International Journal on Advanced Science Engineering Information Technology

Vol.10 (2020) No. 4
ISSN: 2088-5334

Performance Analysis and Validation of Modified Singular Spectrum Analysis based on Simulation Torrential Rainfall Data

Shazlyn Milleana Shaharudin[a], Norhaiza Ahmad[b], Nur Syarafina Mohamed[c] , Nazrina Aziz[d,e]

[a] *Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Tanjung Malim, Perak, Malaysia*
E-mail: shazlyn@fsmt.upsi.edu.my

[b] *Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia*
E-mail: norhaiza@utm.my

[c] *Technical Foundation, Malaysian Institute of Industrial Technology, Universiti Kuala Lumpur, Malaysia*
E-mail: nursyarafina@unikl.edu.my

[d] *School of Quantitative Sciences (SQS), UUM Sintok, Kedah, Malaysia*

[e] *Institute of Strategic Industrial Decision Modelling (ISIDM), Universiti Utara Malaysia, Malaysia*
Email: nazrina@uum.edu.my

***Abstract*— A popular method for time series analysis to extract the components of noise and trend from the time series data is called the singular spectrum analysis (SSA). However, the drawback of SSA is its problem in determining the appropriate window length, *L* for certain data set in confirming patent separation of the components of trend and noise. Another issue that crops up when using SSA is that, over time, the sum of day-to-day rainfall becomes nearly comparable. In this case, disjoints sets of singular values and distinctive series components could essentially be intermixed, resulting in poor separability between trend and noise components. The introduction of modified SSA is to mitigate the problems efficiently. The performance of modified SSA is measured by using w-correlation and RMSE based on simulated data. These results show that the parameter *L = T/5* was suitable to use in short time series rainfall data. It can be proved by the plot of the extracted trend for modified SSA that appears to conform to the original data configuration for time series rainfall however there is the omission of components of noise predominantly for *L = T/5* in detecting the uncharacteristically heavy downpour which could potentially initiate the occurrence of torrential rainfall. In addition, the result shows that average RMSE for reconstructed time series components of modified SSA is much smaller than SSA for each *L*.**

***Keywords*— singular spectrum analysis; trend; simulation; iterative o-ssa; robust sparse k-means; window length; modified singular spectrum analysis.**

I. INTRODUCTION

This study seeks for identifying the patterns of temporal torrential rainfall specifically in Peninsular Malaysia by determining a local time scale in which the torrential rainfall occurrences take place in each specific rainfall station based on its trend [1], [2]. It could be distinguished from the observation of a time series data trend which is depicted based on the time series data shape. Based on the time scale, many previous studies had been conducted to identify the tendency of rainfall [3]. A significant number of these researches had been suggesting the Kendall correlation coefficient for the identification of the time series data trend. Nonetheless, the previously-mentioned method was incapable of performing fittingly as there was unfeasibility in determining the trends type that can be extricated through its processes [4]. Additionally, any noise is not considered as possibly compromising the method applied. This research aims to describe possible trend which is capable of specifying the interval of local time scale in order to ascertain if there is any observation of high extreme rainfall values for the purpose of detecting abnormal and heavy rainfall may be leading to the occurrence of torrential rainfall. A method to identify the local time scale range as can be seen from the trend uses singular spectrum analysis (SSA) as the base. Generally, the depiction of a single time series (univariate case) is provided by a certain SSA. This SSA is specifically converted in eigenvalues and eigenvectors of a trajectory matrix. Eigenvalues is also

known as characteristic values; eigenvectors are known as characteristic vectors. Through the decomposition of its time series eigen and their reconstruction into group selection, the function of SSA is separating the time series data into noise and trend. In SSA, the time series data is separated into an additive component of trend and noise [5].

However, the strength in separating the trend and noise components in SSA can be compromised by several factors. One, the choice of window length to form the trajectory matrix in SSA is important to ensure the signal and noise components have been distinctively divided [6]. Generally, it is required for the window length to be adequately huge but less than the examined time-series data by half [7]. Nevertheless, distinctive behavior which is being observed for a data set could be influential towards the window length choice. Two, over time, the day-to-day amount of torrential rainfall time series data in Peninsular has become nearly comparable. Such circumstance has led to issues of employing SSA in which the singular values coincide. This makes separate singular values set and distinctive series components to precisely be intermixed when using the singular value decomposition (SVD) expansion. Consequently, the time series trends which were extricated from SSA tend to be flattened out as well as not showing a certain type of distinctive configuration. Hence, it is challenging to ascertain the range of the local time scale to estimate the possible occurrence of the torrential rainfalls.

In finding effective solutions, SSA modification for issues alleviation is introduced. Primarily, a new approach in SSA is introduced via the incorporation of the suitable window length for the trajectory matrix and make adjustments on the coinciding singular values that are acquired from the decomposed matrix of time series based on a restricted SVD using iterative oblique SSA (Iterative O-SSA). In addition, it also discriminated the eigenvectors from this iterative procedure by incorporating a guided clustering method, i.e. Robust Sparse k-means (RSk-means) to identify the noise and trend components more objectively.

To assess whether the modified SSA could perform against SSA approaches, both methods employed on simulated time series which replicate the real data, abiding by the distribution of the original torrential time series rainfall data in Peninsular Malaysia. The main aim is to employ simulations with the aim of determining a fitting window length, L in SSA to ensure the two components, which are trend and noise, would be distinctively divided. Additionally, a method which could extract time series components from observed series is called the separability of components. Hence, the separability of modified SSA against SSA is measured using a weighted correlation known as w-correlation. This is conducted in the analysis to distinguish the patterns of temporal torrential rainfall in Peninsular Malaysia.

Section 2 describes the applied rainfall data in current research. Consequently, the methodologies linked to the SSA modification as well as the simulation procedure for modified SSA is shown. Section 3 subsequently delivers the results and discussion for the identification of temporal cluster patterns by employing SSA and modified SSA. In the last section, the conclusion is presented.

II. Materials and Methods

A. Data

The day-to-day rainfall data which was gathered from a number of 75 distinct stations across Peninsular Malaysia had been attained from the Department of Irrigation and Drainage. This research focuses on extreme rainfall occurrence known as torrential rainfall. Based on conditions, certain days which demonstrated torrential criteria had been selected specifically. Hence, the set of criteria was very important as it leads to the threshold establishment so that the factors which constitute a torrential rainfall day in Peninsular Malaysia could be clearly distinguished. For this reason, the most generic threshold employed is the tropical climate zone of 60 mm/day [8]. Its time interval depends on the series of day-to-day maximum rainfall in observation, from the period of 1st of November 1975 until 31st of December 2007, which covered 61 sequential days.

B. Algorithm for SSA

Singular spectrum analysis (SSA) is a method of time series analysis which is a construct of some elements of multivariate geometry, classical time series, multivariate statistics, signal processing and dynamic system. The SSA method functions as a mean of decomposing the initial components to three components, i.e. noise, seasonal and trend [9]. This method comprises two corresponding stages that are known as reconstruction and decomposition stages. During the decomposition stage, two crucial steps could be classified into embedding and singular value decomposition. Meanwhile, there are two other key procedures, that are categorized into grouping and diagonal averaging, in the reconstruction stage. An epigrammatic discussion of SSA method methodology can be found below. The following are the steps in each stage:

1) Stage 1: Decomposition: The steps in the decomposition stage can be divided into two: i) embedding as well as ii) singular value decomposition (SVD). Generally, the objective of this stage is decomposing the series to attain the data of Eigen time series.

- Step I: Embedding. Primarily, the initial step for basic SSA algorithm is transforming a one-dimensional series into a multi-dimensional series to construct the trajectory matrix. This dimension is known as the window length, L. In such, let a one-dimensional time series, $\mathbb{Y}_T = \{y_1, y_2, \dots, y_T\}$ be converted to 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 $K = T - L + 1$.
- Step II: Singular Value Decomposition (SVD). In this study, the SVD of X_i is carried out where $\lambda_1, \cdots, \lambda_L$ are denoted as the eigenvalues of XX^T coordinated according to a declining order ($\lambda_1 \geq \cdots \lambda_L$) and by $U_1, \cdots, U_L$the corresponding eigenvectors. The SVD of X is signified as $X = X_1 + \cdots + X_L$, where $X_i = \sqrt{\lambda_i} U_i V_i^T$ and $V_i = \frac{X^T U_i}{\sqrt{\lambda_i}}$ (if $\lambda_i = 0$ we set $X_i = 0$). The set of $\left(\sqrt{\lambda_i}, U_i, V_i\right)$ is known as the $i - th$ eigentriple (ET) of the matrix X_i, and $\sqrt{\lambda_i}$ is matrix X_i singular values.

2) Stage 2: Reconstruction:

- *Step* I: Grouping. Essentially, splitting the trajectory matrix into several categories depending on the signal trend components and noise components during the step of grouping. Let $I = \{i_1, \cdots, i_p\}$ be a group index $i_1, \cdots, i_p$. where $(p < L)$. Next, the matrix X_I corresponding to the group I could be described as $X_I = X_{i1} + \cdots + X_{ip}$. The set of indices $\{1, \cdots, L\}$ are split into disjoint subsets $I_1, \cdots I_m$ corresponding to the depiction of $X = X_{I1} + \cdots + X_{Im}$.
- Step II: Diagonal averaging. The point of this step is transforming a matrix Z into a Hankel matrix $\mathcal{H}Z$ form that could be transformed into a time series. If z_{ij} denotes an element of matrix Z, then the kth-term of the resulting series is attained through averaging z_{ij} for all i, j such that $i + j = k + 1$. Hankelization $\mathcal{H}Z$ is an optimal procedure, which is the closest to Z in regard to the matrix norm. By employing the Hankelization procedure to $X = X_{I1} + \cdots + X_{Im}$, another expansion is described as $X = \tilde{X}_{I1} + \cdots + \tilde{X}_{Im}$, where $\tilde{X}_{I1} = \mathcal{H}Z$. Such is paralleling to the decomposition of the initial series $Y_T = \{y_1, y_2, \dots, y_T\}$ into a total of m series, $y_t = \sum_{k=1}^{m} \tilde{y}_t^{(k)}$ where $\tilde{Y}_T^{(k)} = (\tilde{y}_1^{(k)}, \dots, \tilde{y}_T^{(k)})$ corresponding to the matrix X_{Ik}.

C. Algorithm for Modified SSA

In this study, two significant approaches are combined for large SSA Modification. The first approach is known as Iterative Oblique SSA (Iterative O-SSA), based on the constrained SVD. This adjusts the singular values gathered from the decomposing time series matrix. The second approach is a Robust Sparse K-means (RSK-means). Therefore, the relevant cluster is identified for the acquired eigenvector from the decomposing time series matrix. The main purpose is to properly break up the components of trends and noise in the time series torrential rainfall data.

1) Step 1: Transfer a one-dimensional time series $\mathbb{Y}_T = \{y_1, y_2, \dots, y_T\}$ to multi-dimensional series with vectors $\mathbb{X}_i = (y_i, \dots, y_{i+L-1})^{\mathrm{T}}$ where $K = T - L + 1$ is lagged vector and window length, *L*. Then, form the trajectory matrix which denoted by $\mathbf{X} = (X_1, \dots, X_K)$.

2) Step 2: Construct a restricted SVD of $\mathbf{X}$ by using Iterative O-SSA in the form

$$\mathbf{X} = \sum_{i=1}^{r} \sigma_i P_i Q_i^T \qquad (1)$$

To calculate the eigenvalues, eigenvector, and principal components.

3) Step 3: Test the principal components from the output in step 2 by using statistical measure, Fisher g-test was to verify which components are significance to be extracted, and it would be used in the grouping step.

4) Step 4: Partition the principal components using RSK-means to perform grouping in obtaining refined matrix decomposition.

5) Step 5: Obtain a reconstructed time series component.

6) Step 6: Test the separability of reconstructed time series component obtained from modified of SSA by using a w-correlation matrix.

D. Robust Sparse k-Means (RSk-means)

In data mining, a well-known method for cluster analysis is k-means clustering. Intentionally, such method subdivides either the components or observations into several groupings. These fit the group with the closest mean. For modifying SSA, we want to use this clustering technique in the grouping step, which means that the SVD leading components extricated are being grouped. However, k-means could not extricate satisfactory partition for the leading components. In fact, its performance is unsuccessful when separating the noises and trends. Consequently, we introduce the efficient Robust Sparse k-means (RSk-means) in the modification of SSA.

Originally, RSk-means had been propositioned by [16in developing a robust clustering which could accomplish various options in the data set. Such an approach combines ideas based on the Sparse k-means and trimmed k-means that used squared Euclidean and weighted squared Euclidean distances. Weighted squared-Euclidean distance is employed for the elimination of noise effects from the partition choice in the data set. Such proposed method aims to generate a double step procedure through the trimmed k-means known as weighted and unweighted trimmed procedures.

The set of trimmed cases is denoted as O_E for weighted squared Euclidean distance and O_F for unweighted square Euclidean distance. The first step in RSk-means method is to eliminate the effect of noise from the selection of a partition. In the second step, trimmed k-means in the non-weighted squared Euclidean distance is proposed to eliminate noise "surviving" from the first step due to small weights assigned to noise components in the previous step. Based on the results of simulation studies, RSk-means shows better performances compared to other algorithms both in the selection of variables as well as the selection of a partition when datasets are contaminated [17].

The main idea in the extension of classical k-means method is to include the trim $\alpha 100\%$ observations with the widest distance to their cluster center. The tuning parameter α controls the sum of trimming. It is adjusted according to the study. RSk-means algorithm could be defined as the following:

1) Trimmed k-means is performed on the weighted data set:

Let

O_E = {Trim cases in the weighted square Euclidean distance}

- Given weights *w* and cluster centers $\mu_1, \dots, \mu_K$ solve

$$\min_{C_1, \dots, C_K} \sum_{k=1}^{K} \sum_{i \epsilon C_k} \sum_{j=1}^{p} \boldsymbol{w_j} (\boldsymbol{x_{i,j}} - \boldsymbol{\mu_{k,j}})^2 \qquad (2)$$

  that is gathered by allocating every point to the cluster with the nearest center by employing the weighted Euclidean square distance.
- Given weights *w* and cluster assignments, the $\alpha 100\%$ components with the biggest distance to their cluster center are trimmed, and the cluster center to the sample mean of the outstanding elements in every cluster is updated.

- The step above is repetitive until the stopping rule is satisfied.

2) Trimmed k-means is performed on the unweighted data set:

Let

O_F = {Trim cases in the unweighted square Euclidean distance}

- From the previous result, the unweighted cluster centers $\tilde{\mu}_k, k = 1, \dots, K$ are computed. For each component, x_i, let $\tilde{d}_i$ be the unweighted distance to its cluster centre, i.e., $\tilde{d}_i = \|x_i - \tilde{\mu}_k\|^2$ where $i \epsilon C_k$.
- The set of trimmed points $O = O_E \cup O_F$ is formed.
- Given cluster partition $C = (C_1, \dots, C_K)$, centres $\mu_1, \dots, \mu_K$ and trimmed points O, a new set of weights w is obtained through the resolving of weights between cluster dissimilarity measure.

$$\max_{\|w\|_2 \le 1, \|w\|_1 \le l_1} \sum_{j=1}^{p} w_j B_j(C_1, \dots, C_K, O) \quad (3)$$

where $B_j(C_1, \dots, C_K, O)$, $1 \le j \le p$, $w_j \ge 0\ \forall j$, is calculated without the components in O. p is denoted as a clustering variable.

E. Fisher G-Test

Determining a significant amount of components for extraction in the SVD method is one of the important parameters when using the SSA method. The extracted components are the targeted output in the grouping stage where the components will split into a number of groupings, and the components will be summed in the form of intergroup. The selected leading components are not consecutive in terms of eigenvalues or scree plot. Factually, to extract the leading components based on the eigenvalues may not direct to the biggest amount of principal components in such data set. This problem can be formally address using statistical significance using Fisher g test. It has two complementary procedures which decide the dominant frequency through the discovery of the peak in the periodogram. This finding uses a graphical device that calls it average periodogram.

Meanwhile, its statistical significance test will be examined via Fisher g-test. The significance components to extract for the reconstruction step is observed at the peak of the frequency in the periodogram and stays in the frequency range. It is statistically substantial based on the Fisher g-statistic results[15]. The purpose of using Fisher g-test is checking the accounted power proportion for the frequency. It is based on the peak in the periodogram that shows if the peak is random. Let $X_1, \dots, X_n$ is an equally-spaced time series, the periodogram comprises the set of points

$$\{(\omega_r, I(\omega_r)) : B = 1, \dots, B\}, \quad B = \lfloor (n-1)/2 \rfloor \quad (4)$$

where $\lfloor \cdot \rfloor$ symbolize as the floor function, $\omega_b = 2\pi b/n$ knows as Fourier frequencies for $b = 1, \dots, B$. Then, the periodogram denoted as $I(\omega)$ is defined as

$$I(\omega) = \frac{1}{n}\left|\sum_{t=1}^{n} X_t e^{-i\omega t}\right|^2 = \frac{1}{n}[\{\sum_{t=1}^{n} X_t \sin(\omega t)\}^2 + \{\sum_{t=1}^{n} X_t \cos(\omega t)\}^2], \quad \omega \epsilon (0, \pi) \quad (5)$$

This equation above can be used to determine the significant components in observed torrential rainfall series data $X_1, \dots, X_n$, in which n is the sample size of the rainfall data. It noted that the periodogram would exhibit a peak at a frequency ω^* when the series of data has significance components with frequencyω^*. When noticing that the periodogram exhibit a peak, Fisher g-test is carried out to determine whether the peak is significant or vice versa. G-statistic is introduced in Fisher g-test to derive a test from analyzing the spectral peak significance.

$$g = \frac{\max\{I(\omega_1), \dots, I(\omega_B)\}}{\sum_{b=1}^{B} I(\omega_b)} \quad (6)$$

H_o = Spectral peak is not statistically significant

H_1 = Spectral peak is statistically significant. Based on the hypothesis testing, huge g values direct to null hypothesis rebuff. The statistical significance test of p-value under the null hypothesis is represented as

$$p \equiv P(g > g^*) = \sum_{\kappa=1}^{K} (-1)^{\kappa-1} \frac{B!}{\kappa!(B-\kappa)!} (1 - \kappa g^*)^{B-1} \quad (7)$$

where K denotes the biggest integer less than $\frac{1}{g^*}$ and g^* represents the observed g-statistic value.

Here, this is the algorithm for determining the significant number of components based on Fisher g-test.

Let $\Omega \subseteq (0, \Omega)$be denoted as frequencies range of interest and let $\overline{\mathbb{C}}_i = \overline{\mathbb{C}}(u_i v_i^T \sqrt{\lambda_i})$ represent as the ith principal component.

Set $S^{(0)} = \phi$ and for $i = 1, \dots, d$

- Step **1**: Obtain the periodogram of $\overline{\mathbb{C}}_i$ and calculate $\omega_i^* = arg^{max}_{\omega \epsilon \{\omega_1, \dots, \omega_R\}} I_i(\omega)$
- Step 2: If $\omega_i^* \epsilon \Omega$, then proceed to next step, or else turn back to step 1 and increase i
- Step 3: Calculate the g-statistic, g_i using equation (4.11) that related with $\overline{\mathbb{C}}_i$
- Step 4: Compute the p-value using equation (4.13) based on resulted from step 3
- Step 5: Set $S^{(i)} = S^{(i-1)} \cup \{i\}$ if $p_i < 0.05$ or set $S^{(i)} = S^{(i-1)}$
- Step 6: Set $S_g = S^{(i)}$ if $i = d$ and stop, otherwise turn back to step 1 and increase i

From the above algorithm, g_i represent the Fisher g-test generated from the periodogram $\overline{\mathbb{C}}_i = \overline{\mathbb{C}}(u_i v_i^T \sqrt{\lambda_i})$, p_i is denoted as p-value and S_g denotes as the grouping set. In summary, Fisher g-test has three complementary procedures:

- Time series data is checked if it has a frequency peak by applying the average periodogram.
- The g-statistic is calculated for every time series data.
- The p-value corresponding to g-statistic is calculated.

F. Simulation Procedure for Modified SSA

The generation of the data sets depends on the probability distributions which replicate a univariate time series torrential rainfall data. For tropical rainfall data, the form of the distributions has been known to be commonly slanted towards the right. Hence, distributions exhibiting such criteria could is considered applicable to model the torrential rainfall [10]. The distributions, namely Gamma, Log-Normal

and Weibull, are selected and compared. Generally, such distributions are employed as prospective candidates for the generation of rainfall data mechanism [11]–[13]. The parameters approximation for every sample of the probability distributions is grounded from the statistical summary of the initial Peninsular Malaysia torrential rainfall data. From the sampled probability distributions, Log-Normal distribution was found as the best fit for univariate time series torrential rainfall data of Peninsula. At significance level $\alpha = 0.01$ if $p - value > \alpha$, this distribution is deemed to be excellent. Hence, it proves the validity of the null hypothesis. For instance, the Log-Normal distribution offers an accurate statistical model for univariate time series rainfall data.

The sample Log-Normal distributions undergo simulations which were categorized by two parameters, mean μ =138.13 and standard deviation σ = 61.54. These were acquired from the 33-year interval, initial series torrential rainfall data in Peninsular Malaysia to assemble a real-valued time series $\mathbb{Y}_T = \{y_1, y_2, \dots, y_T\}$ of length T with $T = 61$ torrential rainfall days based on selected one rainfall station. The length of T is the period grounded on observed series of day-to-day maximum rainfall, beginning from the period of 1st of November 1975 until 31st of December 2007, which covered 61 sequential days.

These parameters were obtained from univariate time series torrential rainfall data that most frequently affected by torrential rainfall events located in Kg. Jabi (Terengganu). To vary the stimulations being tested, there are two distinctive settings taken into account. Primarily, the standard deviation is being set diversely higher and lower than the standard deviation of the first series torrential rainfall data to evaluate what is affected when nearly all of the data variations are preserved. Every data that is produced evidently encompasses values of approximately 60, reflecting the torrential rainfall's threshold of 60mm/day. Next, a number of different of window length, L onto the generated series torrential rainfall data are tested to investigate its effect on extracting its trends.

Every produced data set is afterwards applied to the two approaches, SSA and modified SSA. From both approaches, the outcomes are further contrasted in terms of the plot of the extracted trend to determine which of the extracted trend appears obeying the simulated time series pattern. However, there was an exclusion of noise components. The analysis results may be applicable in detecting the abnormally heavy downpour, which potentially causes torrential rainfall events at a certain setting. In addition, w-correlation is used to assess the separability among the reconstructed time series data for both approaches. Besides that, window length selection is discussed so that the sensitivity of the modified SSA for different L could be examined.

G. Root Mean Square Error (RMSE)

The root means square error (RMSE) is commonly-employed as a standard statistical metric to evaluate model performance in research studies for air quality, meteorology, as well as climate. RMSE presents information on short-term efficiency as a benchmark for the predicated values differences regarding the observed values. The RMSE is given by the following formula.

$$\mathbf{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2/n} \quad (8)$$

where y_i is the observed rainfall, $\tilde{y}_i$ is the predicted rainfall data, and $\bar{y}_i$ indicates the average rainfall data over rainfall station in Peninsular Malaysia.

III. Results and Discussion

The efficacy of modified SSA against SSA approaches compared through the simulated data for the purpose of separating the components of noise and trend in short and skewed time series data. The efficiency of above-mentioned approaches was calculated from the assessment of its weighted correlation, i.e. w-correlation at distinctive window length, L. The w-correlation measures the separability between the reconstructed time series components, that are the components of noise and trend [14]. A number of selections of L, L= $T/2$, $T/5$, $T/10$ and $T/20$, that refers to $L =$ 3,6,12,30 correspondingly for T according to 61 simulated torrential rainfall days are selected. Such scales are designated to befit our short time series data and to form stability for the objective of achieving a fitting sequence of lag vector.

Fig.1 illustrates the average of w-correlation when employing the SSA and modified SSA from 20 simulated data at several window lengths. The plotted red rectangles represent SSA while the blue triangles represent the modified SSA. As shown, the plot demonstrates how the average w-correlation shows a declining pattern since the total of window length declines for each approach. It can be addressed that distinctive window length could affect the components separability. Besides, the modified SSA points out to the lowermost average w-correlation at window length, $T/5$. This means that it makes the greatest separability among the reconstructed components since it has the nearest value to zero.

To identify the local time scale at which torrential rainfall events hit certain locations, the broad structure of the reconstructed time series trend from the simulated time series data is observed. The trend of the reconstructed time series should not be too smooth to identify the peak occurrence of torrential rainfall.

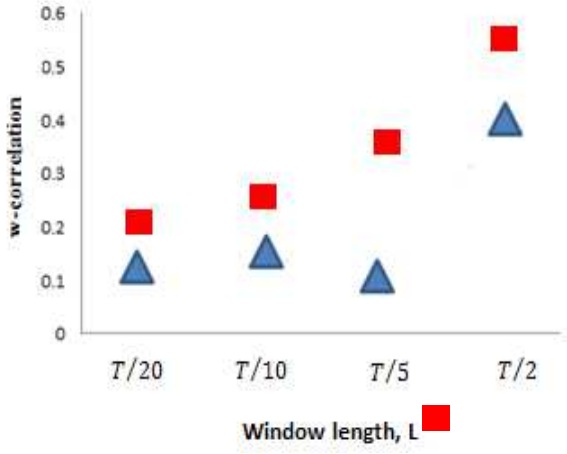


Fig. 1 Effect of w-correlation based on SSA, and modified SSA, using 20 simulated data at different window lengths

In addition, it should still follow the pattern of the simulated data after the exclusion of noise component. As shown in Figure 2 (a) to (d), the plots of simulated time series against reconstructed series data using both SSA and modified SSA approaches are demonstrated. Various window lengths, L is involved.

The denotation for each type of the plotted line is as follows:

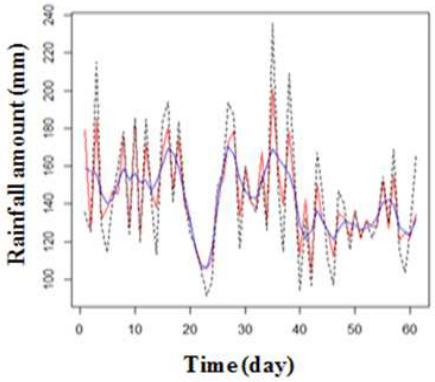
- dashed line = simulated time-series data
- blue line = reconstructed series based on the extracted SSA trend components
- redline = reconstructed series based on extracted modified SSA trend components

According to the plot, both reconstructed time series data approaches seem to abide by the simulated time series patterns, although there is the omission of noises, predominantly for smaller window length. However, modified SSA fares better compared to SSA as window length increases. In particular, as L increases the plotted line of the reconstructed series data from SSA is likely to even out in comparison to the modified SSA. In summary, the analysis of simulated data appears to suggest that $L = T\ /\ 5$ is suitable based on skewed and short time series of rainfall data. By comparison methods, the separability between trends and noise components using two different algorithms is determined. It is clearly shown that the modified SSA can improve the separability and the reconstruction series accuracy.

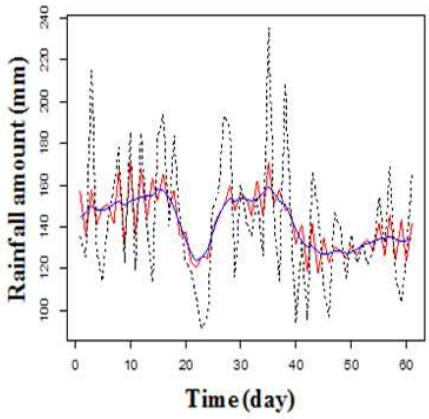
Root mean square error (RMSE) is used for comparing the performance of both methods in reconstructed time series components based on simulated data. Table 1 shows that average RMSE for reconstructed time series components of modified SSA is much smaller than SSA for each L. Additionally, this demonstrates that the modified SSA highlights the least average RMSE at window length, $T/5$.

TABLE I
THE PERFORMANCE OF COMPARISON BETWEEN MODIFIED SSA AND SSA ON SIMULATED DATA BY USING RMSE

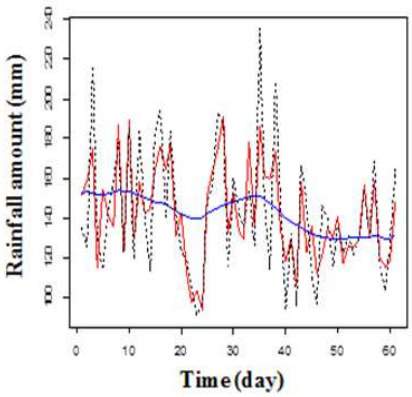
L	Modified SSA	SSA
$T/20$	20.01	24.38
$T/10$	24.47	27.53
$T/5$	**18.54**	29.74
$T/2$	30.75	57.5



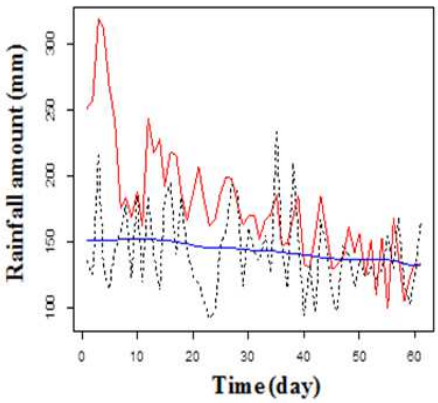
(a) $L = {}^{T}/_{20}$



$(b)\ L = {}^{T}/_{10}$



(c) $L\ = {}^{T}/_{5}$



(d) $L = {}^{T}/_{2}$

Fig.2 Simulated data (dash line), modified SSA (red line) and SSA (blue line) for trends of reconstruction series data with different *L*

For modified SSA, Fisher g-test is used to choose the components for aggregation in the reconstructed series of the data set. The first step to select the components in the time series data through the seeking of the peak in the periodogram. Fig.3 illustrates that the average periodogram shows all the components in time series data exhibit peak and do not shows a flat line.

Then, the statistical significance can be assessed through *g-test* to confirm the result from the plot in average periodogram. The *p-value* corresponding to g-statistic rejected the null hypothesis if the *p-value* is less than the significance level, α. From these results, the *p-value* for all components is less than $\alpha = 0.05$. Thus, the null hypothesis is discarded for all components in which the spectral peak in Fig.3 is statistically significant. Thus, every component in the time series data should be included in the grouping step for reconstructed series components.

IV. Conclusion

A commonly-known time series method to extricate components like trend and noise from the brief time series data is called SSA. The appropriate window length and adjustments on the coinciding singular values attained from the decomposed time series matrix based on a restricted singular value decomposition (SVD) using iterative oblique SSA (Iterative O-SSA) is proposed. In addition, a guided clustering method called Robust Sparse K-means (RSK-means) to discriminate the eigenvectors from this iterative procedure is suggested to identify the components of noise and trend more objectively. To summarize, the analysis of simulated data appears to suggest that $L = T/5$ is suitable based on skewed and short time series of rainfall data.

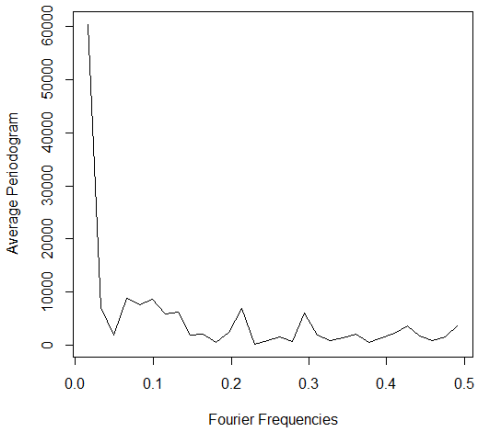


Fig.3 Average periodogram for original series torrential rainfall data in Peninsular Malaysia

It can be proved by the plot of the extracted trend for modified SSA as it seems to be abiding by the pattern of the initial time series rainfall data even though there is the omission of noise components specifically for $L = T/5$ so that the abnormal, heavy downpour could be detected which could trigger the torrential rainfall events. The separability between trends and noise is established through the application of two distinct algorithms for the comparison methods. As evidence, modified SSA is potentially capable of enhancing the accurateness of the reconstruction series and component separability.

Acknowledgement

We gratefully thank the Universiti Pendidikan Sultan Idris especially for the research financing via the 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, 2005.

[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, pp. 1-2, Sept. 2014.

[3] A. 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, pp. 70-78, April 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, pp. 1-5, June 2008.

[5] H. Hu, S. Guo, R. Liu, and P. Wang, "An adaptive singular spectrum analysis method for extracting brain rhythms of electroencephalography," PeerJ, June, 2017. [Online]. Available: DOI 10.7717/peerj.3474.

[6] S.M. Shaharudin, N. Ahmad, F. Yusof, "Effect of window length with singular spectrum analysis in extracting the trend signal on rainfall data," *AIP Conf. Proc*., vol. 1643, pp. 321-326, 2015.

[7] N. Golyandina, V. Nekrutkin, and A. Zhigljavsky, "*Analysis of time series structure:ssa and the related technique*," 1th ed., New York: Chapman &Hall/CRC, 2001.

[8] S.M. Shaharudin, N. Ahmad, F. Yusof, X.Q. Yap, "The comparison of t-mode and pearson correlation matrices in classification of daily rainfall patterns in Peninsular Malaysia," *Matematika*, vol.29, pp. 187-194, 2013.

[9] E. Barton, B. Al-Sarray, S. Chretien, K. Jagan, "Decomposition of dynamical signals into jumps, oscillatory patterns and possible outliers," Mathematics, vol. 6(7), pp.124-137, July 2018.

[10] S.M. Shaharudin, N. Ahmad, N.H. Zainuddin, N.S. Mohamed, "Identification of rainfall patterns on hydrological simulation using robust principal component analysis," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 11, pp. 1162-1167, Sept. 2018.

[11] H-K. Cho, K.P. Bowmanand, and G.R. North, "A comparison of gamma and log-normal distribution for characterizing satellite rain rates from the tropical rainfall measuring mission,"*Journal of Applied Meteorology*, vol. 43, pp.1586-1597, 2004.

[12] R.H. Al-Suhili, and R. Khanbilvardi, "Frequency analysis of the monthly rainfall data at sulaimania region, Iraq," *American Journal of Engineering Research*, vol.3, pp. 212-222, 2014.

[13] N.O.S. Alghazali, and D.A.H. Alawadi, "Fitting statistical distributions of monthly rainfall for some Iraqi stations," *Civil and Environmental Research*, vol. 6, pp. 40-47, 2014.

[14] O.I. Traore, L. Pantera, N. Favretto-Cristini, P. Cristini, S. Viguier-Pla, and P. Vieu, "Structure analysis and denoising using Singular Spectrum Analysis: Application to acoustic emission signals from nuclear safety experiment," *Measurement*, vol.104,pp.78-88, Feb. 2017.

[15] M. B. Priestley, *Spectral Analysis and Time Series*, London: Academic Press., 1981.

[16] Y. Kondo, M. Salibian-Barrera, and R. Zamar, A Robust and Sparse K-means Clustering Algortihm, arXiv:1201.6082v1.

[17] S. M. Shaharudin et al., "Modified singular spectrum analysis in identifynig rainfall trend over Peninsular Malaysia," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 15, No. 1, pp. 283-293, 2019.