sqrt{\lambda_i}$, $P_i = \mathbf{O_L^\dagger} U_i$ and $Q_i = \mathbf{O_R^\dagger} V_i$ is the $(\mathbf{L}, \mathbf{R}) - SVD$ where † denotes pseudo-inverse.
Note that if **L** and **R** are the identity matrices, then oblique SSA coincides with classical SSA, $\sigma_i = \sqrt{\lambda_i}$, $P_i = U_i$ and $Q_i = V_i$.

$(\mathbf{L}, \mathbf{R}) - SVD$ is another SVD approach. Its function to extract the leading components, like classical SVD, is inappropriate in extracting the trend and noise from components. Trends components would be mixed with noise components in reconstruction step when using $(\mathbf{L}, \mathbf{R}) - SVD$ approach. The Iterative oblique SSA (Iterative O-SSA) is introduced to improve the separability between the components.

$\mathbf{X} = \mathbf{X}_{I1} + \dots + \mathbf{X}_{Im}$ is a matrix decomposition obtained from the grouping step in SSA method (refer to 5.5) where each group corresponds to a separated time series component. Let sth group $I = I_s$ be appointed for a redefined decomposition. Denote $\mathbf{Y} = \mathbf{X}_I$, $r = rank$ $\mathbf{Y}$, $\mathbb{Y} = \mathcal{T}^{-1}\mathcal{H}\mathbf{Y}$, the series obtained from **Y** by diagonal averaging.
Algorithm 2: Iterative O-SSA
Input: The matrix Y contains column space of L and row space of R, such a pair (L, R) is consistent with the matrix Y
Output: Redefined series decomposition of $\mathbb{Y} = \widetilde{\mathbb{Y}}^{(1)} + \dots + \widetilde{\mathbb{Y}}^{(l)}$.
a) Construct an (**L**,**R**)-SVD of Y using algorithm 1 in (7)
b) Partition the set $\{1, \dots, r\} = \coprod_{m=1}^{l} J_m$ and form group between component to obtained a refined matrix decomposition $\mathbf{Y} = \mathbf{Y}_{J1} + \dots + \mathbf{Y}_{Ji}$.
c) Obtain a refined series decomposition $\mathbb{Y} = \widetilde{\mathbb{Y}}^{(1)} + \dots + \widetilde{\mathbb{Y}}^{(l)}$, where $\widetilde{\mathbb{Y}}^{(m)} = \mathcal{T}^{-1}\mathcal{H}\mathbf{Y}_{Jm}$.
d) The decomposition of the series $\mathbb{X}$ is obtained, $\mathbb{X} = \widetilde{\mathbb{X}}^{(1)} + \dots + \widetilde{\mathbb{X}}^{(p)}$ where $\widetilde{\mathbb{X}}^{(s)} = \widetilde{\mathbb{Y}}^{(1)} + \dots + \widetilde{\mathbb{Y}}^{(l)}$.

3.5. Robust Sparse K-Means (RSK-means)

K-means clustering is one of the popular methods for cluster analysis in data mining [14]. The purpose of this clustering method is to partition the observations or components into several groups in which observations or components belong to the group with the nearest mean. In the modification of SSA, this clustering technique is chosen for the purpose of grouping step where the leading components extracted from SVD are assembled. Unfortunately, k-means is unable to extract a very good partition for leading components because it performs rather poorly in the separation of the trends and noise components. Therefore, an efficiency method from the extension of k-means approach was introduced which is called the Robust Sparse K-Means (RSK-Means).

RSK-means was initially proposed by [6] to develop a robust clustering method that is also able to perform variable selection in the data set. This approach is based on combining idea from two methods which are sparse K-means and trimmed K-means where both methods use squared Euclidean and weighted squared Euclidean distances. The weighted squared Euclidean distance is used to eliminate the effect of noise from the selection of a partition in the data set. The purpose of this proposed method is to produce a double step procedure by trimmed k-means called weighted and unweighted trimmed procedure.

The steps involved in RSK-means algorithm are as follows:

a) Given cluster centers $\mu_1, \ldots, \mu_K$, assign each point to the cluster with the closest center.
b) Given a cluster assignment, trim the $\alpha 100\%$ observations with largest distance to their cluster centers, and update the cluster centers to the sample mean of the remaining observations in each cluster.

The tuning parameter α regulates the amount of trimming and is selected by the researcher.

3.6. Assessing Separability in Time Series Data

Separability is one of the main concepts in studying SSA which necessitates how well different components of the time series can be separated from each other to allow further analysis to be meaningfully done. According to the literature review [8], separability is an important device when dealing with SSA method in various fields of study. The impact of separability could be resulted in the proper decomposition and proper component extracted. W-correlation is the method to measure the separability between two different reconstructed time series components.

W-correlation is the weighted correlation between the reconstructed time series components and it would give very helpful information for separation and group identification for reconstructed components. The weights reflect the elements of the time series terms into trajectory matrix. It ranges from the absolute values of 0 to 1 where well separated components incline to zero whereas poorly separated components generally incline to one. Besides that, the other function of w-correlation matrix is checking the grouped decomposition between reconstructed components. The formulation of the w-correlation matrix is as follows:

$$\rho_{12}^{w} = \frac{\langle X^{(1)}, X^{(2)} \rangle w}{\|X^{(1)}\| w \|X^{(2)}\| w} \tag{13}$$

where $\|X^{(i)}\| w = \sqrt{\langle X^{(i)}, X^{(i)} \rangle w}, i = 1,2, \langle X^{(1)}, X^{(2)} \rangle w = \sum_{i=0}^{N-1} w_i x_i^{(1)} w_i^{(2)}$ and the weights w_i are defined as follows:
Let $L^* = \min(L, K)$and $K^* = \max(L, K)$. Then,

$$w_i = \begin{cases} i + 1 \; for \; 0 \le i \le L^* - 1, \\ L^* \; for \; L^* \le i \le K^*, \\ T - i \; for \; K^* \le i \le T - 1. \end{cases} \tag{14}$$

w-correlation matrix can be illustrated graphically in the white-black scale where small correlation is shown in white while correlation between the series components are close to 1 is shown in black.

4. RESULTS AND DISCUSSION

The effectiveness of the modified SSA was compared against SSA approaches using the daily torrential rainfall data described in data section to separate trend and noise components in skewed time series data. The effectiveness of these approaches were measured by assessing its weighted correlation i.e. w-correlation at different window length, L. The w-correlation as described in the methodology section measured the separability between the reconstructed time series components, which were the trend and noise components. Several selection of L which were L= $T/2, T/5, T/10$ and $T/20$, which refer to $L = 3,6,12,30$ respectively for T based on 61 daily torrential rainfall days that were chosen. These scales were selected to suit the time series data and to strike a balance to achieve an appropriate sequence of lag vector.

Figure 3 shows the w-correlation using SSA and modified SSA from daily torrential rainfall data at different window lengths. The red rectangle refers to SSA and the blue triangle in the plot refers to the modified SSA. The plot shows that the average w-correlation illustrates a decreasing pattern as the number of window length decrease for both approaches. This implies that different window lengths affect the separability of components. It also shows that modified SSA pointed out to the lowest average w-correlation at window length, L=$T/5$ which shows the strongest separability between the reconstructed components as it is closest to zero.

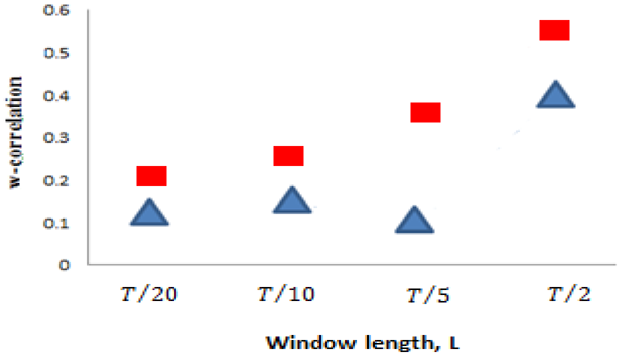
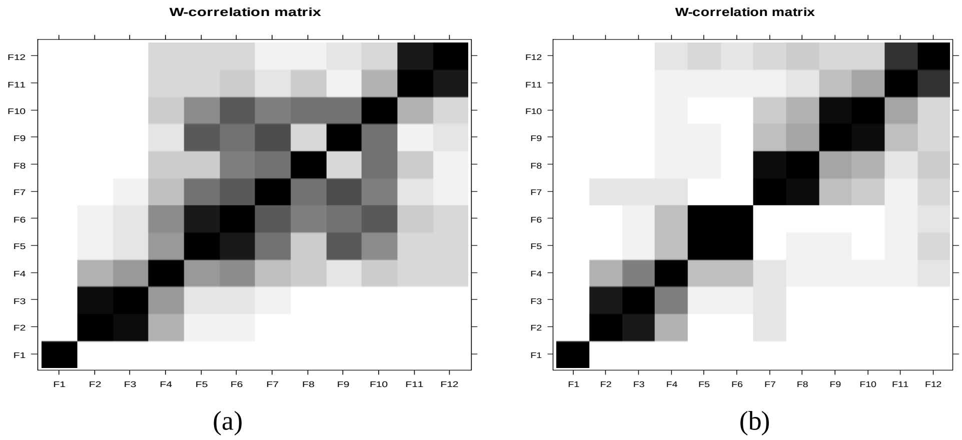


Figure 3. Effect of w-correlation based on SSA, 🟥 and modified SSA, ▲
using daily torrential rainfall data at different window lengths

The graphs in Figure 4 show heatplots of window lengths, L=T/5 based on w-correlations from SSA and modified SSA. As described by [8] the heatplot of w-correlation for the reconstructed components on a grade scale from white to black corresponding to absolute values of correlation from 0 to 1. Large values of correlations between reconstructed components indicated that the components could possibly be gathered into one group and correspond to the same component [9]. The shade of each square in Figure 4 represents the strength of the w-correlation between two components. Figure 4(a) shows that the components tend to be correlated with more other components even if the correlation is light at times. It means that in SSA, trends components are still slightly mixed with the noise components and it was corrected by the modified SSA where it clearly shows in Figure 4(b) the better improvement of separability. For trend extraction, it is important that correlations between trend and noise are close to zero. Here, correlations which employ modified SSA between the reconstructed time series components were small compared to SSA. In this study, the w-correlation is 0.35 for SSA and 0.20 for modified SSA, indicating that modified SSA has strong separability compared to SSA.



(a) (b)

Figure 4. The w-correlation using SSA and modified SSA with $L = T/5$

Figure 5 and Figure 6 show the plot of form of twelve leading eigenvectors. The eigenvector plot could be used to help in choosing the proper group for the components in time series data which was useful in the separation of the trend and noise components. This information will be used for further analysis in grouping step in SSA. In identifying the trend components through eigenvector plot, the slow cycles of the graph with higher frequencies refer to the trend component while the saw-tooth of the graph with low frequencies refer to the noise component. From Figure 5, it clearly shows that SSA did not separate the trend and noise components for this time series, likely due to the lack of strong separability. It can be visualized from Figure 5 that the first component have high frequency but the graph is flattened out. This is the typical situation when trend components are mixed with the noise components and therefore it was corrected by modified SSA. Modified SSA help to move apart decomposition components where it would refined groups by using Robust Sparse K-Means (RSK-means) and grouping results of the components can be supported by eigenvector plot. Figure 5 illustrates the saw-tooth of the graph and low frequencies for the first seven components, thus the components refer to the noise group. Meanwhile the slow cycle of the graph with high frequencies for the 8-12 components refer to the trend group.

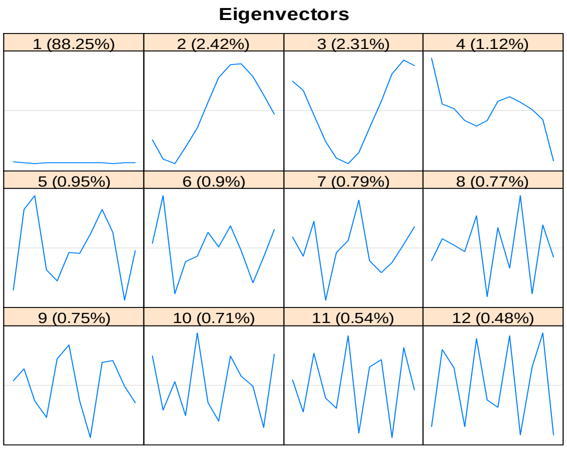


Figure 5. Eigenvectors plot obtained by SSA

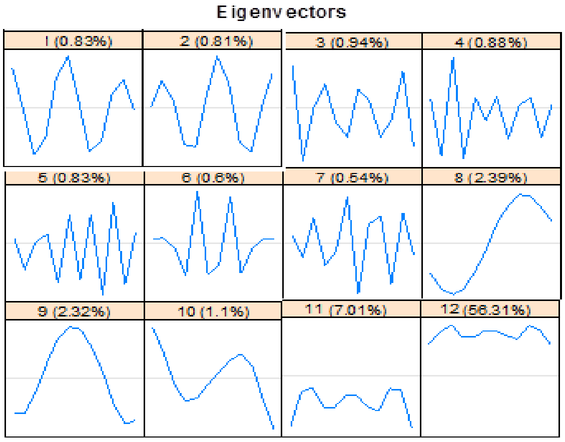


Figure 6. Eigenvectors plot obtained by modified SSA

Figure 6 demonstrates the reconstructed time series components plot from the extracted trend using modified SSA and SSA for a rainfall station where this area typically receives abundant and intense rainfall in Peninsular Malaysia. The trend component of the time series data is used for identifying the range of local time scale at single rainfall station in order to detect the abnormally heavy rainfall that can cause torrential rainfall events. The dashed line in the plot refers to the original time series rainfall data, the blue line refers to the reconstructed series based on extracted trend components from SSA and the red line refers to the reconstructed series based on extracted trend components from modified SSA. From Figure 7, the plotted line of the reconstructed time series components of SSA tend to flatten out compared to modified SSA. Therefore, there was difficulties in determining the range of the local time scale to detect the torrential rainfall events occur. Meanwhile, the plot of the reconstructed time series components of modified SSA appeared to follow the pattern of the original time series rainfall data despite the exclusion of noise components particularly for $L = T/5$. In particular, for the reconstructed component plot in Figure 7, the daily torrential rainfall appear to increase and decrease frequently with peaks occurring in the period between the 20th and 30th day of the time series.

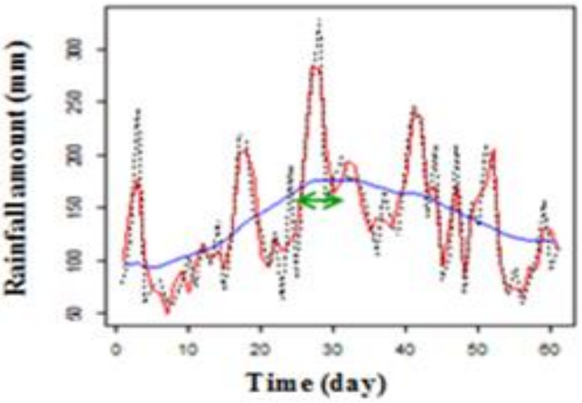


Figure 7. Plot of reconstructed components from extracted trends using SSA and modified SSA.
↔ refer to range of local time scale

5. CONCLUSION

The main purpose of introducing SSA in this study is identifying rainfall trend in order to determine when torrential rainfall events occur at a particular location. This can be detected by observing the trend of a time series data characterized by the shape of time series data. Modified SSA has been proposed to solve the problem in separability components and proper grouping as discussed in Introduction section. A new decomposition approach is proposed in Stage 1 of SSA by incorporating the appropriate window length and an iterative oblique SSA (Iterative O-SSA) which helps to improve the weak separability issue above. This method performs a new decomposition of a part in the SSA which suggests the appropriate window length that suit to time series rainfall data and corrects the eigenvalues in order to avoid their possible mixture or disjoint sets of singular values in the group corresponding to different components. Secondly, Robust Sparse K-means clustering (RSK-means) is introduced in order to guide an effective grouping of decompose eigenvector. The modified SSA is considered as refining of the decomposition and reconstructing obtained by SSA. Based on the results, it clearly stated that window length, $L = T/5$ shows a trend that fits well to the original torrential rainfall time series data in Peninsular Malaysia. In addition, the results of modified SSA shows that it has slightly better performance in reconstructed time series components for small window length. This can be illustrated that the plot of the reconstructed time series data of modified SSA appear to follow the pattern of the original time series rainfall data despite the exclusion of noise components particularly for smaller window length. The results of this analysis can be used to detect the abnormally heavy rainfall that can cause torrential rainfall events at a particular location.

ACKNOWLEDGEMENTS

The authors would like to thank Universiti Pendidikan Sultan Idris for their financial funding through GGPU grant Vote No. 2018-0084-106-01.

REFERENCES

[1] F.W. Githui, A. Opere, and W. Bauwens, “Statistical and trend analysis of rainfall and river discharge: Yala River basin, Kenya,” in *Proc. International Conference of UNESCO*, Netherlands, 2003.

[2] M. Khaleghi, H. Zeinivand, and S. Moradipour, “Rainfall and river discharge trend analysis: a case study of Jajrood watershed, Iran,” *International Bulletin of Water Resources and Development*, vol.2, no.3, 2014.

[3] RA. Mondal, S. Kundu, and A. Mukhopadhyay, “Rainfall trend analysis by Mann-Kendall test: a case study of North-Eastern part of Cuttack district, Orissa. *International Journal of Geology, Earth and Environmental Sciences*, vol. 2, no. 1, pp.70-78, Jan. 2012.

[4] T. Alexandrov, N. Golyandina, and A. Spirov, “Singular spectrum analysis of gene expression profiles of early drosophila embryo: exponential-in-distance patterns,” *Research Letters in Signal Processing*, vol. 2008, no. 2008, pp. 1-5, June 2008.

[5] MAR. Khan, D.S. Poskitt, “Moment tests for window length selection in singular spectrum analysis of short- and long-memory processes,” *Journal of Time Series Analysis*, vol.34, no.2, pp. 141-155, 2013.

[6] RO. Awichi, W. Muller, “Improving SSA predictions by inverse distance weighting,” *REVSTAT*, vol.13, no.1, pp. 105-119, 2013.

[7] N. Itoh, N. Marwan, “An extended singular spectrum transformation (sst) for the investigation of kenyan precipitation data,” *Nonlinear Processes in Geo-physics*, vol. 20, no. 4, pp. 467-481,2013.

[8] N. Golyandina, and A. Shlemov, “Variations of singular spectrum analysis for separability improvement: non-orthogonal decompositions of time series,” *Statistics and Its Interface*, vol. 8, no.3, pp. 277-294, Feb. 2014.

[9] H. Hassani, “Development of the theoretical and methodological aspects of the singular spectrum analysis and its application for analysis and forecasting of economics data”. Cardiff, Wales: Cardiff University, 2009.

[10] FJ. Alonso, DR. Salgado, J. Cuadrado,P. Pintado, “Automatic smoothing of raw kinematics signals using ssa and cluster analysis,” *Euromech Solid Mechanics Conference Lisbon*, Portugal, pp. 1-9, September 2009.

[11] YS. Maslennikova, VV. Bochkarev, “20 training algorithm for neuro-fuzzy based on singular spectrum analysis,” *World Conference on Information Technology*, Barcelona, Spain, pp. 606-610, November 2012.

[12] SM. Shaharudin, N. Ahmad, NH. Zainuddin, NS. Mohamed, “Identification of rainfall patterns on hydrological simulation using robust principal component analysis*,” Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*,” vol. 5, no. 3, pp. 401-408, January 2017.

[13] SM. Shaharudin, N. Ahmad, Fadhilah Yusof, “Effect of window length with singular spectrum analysis in extracting the trend signal on rainfall data,”*AIP Conf. Proc.*, vol. 1643, no. 321, pp. 321-326, 2015.

[14] AP. Putra, A. Buono, BP. Silalahi, “ Modeling singular value decomposition and k-means of core image in classification of potential nickel,” *TELKOMNIKA Indonesian Journal of Telecommunication, Computing, Electronics and Control*,” vol. 13, no.3, pp. 561-567, March 2016.

[15] Y. Kondo, Robustification of the Sparse K-Means Clustering Algorithm. West Mall, Vancouver: The University of British Columbia, 2009.

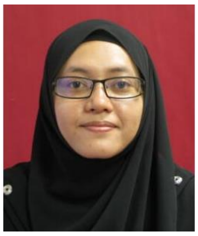
[16] Y. Kondo, M. Salibian-Barrera, and R. Zamar, "A robust and sparse k-means clustering algorithm," arXiv:1201.6082v1, pp. 1-20, Jan. 2012.

[17] N. Golyandina, and A. Korobeynikov, "Basic singular spectrum analysis and forecasting with R," *Computational Statistics and Data Analysis*, vol. 71, pp. 934- 954, Apr. 2013.

BIOGRAPHIES OF AUTHORS

Shazlyn Milleana was born in Johor Bahru, Malaysia, in 1988. She is a senior lecturer at the Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris (UPSI). She graduated with a bachelor science degree in Industrial Mathematics from Universiti Teknologi Malaysia, in 2010. Upon graduation, she began her career as an Executive in banking institution. In the following year, she received an offer to continue her study as a fast-track PhD student at the same university. During her PhD journey, she developed an interest in multivariate analysis, specifically in finding patterns which deals with big data. Her research focuses on the area of dimension reduction methods specifically in climate informatics which involves analysis on huge climate-related datasets based on techniques in Data Mining. She had published her research in Scopus indexed journal and presented her work in various local and international conferences. She completed her PhD thesis at the end of 2016 and was conferred a doctorate degree in 2017.

Norhaiza is a senior lecturer at the Department of Mathematical Sciences, Faculty of Science, UniversitiTeknologi Malaysia (UTM). She graduated with an honors degree in Mathematics, Statistics and Operational research from the University of Manchester, in 1996. She joined UTM in August 2000. In the following year, she continued her studies at the University of Sheffield for her masters degree. In Sheffield, she developed an interest in multivariate analysis, specifically in finding patterns which lead her to pursue a PhD degree at the University of Kent. She completed her PhD thesis at the end of 2007 and was conferred a doctorate degree in 2008. Finding patterns in any data have always been her research interests. She started the interests in profiling data – finding statistically distinctive and significant groups and features in the object of interest whilst at Sheffield. Currently, her research interests revolve around hydroinformatics particularly in investigating the streamflow variability of the local rivers.



Nurul Hila was born in Kelantan, Malaysia, in 1987. She is a senior lecturer at the Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris (UPSI). She graduated with a bachelor science d`egree in Financial Mathematics from Universiti Malaysia Terengganu, in 2010. In the following year, she continued her studies at the Universiti Malaysia Terengganu for her masters degree and in 2016 she managed to complete her Ph.D at the same university. During her PhD journey, she developed an interest in hybrid modelling, specifically double bootstrap on control chart. Her research focuses on the volatility point of sukuk (Islamic certificate) investment. She had published her research in Scopus indexed journal and presented her work in various local and international conferences.