ISSN 2278-3091
Volume 9, No.1.4, 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse2691.42020.pdf
https://doi.org/10.30534/ijatcse/2020/2691.42020

Predictive Modelling of Covid-19 Cases in Malaysia based on Recurrent Forecasting-Singular Spectrum Analysis Approach

Shazlyn Milleana Shaharudin[1], Shuhaida Ismail[2], Mou Leong Tan[3], Nur Syarafina Mohamed[4] and Nurul AininaFilzaSulaiman[5]
[1,5]Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Malaysia
[2]Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia, Malaysia
[3]Geography Section, School of Humanities, UniversitiSains Malaysia, Malaysia
[4]Malaysian Institute of Industrial Technology, Universiti Kuala Lumpur, Pasir Gudang, Johor, Malaysia
[1]shazlyn@fsmt.upsi.edu.my
[2]shuhaida@uthm.edu.my
[3]mouleong@usm.my
[4]nursyarafina@unikl.edu.my
[5]aininafilza@gmail.com

ABSTRACT

Covid-19 (novel coronavirus) was discovered in Wuhan, China in December 2019, and has since affected millions' lives worldwide. By 10th April 2020, Malaysia reported more than 4,000 outbreak cases, the highest in Southeast Asia. Recently, a forecasting model was developed to measure and predict daily Covid-19 cases in Malaysia for the coming 10 days using previously-confirmed cases. A Singular Spectrum Analysis-based forecasting model that discriminates noise in a time series trend is introduced. The key concept of the proposed model, RF-SSA, is improving the efficiency of recurrent SSA by establishing L and r parameters via several tests. The RF-SSA model assessment is based on the World Health Organization's official Covid-19 data to predict the daily confirmed cases after 10th April until 20th April, 2020. These results show that the parameter L= 4 ($T/20$) for RF-SSA model was suitable for short time series outbreak data and the appropriate number of eigentriples to obtain is important as it influences the forecasting result. Evidently, the RF-SSA has over-forecasted the cases by 0.36%. This indicates RF-SSA's competence to predict the impending number of Covid-19 cases. Nevertheless, enhanced RF-SSA algorithm should to be developed for higher effectivity in capturing any extreme data changes.

Key words: Covid-19; Singular Spectrum Analysis (SSA); Recurrent Forecasting (RF); forecasting, trend, window length; eigentriples

1.INTRODUCTION

In 2020, Malaysia witnessed the outbreak of a virus called Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) or Covid-19 that is highly infectious towards human's respiratory system, hepatic system, gastrointestinal system, as well as neurological disorders. This virus could be spread between humans as well as livestock, and many types of wild animals like birds, bats, and mice [1–2]. Corona virus is correlated with the family of MERS and SARS virus [3]. Belonging to the coronavirus family, this novel virus type is accountable as a cause for mild to moderate colds. SARS-CoV-2 could cause severe acute respiratory illnesses which result in fatality for various cases. As described by [4], the symptoms of Covid-19 are cough, fever, nose congestion, shortness of breath, and occasionally, diarrhea. In Malaysia, the virus started to spread swiftly by the end of January 2020. Since then, the Crisis Preparedness Response Centre (CPRC) of Malaysia's Ministry of Health began to record and report the cases. These Covid-19 statistics updates regarding the total of active cases, recoveries and casualties could be attained daily from the health ministry website.

The worst scenario leading from SARS-CoV-2 infection to individuals is fatality. Conversely, there has been lack of information on the mechanism of the spread of the virus or how it affects a patient. The Centers for Disease Control and Prevention or known as CDC had verifiedthat Covid-19 is transmittable from humantohuman on the 30th of January 2020. As noted by the CDC, Covid-19 could spread via air, close contact with infected patients, as well as contact with any surfaces or objects which has the particles of the virus. The common incubation period of COVID-19 is within the range of 2 and 14 days, or longer,with the average of 5 days [5-6]. Previously, Zhao et al. [7] had suggested a mathematical model for the purpose of approximating the actual total of Covid-19 cases, including the unreported ones, around the first fortnight in January 2020. It had been deduced that the unreported cases count was a total of 469 between 1 and 15th of January 2020. Additionally, the estimation of cases from 17th of January 2020 onwards revealed that the case numbers astonishingly encountered a 21-fold upsurge. [8]

forecasted that towards the end of February, China's pandemic will hit the highest point and only halt by the end of April by combining the SEIR model and the approach of the machine-learning artificial intelligence (AI). Subsequently, Tang et al. [9] suggested a mathematical model which is capable of estimating the risk of Covid-19 transmission. Based on this, the potential number of the basic reproduction was determined as 6.47. It had also forecasted a seven-day confirmed cases total for the time interval of 23rd to 29th of January 2020. Consequently, the estimated peak was after two weeks from the initial date of 23rdof January 2020. As stated in [10], in order to estimate the prolonged Covid-19 human-to-human transmissions, the data for 47 patients had been observed. It was pointed out by the author that the transmission rate is 0.4. However, if the duration between the symptom detection and the patient hospitalization was halved from the tested study data, the transmission rate could be reduced to 0.012. From the study of [11], an estimation of SIR model was exhibited for the Covid-19 outbreak in Malaysia for predict short-term daily Covid-19 cases. It was pointed out by the author that the transmission rate is 0.22 with the assumption that an infectious person will be able to infect or spread the disease to 1 person (on average) in four days' time. The transmission rate which corresponds to a hypothesized scenario whereby an individual will infect another individual within a 4-day interval, should not be taken lightly. Furthermore, a one-to-one transmission on an interval of 4 days can be seen to be rather conservative.

In this paper, the prediction model was developed based on Singular Spectrum Analysis (SSA) which called it as Recurrent Forecasting (RF-SSA) to predict the new daily confirmed Covid-19 cases for a short-term period. Recently, the progress of singular spectrum analysis (SSA) becomes anengaging alternative for which, it is capable to reduce noise substantively, deal with the trend components, and reveal the temporal structure of the data excluding the preliminary manipulation [12]. SSA was used in this study as a base approach for developing the forecasting model. Generally, SSA specifies a representation of a univariate time series which is transformed in terms of the eigenvalues and eigenvectors of a trajectory matrix. This SSA method is a multidimensional analogue of a principal component analysis adapted to time series. The function of SSA is separating the time series data into categories of trend, seasonal and noise through the decomposition of its time series eigen and their reconstruction into a group selection [13]. However, the separation of the components in this approach depends on the parameter choice which is the selection of window length, L to form the trajectory matrix and identifying the number of leading components, r based on eigenvector plot [14]. This separation is very important in this model to ensure that the trend, seasonal and noise components are easily separated. Principally, this study's objectives are finding the best option of window lengths of L and r, as well as forecasting daily Covid-19 cases by employing the RF-SSA algorithm. Till date, the usage of SSA models is rare in the analysis of epidemiological data. The advantage using this model compared to other is its convenience andit requires nocondition of models of time series and trend. It also permits the extraction of the trendswith the occurrence of noise and oscillations, andonly two parameters are needed to identify in order to get the accuracy and flexibility for prediction outcome [15].

The following sections will discuss more detailed description of the data, specifically in section two, followed by several sections namely the methodology, results and discussions and finally, the conclusion.

2.DATA

Daily Corona virus Disease (Covid-19) prevalence data from 25th of January 2020 until 10th of April 2020 were collected from the records of the Ministry of Health Malaysia. As this Covid-19 is a newly-found virus; therefore, there is no available Covid-19 data from the previous year. The suspected Covid-19 cases were diagnosed by Reverse Transcription Polymerase Chain Reaction (RT-PCR) technique and confirmed as Covid-19 case-counts. All fully-anonymized, laboratory-confirmed cases were abstracted on Covid-19 in which 4,346 cases represented Covid-19 infection in 16 states in Malaysia as recorded by the Ministry of Health Malaysia.

Figure 1 illustrates the total positive cases for Covid-19. The figure shows a significant spike in the number of positive cases which is a result of the 2nd wave of Covid-19 pandemic in Malaysia. With this substantial number, Malaysia Government has announced the Movement Control Order (MCO) which took place on 18th of March until 31st of March 2020. The MCO was later extended till to the 4th phase.

Figure 2 illustrates the observed numbers cases for Covid-19 for the last 77 days in Malaysia. The Ministry of Health (MOH) had categorized four zones of Covid-19 areas in Malaysia according to the areal cases number. According to the National Security Council (MKN), the four zones are i) green zone for areas with no positive case, (ii) yellow zone for areas with one to 20 positive cases, (iii) orange zone for areas with 21 to 40 positive cases and (iv) red zone for areas with more than 40 positive cases [16].

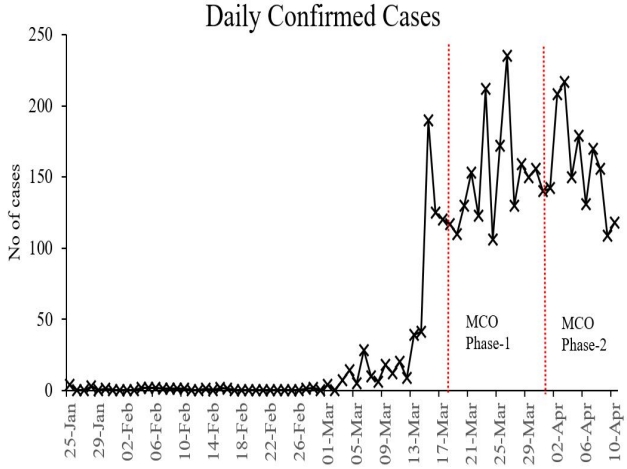


Figure 1: Covid-19 Daily confirmed cases in Malaysia from 25th Jan 2020 until 10th April

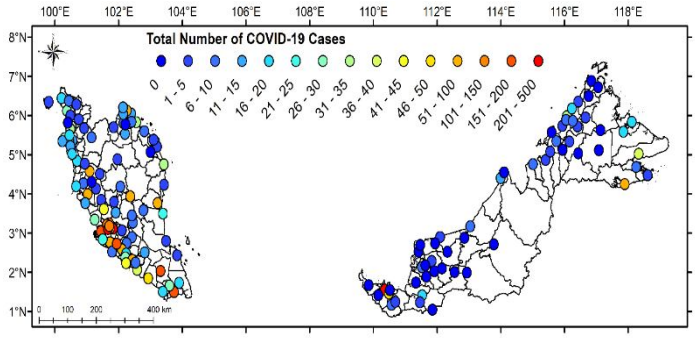


Figure 2: State classification according to number of Covid-19 cases in Malaysia.

3.MATERIALS AND METHODS

This section explains the specifics of Singular Spectrum Analysis model and its components.

A. Singular Spectrum Analysis (SSA) Model

Singular spectrum analysis (SSA) is a model-free approach which can be applied to all types of data, whether it is gaussian or non-gaussian, linear or nonlinear and stationary or non-stationary [17]. Daily Covid-19 data can be decomposed into a number of additive components via SSA which could be defined in the forms of trend, seasonal and noise components [18]. The possible application areas of SSA are diverse [19-20]. SSA comprises two complementary stages known as the stages of decomposition and reconstruction [21].

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 the eigen time series data.

Step I: Embedding. The first step in basic SSA algorithm is embedding step which refer to constructing a one dimensional series i.e. univariate vector, $\mathbb{Y}_T = \{y_1, y_2, \dots, y_T\}$ to a multidimensional series contain in a matrix, $\mathbf{X} = (X_1, \dots, X_K)$called the trajectory matrix as shown in Equation (5.1). The rows and columns of **X**are subseries of the original one-dimensional time series data. The dimension of the trajectory matrix is called the window length, L which ranges from $2 \le L \le {}^{T}/_{2}$. The columns $X_1, \dots, X_K$ of the trajectory matrix, **X** are called lagged vectors, $K = T - L + 1$.

$$\mathbf{X} = (X_1, \dots, X_K)\begin{pmatrix} y_1 y_2 y_3 & \cdots & y_K \\ y_2 y_3 y_4 & \cdots & y_{K+1} \\ y_3 y_4 y_5 & \cdots & y_{K+2} \\ \vdots & \vdots & \ddots & \vdots \\ y_L y_{L+1} y_{L+2} & \cdots & y_T \end{pmatrix} \quad (1)$$

Step II: Singular Value Decomposition (SVD).In the second step, trajectory matrix inStep Iis decomposed to obtain its eigen time series based on their singular values using Singular Value Decomposition (SVD). The SVD of the trajectory matrix, **X** is represented as

$$\mathbf{X} = U^T \Sigma V \quad (2)$$

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 known as left singular vectors, v_i are the right singular vectors whileσ_i are the singular values. Let $\mathbf{S} = \mathbf{X}\mathbf{X}^\mathbf{T}$ where the singular values be arranged in descending order such that ($\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_L$). Let $d = max\{i, \text{such that} \sigma_i > 0\}$. $V_i = X^T U_i / \sqrt{\sigma_i}$ $(i = 1, \dots, d)$, then, the SVD of the trajectory matrix **X** can be written as

$$\mathbf{X} = \mathbf{X_1} + \cdots + \mathbf{X_d} \quad (3)$$

where$\mathbf{X_i} = \sigma_i \mu_i v_i^T$. Note that, the matrices of $\mathbf{X}_i$ are called elementary matrices if $\mathbf{X_i}$ has rank one. The collection (σ_i, u_i, v_i) is identified as 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 the grouping step, the trajectory matrix is split into two groups in reference to the trend and noise components. The indices set $\{1, \dots, L\}$ is segregated into m disjoint subsets $I_1, \dots, I_m$, conforming to dividing the elementary matrices into m groups. Set $I = \{i_1, \dots, i_p\}$, then the resultant matrix $\mathbf{X}_I$ is defined as

$$\mathbf{X}_I = \mathbf{X}_{i1} + \cdots + \mathbf{X}_{ip} \quad (4)$$

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

$$\mathbf{X} = \mathbf{X}_{I1} + \cdots + \mathbf{X}_{Im} \quad (5)$$

where a number of m resultant matrices represents the trajectory matrix. The range of the sets $I = I_1, \dots, I_m$ is called eigentriple grouping.

Step 2: Diagonal averaging. The concluding step in SSA transfigures into a different series of length T from each matrix of the grouped decomposition (5.5).

Let $\mathbf{Z}$ be an $L \times K$ matrix with elements $z_{ij}, 1 \leq i \leq L, 1 \leq j \leq K$. 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 making the diagonal averaging, we transfer the matrix $\mathbf{Z}$ into the$z_1, \dots, z_T$ using the formula

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

Diagonal averaging in Equation (5.6) applied to a resultant matrix $\mathbf{X}_{Ik}$ produced reconstructed series $\tilde{\mathbb{Y}}_T^{(k)} = (\tilde{y}_1^{(k)}, \dots, \tilde{y}_T^{(k)})$. Hence, the initial series $\mathbb{Y}_T = \{y_1, y_2, \dots, y_T\}$ decomposes to a figure of m reconstructed series, $y_t = \sum_{k=1}^{m} \tilde{y}_t^{(k)}$. The outcome of the elementary grouping is the reconstructed series that will be known as the elementary reconstructed series.

B. Forecasting with SSA Model

In making the SSA forecasting, a fundamental condition is that the time series satisfies a linear recurrent formula (LRF). A time series $Y_T = (y_1, \dots, y_T)$satisfies LRF of order d if:

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_d y_{t-d`}, t = d + 1, \dots, T \quad (7)$$

In this study, Recurrent SSA is used for forecasting purpose since it is a popular approach when predicting data [22-23]. These algorithms described as follows and further details can be found in [24].

Let us assume that U_j^{∇} is the vector of the first $L-1$ components of the eigenvector U_j and π_j is the last component of $U_j (j = 1, \dots, r)$. Denoting $v^2 = \sum_{j-1}^{r} \pi_j^2$ we define the coefficient vector $\Re$ as:

$$\Re = \frac{1}{1-v^2}\sum_{j=1}^{r} \pi_j U_j^{\nabla} \quad (8)$$

In consideration of the prior notation, the recurrent SSA (RSSA) forecasts($\hat{y}_{T+1}, \dots, \hat{y}_{T+M}$) can be attained by

$$\hat{y}_i = \begin{cases} \tilde{y}_i, & i = 1, \dots, T \\ \Re^T Z_i, & i = T+1, \dots, T+M \end{cases} \quad (9)$$

where, $Z_i = [\hat{y}_{i-L+1}, \dots, \hat{y}_{i-1}]^T$ and $\tilde{y}_1, \dots, \tilde{y}_T$, are the reconstructed time series values and be attained from 4th step as above.

C. SSA Parameter Selection

The trend extraction from the original time series data depends on the selection of window length, L to form the trajectory matrix in SSA.An improper values selection for the parameter L yield unfinished reconstruction hence could potentially bring about misleading forecasting results. According to [25], L should be large enough but not greater than half of the number of observations understudy at $\frac{T}{2}$. However, the appropriate selection of window length is dependent on the current problems as well as the structure of the time series data [26]. Generally, there is no rough guide to determine the proper L in data set. In addition, the separability conditions for shorter time series could be restrictive due to the singular value decomposition properties used in estimating the signal component in SSA. Therefore, in this study, several L, $\frac{T}{2}, \frac{T}{5}, \frac{T}{10}, \frac{T}{20}$ were investigate on Covid-19 data based on performance error which is root mean square error (RMSE).

Another parameter that need to be considered when using SSA approach is the amount of eigentriples employed for the reconstruction r by using eigenvector plot. This plot reflects the eigenvector of the SVD of the trajectory matrix for the time series data. Inspect the one-dimensional graphs of eigenvectors where it would help to identify the trend components. Note that the trend has complex form when the trend and noise components were not properly distinguished. It is highly possible that a lack of separability caused the presence of the mix-up between the components. This information can be used as a guideline to identify proper grouping for the component's separation of the trend and noise appropriately. Apart from that, it could also reflect a connection between the stages of decomposition and reconstruction.

D. Evaluating Separability in Time Series Data

A key concept when studying SSA is the separability. It dictates the extent to whichdistinctive time series components are distinguishablebetween one another so that further analysis could be conducted significantly. Based on [27] literature review, when working with SSA method in numerous study fields, separability becomes a vital mean. The separability impact possibly results in the appropriate decomposition and extraction of component. W-correlation is ameansto measure the separability of two distinctive components of reconstructed time series.

W-correlation is known to be the weighted correlation among the components of reconstructed time series which provides highly useful knowledge for the separation and identification of group for reconstructed components [28]. The elements of the time series terms are indicated by the weights into trajectory matrix. These would range from the absolute values of 0 to 1. Components that separate properly would have notable tendency towards zero whereas components which did not basically would slant towards one. Furthermore, w-correlation matrix also checks the

grouped decomposition among the reconstructed components. The following is the w-correlation matrix formulation:

$$\rho_{12}^{w} = \frac{\langle X^{(1)}, X^{(2)} \rangle w}{\|X^{(1)}\| w \|X^{(2)}\| w} \quad (10)$$

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,

$$w_i = \begin{cases} i+1 & for\ 0 \le i \le L^* - 1, \\ L^* & for\ L^* \le i \le K^*, \\ T - i & for\ K^* \le i \le T - 1. \end{cases} \quad (11)$$

The graphic illustration of w-correlation could be made in the white-black scale where white represents small correlation while black represents correlation between the series components which are close to 1.

4.RESULTS AND DISCUSSION

E. Decomposition and Reconstruction

The first stage in this study is decomposing Covid-19 data into components, facilitated by the SSA model. This decomposition by SSA requires identifying the parameter pair (s, L). The choice of L represents a compromise between information content and statistical confidence. A fitting L value couldcoherently determine the distinct oscillations unknown in the initial signal.

The performance of the SSA results were determined by evaluating its weighted correlation i.e. w-correlation at distinct window length, *L*. The w-correlation as explained in the methodology section calculated the separability among the reconstructed time series components of trend, seasonal and noise. A number of selections of L that were $L = T/2, T/5, T/10$ and $T/20$, that represent L =4,8,15,38respectively for T based on 75 daily cases on Covid-19 data were selected. These scales were chosen to fit the data of the time series as well as striking a balance in achieving a proper lag vector sequence.

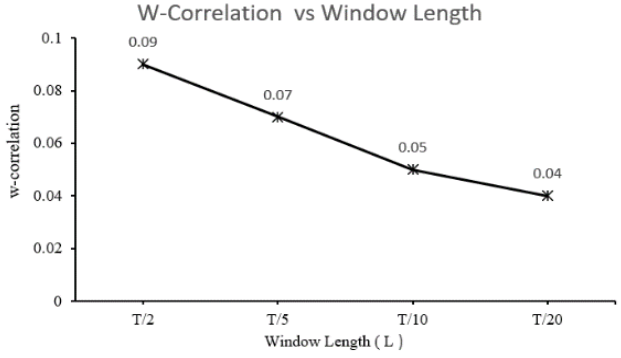


Figure 3: Effect of w-correlation based on SSA using Covid-19 data at different window lengths

Figure 3 displays the w-correlation through the SSA from day-to-day cases of Covid-19 data at different window lengths. As can be seen from the plot, the w-correlation shows a declining trend as the total of window length declines for SSA approach. The correlations among trend and other components need to be near to zero for the extraction of trend. This means that distinct window lengths have a certain effect on the component's separability. Moreover, it also displays that SSA directed to the least w-correlation at window length, L= $T/20$ indicating the clearest separability among the reconstructed components since it is the nearest to zero.

The graphs in Figure 3(a – d) illustrated that the heat-plot of different window lengths, L according to the w-correlations from SSA approach. Based on [29], the heatplot of w-correlation for the reconstructed components on a white-to-black grading scale that corresponds to the range of correlation of 0 to 1. Great correlations values absolute values between reconstructed components showed possible components gathering into a group while corresponding to the same component. As illustrated in Figure 3, the strength of the w-correlation between two components is represented by the shade of each square. Meanwhile, Figure 3 (a-c) indicates how the components have the tendency of being correlated to more other components although the correlation is occasionally subtle. Subsequently, this denotes that the components of trends are still, to some extent, mixed with the noise and seasonal components in SSA and it was rectified by the small window length, $L = 6$ that is evidently demonstrated in Figure 3(d) for better improvement of separability.

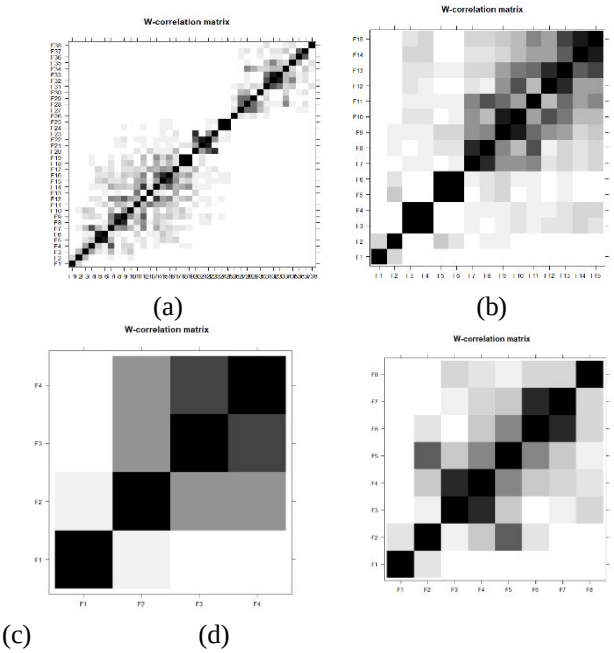


Figure 4: The w-correlation plot using SSA with different (a) $L = 38$, (b)$L = 15$, (c) L=8, (d) L=4.

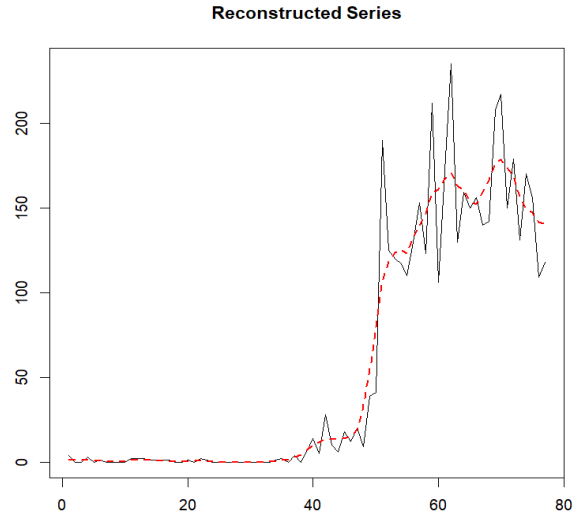
Root mean square error (RMSE) is used for evaluating the performance of L. Table 1 presents the reconstructed time series components at $L = T/20$ which has the

smallest RMSE compared to other L. It is noted that higher values of RMSE is obtained in this study due to high model variance when number of samples is small [30]. In summary, the analysis of daily Covid-19 cases data appears to suggest that $L = T/20$ is suitable based on short time series of the outbreak data.

Table 1: The performance of comparison prediction model based on SSA for several L.

Window Length, L	RMSE
$T/2 = 38$	33
$T/5 = 15$	29
$T/10 = 8$	24
$T/20 = 4$	20

Figure 5 shows the plot of form of four leading eigenvectors. The eigenvector plot is helpful when



selecting the suitable group for the time series data components, particularly in the separation of the trend, seasonal and noise components. This information is applicable for more detailed analysis in the RF-SSA grouping step. To identify the trend components through eigenvector plot, the trend component and the seasonality component consists of sine waves are denoted by the slow cycles as demonstrated in the graph in terms of higher frequencies while the noise component is denoted by the saw-tooth of the graph with lower frequencies. The leading eigenvector has nearly continual coordinates and hence, corresponding to a pure smoothing by the Bartlett filter [31-32]. The reconstruction result by each of the four eigentriples is displayed in Figure 5. The two figures substantiate how the first and second eigentriple are compatible to the trend, while the other eigentriples consist of noise components and hence are trend-irrelevant. In addition, it confirmed that Covid-19 data in Malaysia are not influenced by the seasonality since both figures not illustrated by the periodogram.

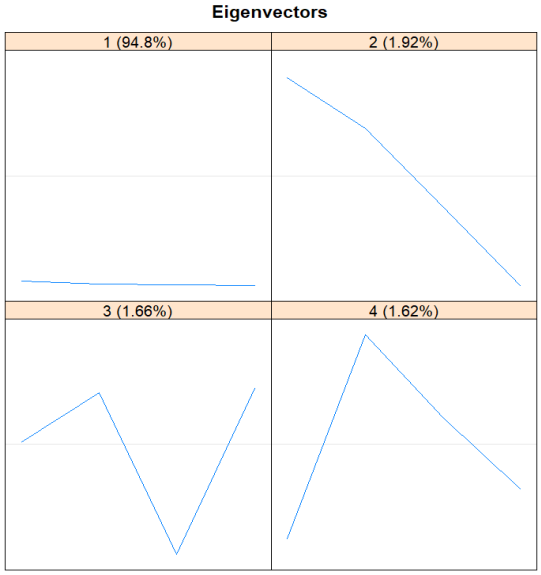


Figure 5: Eigenvector plot obtained by SSA.

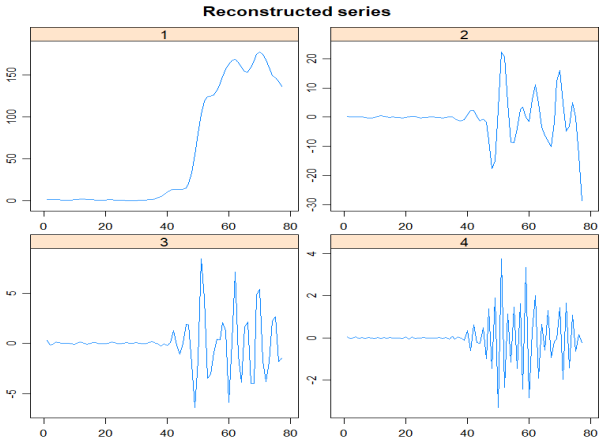
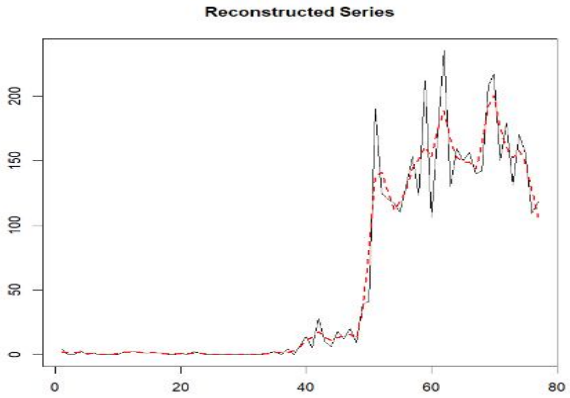


Figure 6: First stage: elementary reconstructed series (L=4)

(a)



(b)

Figure 7: Plot of daily Covid-19 cases of reconstructed components from extracted trends using SSA at (a)L=4, ET 1 (b) L=4, ET 2

Figure 7 demonstrates the components of reconstructed time series plot from the trend that was extricated through RF-SSA for daily Covid-19 cases in Malaysia meanwhile Figure 8 exemplifies the reconstructed time series components plot from the extracted trend using RF-SSA for cumulative Covid-19 cases in Malaysia. The trend component of the time series data is employed in observing the occurrence of the cases trend and pattern as it was randomly-tabulated as per daily cases, as illustrated in Figure 7 and cumulative cases in Figure 8. The trend from Figure 7-8 (a) and Figure 6 (a) are precisely generated by a leading eigentriple, coinciding with the first reconstructed component in Figure 6. Meanwhile, the trend from Figure 7-8 (b) are precisely generated by both leading eigentriples, coinciding with the first and second reconstructed component in Figure 6. The straight and dashed plot linescorrespondingly refer to the original time series Covid-19 data and the reconstructed series according to the extracted trend components from SSA.The plots of the reconstructed time series components that were produced by both leading eigentriple are abide by the of the original time series rainfall data even though there is noise components omission specifically for $L = 4$ for both daily and cumulative Covid-19 cases in Malaysia.

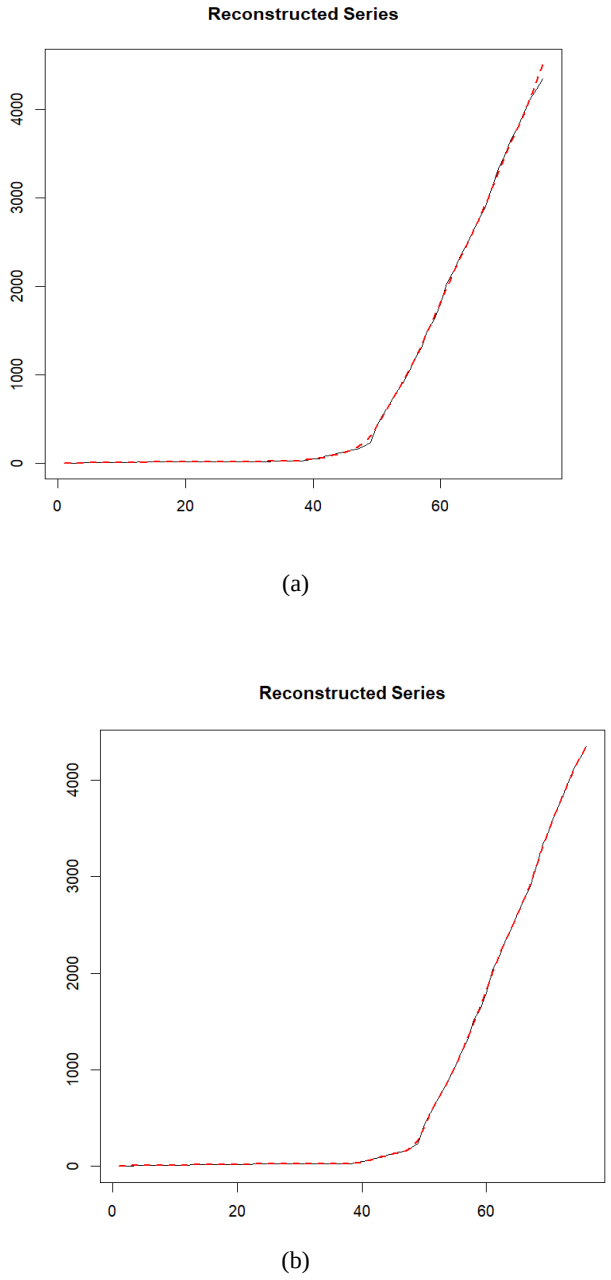


Figure 8: Plot of daily Covid-19 cases reconstructed components from extracted trends using SSA at (a) L=4, ET1 (b) L=4, ET2

F. Forecasting Daily Covid-19 Cases using SSA

As mentioned in the previous section, Malaysian daily Covid-19 cases were forecasted using SSA model. The SSA forecasting algorithm known as recurrent forecasting was used to forecast future cases starting from 11th of April 2020 to 20th of April 2020. At the time this experiment was conducted, the historical cases from 25th of January 2020 until 10th of April 2020 were used and the future 10-days ahead of covid-19 cases were predicted accordingly. Figure 9 illustrates the confirmed cases from 25th of January 2020 to 10th of April 2020 and the forecasted daily cases until 20th of April 2020.

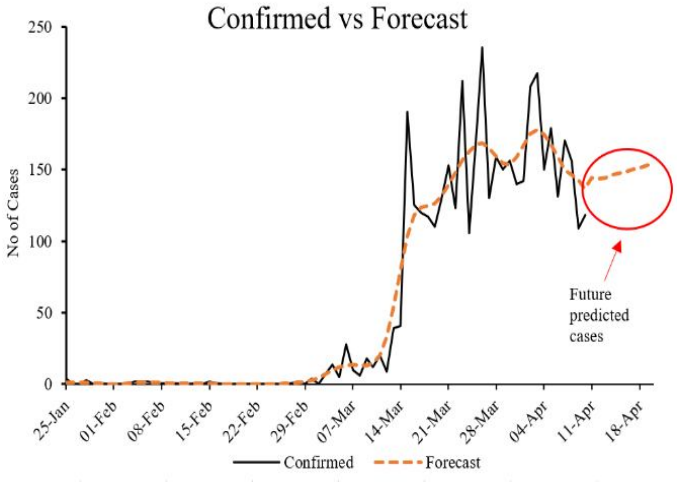


Figure 9: New confirmed and prediction cases of Covid-19 in Malaysia

As the number of daily cases of Covid-19 was small, Figure 8 displays a noticeable but faint decreasing pattern from 27th of March 2020 onwards. One of the contributing factors for slight decreasing trend was due to the Movement Control Order (MCO) announced by the Malaysian Government which took place on 18th of March to 31st of March 2020. The prediction plot using RF-SSA as depicted in Figure 9 shows that there was a general pattern of nonlinear increasing trend in the newly-confirmed daily Covid-19 cases in Malaysia. The projection and estimate daily cases of Covid-19 obtained are impacted by the

definition of the case reported to CPRC daily, the large number of pending result test daily were definitely influential to a non-consistent increase in the number of confirmed cases. The evidence for the prediction cases increase is supported by several of the biggest clusters identified by the Ministry of Health Malaysia such as Seri PetalingTabligh Cluster, Wedding Kenduri in Bandar BaruBangi, Seri Petaling Sub-Cluster in Rembau, Italy Cluster in Kuching, Sarawak and Church Fellowship Cluster in Sarawak. Furthermore, the new confirmed cases will be extremely spiking as the target of biology sample taken were directly from highly susceptible infected population.

Moreover, the experimental result showed that the RF-SSA obtained Mean Absolute Error (MAE) of 10.378, and 19.9 for Root Mean Square Error (RMSE). Meanwhile, Pearson Correlation (*r*) of 0.96, close to 1.0, indicates that the model has good correlation between the confirmed and predicted cases. Finally, the Mean Forecast Error (MFE) shows that the RF-SSA algorithm has over-forecasted daily Covid-19 cases by 0.36%. The 10-days ahead prediction of Covid-19 cases is shown in Figure 9.

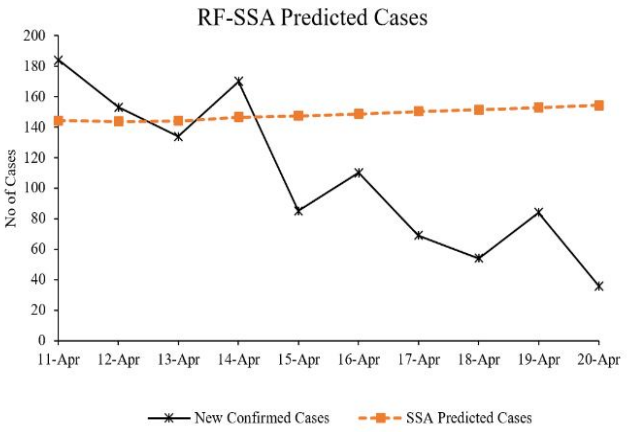


Figure 10: RF-SSA 10-days Ahead vs New Confirmed Cases for Covid-19 in Malaysia

Above figure shows the predicted values of 10-days ahead using RF-SSA algorithm against new confirmed cases of Covid-19 in Malaysia. Despite an encouraging statistical finding based from the historical data, and lower over-forecast value, the RF-SSA is unable to capture the sudden drop in Covid-19 cases, which is considered to have never happened before. The sudden drop seen on this pandemic case is highly likely due to the MCO that has been extended to phase-4, planned to end by this coming 12th of May 2020.

5.CONCLUSION

This paper studies the applicability of RF-SSA model in predicting the Covid-19 cases in Malaysia. The application of this model is specifically advantageous for the health authorities in terms of flattening the curve by preparing a prompt and efficient strategies. Moreover, this model allows the health authorities to comprehend the outbreak pattern better. It was found that the pattern follows the RF-SSA model which can be applied to forecast the outbreak cases growth pattern in Malaysia. By using this model, the selections of the parameter are the choice of window length, L and the total of eigentriples employed for reconstruction, *r*. These results show that the parameter L= 4 ($T/20$) was suitable to use in short time series outbreak data and the appropriate number of eigentriples to obtain is important which will give effect on the forecasting result. Overall, the results showed that the RF-SSA model is able to forecast this pandemic with reasonable accuracy as the model has over-forecasted by 0.36% with high correlation values between confirmed and predicted cases. However, RF-SSA model is unable to capture the sudden drop in Covid-19 cases, likely due to the MCO which has been extended to 12th May 2020. To improve the accuracy of the model, more information is required to have a better prediction for Covid-19 cases for a long period. In the meantime, case definition and data collection must be maintained in real time to improve the RF-SSA for further study. It is suggested that the RF-SSA model is enhanced in order for the model to be able to capture sudden and rapid changes in the dataset.

ACKNOWLEDGEMENTS

This research has been carried out under Fundamental Research Grants Scheme 2019-0132-103-02 (FRGS/1/2019/STG06/UPSI/02/4) provided by Ministry of Education of Malaysia.

REFERENCES

1. Y. Chen, Q. Liu, D. Guo. **Emerging coronaviruses: Genome structure, replication, and pathogenesis**. *J. Med. Virol*. 2020. [CrossRef]
2. X.Y Ge, J.L Li, X.L. Yang, A.A. Chmura, G. Zhu, J.H. Epstein, J.K.Mazet, B. Hu, W. Zhang, C. Peng, et al. **Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor**. *Nature*, vol. 503, pp. 535–538, 2013. [CrossRef]
3. K. Dheeraj.**Analysing COVID-19 news impact on social media aggregation**. *International Journal of Advanced Trends in Computer Science and Engineering*, vol.9, no. 3, pp. 2848–2855, 2020. https://doi.org/10.30534/ijatcse/2020/56932020
4. Coronavirus Website - Ministry of Health. URL http://www.moh.gov.my/index.php accessed on 3rd April 2020.
5. S.A. Lauer, K.H. Grantz, Q. Bi, F.K. Jones, Q. Zheng, H.R. Meredith, A. S. Azman, N.G. Reich, J.Lessler. **The Incubation Period of Coronavirus Disease 2019 (COVID-19) From PubliclyReported Confirmed Cases: Estimation and Application.***Ann Intern Med*., 2020.https://doi.org/10.7326/M20-0504.
6. O. Ghorbel, R. Ayedi, H. Ben Chikha, O. Shehin,& M. Frikha.**Design of a smart medical bracelet prototype for COVID-19 based on wireless sensor networks**. *International Journal of Advanced Trends in Computer Science and Engineering*, vol.9, no.3, pp. 2684–2688,

2020.https://doi.org/10.30534/ijatcse/2020/30932020

7. S. Zhao, S.S. Musa, Q. Lin, J. Ran, G. Yang, W. Wang, Y. Lou, L. Yang, D. Gao, D. He and et al. **Estimating the Unreported Number of Novel Coronavirus (2019-nCoV) Cases in China in the First Half of January 2020: A Data-Driven Modelling Analysis of the Early Outbreak**. *J. Clin. Med.*, vol. 9, pp. 388, 2020. [CrossRef]
8. Z. Yang, Z. Zeng, K. Wang, S.S. Wong, W. Liang, M. Zanin, J. He. **Modified SEIR and AI Prediction of the Epidemics Trend of COVID-19 in China under Public Health interventions**. *Journal of Thoracic Disease***,** vol.12, pp.165, 2020.
9. B. Tang, X. Wang, Q. Li, N.L. Bragazzi, S. Tang, Y. Xiao, J. Wu. **Estimation of the Transmission Risk of the 2019-nCoV and Its Implication for Public Health Interventions**. *J. Clin. Med.*, vol. 9, pp. 462, 2020. [CrossRef]
10. R.N. Thompson. **Novel Coronavirus Outbreak in Wuhan, China, 2020: Intense Surveillance Is Vital for Preventing Sustained Transmission in New Locations**. *J. Clin. Med.*, vol. 9, pp. 498, 2020. [CrossRef] [PubMed]
11. M.R.K. Ariffin and et al**. Malaysian Covid-19 Outbreak Data Analysis and Prediction.** *Institute for Mathematical Research.* 2020. http://einspem.upm.edu.my/covid19maths/file/Report_001%20v13.pdf.
12. N. Golyandina and A. Zhigljavsky.***Basic SSA. In Singular Spectrum Analysis for Time Series***; Springer: Berlin/Heidelberg, Germany, 2013, pp. 11–70.
13. S.M. Shaharudin, N. Ahmad, N.H. Zainuddin. **Modified Singular Spectrum Analysis in Identifying Rainfall Trend over Peninsular Malaysia**. *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 15, pp. 283, 2019.
14. S.M. Shaharudin, N. Ahmad, F. Yusof. **Effect of window length with singular spectrum analysis in extracting the trend signal of rainfall data**. *Aip Proceedings*, 2015, pp. 321.
15. T. Alexandroy.**A Method of Trend Extraction using Singular Spectrum Analysis**. *REVSTAT-Statistical Journal*., vol. 7, pp. 1, 2009.
16. Coronavirus Website - Ministry of Health. URL https://kpkesihatan.com/ accessed on 3rd April 2020.
17. C. Deng.**Time Series Decomposition using Singular Spectrum Analysis**. Master, East Tennessee State University, 2014.
18. M. Biabanaki, S.S. Eslamian, J.A. Koupai, J. Canon, G. Boni, M. Gheysari. **A principal components/singular spectrum analysis approach to enso and pdo influences on rainfall in West of Iran**. *Hydrology Research*, vol. 45, pp. 250-262, 2014.
19. L.J. Rodriguez-Aragon,Zhiglkavsky, **A. Singular Spectrum Analysis for Image Processiong**. *Statistics and Its Interface*, vol. 3, pp. 419-426, 2010.
20. T. Alexandrov, N. Golyandina, A. Spirov.**Singular Spectrum Analysis of Gene Expression Profiles of Early Drosophila Embryo: Exponential-in-Distance Patterns**. *Research Letters in Signal Processing*, Article ID 825758, vol. 18, 2008. doi:10.1155/2008/825758.
21. M.D. Carvalho and A. Rua. **Real-Time Nowcasting the US Output GAP: Singular Spectrum Analysis at Work**. *ISBN 978-989-678-304-4*. Lisboa, Portugal: Banco De Portugal, 2014.
22. D. Danilov. **Principal components in time series forecast**, *Journal of Computational and Graphical Statistics.*, vol. 6, pp.112–121, 1997.
23. D. Danilov and A. Zhigljavsky.**The Caterpillar method for time series forecasting**, in *Principal Components of Time Series: The Caterpillar method;*University of St. Petersburg, St. Petersburg. Russian, 1997, pp. 73–104.
24. N. Golyandina, V. Nekrutkin, A. Zhigljavsky. ***Analysis of Time Series Structure: SSA and related techniques***, Chapman & Hall/CRC, New York–London, 2001.
25. R.A. Mondal, S. Kundu, Mukhopadhyay. **A. Rainfall trend analysis by Mann-Kendall test: a case study of North-Eastern part of Cuttack district, Orissa**. *International Journal of Geology*, vol. 2, pp.70-78, 2012.
26. F.J. Alonso, D.R. Salgado, J. Cuadrado, P. Pintado. **Automatic smoothing of raw kinematics signals using ssa and cluster analysis**. *Euromech Solid Mechanics Conference Lisbon*, pp. 1-9, 2009.
27. N. Golyandina, A. Shlemov, A. **Variations of singular spectrum analysis for separability improvement: Non-orthogonal decompositions of time series**. *Statistics and Its Interface*., vol. 8, pp. 277-294, 2014.
28. N.E. Golyandina, A. Korobeynikoy.**Basic singular spectrum analysis forecasting with R**. *Computational Statistics & Data Analysis.*, vol. 71,pp. 934-954, 2014.
29. H. Hassani.**Singular Spectrum Analysis: Methodology and Comparison**. *Journal of Data Science.*, vol. 5, pp. 239-257, 2007.
30. B. Boehmke, B. Greenwell. **Hands-On machine learning with R**. Broken Sound Parkway NW: Taylor & Francis.
31. N. Golyandina, V. Nekrutkin, A. Zhigljavsky. **Analysis of Time Series Structure: SSA and related techniques**. New York – London: Chapman Hall/CRC, 2001.
32. R. Mahmoudvand, T. Rodrigues. **Forecasting mortality rate by singular spectrum analysis**. *REVSTAT-Statistical Journal*., vol. 13, pp. 193-206, 2015.