Sains Malaysiana 50(4)(2021): 1131-1142
http://doi.org/10.17576/jsm-2021-5004-23

Prediction of Epidemic Trends in COVID-19 with Mann-Kendall and Recurrent Forecasting-Singular Spectrum Analysis

(Ramalan Kecenderungan Wabak pada Covid-19 dengan Mann-Kendall dan Ramalan Berulang-Analisis Spektrum Tunggal)

SHAZLYN MILLEANA SHAHARUDIN*, SHUHAIDA ISMAIL, MOHD SAIFUL SAMSUDIN, AZMAN AZID, MOU LEONG TAN & MUHAMAD AFDAL AHMAD BASRI

ABSTRACT

Novel coronavirus also known as COVID-19 was first discovered in Wuhan, China by end of 2019. Since then, the virus has claimed millions of lives worldwide. In 29th April 2020, there were more than 5,000 outbreak cases in Malaysia as reported by the Ministry of Health Malaysia (MOHE). This study aims to evaluate the trend analysis of the COVID-19 outbreak using Mann-Kendall test, and predict the future cases of COVID-19 in Malaysia using Recurrent Forecasting-Singular Spectrum Analysis (RF-SSA) model. The RF-SSA 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 RF-SSA model assessment is based on the World Health Organization (WHO) official COVID-19 data to predict the daily confirmed cases after 29th April until 9th May, 2020. The preliminary results of Mann-Kendall test showed a declining trend pattern for new cases during Restricted Movement Order (RMO) 3 compared to RMO1, RMO2 and RMO4, with a dramatic increase in the COVID-19 outbreak during RMO1. Overall, 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. The proposed model predicted that Malaysia would hit single digit in daily confirmed cased of COVID-19 by early-June 2020. These findings have proven the capability of RF-SSA model in apprehending the trend and predict the cases of COVID-19 with high accuracy. Nevertheless, enhanced RF-SSA algorithm should to be developed for higher effectivity in capturing any extreme data changes.

Keywords: COVID-19; forecasting; Mann-Kendall test; recurrent forecasting (RF); singular spectrum analysis (SSA)

ABSTRAK

Koronavirus baru juga dikenali sebagai COVID-19 telah dilaporkan pertama kali di Wuhan, China pada akhir 2019. Sejak itu, virus tersebut telah meragut berjuta-juta nyawa di seluruh dunia. Pada 29 April 2020, terdapat lebih daripada 5,000 kes wabak di Malaysia seperti yang dilaporkan oleh Kementerian Kesihatan Malaysia (KKM). Kajian ini bertujuan untuk menilai analisis tren wabak COVID-19 menggunakan ujian Mann-Kendall dan meramalkan kes COVID-19 yang akan datang di Malaysia menggunakan model Ramalan Berulang-Analisis Spektrum Tunggal (RF-SSA). Model RF-SSA dibangunkan untuk mengukur dan meramalkan kes COVID-19 harian di Malaysia selama 10 hari akan datang dengan menggunakan kes yang telah disahkan sebelumnya. Model ramalan berdasarkan Analisis Spektrum Tunggal yang membezakan kebisingan dalam aliran siri masa diperkenalkan. Penilaian model RF-SSA berdasarkan data rasmi COVID-19 oleh Organisasi Kesihatan Dunia (WHO) untuk meramalkan kes-kes yang diselesaikan setiap hari selepas 29 April hingga 9 Mei 2020. Hasil awal ujian Mann-Kendall menunjukkan penurunan corak tren untuk kes baru semasa Perintah Kawalan Pergerakan (RMO) 3 berbanding RMO1, RMO2 dan RMO4, dengan peningkatan mendadak wabak COVID-19 semasa RMO1. Secara keseluruhan, RF-SSA telah meramalkan kes sebanyak 0.36%. Ini menunjukkan kecekapan RF-SSA untuk meramalkan jumlah kes COVID-19 yang akan datang. Model yang dicadangkan juga meramalkan bahawa Malaysia akan mencapai angka satu digit dalam kes COVID-19 yang disahkan setiap hari pada awal Jun 2020. Penemuan ini telah membuktikan kemampuan model RF-SSA dalam menangkap tren dan meramalkan kes COVID-19 dengan ketepatan tinggi. Walaupun begitu, algoritma RF-SSA harus dipertingkatkan untuk keberkesanan yang lebih tinggi dalam menangkap perubahan data yang melampau.

Kata kunci: Analisis spektrum tunggal (SSA); COVID-19; ramalan; ramalan berulang (RF); ujian Mann-Kendall

Introduction

The World Health Organization (WHO) reported the severe acute respiratory coronavirus-2 syndrome known as COVID-19 in late December 2019. COVID-19 sequence similarity scores with Bat SARS-like, SARS-CoV, and MERS-CoV were based on approximately 99, 96, and 50%, respectively (Kannan et al. 2020). Malaysian National Security Council (NSC) (2020), showed that Malaysia faces the first COVID-19 outbreak on 25th January 2020. The number of cases has since increased, especially in March 2020 and April 2020. This rise in cases of COVID-19 outbreaks in Malaysia has urged several actions to be taken, including the installation of an investigation system for the immediate detection of cases; rapid diagnosis; immediate isolation of cases and rigorous tracking; and the quarantining of close contacts of those positively tested with COVID-19 (Abdullah et al. 2020).

Since then, the Crisis Preparedness Response Center (CPRC) of the Ministry of Health Malaysia has begun to record and report cases daily. The ministry website offers regular information on new COVID-19 confirmed incidents, recoveries and deaths. Thus, YAB Tan Sri Muhyiddin Hj Mohd Yassin, Malaysia's Prime Minister made a press release, the Malaysian government agreed to enforce a nationwide first step Restricted Movement Order (RMO) from 18th March to 31st March. This order is enforced under the Control and Prevention of Infectious Diseases Act 1988 and the Police Act 1967, with the aim of isolating the source of the outbreak of COVID-19. According to Malaysian National Security Council (NSC) (2020), a number of activities, including operating, are not allowed during RMO, except for essential services. RMO deployed in four stages: Phase I (18 March to 31 March), Phase II (1 April 2020 to 14 April 2020), Phase III (15 April 2020 to 28 April 2020) and Phase IV (29 April 2020 to 12 May 2020). RMO was made to prohibit the citizen movement and mass assembly nationwide would include all religious, sports, social and cultural activities.

Several predictive and modelling researches was conducted worldwide on COVID-19. Zhao et al. (2020) suggested a statistical model for estimating the actual number of COVID-19 cases not identified in the first half of January 2020. Nishiura et al. (2020) proposed an estimated COVID-19 infection rate model in Wuhan, China using data from 565 Wuhan-evacuated Japanese people between 29th and 31st January 2020. They conclude that prevalence is calculated at 9.5%, and the risk of death is 0.3 to 0.6%. However, the number of Japanese people evacuated from Wuhan is limited and inadequate to estimate infection and death. Tang (2020) proposed a mathematical model to estimate COVID-19 transmission likelihood. They concluded the basic number of reproductions could be 6.47. They also expected confirmed cases within 7 days (23rd January - 29th January, 2020). Therefore, they predicted hitting the plateau after two weeks (as of 23rd January, 2020). In Thompson (2020), data from 47 patients were used to estimate continuous human-to-human transmission of COVID-19. The author concluded that the transmission is 0.4, and if the duration of the hospitalization symptom is half the data tested, the transmission is just 0.012. The study of Muhammad Rezal et al. (2020) provided an approximation of the Susceptible Infected Recovered (SIR) Model for the COVID-19 outbreak in Malaysia for short-term COVID-19 cases. The author pointed out that the transmission rate is 0.22%, meaning an infectious individual can transmit or spread the disease to 1 person (on average) in four days. The transmission rate corresponding to a potential scenario that a person would infect another person within 4 days should not be taken lightly. Therefore, a one-to-one transmission at 4-day intervals can be considered very conservative.

This paper established trend analysis and prediction model to predict the new regularly reported cases of COVID-19 in Malaysia for a few months. Mann-Kendall test was widely used to classify significant trends in time series (Hamzah et al. 2017; Samsudin et al. 2017). The identical distribution of Mann-Kendall's Test for a n independent series and data $(x_1,x_2,\ldots,x_n)$ suggests the previous acceptance of the null hypothesis H_0 (non-trend). Using the Mann-Kendall 's Test, it is possible to evaluate the presence of an increasing or decreasing trend, but it is difficult to quantify because the tool, Sen's Slope Estimator developed by Sen (1968), allows to calculate the slope of the regression line for each parameter at any time without the influence of outliers (Bouza-Deaño et al. 2008). In this research, Singular Spectrum Analysis (SSA) was used as a basis approach to forecasting model development. In general, SSA provides a representation of a univariate time series that is transformed in terms of the eigenvalue values and the eigenvectors of the trajectory matrix. SSA is useful to separate time series data into trend, seasonal, and noise by decomposing its time series eigen and reconstructing it into group selection. However, separating the components in this approach depends on the selection of parameters which is the choice of window length, L forming the trajectory matrix and identifying the number of leading components, r based on eigenvector plot. This separation is very important to ensure that trend, seasonal and noise components are easily separated. Hence, in this study, different lengths of L and r selection were tested on COVID-19 data to investigate its effect on component extraction. Details of the research will be addressed in the

next sections where the data obtained in this review will be discussed in section two, accompanied by methods, conclusions and discussions, and the final part will be the conclusion. A Singular Spectrum Analysis-based forecasting model is introduced, discriminating noise in a time-series trend. The Recurrent Forecasting-SSA (RF-SSA) model assessment is based on the official data of the World Health Organization (WHO) COVID-19 to predict daily confirmed cases from 29th April to 9th May 2020. The aims are to gain a better understanding of the trend in the situation and recovery phase over the duration of RMOs and to calculate and forecast COVID-19 cases in Malaysia for the next 10 days using previously reported cases.

DATA DESCRIPTION

Daily Coronavirus Disease 2019 (COVID-19) prevalence data were collected from 25th January 2020 to 29th April 2020 as reported on the official website of the Ministry of Health of Malaysia. As this COVID-19 is a newly discovered virus, no previous year data is available. Reverse Transcription Polymerase Chain Reaction (RT-PCR) diagnosed the suspected COVID-19 cases and confirmed as COVID-19 victim. All fully anonymised, laboratory-confirmed cases of COVID-19, in which 5,945 cases of COVID-19 infection have been reported by the Ministry of Health in 16 states in Malaysia.

Figure 1 indicates Malaysia's total positive COVID-19 cases. This shows a large rise in the number of positive cases related to the second wave of COVID-19 pandemic in Malaysia. With this substantial amount, Malaysia's government declared a Restriction Movement Order (RMO)/Movement Control Order (MCO) from 18th March to 31st March 2020. RMO was later expanded to phase four.

Figure 2 shows Malaysia's number of cases reported for COVID-19 over the last 96 days. The Ministry of Health (MOH) has categorized by area number four COVID-19 areas in Malaysia. According to the National Security Council (MKN), the four zones are: green zone for non-positive areas, yellow zone for areas with 1 to 20 positive cases, orange zone for areas with 21 to 40 positive cases, and red zone for areas with more than 40 positive cases (Kannan et al. 2020).

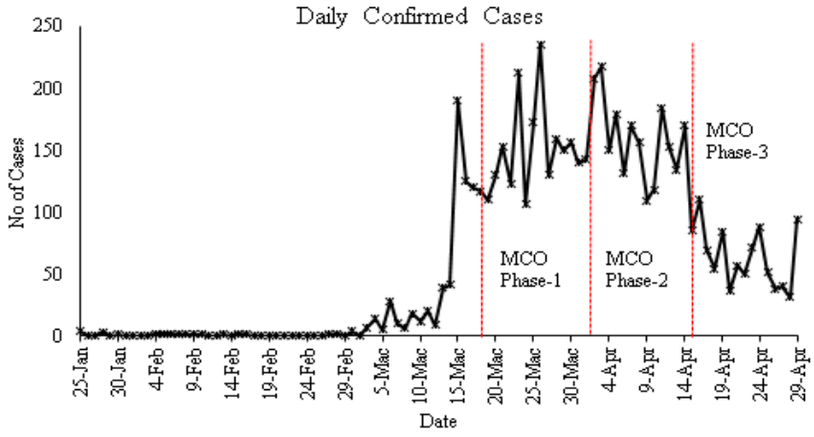


FIGURE 1. COVID-19 Daily confirmed cases in Malaysia from 25th January 2020 until 29th April 2020

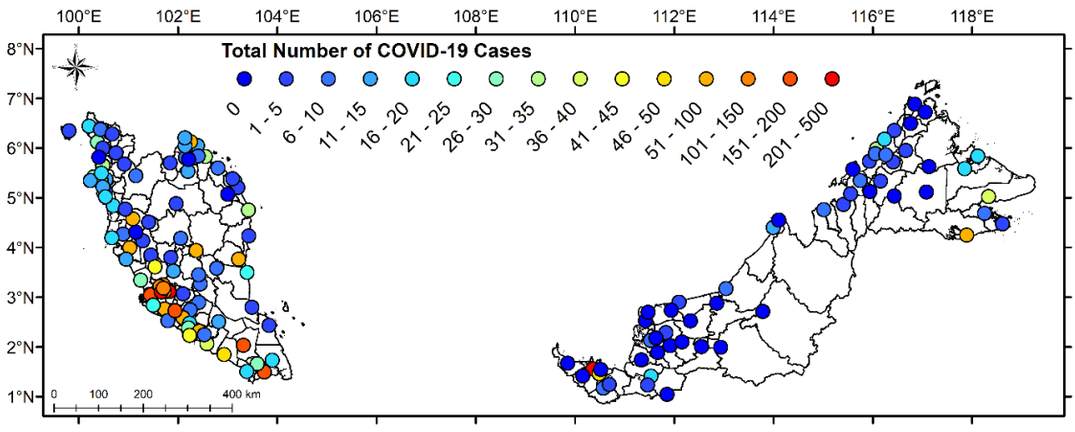


FIGURE 2. State classification according to number of COVID-19 cases in Malaysia

METHODS

This section explains Trend Analysis specifics using Mann-Kendall Test, Singular Spectrum Analysis Model, and its components. The Mann-Kendall test was performed using COVID-19 outbreak data during Malaysia's RMO process. The method is based on correlating observed variables with their time series. Usually, Mann-Kendall's non-parametric statistical test was used to assess the significance of a site trend (Mann 1945). Statistics of the Mann-Kendall, S is defined as:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sgn\,(X_o - X_i) \tag{1}$$

where X_o are the sequential data values, n is the length of the data set, and

$$sgn\,(\theta) = \begin{cases} 1 & for \quad \theta > 0 \\ 0 & for \quad \theta = 0 \\ -1 & for \quad \theta < 0 \end{cases} \tag{2}$$

Kendall (1975) has observed that, when $n > 8$, the statistic S is approximately normally distributed with the mean and variance given by:

$$E[S] = 0 \tag{3}$$

$$Var(S) = \frac{n\,(n-1)(2n+5) - \sum_{i=1}^{n} t_i i(i-1)(2i+5)}{18} \tag{4}$$

where t_i is the number of ties of extent i. The standardized test statistic Z is computed by:

$$Z = \begin{cases} \frac{S-1}{\sqrt{Var(S)}} & for \qquad S > 0 \\ 0 & for \quad S = 0 \; fracS + 1\sqrt{Var(S)} \; for\, S \leq 0 \end{cases} \tag{5}$$

Under the null hypothesis of no trend, the standardized Mann-Kendall statistic Z follows a standard normal distribution with mean zero and variance one. A positive Z value indicates an upward trend, while a negative one indicates a downward trend. The p-value (probability value p) of the Mann-Kendall statistic S sample data can be estimated using the normal cumulative distribution function:

$$P_{lm} = \Pr\,(r_{t+1} = m | r_{t+1} = 1), plm \geq 0, \sum_{m=1}^{M} Plm = \tag{6}$$

where

$$\Phi(|Z|) = \frac{1}{2\pi} \int_0^{|Z|} e^{\frac{-t^2}{2}} dt \tag{7}$$

If the p-value is small enough, the trend is possibly due to random sampling. At the significance level of 0.05, if $p < 0.05$, the current trend is assessed as statistically significant.

If a linear trend exists, the true slope can be estimated by (a) computing the slope's least square estimate, or (b) linear regression methods. However, (b) can deviate greatly from the true slope if the data set contains gross errors or outliers. Sen (1968) developed a method called Sen's method which is not greatly affected by gross data errors or outliers and can be calculated when data is missing. This test is close to the Mann-Kendall test (Kendall 1975).

To obtain the Sen's slope estimator, it is first necessary to calculate the N' slope estimates, Q, as:

$$Q = \frac{x_{i\prime} - x_i}{i' - 1} \tag{8}$$

where x'_i and x_i are data values at times i' and i, respectively, and where $i' > i$; N' is the number of data pairs for which $i' > i$ is used. Q's median of these N' values is Sen's slope estimator. If there is only one datum for each period of time, then

$$N' = \frac{n(n-1)}{2} \tag{9}$$

The results of the Mann-Kendall trend test are then interpreted. Test processing with p-values < 0.05 indicates that there is a significant difference for that particular test. If the Sen's slope shows a positive value, there's an upward trend and *vice versa*. For the test showing the p-value > 0.05, there is no significant difference for the parameter.

SINGULAR SPECTRUM ANALYSIS (SSA) MODEL

Singular Spectrum Analysis (SSA) is a model-free method that can be applied to all data forms, whether gussian or non-gussian, linear or non-linear, stationary or non-stationary (Shaharudin et al. 2020). Daily COVID-19 data can be decomposed into a number of additive components *via* SSA that can be defined in trend, seasonal, and noise components (Shaharudin et al. 2019; Suhartono et al. 2019). Possible SSA implementations are diverse (Alexandrov et al. 2008; Chau & Wu 2010; Rodriguez-Aragon & Zhiglkavsky 2010). SSA comprises two complementary stages of decomposition and reconstruction (Carvalho & Rua 2014).

STAGE 1: DECOMPOSITION

In the decomposition stage, two steps are embedding and singular value decomposition (SVD). Generally, 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 (10). 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 of the trajectory matrix, **X** are called lagged vectors, $K = T - L + 1$.

$$\mathrm{X} = (X_1, \dots, X_K) = \begin{pmatrix} y_1 & y_2 & y_3 & \cdots & y_K \\ y_2 & y_3 & y_4 & \cdots & y_{K+1} \\ y_3 & y_4 & y_5 & \cdots & y_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & y_{L+2} & \cdots & y_T \end{pmatrix} \quad (10)$$

Step II: Singular Value Decomposition (SVD). In the second step, trajectory matrix in Step I is 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

$$\mathrm{X} = U^T \Sigma V \quad (11)$$

where $U = (u_1, \dots, u_L)$ is an $L \times L$ orthogonal matrix, $V = (v_1, \dots, v_k)$ is a orthogonal matrix and Σ is an $L \times K$ diagonal matrix with nonnegative real diagonal entries $\Sigma_{ii} = \sigma_i$ for $i = 1, \dots, L$. The vectors u_i are called left singular vectors, the v_i are the right singular vectors and the σ_i are the singular values. Let $\mathrm{S} = \mathrm{XX}^{\mathrm{T}}$ where the singular values be arranged in descending order such that ($\sigma_1 \ge \sigma_2 \ge \dots \ge \sigma_L$). Let $d = max\{\mathrm{i}, \text{ such that } \sigma_i > 0\}$. $V_i = X^T U_i / \sqrt{\sigma_i}\ (i = 1, \dots, d)$, then, the SVD of the trajectory matrix X can be written as

$$\mathrm{X} = \mathrm{X}_1 + \cdots + \mathrm{X}_\mathrm{d} \quad (12)$$

where $X_i = \sigma_i \mu_i v_i^T$. Note that, the matrices of X_i are called elementary matrices if X_i has rank one. The collection (σ_i, μ_i, v_{ij}) is called the eigentriple of the SVD.

STAGE 2: RECONSTRUCTION

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

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

$$\mathrm{X}_I = \mathrm{X}_{i1} + \cdots + \mathrm{X}_{ip} \quad (13)$$

The resultant matrices are computed for $I = I_1,\dots, I_m$ and substituted in (14). The expansion is defined as

$$\mathrm{X} = \mathrm{X}_{I1} + \cdots + \mathrm{X}_{Im} \quad (14)$$

where the trajectory matrix is represented as a total of resultant matrices. The selection of the sets $I = I_1,\dots, I_m$ is known as eigentriple grouping.

Step 2: Diagonal averaging. Final step in SSA transforms each matrix of the grouped decomposition (14) into a new series of length T.

Let Z be an $L \times K$ matrix with elements $z_{ij}, 1 \le i \le L, 1 \le j \le K$. Set $L^{\wedge *} = \min(L, K)$, $K^{\wedge *} = \max(L, K)$ and $N = L + K - 1$. Let $zij^{\wedge *} = z_{ij}$ *if* $L < K$ and $z_{ij}^{\ *} = z_{ji}$. By making the diagonal averaging, we transfer the matrix Z into the $z_1, \dots, z_T$ using the formula

$$z_k \begin{cases} \frac{1}{k}\sum_{m=1}^{k} z^*_{m,k-m+1} & 1 \le k < L^* \\ \frac{1}{L^*}\sum_{m=1}^{L^*} z^*_{m,k-m+1} & L^* \le k \le K^* \\ \frac{1}{T-k+1}\sum_{m=k-k^*+1}^{T-K^*+1} z^*_{m,k-m+1} & K^* < k \le N \end{cases} \quad (15)$$

Diagonal averaging in (15) applied to a resultant matrix X_{Ik} produced reconstructed series $\tilde{\mathbb{Y}}_T^{(k)} = (\tilde{y}_1^{(k)}, \dots, \tilde{y}_T^{(k)})$. Hence, the initial series $\mathbb{Y}_T = \{y_1, y_2, \dots, y_T\}$ is decomposed into a sum of m reconstructed series, $y_t = \sum_{k=1}^{m} \tilde{y}_t^{(}$. The reconstructed series produced by the elementary grouping will be called elementary reconstructed series.

FORECASTING WITH SSA MODEL

In making the SSA forecasting, a fundamental condition is that the time series satisfies a linear recurrent relation (LRR). A time series $Y_T = (y_1, \dots, y_T)$ satisfies LRR of order d if there exists the coefficients $a_1, \dots, a_d$ such that:

$$y_{i+d} = \sum_{k=1}^{d} a_k y_{i+d-k}, \qquad 1 \le i \le N - d,\ a_d \ne 0, d < N \quad (16)$$

In this study, Recurrent SSA is used for forecasting purpose since it is a popular approach when predicting data (Danilov 1997). SSA forecasting is based on the multiplication of a weight that obtained from the eigenvector.

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,..., r). Denoting $v^2 = \sum_{j-1}^{r} \pi_j^2$ we define the coefficient vector as:

$$\Re = \frac{1}{1-v^2} \sum_{j=1}^{r} \pi_j \; U_j^{\nabla} \tag{17}$$

The recurrent SSA forecasting algorithm can be presented as follows.

1. The time series $Z_{N+h} = \{z_1, \ldots, z_{N+h}\}$ is defined by

$$z_i = \begin{cases} \tilde{y}_i, & i = 1, \ldots, T \\ \sum_{j=1}^{L-1} a_j z_{i-j}, & i = T+1, \ldots, T+h \end{cases} \tag{18}$$

2. The numbers $Z_{T+1}, \ldots, z_{N+h}$ are the h step ahead recurrent forecasts.

These algorithms described as follows and further details can be found in Hossein et al. (2017).

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 Mondal et al. (2012), L should be large enough but not greater than half of the number of observations understudy at $T/2$. However, the appropriate selection of window length is dependent on the current problems as well as the structure of the time series data (Alonso et al. 2009). Generally, there is no rough guide to determine the proper L in data set (Shaharudin et al. 2015). 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, which are $T/2, T/5, T/10, T/20$ were investigated on COVID-19 data based on performance error which is Root Mean Square Error (RMSE).

Another parameter that needs 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 for proper grouping of trend and noise separation by the component. Besides, it could also reflect a connection between decomposition and reconstruction stages.

RESULTS AND DISCUSSION

The Mann-Kendall Test was used to study the 77-day trends from 18th March 2020 to 12th May 2020 of COVID-19 cases in Malaysia. This study shows either an upward trend, a downward trend or no (NT) trend. When a trend was detected (positive or negative), quantified levels of the above-mentioned trend were determined using a Sen slope with a p-value of 0.05. Table 1 shows clearly that during the RMO period, new cases of RMO3 showed a declining trend pattern compared to RMO1, RMO2, and RMO4. During Phase I RMO, the outbreak of COVID-19 has increased dramatically across the country since 15th March 2020. On the basis of the investigation, the majority of these additional cases relate to the Seri Petaling Cluster (Ministry of Health Malaysia (MOH), Malaysia 2020). Figure 3 shows statistically downward trends following the implementation of RMO Phase II, RMO Phase III, and RMO Phase IV.

As mentioned in the previous section, the Malaysian daily COVID-19 cases were predicted using the SSA model. The SSA predicting algorithm known as Recurrent Forecasting was used to predict future cases from 29th April 2020 to 9th May 2020. At the time of this experiment, historical cases were used from 25th January 2020 to 29th April 2020 and future 10 days ahead of COVID-19 cases were predicted accordingly. Figure 6 shows the confirmed cases from 25th January 2020 to 29th April 2020 and the forecasted daily cases until 9th May 2020.

The first stage of this study is the decomposition of COVID-19 data into components facilitated by the SSA model. This decomposition by SSA requires the identification of the parameter pair. The choice of L is a compromise between information content and statistical confidence (Hassani 2007). The value of the appropriate should be able to clearly resolve the various oscillations hidden in the original signal.

The performance of the SSA results was determined by evaluating its weighted correlation, i.e. w-correlation at distinct window length, The W-correlation, as explained in the Methods section, calculated the separation between trend, seasonal and noise components of the reconstructed time series. A number of L selections which were L =

$T/2$, $T/5$, $T/10$ and $T/20$, representing $L = 48$, 19, 10 and 5, respectively, for T based on 96 daily COVID-19 data cases, were selected. These scales were chosen to fit the time series data as well as to strike a balance in order to achieve a proper lag vector sequence.

Figure 4 shows the w-correlation of day-to-day COVID-19 data with different window lengths using SSA. As can be seen from the plot, the W-correlation shows a declining pattern as the total window length declines for the SSA approach. The correlations between trends and other components need to be closed to zero for trend extraction. This means that distinct window lengths have a certain effect on the separability of the component. It also shows that SSA is directed to the lowest w-correlation at window length, $L = T/20$ indicating the strongest separability between the reconstructed components as it is nearest to zero.

Root mean square error (RMSE) is used to evaluate the performance of L. Table 2 presents the components of the reconstructed time series at $L = T/20$, which has the smallest RMSE compared to the other L. It is noted that higher RMSE values are obtained in this study due to high model variance when the number of samples is small (Boehmke & Greenwell 2019). In summary, the analysis of daily COVID-19 case data appears to suggest that $L = T/20$ is suitable on the basis of a short time series of outbreak data.

Figure 5 shows the components of the reconstructed time series plot based on two eigentriple (ET) from the trend of RF-SSA for daily COVID-19 cases in Malaysia. Reconstructed series is a new set of data formed from original data, clear from noise. It is a very important part of SSA to ensure that the forecasting results obtained are more precise and accurate (Hassani & Zhigljavsky 2009). The trend component of time series data is used to observe the occurrence of the trend and pattern of cases as randomly tabulated as per day cases, as shown in Figure 5. The trend in Figure 5(a) and Figure 5(b) is precisely generated by the leading eigentriple, coinciding with the first reconstructed component in Figure 5. In the meantime, the trend in Figure 5(b) is generated precisely by the two leading eigentriples, coinciding with the first and second reconstructed components in Figure 5. The straight and dashed lines in the plot apply, respectively, to the original time series COVID-19 data and the reconstructed series based on SSA's extracted trend components. The plot of the reconstructed time series components produced by both leading eigentriple follows the original COVID-19 data, although there is noise component omission specifically for $L = 5$ in Malaysia for daily COVID-19 cases.

Detailed analysis showed that the RF-SSA obtained a Mean Absolute Error (MAE) of 11.00 and 19.12 for Root Mean Square Error (RMSE). Meanwhile, Pearson correlation (r) of 0.96 near 1.0 indicates a good correlation between confirmed and predicted cases. Finally, the Mean Forecast Error (MFE) shows that the RF-SSA algorithm tends to over-predict COVID-19 cases by 2.8%. Figure 6 also showed that RF-SSA model predicted the daily verified COVID-19 cased in Malaysia would hit one digit by early-June 2020.

As COVID-19's number of daily cases was small, Figure 6 shows a noticeable but vulnerable decreasing pattern from 27th March 2020. One of the contributing factors to the slightly decreasing trend was the Movement Control Order (MCO) declared by the Malaysian government on 18th March to 31st March 2020. The prediction plot using RF-SSA in Figure 6, showed a general pattern of nonlinear rising trend in the daily confirmed COVID-19 cases in Malaysia.

The prediction and estimate of daily cases of COVID-19 obtained was influenced by the case description reported to CPRC daily, the large number of pending outcome test daily was certainly influential to a non-consistent increase in the number of confirmed cases. Several of the largest clusters found by the Ministry of Health Malaysia, such as Seri Petaling Tabligh Cluster, Wedding Reception in Bandar Baru Bangi, Seri Petaling Sub-Cluster in Rembau, Italy Cluster in Kuching, Sarawak and Church Fellowship Cluster in Sarawak, support the prediction case increase.

Although more data are needed to have detailed prediction, to date, virus spread has decreased, and the number of daily deaths has decreased steadily. The number of reported cases, however, has yet to hit one digit. Limitation in this research is discussed and should be emphasized when using the RF-SSA model, particularly pandemic data in Malaysia. First, RF-SSA model works best when data shows a stable or consistent pattern over time with a minimum outlier. It will help to obtain accurate and accurate results for future predictive cases. Next, the sudden increase in data will result in a low performance of the forecasting results using this predictive RF-SSA model. After that, the RF-SSA model is mainly used to project future values using historical time series data for short-term forecasting. Lastly, the recurrent approach is a better contender than the vector approach for forecasting SSA data in short and medium time series. However, under such scenarios, users should also evaluate the performance of the SSA forecasting approach on their data for a complete picture.

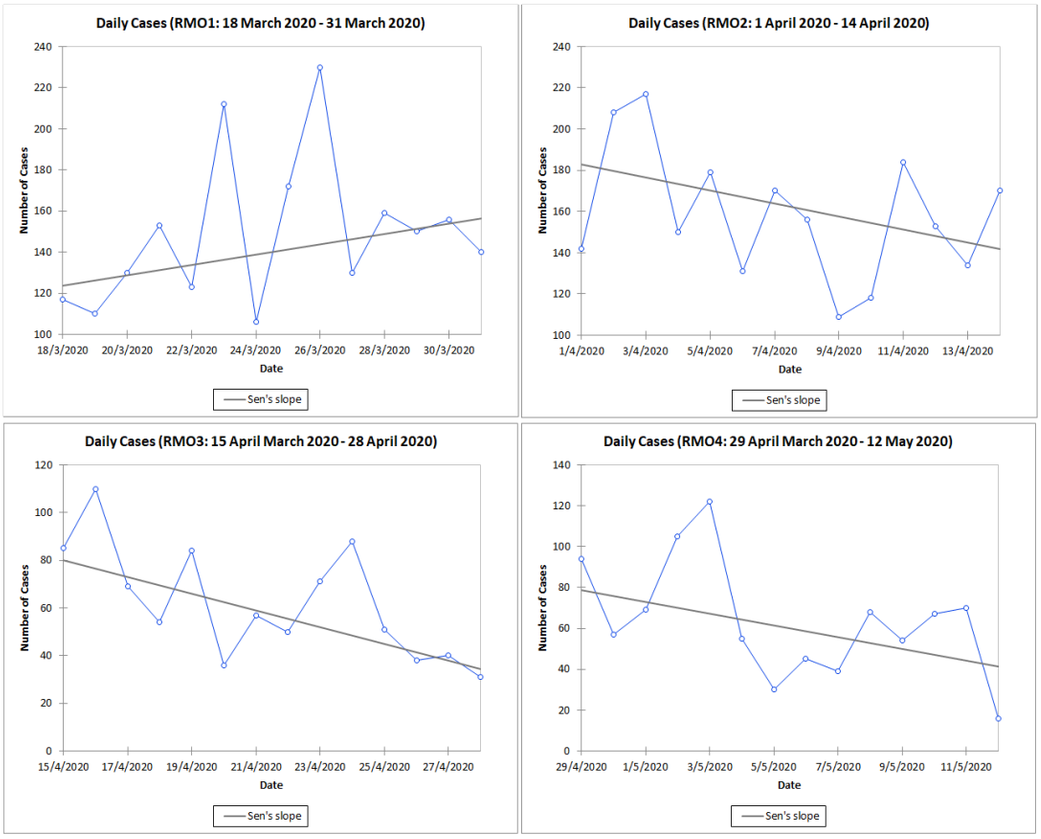


FIGURE 3. Comparison of Sen's slope between New Cases during 4 RMOs

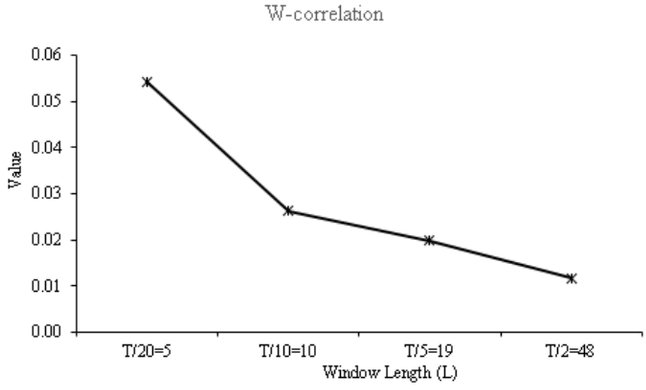
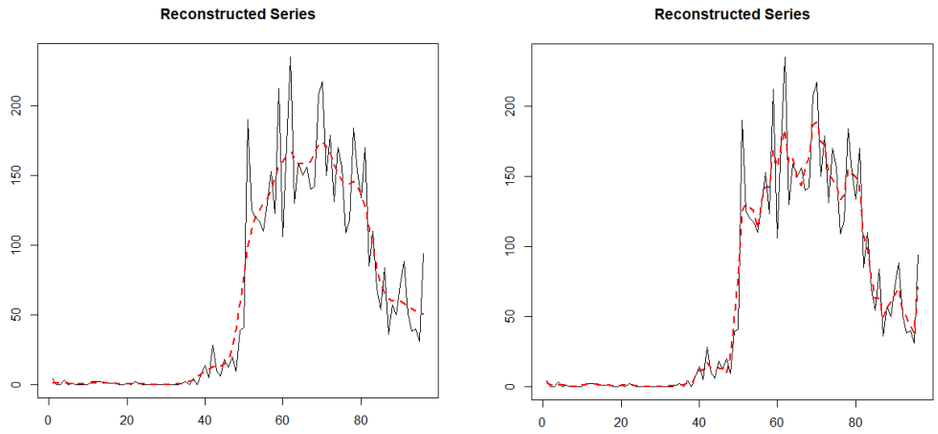


FIGURE 4. Effect of w-correlation based on SSA using COVID-19 data at different window lengths



(a) (b)

FIGURE 5. Plot of daily COVID-19 cases of reconstructed components from extracted trends using SSA at (a) $L = 5$, ET 1 (b) $L = 5$, ET 2

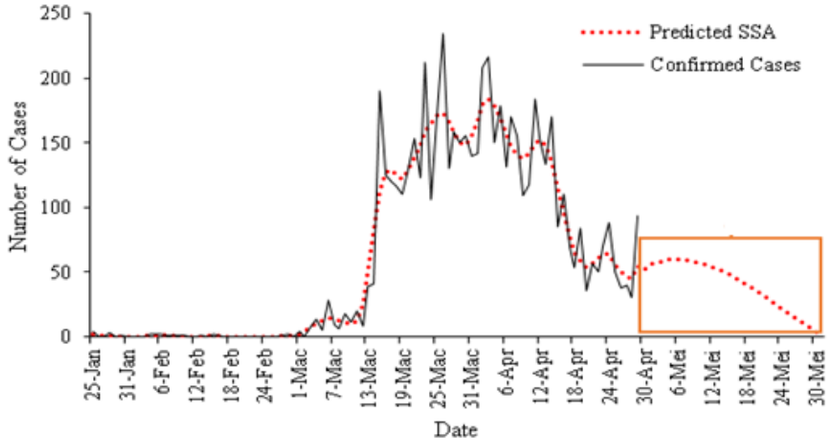


FIGURE 6. Confirmed cases versus predicted RF-SSA of COVID-19 in Malaysia

TABLE 1. Summary of Mann-Kendall test value for new cases in Malaysia during RMO

TEST	RMO	Kendall's tau	*p*-value	Sen's slope	Trend
CASES	RMO1	0.2652	0.2073	2.5000	NT
	RMO2	-0.1768	0.4108	-3.1667	NT
	RMO3	-0.4725	**0.0215**	-3.5000	↓
	RMO4	-0.2747	0.1889	-2.8889	NT

TABLE 2. The performance of comparison prediction model based on SSA for several

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

CONCLUSION

Currently, confirmed cases reported on a daily basis show a declining and plateauing trend. As the number of COVID-19 cases recovered increased, this led to a decrease in active cases during Phase IV of the RMO. This contributed to the flattening of the curve and the nation is now entering recovery. This statement supported the outcome of the Mann-Kendall test of this study, which showed downward trends in all cases and new cases in the states. In addition, this paper studies the applicability of the RF-SSA model to the prediction of COVID-19 cases in Malaysia. The application of this model is particularly advantageous for the health authorities in terms of flattening the curve by preparing a timely and effective strategy. In addition, this model allows health authorities to better understand the pattern of the outbreak. It was found that the pattern follows the RF-SSA model that can be used to predict the growth pattern of outbreak cases in Malaysia. Using this model, the selection of the parameter is the choice of the length of the window, *L* and the total number of eigentriples employed for reconstruction, *r*. These results show that the parameter $L = 5$ (T / 20) was suitable for use in short time series outbreak data and that it is important to obtain an appropriate number of eigentriples which will have an effect on the forecasting result. Overall, the results showed that the RF-SSA model was able to predict this pandemic with reasonable accuracy as the model over-forecasted by 0.36% with high correlation values between confirmed and predicted cases. However, the RF-SSA model is not capable of capturing the sudden drop in COVID-19 cases, likely due to the RMO, which was extended to 12th May 2020. In order to improve the accuracy of the model, more information is needed to better predict COVID-19 cases over a long period of time. In the meantime, case definition and data collection must be maintained in real time in order to improve the RF-SSA for further study. It is suggested that the RF-SSA model be enhanced so that the model can capture sudden and rapid changes in the data set.

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 the Ministry of Education, Malaysia.

REFERENCES

Abdullah, S., Mansor, A.A., Napi, N.N.L.M., Mansor, W.N.W., Ahmed, A.N., Ismail, M. & Ramly, Z.T.A. 2020. Air quality status during 2020 Malaysia Movement Control Order (RMO) due to 2019 novel coronavirus (2019-nCoV) pandemic. *Science of The Total Environment* 729: 139022.

Alexandrov, T., Golyandina, N. & Spirov, A. 2008. Singular spectrum analysis of gene expression profiles of early drosophila embryo: Exponential-in-distance patterns. *Research Letters in Signal Processing* 2008: Article ID. 825758.

Alonso, F.J., Salgado, D.R., Cuadrado, J. & Pintado, P. 2009. Automatic smoothing of raw kinematics signals using SSA and cluster analysis. *Euromech Solid Mechanics Conference Lisbon*. pp. 1-9.

Boehmke, B. & Greenwell, B 2019. *Hands-On Machine Learning with R. Broken Sound*. Parkway NW: Taylor & Francis. pp. 1-15.

Bouza-Deaño, R., Ternero-Rodríguez, M. & Fernández-Espinosa, A.J. 2008. Trend study and assessment of surface water quality in the Ebro River (Spain). *Journal of Hydrology* 361(3-4): 227-239.

Carvalho, M.D. & Rua, A. 2014. *Real-Time Nowcasting the US Output GAP: Singular Spectrum Analysis at Work.* Portugal: Banco De Portugal.

Chau, K.W. & Wu, C.L. 2010. Hybrid model coupled with singular spectrum analysis for daily rainfall prediction. *Journal of Hydroinformatics* 12(4): 458-473.

Danilov, D. 1997. The Caterpillar method for time series forecasting. In *Principal Components of Time Series: The Caterpillar Method.* Russian: University of St. Petersburg.

Gilbert, R.O. 1987. *Statistical Methods for Environmental Pollution Monitoring*. New York: John Wiley & Sons. pp. 23-52.

Hamzah, F.M., Saimi, F.M. & Jaafar, O. 2017. Identifying the monotonic trend in climate change parameter in Kluang and Senai, Johor, Malaysia. *Sains Malaysiana* 46(10): 1735-1741.

Hassani, H. 2007. Singular spectrum analysis: Methodology and comparison. *Journal of Data Science* 5: 239-257.

Hassani, H. & Zhigljavsky, A. 2009. Singular spectrum analysis: Methodology and application to economics data. *Journal of Systems Science and Complexity* 22(3): 372-394.

Hossein, H., Mahdi, K. & Masoud, Y. 2017. An improved SSA forecasting result based on a filtered recurrent forecasting algorithm. *Statistics/Theory of Signals* 355(9): 1026-1036.

Kannan, S., Ali, P.S.S., Sheeza, A. & Hemalatha, K. 2020. COVID-19 (Novel Coronavirus 2019)-recent trends. *European Review for Medical and Pharmacological Sciences* 24(4): 2006-2011.

Kendall, M.G. 1975. *Rank Correlation Measures*. London: Charles Griffin. pp. 11-36.

Malaysian National Security Council (NSC). 2020. *Movement Control Order (RMO).* https://www.mkn.gov.my/web/ms/COVID-19/. Accessed on 13 May 2020.

Mann, H.B. 1945. Nonparametric tests against trend. *Journal of the Econometric Society* 13(3): 245-259.

Malaysia Ministry of Health Malaysia (MOH). 2020. *Press Statement Updates on The Coronavirus Disease 2019 (COVID-19) Situation in Malaysia.* https://www.moh.gov.my/index.php/pages/view/2019-ncov-wuhan-kenyataan-akhbar Accessed on 13 May 2020.

Ministry of Health Malaysia. 2019. *Coronavirus Website*. http:/covid19.moh.gov.my/.

Mondal, R.A., Kundu, S. & Mukhopadhyay, A. 2012. Rainfall trend analysis by Mann-Kendall test: A case study of North-Eastern part of Cuttack district, Orissa. *International Journal of Geology* 2: 70-78.

Muhammad Rezal Kamel Ariffin, Kathiresan Gopal, Isthrinayagy Krishnarajah, Iszuanie Syafidza Che Ilias, Mohd Bakri Adam, Noraishah Mohammad Sham, Jayanthi Arasan, Nur Haizum Abd Rahman & Nur Sumirah Mohd Dom. 2020. Malaysian COVID-19 Outbreak Data Analysis and Prediction. *Institute for Mathematical Research.* http://einspem.upm.edu.my/covid19maths/file/Report_001%20v13.pdf.

Nishiura, H., Kobayashi, T., Yang, Y., Hayashi, K., Miyama, T., Kinoshita, R., Linton, N.M., Jung, S.M., Yuan, B., Suzuki, A. & Akhmetzhanov, A. 2020. The rate of under ascertainment of novel coronavirus (2019-nCoV) infection: Estimation using Japanese passengers data on evacuation flights. *Journal of Clinical Medicine* 9(2): 1-3.

Rodriguez-Aragon, L.J. & Zhiglkavsky, A. 2010. Singular spectrum analysis for image processing. *Statistics and Its Interface* 3(3): 419-426.

Samsudin, M.S., Khalit, S.I., Juahir, H., Nasir, M., Fahmi, M., Kamarudin, M.K.A. & Lananan, F. 2017. Application of Mann-Kendall in analyzing water quality data trend at Perlis River, Malaysia. *International Journal on Advanced Science, Engineering and Information Technology* 7(1): 78-85.

Sen, P.K. 1968. Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association* 63(324): 1379-1389.

Shaharudin, S.M., Ahmad, N., Mohamed, N.S. & Aziz, N. 2020. Performance analysis and validation of modified singular spectrum analysis based on simulation torrential rainfall data. *International Journal on Advanced Science Engineering Information Technology* 10(4): 1450-1456.

Shaharudin, S.M., Ahmad, N. & Zainuddin, N.H. 2019. Modified singular spectrum analysis in identifying rainfall trend over Peninsular Malaysia. *Indonesian Journal of Electrical Engineering and Computer Science* 15(1): 283-293.

Shaharudin, S.M., Ahmad, N. & Yusof, F. 2015. Effect of window length with singular spectrum analysis in extracting the trend signal of rainfall data. *AIP Proceedings* 1643(1): 321-326.

Suhartono, Ashari, D.E., Prastyo, D.D., Kuswanto, H. & Lee, M.H. 2019. Deep neural network for forecasting inflow and outflow in Indonesia. *Sains Malaysiana* 48(8): 1787-1798.

Tang, B., Wang, X., Li, Q., Bragazzi, N.L., Tang, S., Xiao, Y. & Wu, J. 2020. Estimation of the transmission risk of the 2019-nCoV and its implication for public health interventions. *Journal of Clinical Medicine* 9(2): 462-465.

Thompson, R.N. 2020. Novel coronavirus outbreak in Wuhan, China, 2020: Intense surveillance is vital for preventing sustained transmission in new locations. *Journal of Clinical Medicine* 9(2): 498-505.

Zhao, S., Musa, S.S., Lin, Q., Ran, J., Yang, G., Wang, W., Lou, Y., Yang, L., Gao, D., He, D. & Wang, M.H. 2020. 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. *Journal of Clinical Medicine* 9(2): 388-394.

Shazlyn Milleana Shaharudin* & Muhamad Afdal Ahmad Basri
Department of Mathematics
Faculty of Science and Mathematics
Universiti Pendidikan Sultan Idris
35900 Tanjung Malim, Perak Darul Ridzuan
Malaysia

Shuhaida Ismail
Data Analytics, Sciences & Modelling (DASM)
Department of Mathematics and Statistics
Faculty of Applied Sciences and Technology
Universiti Tun Hussein Onn Malaysia
86400 Batu Pahat, Johor Darul Takzim
Malaysia

Mohd Saiful Samsudin
Faculty Business and Entrepreneurship
Universiti Malaysia Kelantan
Kampus Kota
Karung Berkunci 36 Pangkalan Chepa
16100 Kota Bharu, Kelantan Darul Naim
Malaysia

Azman Azid
Faculty of Bioresources and Food Industry
Universiti Sultan Zainal Abidin
Besut Campus
22200 Besut, Terengganu Darul Iman
Malaysia

Mou Leong Tan
GeoInformatic Unit
Geography Section
School of Humanities
Universiti Sains Malaysia
11800 Pulau Pinang
Malaysia

*Corresponding author; email: shazlyn@fsmt.upsi.edu.my

Received: 18 June 2020
Accepted: 8 September 2020