frontiers in Public Health

ORIGINAL RESEARCH
published: 14 June 2021
doi: 10.3389/fpubh.2021.604093

Short-Term Forecasting of Daily Confirmed COVID-19 Cases in Malaysia Using RF-SSA Model

Shazlyn Milleana Shaharudin[1]*, *Shuhaida Ismail*[2], *Noor Artika Hassan*[3], *Mou Leong Tan*[4] *and Nurul Ainina Filza Sulaiman*[1]

[1] Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Tanjung Malim, Malaysia, [2] Data Analytics, Sciences & Modelling (DASM), Department of Mathematics & Statistics, Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Malaysia, [3] Department of Community Medicine, Kulliyyah of Medicine, International Islamic University Malaysia, Kuantan, Malaysia, [4] Geoinformatic Unit, Geography Section, School of Humanities, Universiti Sains Malaysia, Gelugor, Malaysia

Novel coronavirus (COVID-19) was discovered in Wuhan, China in December 2019, and has affected millions of lives worldwide. On 29th April 2020, Malaysia reported more than 5,000 COVID-19 cases; the second highest in the Southeast Asian region after Singapore. Recently, a forecasting model was developed to measure and predict COVID-19 cases in Malaysia on daily basis for the next 10 days using previously-confirmed cases. A Recurrent Forecasting-Singular Spectrum Analysis (RF-SSA) is proposed by establishing L and *ET* parameters via several tests. The advantage of using this forecasting model is it would discriminate noise in a time series trend and produce significant forecasting results. The RF-SSA model assessment was based on the official COVID-19 data released by the World Health Organization (WHO) to predict daily confirmed cases between 30th April and 31st May, 2020. These results revealed that parameter $L = 5$ (T/20) for the RF-SSA model was indeed suitable for short-time series outbreak data, while the appropriate number of eigentriples was integral as it influenced the forecasting results. Evidently, the RF-SSA had over-forecasted the cases by 0.36%. This signifies the competence of RF-SSA in predicting the impending number of COVID-19 cases. Nonetheless, an enhanced RF-SSA algorithm should be developed for higher effectivity of capturing any extreme data changes.

Keywords: COVID-19, eigentriples, forecasting, recurrent forecasting, singular spectrum analysis, trend, window length

OPEN ACCESS

Edited by:
Catherine Ropert, Federal University of Minas Gerais, Brazil

Reviewed by:
Viviana Mariani, Pontifical Catholic University of Parana, Brazil
Ma Khan, HITEC University, Pakistan
Mohammed A. a Al-qaness, Wuhan University, China

***Correspondence:**
Shazlyn Milleana Shaharudin
shazlyn@fsmt.upsi.edu.my

Specialty section:
This article was submitted to Infectious Diseases - Surveillance, Prevention and Treatment, a section of the journal Frontiers in Public Health

Received: *08 September 2020*
Accepted: *28 April 2021*
Published: *14 June 2021*

Citation:
Shaharudin SM, Ismail S, Hassan NA, Tan ML and Sulaiman NAF (2021) Short-Term Forecasting of Daily Confirmed COVID-19 Cases in Malaysia Using RF-SSA Model. Front. Public Health 9:604093. doi: 10.3389/fpubh.2021.604093

INTRODUCTION

In 2020, Malaysia has witnessed the outbreak of a virus called Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) or COVID-19 that is highly infectious to human's respiratory system, hepatic system, gastrointestinal system, and neurological disorders. This virus can spread between humans, livestock, and wild animals, such as birds, bats, and mice (1, 2). Belonging to the coronavirus family, this novel virus type is accountable as a cause for mild to moderate colds. The SARS-CoV-2 may cause severe acute respiratory illnesses that result in fatality for various cases. The symptoms of COVID-19 are cough, fever, nose congestion, shortness of breath, and occasionally, diarrhea (3). 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 (MOH) has begun recording and reporting the cases. The COVID-19 statistics is updated based on the total active cases, recoveries, and casualties attained daily from the MOH website.

The worst scenario of SARS-CoV-2 infection to individuals is fatality. Nevertheless, information on the mechanism of the spread of the virus or how it affects a patient seems to be in scarcity. The Centres for Disease Control and Prevention (CDC) has verified the COVID-19 human-to-human transmission on 30th January 2020. As noted by the CDC, COVID-19 can spread via droplet, close contact with infected patients, and contact with surfaces or objects that has the particles of the virus. It has been stipulated that 2–14 days or longer as the incubation period of COVID-19 with 5 days on average (4).

As the impact of this virus is severe, therefore it is important to be able to detect the pattern and forecast the spread of confirmed cases is very crucial. For an instance, Zhao et al. (5) had proposed a mathematical model to approximate the actual COVID-19 cases, including those unreported, for the first half of January 2020. It was deduced that the unreported cases count was 469 between 1st and 15th January 2020. Next, the estimation of cases from 17th January 2020 onwards revealed that the case numbers astonishingly encountered a 21-fold upsurge. This epidemic was predicted to reach its peak in late February and subside by late April based on the SEIR model combined with a machine-learning artificial intelligence (AI) method (6). Subsequently, Tang et al. (7) prescribed a mathematical model that could estimate the risk of COVID-19 transmission. Based on this, the potential number of the basic reproduction was 6.47. It also forecasted the total of 7 day confirmed cases with 23rd−29th January 2020 time interval. Consequently, the estimated peak was after 2 weeks from the initial date of 23rd January 2020.

In order to estimate the prolonged COVID-19 human-to-human transmission, data obtained from 47 patients were analyzed and resulted in a transmission rate of 0.4 (8). If the duration between the symptom detection and the patient hospitalization was halved from the tested study data, the transmission rate could reduce to 0.012. In another study, an estimation of SIR model was exhibited for the COVID-19 outbreak in Malaysia to predict the short-term daily COVID-19 cases (9). The study reported a transmission rate of 0.22 by considering that an infected individual can spread the virus to another individual within 4 days. This human-to-human transmission rate of 4 days should be highly considered, or even viewed as conservative.

Furthermore, various researchers have employed Box-Jenkins time series analysis model in predicting future cases of COVID-19 (10–12). For an instance, Rauf and Hannah (12) found out ARIMA (2, 2, 2) model produced the most accurate results compare to others for cases in India. Meanwhile, Jibrin et al. (11) recommended that the Autoregressive Fractional Integral Moving Average (ARFIMA) model should be used for further analysis of daily COVID-19 new cases. Rauf and Hannah (12) found an upward trend of the spread of COVID-19 in Nigeria based on ARIMA (1,1,0) model and more. According to Jianxi (13), the developed predictive model of COVID-19 cases must be considered on several factors such as intertwined human, social, and political factors. Due to that, predictive monitoring paradigm was proposed, which synthesized the prediction and monitoring of the daily COVID-19 cases in the study area. Another forecasting method to predict COVID-19 cases is based on machine learning approaches (14–17). Jianxi (13) stated that the hybridization model of machine learning approaches produces better performances in predicting cumulative COVID-19 cases with high daily incidence. In addition, the climatic variables were employed as inputs for proposed forecasting machine learning models.

Most of the previous studies focuses on the forecasting of future cases COVID-19. However, the analysis of this pandemic pattern is equally important. The proposed method suggested by Yogesh (18) considered the trend of new cases of COVID-19 in developing forecasting model. Nevertheless, this model didn't ensure that the trend and noise components in the data were clearly separated before the forecasting values were generated. The suitable analytical tools to assess the global change pattern with uncertainty metrics seem to be rather limited and seldom applied systematically, as it is often presented as an operational pattern worldwide. Systematically tracking and observing the infectious disease in a specific population and presented chronologically at high temporal resolution can lead to a modern and sophisticated methodology to perform in-depth data analysis. Hence, suitable analytical methods for time series data may be used if cases of health outcomes are assembled and aggregated with time units (e.g., weekly or daily basis).

Singular Spectrum Analysis (SSA) is a superb and effective alternative to address trend components, substantially minimize noise, and unravel the temporal structure of data minus preliminary manipulation (19). Generally, SSA represents univariate time series transformed into eigenvectors and eigenvalues of any trajectory matrix. The SSA refers to a multidimensional analog of principal component analysis (PCA), which is transformed into time series. One function of the SSA is to separate the time series data into noise, trend, and seasonal categories by decomposing the time series eigen, and later, reconstructing them into group selection (20).

The SSA, essentially, transforms a single dimension time series into trajectories with multiple dimensions via PCA [Singular Value Decomposition (SVD)], as well as reconstruction (approximation) of chosen Principal Components. However, the separation of the components in this approach depends on the parameters, which is the selection of window length, L, to form trajectory matrix and identifying the number of leading eigentriples (ET), based on eigenvector plot (21). This separation is crucial in this model to ensure that the trend, seasonal, and noise components are easily separated.

Although SSA lacks parametric description and highly relies on the length of time series, these flexible SSA models can recreate the asymmetric shapes of a trend, hence allowing better prediction of seasonal peaks than can harmonic models. This model, when compared to others, is easy to use, dismisses specification of models of time series and trend, enables extraction of trend in the presence of noise and oscillations,

and involves only two parameters to determine the accuracy and flexibility in predicting outcomes (22).

As the SSA models are seldom used to assess epidemiological data, this study is set to introduce the SSA model based on combining forecasting elements of time series analysis known as Recurrent Forecasting-Singular Spectrum Analysis (RF-SSA). To ensure that this developed model produces significant forecasting results, the selection of the parameter for this model, which are the window length, L and the amount of leading eigentriples used, ET, was identified using several tests. The SSA was used in this study as a base approach to build the forecasting model. The next sections describe the data in detail, followed by several sections that present the methodology, the results and discussion, and finally, the conclusion.

DATA

Daily COVID-19 prevalence data from 25th January to 29th April 2020 were gathered from MOH records. As this COVID-19 is a newly-founded virus; no COVID-19 data was available from the previous year. The suspected COVID-19 cases were diagnosed by using the Reverse Transcription Polymerase Chain Reaction (RT-PCR) technique and were confirmed as COVID-19 case-counts. All fully anonymized, laboratory-confirmed cases were abstracted on COVID-19, in which 5,945 cases represented COVID-19 infections in all 13 states and 3 federal territories in Malaysia, as recorded by MOH.

Figure 1 illustrates the total positive cases for COVID-19. The figure displays a significant spike in the number of positive cases that resulted in the 2nd wave of COVID-19 pandemic in Malaysia. With this substantial number, the Malaysian Government had announced a Movement Control Order (MCO) that took place from 18th to 31st March 2020. The MCO was later extended to the 4th phase.

Figure 2 portrays the observed number of cases for COVID-19 for the last 96 days in Malaysia. The MOH had categorized four zones of COVID-19 areas in Malaysia based on 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 (23).

The projection and estimate daily cases of COVID-19 obtained were impacted by the definition of the case reported to CPRC daily, whereby a large number of pending result test daily was definitely influential to a non-consistent increase in the number of confirmed cases. The increased prediction cases are supported by several of the biggest clusters identified by the MOH, such as Seri Petaling Tabligh Cluster, Wedding Kenduri in Bandar Baru Bangi, Seri Petaling Sub-Cluster in Rembau, Italy Cluster in Kuching, and Church Fellowship Cluster in Sarawak. The new confirmed cases were extremely spiking as the target of biology samples were taken directly from highly susceptible infected population.

MATERIALS AND METHODS

This section elaborates on the specifics of SSA model and its components.

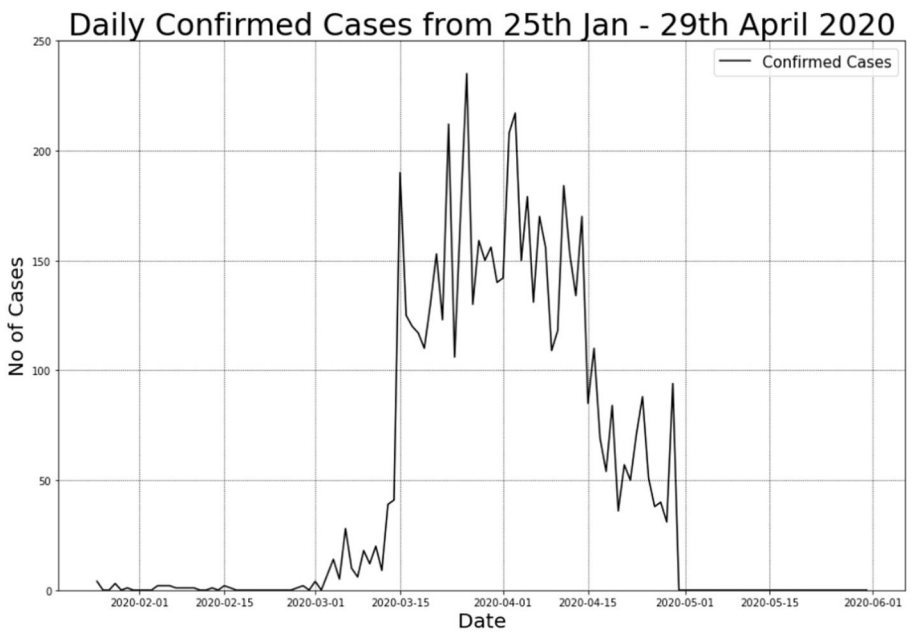


FIGURE 1 | COVID-19 daily confirmed cases in Malaysia from 25th January to 29th April 2020.

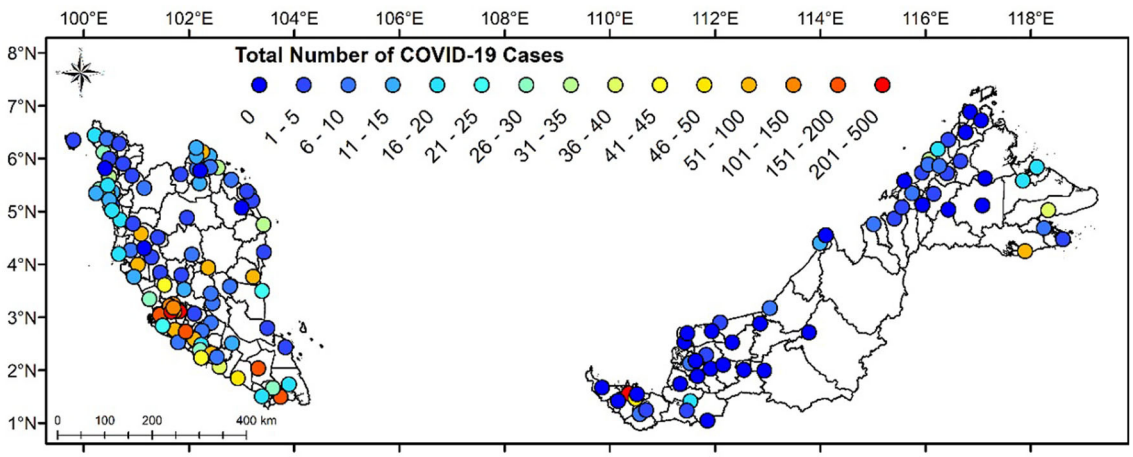


FIGURE 2 | State classification based on number of COVID-19 cases in Malaysia from 25th January to 29th April 2020.

Singular Spectrum Analysis (SSA) Model

The SSA is a model-free approach that can be applied to all types of data, regardless of Gaussian or non-Gaussian, linear or non-linear, and stationary or non-stationary (24). The daily COVID-19 data can be decomposed into several additive components via SSA, which could be defined in the forms of trend, seasonal, and noise components (25). The possible application areas of SSA are diverse (26–28). The SSA is composed of two complementary stages, known as the stages of decomposition and reconstruction (29).

Stage 1: Decomposition

The two steps in the decomposition stage are embedding and SVD. This stage decomposes the series to obtain eigen time series data.

Step I: Embedding. The first step in basic SSA algorithm is embedding, which refers to constructing the original time series into a sequence of lagged vector of size window length, L by forming lagged vectors, $K = T - L + 1$ of size L. $X_i = (x_i, \ldots, x_{i+L-1})^T \ (1 \leq i \leq K)$.

The trajectory matrix of the series $\mathbb{X}$ is

$$X = (X_1, \ldots, X_K) = \left(x_{ij}\right)_{i,j=1}^{L,K} = \begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_K \\ x_2 & x_3 & x_4 & \cdots & x_{K+1} \\ x_3 & x_4 & x_5 & \ldots & x_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & x_{L+2} & \cdots & x_T \end{pmatrix} \quad (1)$$

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

Step II: Singular Value Decomposition (SVD). In the second step, the trajectory matrix in Step I is decomposed to obtain the eigen time series based on their singular values using SVD. The following represents the SVD of the trajectory matrix, X_i where $\lambda_1, \ldots, \lambda_L$ are denoted as the eigenvalues of XX^T where singular values are arranged in a descending order such that $(\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_L)$ and by $U_1, \ldots, U_L$the corresponding eigenvectors. The SVD of X can be represented 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 called the $i - th$ eigentriple (ET) of the matrix X_i, and $\sqrt{\lambda_i}$ are the singular values of the matrix X_i.

Stage 2: Reconstruction

Grouping and diagonal averaging are the two steps in the reconstruction phase. Here, the original series are reconstructed for further analysis, including forecasting.

Step 1: Grouping. Here, the trajectory matrix is divided into dual groups—trend, seasonal and noise components. Upon setting $I = \left\{i_1, \ldots, i_p\right\}$ be a group indices, $i_1, \ldots, i_p$ where $\left(p < L\right)$. Then the matrix $\mathbf{X}_I$ corresponding to the group I is defined $\mathbf{X}_I = \mathbf{X}_{i1} + \ldots + \mathbf{X}_{ip}$. The indices set $\{1, \ldots, L\}$ is divided into m disjoint subsets; $I_1, \ldots, I_m$, based on the division of elementary matrices into groups of m. The retrieved matrices are calculated for $I = I_1, \ldots, I_m$ which called is eigentriple grouping corresponding to the representation of $\mathbf{X} = \mathbf{X}_{I1} + \ldots + \mathbf{X}_{Im}$.

Step 2: Diagonal averaging. The last step in SSA refers to the transformation of each matrix in the grouped decomposition into new series of length, T.

- Let **Z** be $L \times K$ matrix with z_{ij}, $1 \leq i \leq L$elements, $1 \leq j \leq K$. 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. With diagonal averaging, matrix **Z** is transferred into $z_1, \ldots, z_T$ based on the

following formula:

$$z_k \begin{cases} \frac{1}{k}\sum_{m=1}^{k} z^*_{m,k-m+1} & 1 \le k < L^* \\ \frac{1}{L^*}\sum_{m=1}^{L^*} z^*_{m,k-m+1} & L^* \le k \le K^* \\ \frac{1}{T-k+1}\sum_{m=k-k^*+1}^{T-K^*+1} z^*_{m,\ k-m+1} & K^* < k \le N \end{cases} \quad (2)$$

- Upon applying the diagonal averaging in equation above to the resultant matrix, $\mathbf{X}_{Ik}$, reconstructed series of $\tilde{\mathbb{Y}}_T^{(k)} = (\tilde{y}_1^{(k)}, \ldots, \tilde{y}_T^{(k)})$ is produced. The initial series of $\mathbb{Y}_T = \{y_1, y_2, \ldots, y_T\}$ is decomposed into the total of m reconstructed series, $y_t = \sum_{k=1}^{m} \tilde{y}_t^{(k)}$. The reconstructed series generated by elementary grouping refers to 'elementary reconstructed series.

Stage 3: Forecasting

To perform SSA forecasting, the time series should satisfy the linear recurrent formula (LRF). Time series $Y_T = (y_1, \ldots, y_T)$ satisfies LRF of order d if:

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_d y_{t-d}, \qquad t = d+1, \ldots, T \quad (3)$$

In this study, Recurrent SSA (RSSA) was used for forecasting purpose because it is a popular approach to predict data (30, 31). The algorithms described below are detailed in Golyandina et al. (32).

Let us assume that U_j^{∇} is the vector of the first $L-1$ components of eigenvector U_j, while π_j is the last component of $U_j (j = 1, \ldots, r)$. Denoting $\nu^2 = \sum_{j-1}^{r} \pi_j^2$, coefficient vector $\Re$ is defined as follows:

$$\Re = \frac{1}{1-\nu^2} \sum_{j=1}^{r} \pi_j \, U_j^{\nabla} \quad (4)$$

Upon considering the prior notation, the forecast of RSSA $(\hat{y}_{T+1}, \ldots, \hat{y}_{T+M})$ can be attained by

$$\hat{y}_i = \begin{cases} \tilde{y}_i, & i = 1, \ldots, T \\ \Re^T Z_i, & i = T+1, \ldots, T+M \end{cases} \quad (5)$$

where, $Z_i = \left[\hat{y}_{i-L+1}, \ldots, \hat{y}_{i-1}\right]^T$ and $\tilde{y}_1, \ldots, \tilde{y}_T$, are the values of reconstructed time series (noise reduced series).

SSA Parameter Selection

Extraction of trend from the original time series data relies on the window length, L, to form the trajectory matrix in SSA. Improper values selection for parameter L may yield unfinished reconstruction, which may potentially mislead the forecasting results. It has been stipulated that L should be large enough, but not greater than half of the number of observations understudy at $\frac{T}{2}$ (33). The appropriate window length selection depends on the structure of time series data and the current problems (34). Generally, there is no guide to determine the proper L in a dataset. The separability conditions for shorter time series may be restrictive due to the SVD properties used in estimating the signal component in SSA. Therefore, in this study, several L namely $\frac{T}{2}, \frac{T}{5}, \frac{T}{10}, \frac{T}{20}$, were investigated on COVID-19 data based on performance error, which refers to Root Mean Square Error (RMSE).

Another parameter to be considered when using the SSA approach is the amount of leading *ET*by inspecting the eigenvector plot. This plot is the eigenvector of the SVD of trajectory matrix for time series data. The one-dimensional graphs of eigenvectors were inspected to identify the trend components. The trend has a complex form when both the trend and noise components were not properly distinguished. It is highly possible that lack of separability caused the mix-up between the components. This information may serve as a guideline to identify proper grouping for component separation of the trend and noise appropriately. This reflects a link between the stages of decomposition and reconstruction.

Evaluating Separabality in Time Series Data

A key concept when studying SSA is separability, which signifies how the varied components of time series may be differentiated from each other to enable further analysis. When working with SSA method in numerous study fields, separability becomes a vital mean (35). The separability impact can result in appropriate decomposition and component extraction. The *w-correlation* technique measures the separability between two distinct components of the reconstructed time series.

The *w-correlation* reflects the weighted correlation among components of reconstructed time series that offers highly useful information to both separate and identify groups for the reconstructed components (36). The elements of the time series terms are indicated by the weights into trajectory matrix. This ranges between 0 and 1, whereby components that are well-separated slant toward 0, whereas slant toward 1 for otherwise. The *w-correlation* matrix looks into grouped decomposition among the reconstructed components. The matrix formulation of *w-correlation* is as follows:

$$\rho_{12}^{w} = \frac{\left\langle X^{(1)}, X^{(2)} \right\rangle w}{\left\| X^{(1)} \right\| w \left\| X^{(2)} \right\| w} \quad (6)$$

where $\left\| X^{(i)} \right\| w = \sqrt{\left\langle X^{(i)}, X^{(i)} \right\rangle w}$, $i = 1, 2$, $\left\langle X^{(1)}, X^{(2)} \right\rangle w = \sum_{i=0}^{N-1} w_i x_i^{(1)} w_i^{(2)}$, and weights w_i are defined below:

Let $L^* = \min(L, K)$ and $K^* = \max(L, K)$. As a result,

$$w_i = \begin{cases} i+1 & for\ 0 \le i \le L^* - 1, \\ L^* & for\ L^* \le i \le K^*, \\ T - i & for\ K^* \le i \le T - 1. \end{cases} \quad (7)$$

The graphic illustration of *w-correlation* is composed of white-black scale, whereby white represents correlation that is small, whereas black denotes correlation between the series components near to value 1.

Evaluation Performances

In this study, four types of evaluation performances are applied to evaluate the accuracy of the predicted output for the

forecasting models. The measurements used in this study are Mean Absolute Error (MAE), Mean Forecast Error (MFE), and Root Mean Square Error (RMSE), whereby, the best model is selected based on the smallest values for that measurements. Meanwhile, the Pearson Correlation Coefficient (*r)* value is based on a range from +1 to −1. A value of *r* that close to +1 or −1 indicated that the two observed variables are related to each other. Concurrently, a value of 0 indicates that there is no association between two observed variables. The equations for each of the evaluation performances are shown

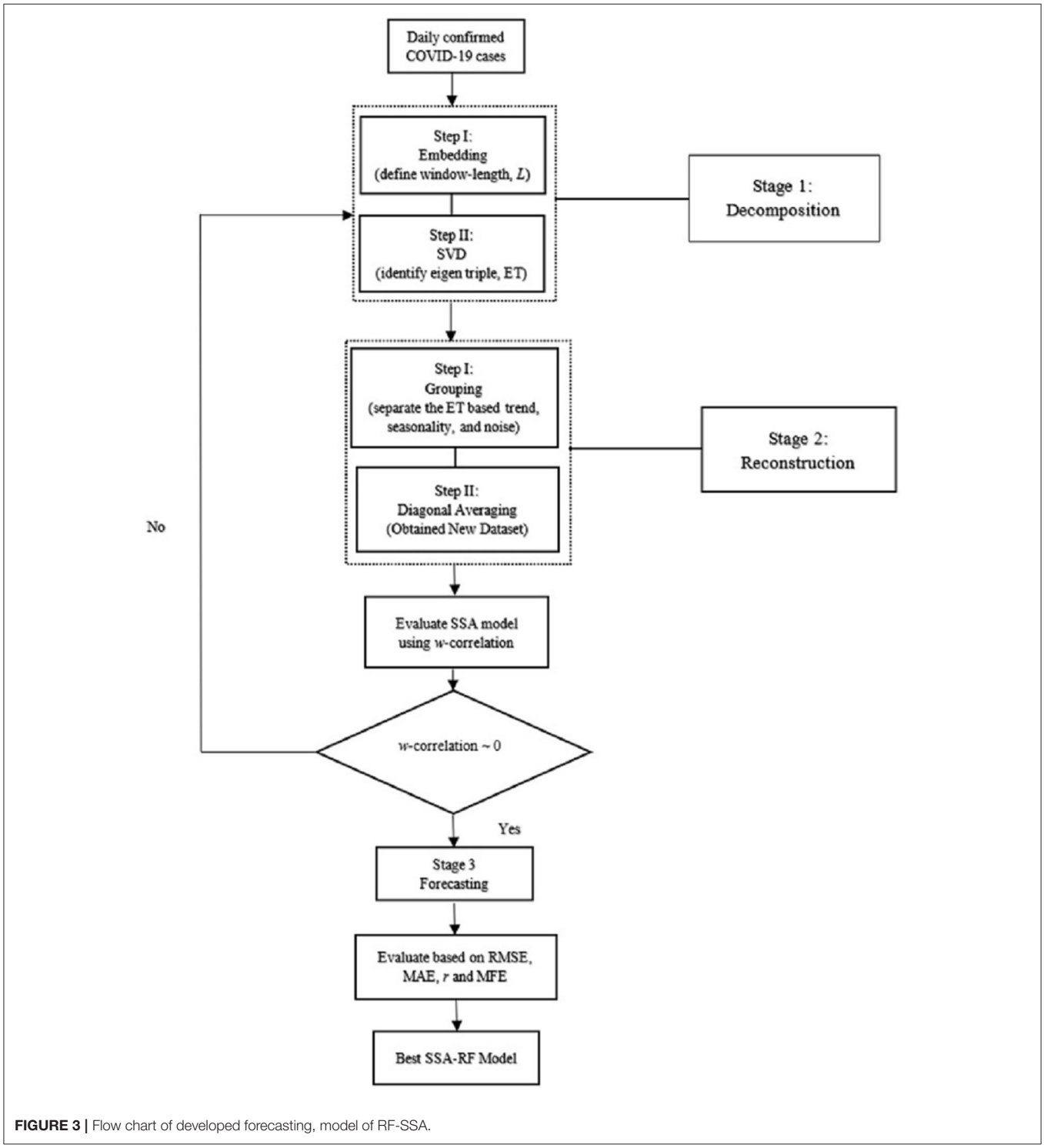


FIGURE 3 | Flow chart of developed forecasting, model of RF-SSA.

as follows:

$$MAE = n^{-1}\left[\sum_{i=1}^{n}|y_t - \hat{y}|\right] \quad (8)$$

$$MFE = n^{-1}\left[\sum_{i=1}^{n}(y_t - \hat{y})\right] \quad (9)$$

$$RMSE = n^{-2}\left[\sum_{i=1}^{n}(y_t - \hat{y})^2\right]^{-0.5} \quad (10)$$

$$r = \frac{n(\sum_{i=1}^{n} x_t y_t) - (\sum_{i=1}^{n} x_t)(\sum_{i=1}^{n} y_t)}{\sqrt{\left[n(\sum_{i=1}^{n} x_t{}^2) - (\sum_{i=1}^{n} x_t)^2\right]\left[n(\sum_{i=1}^{n} y_t{}^2) - (\sum_{i=1}^{n} y_t)^2\right]}} \quad (11)$$

where y_t is the actual values at time t; y_t is the predicted values at time t; n is the number of observations. Flow chart of developed forecasting model based on SSA as shown in **Figure 3**.

RESULTS AND DISCUSSION

Decomposition and Reconstruction

In the initial stage of this study, COVID-19 data were decomposed into components by using the SSA model, which required identification of (L, ET) parameter pair. Here, L denotes the compromise between statistical confidence and information. The suitable L value should resolve the varied oscillations embedded in the original signal.

The performance of the SSA results was determined by assessing the *w-correlation* at distinct window length, L. The *w-correlation* calculated the separability among noise, trend, and seasonal (components of reconstructed time series). Here, $L = T/2$, $T/5$, $T/10$, and $\frac{T}{20}$, which represent $L = 48$, 19, 10, and 5, respectively, for T based on 96 daily cases on COVID-19 data had been selected. The scales were selected to fit the data of the time series, apart from striking a balance to achieve a proper lag vector sequence.

In **Figure 4**, the *w-correlation* is presented based on SSA using daily cases of COVID-19 data at varying window lengths. The *w-correlation* displayed a declining trend when the total window length declined for SSA approach. The correlations among trend and other components should be close to zero for extraction of trend. This means; the distinct window lengths have an impact on the component's separability. Besides, the SSA was directed to the lowest *w-correlation* at $L = T/20$; signifying the best separability among the reconstructed components as it was the closest to zero.

The graphs in **Figures 5A–D** illustrate the heat-plot of different window lengths, L, based on *w-correlations* using the SSA approach. The heat-plot of *w-correlation* for the reconstructed components based on white-black scale ranges between 0 and 1 (37). Huge correlation values among the reconstructed components exhibited the possibility of the components to form a group while corresponding to the same component. As illustrated in **Figure 5**, the shade of each square represents the *w-correlation* strength between two components. Meanwhile, **Figures 5A–C** portrays the tendency of the components to form correlation with other components despite signifying weak correlation. 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, $L = 5$, which is evidently demonstrated in **Figure 5D** for better separability.

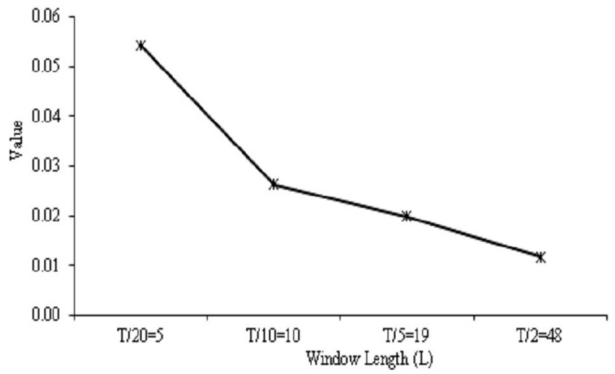


FIGURE 4 | Effect of *w*-correlation based on SSA using COVID-19 data at varied window lengths.

TABLE 1 | Comparison of Singular Spectrum Analysis Prediction Performance for Several Window Length (L).

Window Length, L	RMSE
$T/2 = 48$	29.51
$T/5 = 19$	29.67
$T/10 = 10$	23.97
$T/20 = 5$	19.12

Table 1 presents the reconstructed time series components varied window length. The lowest RMSE was observed from $L = \frac{T}{20}$, which had the smallest value amongst other L, indicating its suitability based on short-time series of the outbreak data. Meanwhile, the high RMSE values were reported in this study due to the high model variance for small sample set.

The plot of five main eigenvectors is displayed in **Figure 6**. Such plot is beneficial to choose an appropriate group for the components of time series data, especially to separate the components of noise, trend, and seasonal. The retrieved information may be further analyzed in the step of grouping in RF-SSA. The component of trend was identified from eigenvector plot, in which seasonal and trend components have sine waves indicated by the slow cycles found in the graph (high frequency). Meanwhile, the component of noise was represented by the saw-tooth found in the graph (low frequency). The leading eigenvector has nearly continual coordinates, thus corresponding to a pure smoothing by Bartlett filter (38, 39). The reconstruction result by each of the five ET is presented in **Figure 7**. The two figures verified the compatibility of the first and second ET with

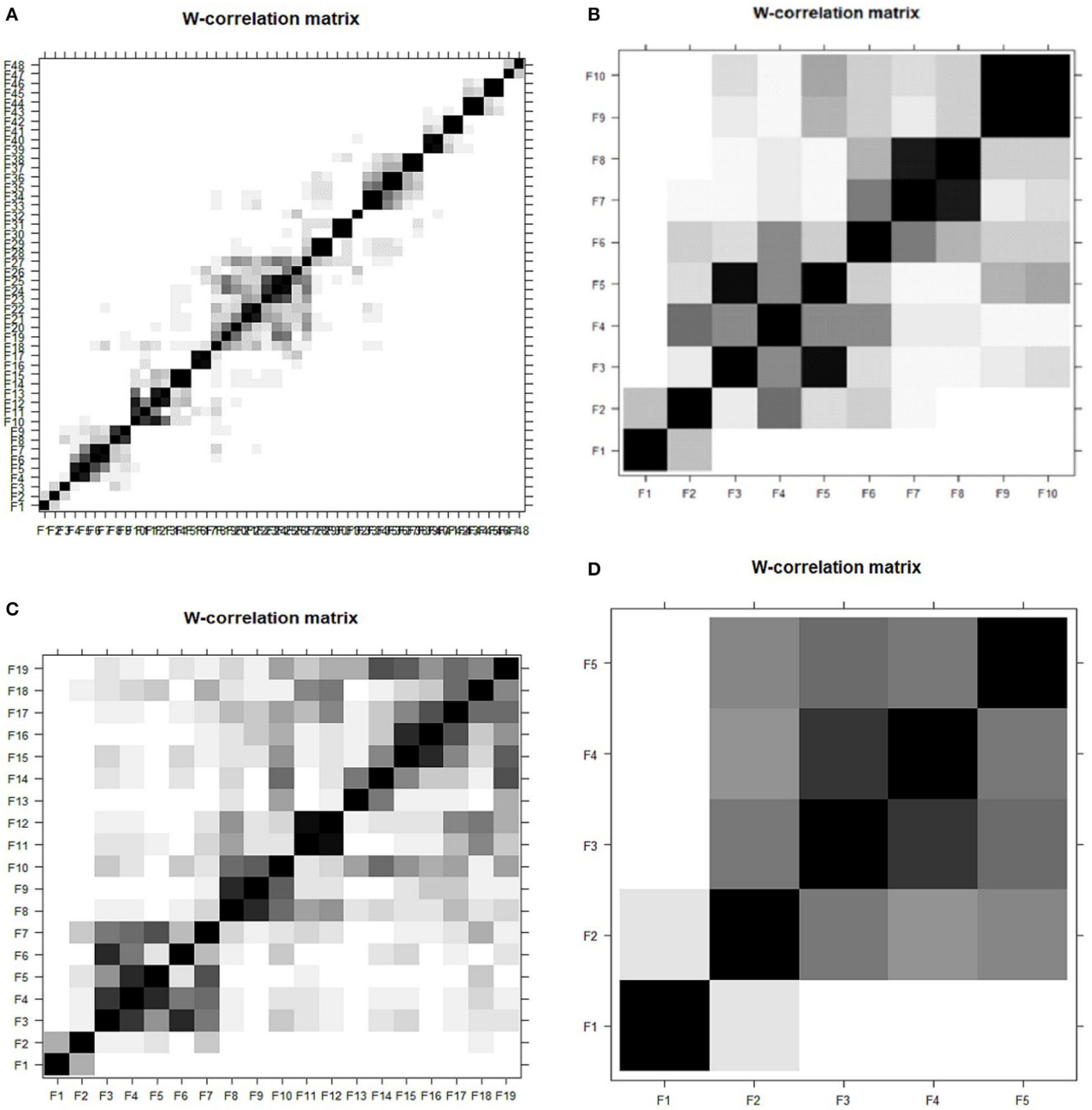


FIGURE 5 | (A–D) *w-correlation* plot using SSA with varied windows length **(A)** $L = 48$ **(B)** $L = 19$ **(C)** $L = 10$ **(D)** $L = 5$.

the trend, whereas the remaining *ET* had the noise component, thus irrelevant to trend.

Figure 8 demonstrates the components of the reconstructed time series plot from the trend extracted via RF-SSA for daily COVID-19 cases in Malaysia. The reconstructed series is the new dataset derived from the original data, which is clear from noise. It is a crucial aspect in SSA to ensure that the forecasting results are precise and accurate (40). The component of trend in the time series data was used to observe the occurrence of the cases trend and pattern, as it was randomly-tabulated as per daily cases (see **Figure 8**). In **Figures 8A**, **7**, the trend was precisely generated by a leading *ET*, which coincided with the initial reconstructed component exhibited in **Figure 8**. The trend in **Figure 8B** was precisely generated by both leading *ET*, which coincided with the first and second reconstructed components shown in **Figure 8**. The dashed and straight lines on the plot denote the reconstructed series based on the extracted trend component from SSA and the COVID-19 original time

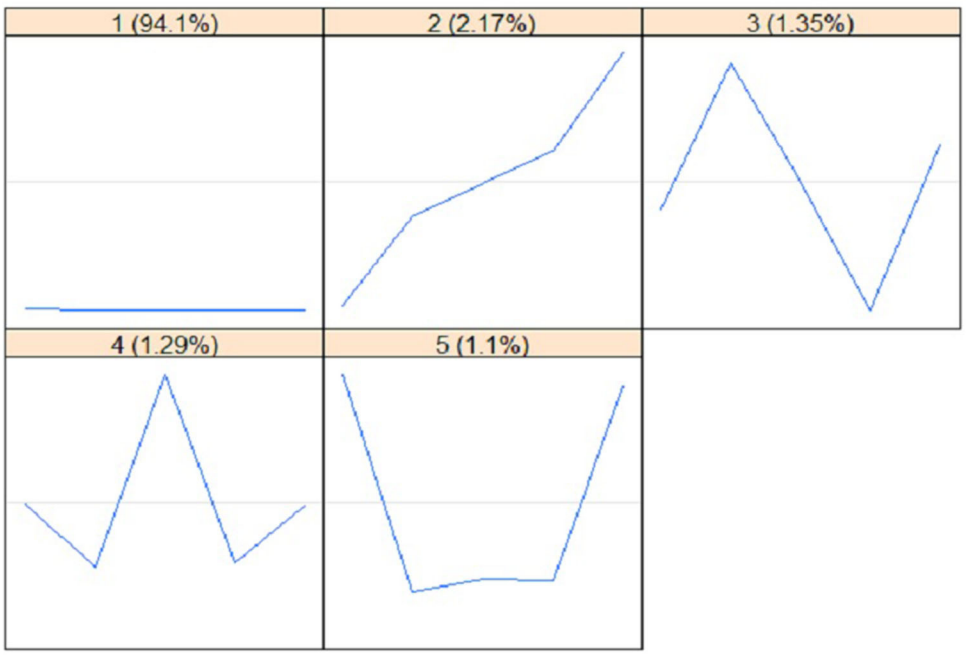


FIGURE 6 | Eigenvectors Plot using Singular Spectrum Analysis.

series data, respectively. The plot of reconstructed time series components, produced by both leading *ET*, abides by the original COVID-19 data although noise component was omitted for $L = 5$ for daily COVID-19 cases in Malaysia.

For proper identification of seasonal series components, the graph of eigenvalues and scatterplots of eigenvectors were applied. In order to determine the seasonal series components using eigenvalues plot, several steps were produced by approximately equal eigenvalues. **Figure 9** portrays the plot of the logarithms of the five singular values for the COVID-19 cases in Malaysia. It clearly showed that no step produced by approximately equal eigenvalues that corresponded to a sine wave. The scatterplot of eigenvectors displays the regular polygons yielded by a pair of eigenvectors to demonstrate that the series components have produced seasonality components. Based on **Figure 10**, no pair of eigenvectors produced regular polygons. This confirmed that the COVID-19 data in Malaysia were not influenced by the seasonality since both figures did not have sine wave.

Forecasting Daily COVID-19 Cases Using SSA-RF

As mentioned in the previous section, the daily COVID-19 cases in Malaysia were first decomposed and reconstructed using SSA model. The next step in this study is to predict the future cases of COVID-19 in Malaysia. In this stage, an SSA forecasting algorithm known as Recurrent Forecasting were used accordingly. From hereafter, the model are known as SSA-RF. **Table 2** presents the summary statistics from the experiment analysis of SSA-RF at several windows length.

Looking at **Table 2**, it is apparent that the best performances can be obtained from $L = 5$ that has the lowest MAE of 11.2549 with the highest r of 0.9619, indicating superb correlation between confirmed and predicted cases. Moreover, the MFE shows that the SSA-RF algorithm with L = 5, tends to under-forecast daily COVID-19 cases by 0.1920%. Meanwhile, the second-best model is observed from SSA-RF with L = 10 where RMSE is 23.9652, MAE of 14.8890, r of 0.9402 with MFE of 0.0067%. Meanwhile, $L = 19$ and $L = 48$ has the worst performances among all models whereby MAE and r for both models are 19.3706 and 0.9086, respectively. Furthermore, MFE statistical results showed that both models are over-forecast by 2.82%. Visual inspection on these models performances are presented in **Figures 11A–D**.

Based on **Figures 11A–D**, it is a clear indication that SSA-RF models able to capture general pattern of non-linear increasing trend of daily confirmed cases of COVID-19 in Malaysia. Detailed analysis from **Figure 11A** found out that model with $L = 5$ performed better that other models whereby the model able to follow the actual pattern of daily confirmed cases of COVID-19. Meanwhile, as can be seen from **Figures 11B–D**, other models which are $L = 10$, $L = 19$, and $L = 48$ unable to follow the actual pattern of the observed data. This is a clear indication that the models performed poorly as compared to $L = 5$ model.

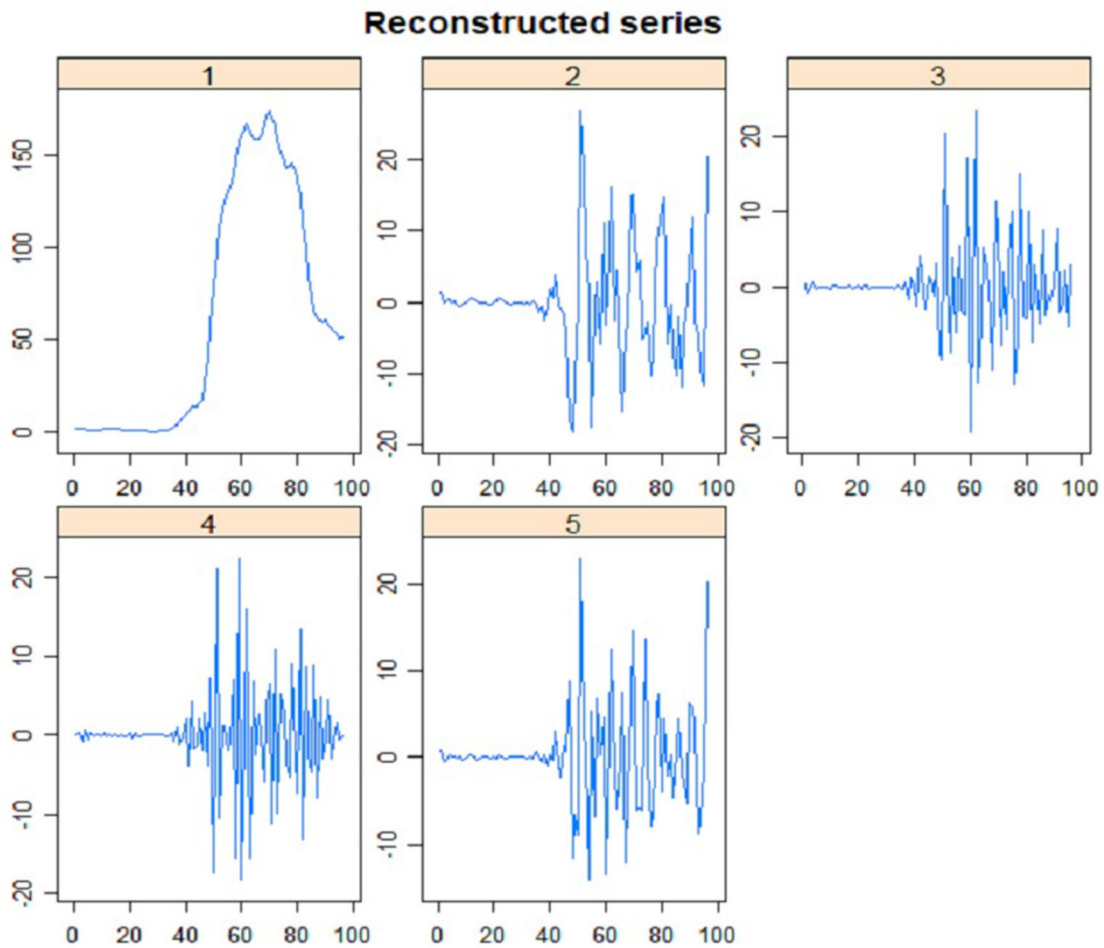
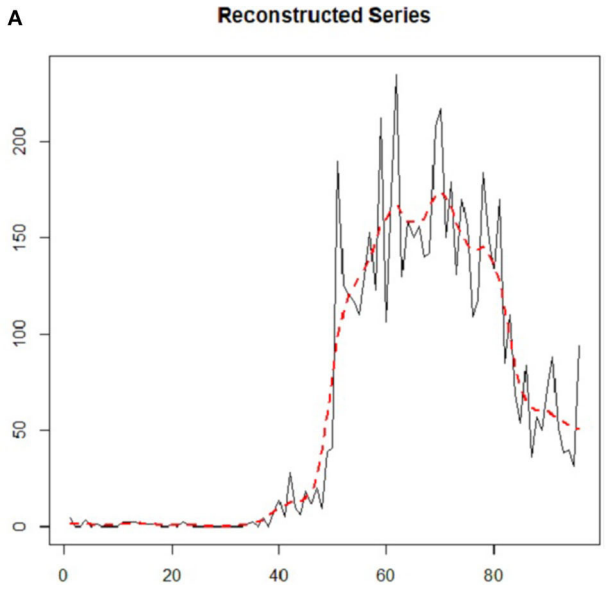


FIGURE 7 | First stage: Elementary Reconstructed Series (L = 5).

Next, the SSA-RF models were used to predict future cases starting from 30th April to 31st May 2020. At the time of this study, the historical cases from 25th January to 29th April 2020 were used and the future 32 days ahead of COVID-19 cases had been predicted accordingly. **Figures 11A–D** illustrates the confirmed cases from 25th January to 29th April 2020 and the forecasted daily cases until 31st May 2020. It is worth noting that the figures display a noticeable but faint decreasing pattern from 5th April 2020 onwards. One of the contributing factors for the decreasing trend was due to the MCO announced by the Malaysian Government which took place on 18th March 2020. The above figures also illustrate the predicted values of 32 day ahead using SSA-RF algorithm against confirmed cases of COVID-19 in Malaysia. Despite the encouraging statistical finding based from the historical data and lower under-forecast value; the SSA-RF models failed to capture the sudden drop in the COVID-19 cases, which is considered to have never happened before. This sudden drop was highly likely due to the MCO that was extended to phase-4, which ended on 12th May 2020.

During the MCO, Malaysians were advised to stay at home as much as possible to minimize the spread of further COVID-19 infections. All schools and most workplaces were closed, and they were directed to work from home except for essential services. Traveling ban, restriction movement order including interstate movement, restriction on gatherings, and public transport closure were imposed strictly by the government. Active case detection was continued, followed by isolation of the cases, and the close contacts were tested and quarantined to further curb the spread of COVID-19. All these actions successfully plateaued and reduced the number of COVID-19 cases (**Figures 11A–D**). In addition, the cases were reduced due to the incubation period of the virus between 2 to 14 days, and the recent findings from WHO has stated that after 5–10 days of the infection, the infected individual starts to gradually produce neutralizing antibodies which will decrease the risk of transmission to others (41, 42). WHO has also reported three research that found the inability of SARS-CoV-2 virus to be cultured after 7–9 days of onset of symptoms (43, 44). From all the latest findings, WHO has concluded that after 14 days, the patients are not likely to be



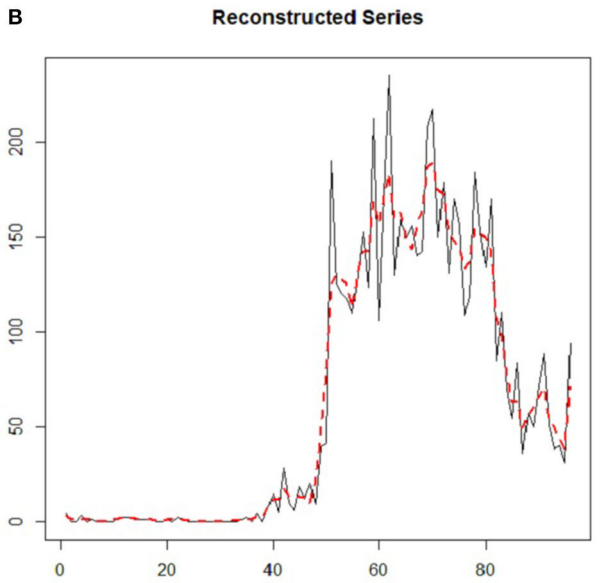


FIGURE 8 | (A,B) Daily COVID-19 cases of reconstructed components from extracted trends using SSA at **(A)** L = 5, ET 1 **(B)** *L = 5,* ET 2.

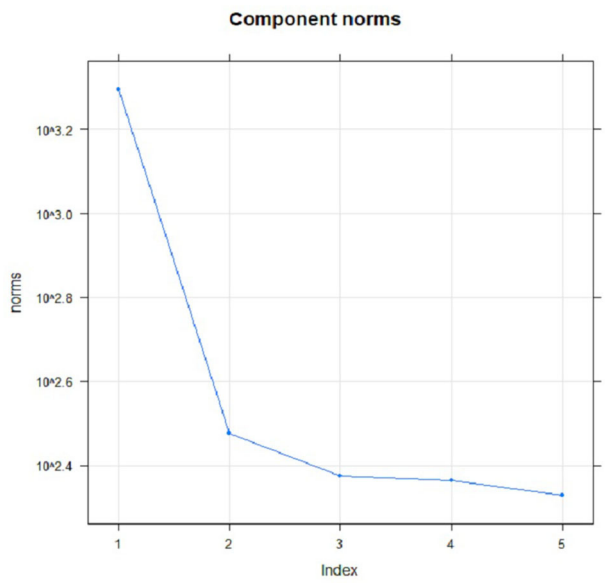


FIGURE 9 | Logarithms of five eigenvalues.

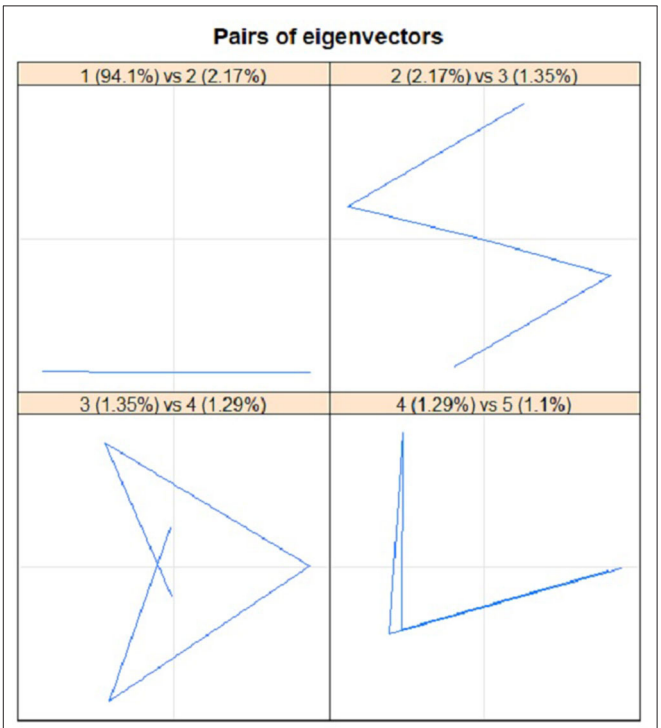


FIGURE 10 | Plots of eigenvectors (EV) pairs: 1-EV and 2-EV, 2-EV and 3-EV, 3-EV and 4-EV, as well as 4-EV and 5-EV for COVID-19 cases.

infectious (45). The government's decision to extend the MCO up to 12th May had successfully plateaued and reduced the curve as it provides sufficient time to break the virus transmission.

Furthermore, the figures showed that different window length suggested a different forecasted value of future cases. For an instance, SSA-RF with L = 48. Nineteen and 10 predicted that there will be insignificant changes in the number of future cases, while SSA-RF with L = 5 showed there will be a significant drop in the future cases. Other than that, the model also suggested that Malaysia will reach single digit in COVID-19 cases by early June 2020. However, the model unable to predict the date for total eradication of COVID-19 cases. This is consistent with WHO which indicated that this virus will not be eradicated even after the vaccine is found. It might persist to be endemic in

certain countries and will need cooperation on a global scale and leveraging tools such as contact tracing and disease surveillance to defeat COVID-19.

Limitation of SSA-RF Model

Some limitations of this study, which should be emphasized when using the SSA-RF model in assessing the pandemic data in Malaysia, are as follows:

- The SSA-RF model works best when the data exhibit a stable or consistent pattern over time with a minimum amount of outlier. This can help to obtain accurate and precise results for future predictive cases.
- The sudden spike in data leads to low performance of forecasting results using this predictive SSA-RF model.
- The SSA-RF model is mainly used to project future values using historical time series data for short-term forecast.
- Recurrent forecasting approach is a better contender than vector approach for forecasting both short and medium time series data of SSA. However, under such scenarios, it is advisable that users also evaluate the performance of forecasting SSA approach on their data to arrive at a complete picture.
- Although SSA able to capture the pattern of the Coronavirus COVID-19 cases, however, its ability in predicting the cases accurately is still need to be investigated further.
- Different observed behavior of a dataset might influence the selection of window length.
- This model did not take into account the effect of incubation period in transmission of the virus, the effect of the government measures to curb the spread of COVID-19.

TABLE 2 | SSA-RF Prediction Performance Several Window Length (*L*).

L	MAE	*r*	MSE	
$T/2 = 48$	19.3706	0.9086	−2.8249	Over-forecast
$T/5 = 19$	19.3706	0.9086	−2.8249	Over-forecast
$T/10 = 10$	14.8890	0.9402	0.0067	Under-forecast
$T/20 = 5$	11.2549	0.9619	0.1920	Under-forecast

CONCLUSION

This study assessed the applicability of SSA-RF 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 devising prompt and effective strategies. This model allows the health authorities to comprehend the outbreak pattern better. The pattern retrieved from the SSA-RF model can be applied to forecast the outbreak cases growth pattern in Malaysia. The parameters used in this model were window length, L, and the total of *ET* employed for reconstruction, r. The results revealed that parameter $L = 5$ ($T/20$) was suitable for short time series outbreak data and the appropriate number of leading *ET* s to obtain was crucial as it affected the forecasting outcomes. Overall, the results showed that the SSA-RF model could forecast this pandemic

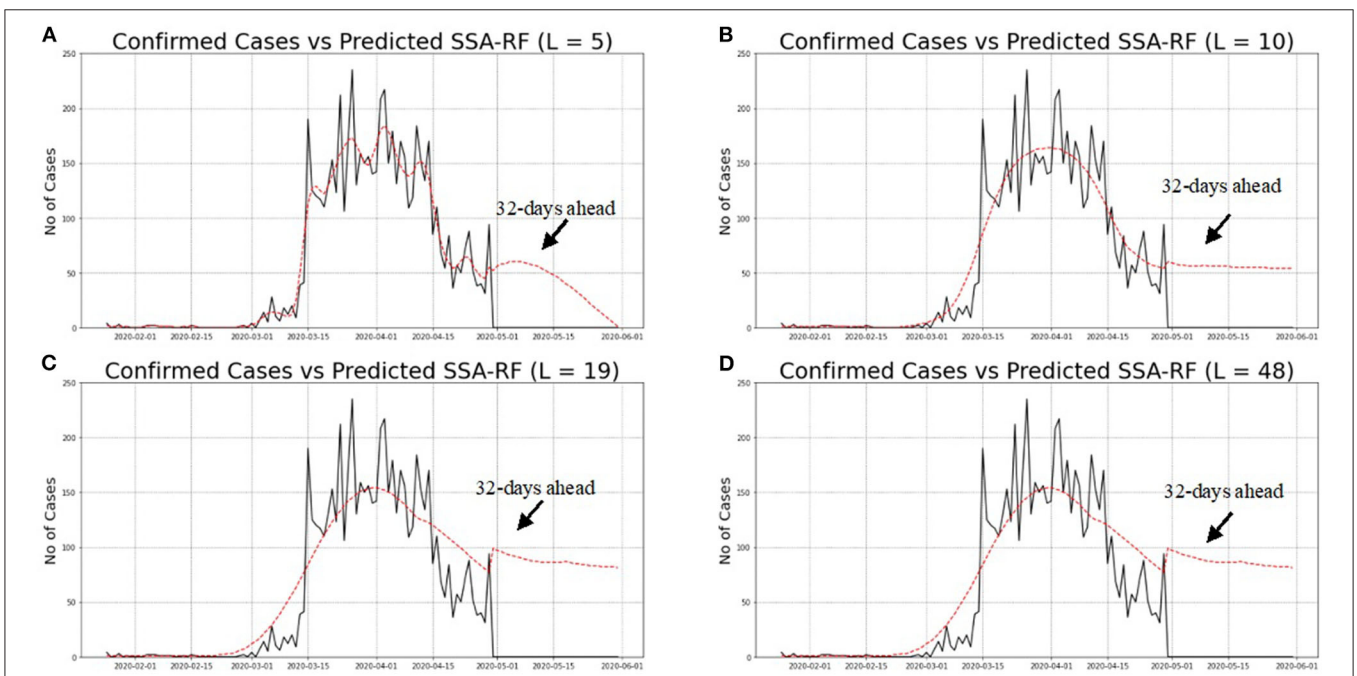


FIGURE 11 | (A–D) Predicted SSA-RF and confirmed cases of COVID-19 in Malaysia for Various Windows Length *(L)*.

with reasonable accuracy as the model had under-forecasted by 0.1920% with high correlation values between confirmed and predicted cases. Nevertheless, the SSA-RF model failed to capture the sudden drop in COVID-19 cases, likely due to the MCO that was extended to 12th May 2020. In order to improve the accuracy of the model, more information is required to better predict the COVID-19 cases for a long period. In the meantime, case definition and data collection must be maintained in real-time to enhance the RF-SSA for further evaluation. It is suggested that the SSA-RF model is enhanced to enable the model to capture sudden and rapid changes in the dataset.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

SS and SI conceived the presented idea, developed the theory, and performed the computations. NH, MT, and NS verified the analytical methods and supervised the findings of this work. All authors discussed the results and contributed to the final manuscript.

FUNDING

The authors would like to thank the Ministry of Higher Education Malaysia (MOHE) for supporting this research under Fundamental Research Grant Scheme Vot No. 2019-0132-103-02 (FRGS/1/2019/STG06/UPSI/02/4) and partially sponsored by Vot No. FRGS/1/2018/STG06/UTHM/03/3.

REFERENCES

1. Chen Y, Liu Q, Guo D. Emerging coronaviruses: genome structure, replication, and pathogenesis. *J Med Virol*. (2020) 92:2249. doi: 10.1002/jmv.26234
2. Ge XY, Li JL, Yang XL, Chmura AA, Zhu G, Epstein JH, et al. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature*. (2013) 503:535–8. doi: 10.1038/nature12711
3. Coronavirus Website - Ministry of Health (2020). Available online at: http://www.moh.gov.my/index.php (accessed April 3, 2020).
4. Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Ann Intern Med*. (2020) 172:577–82. doi: 10.7326/M20-0504
5. Zhao S, Musa SS, Lin Q, Ran J, Yang G, Wang W, 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*. (2020) 9:388. doi: 10.3390/jcm9020388
6. Yang Z, Zeng Z, Wang K, Wong SS, Liang W, Zanin M, et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *J Thorac Dis*. (2020) 12:165. doi: 10.21037/jtd.2020.02.64
7. Tang B, Wang X, Li Q, Bragazzi NL, Tang S, Xiao Y, et al. estimation of the transmission risk of the 2019-nCoV and its implication for public health interventions. *J Clin Med*. (2020) 9:462. doi: 10.3390/jcm9020462
8. Thompson RN. Novel coronavirus outbreak in Wuhan, China, 2020: intense surveillance is vital for preventing sustained transmission in new locations. *J Clin Med*. (2020) 9:498. doi: 10.3390/jcm9020498
9. Ariffin MRK, et al. *Malaysian COVID-19 Outbreak Data Analysis and Prediction*. Institute for Mathematical Research (2020). Available online at: http://einspem.upm.edu.my/covid19maths/file/Report_001%20v13.pdf
10. Yemane AG, Daniel A. Trend analysis and forecasting the spread of COVID-19 pandemic in ethiopia using box-jenkins modeling procedure. *Int J Gen Med*. (2021) 2021:1485–98. doi: 10.2147/IJGM.S306250
11. Da HL, Youn SK, Young YK, Kwang YS, In HC. Forecasting COVID-19 confirmed cases usng empirical data analysis in korea. *Healthcare (Basel)*. (2021) 9:254. doi: 10.3390/healthcare9030254
12. Das RC. Forecasting incidences of COVID-19 using Box-Jenkins method for the period July 12-Septembert 11, 2020: A study on highly affected countries. *Chaos Solitons Fractals*. (2020) 140:1–14. doi: 10.1016/j.chaos.2020.110248
13. Jianxi L. Forecasting COVID-19 pandemic: unknown unknowns and predictive monitoring. *Technol Forecast Soc Change*. (2021) 166:1–4. doi: 10.1016/j.techfore.2021.120602
14. Ramon Gomes da S, Matheus Henrique Dal Molin R, Viviana Cocco M, Leandro dos Santos C. Forecasting Brazillian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables. *Chaos Solitons Fractals*. (2020) 139:1–13. doi: 10.1016/j.chaos.2020.110027
15. Rauf HT, Lali MIU, Khan MA, Kadry S, Alolaiyan H, Razaq A, et al. Time series forecasting of COVID-19 transmission in Asia Pacific countries using deep neural networks. *Pers Ubiquitous Comput*. (2021) 10:1–18. doi: 10.1007/s00779-020-01494-0
16. Muhammad Attique K, Seifedine K, Yu-Dong Z, Tallha A, Muhammad S, Amjad R, et al. Prediction of COVID-19- pneumonia based on selected deep features and one class kernel extreme learning machine. *Comp Electr Eng*. (2021) 90:1–18. doi: 10.1016/j.compeleceng.2020.106960
17. Matheus Henrique Dal Molin R, Roman Gomes da S, Viviana Cocco M, Leandro dos Santos C. Short-term forecasting COVID-19 cumulative confirmed cases: perspectives for Brazil. *Chaos Solitons Fractals*. (2020) 135:1–10. doi: 10.1016/j.chaos.2020.109853
18. Yogesh G. Transfer learning for COVID-19 cases and deaths using LSTM network. *ISA Transac*. (2020). doi: 10.1016/j.isatra.2020.12.057
19. Golyandina N Zhigljavsky A. Basic SSA. In: *Singular Spectrum Analysis for Time Series*. Berlin; Heidelberg: Springer (2013). pp. 11–70.
20. Shaharudin SM, Ahmad N, Zainuddin NH. Modified singular spectrum analysis in identifying rainfall trend over Peninsular Malaysia. *Indonesian J Electr Eng Comp Sci*. (2019) 15:283. doi: 10.11591/ijeecs.v15.i1.pp283-293
21. Shaharudin SM, Ahmad N, Yusof F. Effect of window length with singular spectrum analysis in extracting the trend signal of rainfall data. *Aip Proc*. (2015) 1643:321. doi: 10.1063/1.4907462
22. Fuad MFM, Shaharudin SM, Ismail S, Samsudin NAM, Zulfikri MF. Comparison of singular spectrum analysis forecasting algorithms for student's academic performance during COVID-19 outbreak. *IJATEE*. (2021) 8:178–89. doi: 10.19101/IJATEE.2020.S1762138
23. Coronavirus Website - Ministry of Health (2020). Available online at: https://kpkesihatan.com/ (accessed April 3, 2020).
24. Deng C. *Time Series Decomposition using Singular Spectrum Analysis*. Master, East Tennessee State University (2014).
25. Biabanaki M, Eslamian SS, Koupai JA, Canon J, Boni G, Gheysari M. A principal components/singular spectrum analysis approach to enso and pdo influences on rainfall in West of Iran. *Hydrol Res*. (2014) 45:250–62. doi: 10.2166/nh.2013.166
26. Rodriguez-Aragon LJ Zhiglkavsky A. Singular spectrum analysis for image processiong. *Stat Interface*. (2010) 3:419–26. doi: 10.4310/SII.2010.v3.n3.a14
27. Chau KW, Wu CL. A hybrid model coupled with singular spectrum analysis for daily rainfall prediction. *J Hydroinformat*. (2010) 12:458–73. doi: 10.2166/hydro.2010.032
28. Alexandrov T, Golyandina N, Spirov A. Singular spectrum analysis of gene expression profiles of early drosophila embryo: exponential-in-distance patterns. *Res Lett Signal Proc*. (2008) 2008:825758. doi: 10.1155/2008/825758
29. Carvalho MD Rua A. *Real-Time Nowcasting the US Output GAP: Singular Spectrum Analysis at Work*. Lisboa: Banco De Portugal (2014) ISBN 978-989-678-304-4.

30. Danilov D. Principal components in time series forecast. *J Comput Graph Stat*. (1997) 6:112–21. doi: 10.1080/10618600.1997.10474730
31. Danilov D. The Caterpillar method for time series forecasting. In: Danilov D, Zhigljavsky A, editors. *Principal Components of Time Series: The Caterpillar Method*. St. Petersburg: University of St. Petersburg (1997). p. 73–104.
32. Golyandina N, Nekrutkin V, Zhigljavsky A. *Analysis of Time Series Structure: SSA and Related Techniques*. New York, NY: Chapman & Hall/CRC (2001).
33. Shaharudin SM, Ismail S, Samsudin MS, Azid A, Tan ML, Basri MAA. Prediction of epidemic trends in COVID-19 with mann-kendall and recurrent forecasting-singular spectrum analysis. *Sains Malays*. (2021) 50:1131–42. doi: 10.17576/jsm-2021-5004-23
34. Alonso FJ, Salgado DR, Cuadrado J, Pintado P. Automatic smoothing of raw kinematics signals using SSA andcluster analysis. In: *Euromech Solid Mechanics Conference*. Lisbon (2009). p. 1–9.
35. Golyandina N, Shlemov A. Variations of singular spectrum analysis for separability improvement: non-orthogonal decompositions of time series. *Stat Interface*. (2014) 8:277–94. doi: 10.4310/SII.2015.v8.n3.a3
36. Golyandina NE, Korobeynikov A. Basic singular spectrum analysis forecasting with R. *Comput Stat Data Anal*. (2014) 71:934–54. doi: 10.1016/j.csda.2013.04.009
37. Hassani H. Singular spectrum analysis: methodology and comparison. *J Data Sci*. (2007) 5:239–57. Available online at: https://mpra.ub.uni-muenchen.de/4991/
38. Golyandina N, Nekrutkin V, Zhigljavsky A. *Analysis of Time Series Structure: SSA and Related Techniques*. New York, NY; London: Chapman Hall/CRC (2001).
39. Mahmoudvand R, Konstantinides D, Rodrigues PC. *Forecasting Mortality Rate by Multivariate Singular Spectrum Analysis*. John Wiley & Sons, Ltd. (2017) 33:717–32. doi: 10.1002/asmb.2274
40. Hassani H, Zhigljavsky A. Singular spectrum analysis: methodology and application to economics data. *J Syst Sci Complex*. (2009) 22:372. doi: 10.1007/s11424-009-9171-9
41. Wolfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Müller MA, et al. Virological assessment of hospitalized patients with COVID-19. *Nature*. (2020) 581:465–9. doi: 10.1038/s41586-020-2196-x
42. Atkinson B, Petersen E. SARS-CoV-2 shedding and infectivity. *Lancet*. (2020) 395:1339–40. doi: 10.1016/S0140-6736(20)30868-0
43. Bullard J, Dusk K, Funk D, Strong JE, Alexander D, Garnett L, et al. Predicting infectious SARS-CoV-2 from diagnostic samples. *Clin Infect Dis*. (2020) 71:2663–6. doi: 10.1093/cid/ciaa638
44. Peng Z, Xing-Lou Y, Xian-Guang W, Ben H, Lei Z, Wei Z, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. (2020) 579:270–3. doi: 10.1038/s41586-020-2012-7
45. Centers for Disease Control and Prevention, Coronavirus Disease 2019 (COVID-19). *Symptom-Based Strategy to Discontinue Isolation for Persons With COVID-19*. (2020). Available online at: https://www.who.int/news-room/commentaries/detail/criteria-for-releasing-COVID-19-patients-from-isolation (accessed June 12, 2020).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.