Application of Improved Chaotic Method in Determining Number of *k*-Nearest Neighbor for CO Data Series

Ahmad Basri Ruslan, Nor Zila Abd Hamid

Abstract:** This study is designed to i) apply chaotic approach in predicting Carbon Monoxide (CO) data series and ii) improve the method in determining number of k–nearest neighbor. Chaotic approach is one alternative approach to predict any data series. Prediction through chaotic approach is made after three important parameters which are delay time τ, embedding dimension m and numbers of nearest neighbor k were determined. Therefore, the chaotic approach is applied. In this study, predictions are done to Carbon Monoxide time series observed at Shah Alam in Malaysia. Parameters τ and m are determined through average mutual information and Cao method respectively. While for k, most of the past researches frequently used try and error method. In this study an improvement of the method in determining the number of k is introduced. This improved method is done through plotting the graph of k versus the correlation coefficient (cc) of prediction model. Parameter cc is obtained through the prediction of data series using local mean approximation method (LMAM), local linear approximation method (LLAM) and improved local linear approximation method (ILLAM). Result shows that the cc value of LMAM is 0.9821 with k = 7, LLAM is 0.9873 with k = 3 and ILLAM is 0.9913 with k = 13. Therefore, the improved methods suggest that the optimal value of k is ranged from $3 \leq k \leq 13$. It is hoped that the improved method can be used for future research in developing a better prediction model for chaotic data series.

***Index Terms:** chaotic approach, carbon monoxide, data series, k-nearest neighbor.*

I. INTRODUCTION

Nowadays, human activities lead to the environmental pollutions. It can be water pollutions, soil pollutions, sound pollutions, light pollutions or air pollutions, and it relates to human activities. As for air pollutions, the incomplete combustion of fuel, coal and gas waste from vehicles can lead to emission of Carbon Monoxide (CO) [1]. An excess inhale and breath with CO may lead to several respiratory diseases. It can cause death when people inhale a high concentration of CO for a long time period. Hence, the development of prediction models of the CO concentration time series is a must.

Data series can be classified into two which are deterministic and random [2], [3]. The dynamic of the deterministic time series moves from an initial condition and pass through its trajectory possibly can be traced. Then, a deterministic time series can be predicted. However, a non-deterministic time series cannot be predicted and it is random. Due to the sensitivity on the initial conditions, only short-term prediction is allowed [3]. There are various methods to detect whether the chosen time series is chaotic or not. Previously [4], [5] used phase space plot to detect the presence of chaotic behavior in ozone (O_3) concentration time series. Chaotic approach is not limited to O_3 time series, several studies had been done on river flow modeling by [6] and PM_{10} modeling by [7]. In this study, Cao method and phase space plot are used to detect the presence of chaotic behavior on CO concentration time series.

After the chaotic behavior had been detected, three important parameters need to be determined which are delay time τ, embedding dimension *m* and numbers of nearest neighbor *k* for prediction. The choice of the value τ is important in order to fully capture the structure of attractor [8]. If τ is too small, then the vectors of the phase space are not independent and results in losing the attractors characteristics. But, if τ is too large, the different coordinates may be the almost behaviorally uncorrelated and cause a loss of information from the original system [9]. Several study previously used $\tau = 1$ and the results of their predictions are close to the observed one by measuring with the *cc* near to 1 [10], [11]. *m* will be computed through Cao method for an optimal value of *m* [4]. However, parameter *k* is commonly chosen by try and error. Certain value had been used such as $k = 200$ [5], $k = 100$ [12], $k = 50$ [13]. In previous studies, [6], [14] used $k = 2m$. Recent studies, [15] in her thesis, she varies the number of *k* which are k = 5, 10, 20, 30, 50, 100, 200, 1000, 2000. [16] earlier states that, a smaller value of *k* gives an excellent results. It is not clearly stated what is that small number really is. In this paper, an improved method in determining the value of *k* will be tested from $1 \leq k \leq 200$. This improved method is done through plotting the graph of *k* versus the correlation coefficient (*cc*) of prediction model.

Three methods will be used in this study which are local mean approximation method (LMAM), local linear approximation method (LLAM) introduced by [17] and improved local linear approximation method (ILLAM). ILLAM has been introduced by [5] and this method gives the best performance in predicting.

Manuscript published on 30 September 2019.
***** Correspondence Author (s)

Ahmad Basri Ruslan, Faculty of Science and Mathematics, Departments of Mathematics, Sultan Idris Education University, Tanjong Malim, Perak, Malaysia.

Nor Zila Abd Hamid, Faculty of Science and Mathematics, Departments of Mathematics, Sultan Idris Education University, Tanjong Malim, Perak, Malaysia.

The contributions of this study are to introduce the phase space plot and Cao method for detecting the presence of chaotic behavior, for the first time on CO concentration time series. Besides, this study believes that there is an optimal number of *k* for each prediction model used that gives the best *cc* will be recorded.

II. DATA SERIES

CO time series is observed in Shah Alam. It is a highly populated area in Selangor (Fig. 1). Shah Alam is a city and the state capital of Selangor, Malaysia. There are a lot of attractions that make Shah Alam become one of the most visited places every year such as recreational, sports, education, vacation and many more. This study is the first study in Malaysia using chaotic approach for analyzing the CO time series. Therefore, the time series observed at the benchmark stations are used.

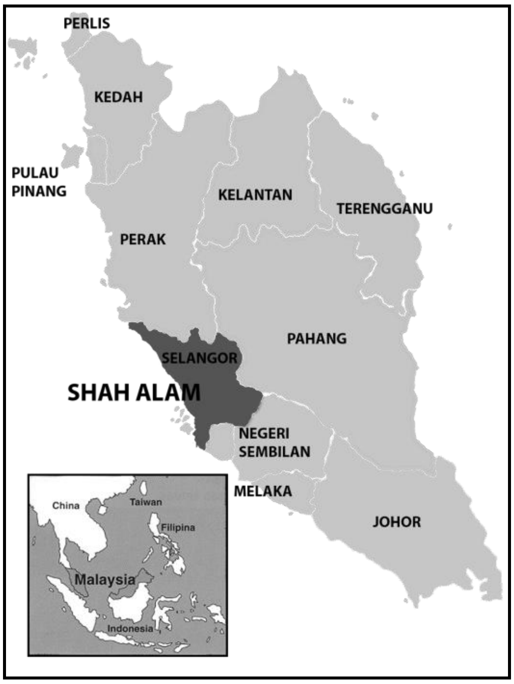

Fig. 1. Map on the location of Shah Alam in Selangor

The data series of CO concentration is observed per hour for one month starting on June 1, 2014 up until June 30, 2014. One month time series period are 30 days consist of 720 hours. The time series is recorded in ppm (part per million) unit and arranged in the scalar form of one-dimensional vector X with

$$X = \{x_1, x_2, x_3, x_4, \dots, x_L\} \tag{1}$$

L is the total number of hours. In this study L = 720. CO concentration time series are divided into two parts of time series. The first series is a training part while the other one is a part to see the performance of the prediction models. Training part is the time series of 552 hours.

$$X_{train} = \{x_1, x_2, x_3, x_4, \dots, x_{552}\} \tag{2}$$

and the next 168 hours,

$$x_{n+1} = Ax_n + B \tag{3}$$

are used for the test parts. Overall time series for 720 hours observed in Shah Alam is illustrated in Fig. 2. As for the statistic details for the CO concentration time series is listed in Table I.

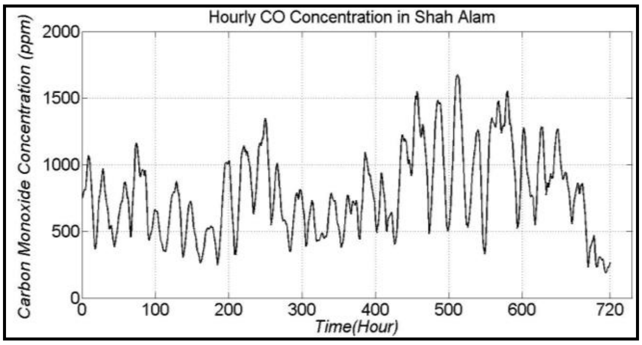

Fig. 2. Hourly CO concentration (ppm) time series.

Table- I: Statistics Details

Statistics	Value
Maximum	1676
Minimum	188.7
Mean	777.5
Median	743.3
Mod	710
Std. Deviation	318.8

III. CHAOTIC NATURE

A. Phase Space Plot

The training part of observed hourly CO concentration was recorded in one-dimensional vector as shown at (2). With $\{x_t\}$ is refer to CO concentration time series at t hour, it has to be shifted or transformed into two-dimensional vector of $\{x_t, x_{t+1}\}$ before building the phase space. However, the parameter of delay time τ needs to be determined. τ is the time interval value to reflect the structure of phase space of the time series. τ can be determined by several methods. Average mutual information and autocorrelation functions were used to determined τ. Previous study such as [10] and [11] using τ =1, as the prediction model gave an excellent results. Since this is the first time where LMAM, LLAM and ILLAM are implemented to CO time series in Malaysia, hence τ = 1 is used.

Since τ = 1 had been set to be used, the phase space plot is built. The existence of an accumulated form called as strange attractor by [18] shows that the nature of the time series is chaotic [19]. Fig. 3 is the phase space plot of a time series of (2) with τ = 1.

It is clearly can be seen that there exists an attractor where most of the points converge towards it. Thus, the observed hourly CO time series is chaotic in nature with τ = 1.

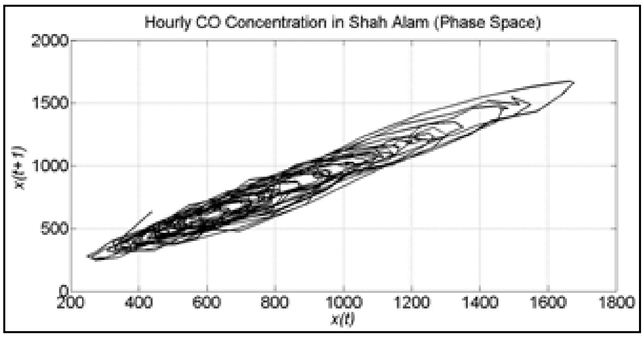

Fig. 3. Phase space plot

B. Cao Method

The series of X_{train} is reconstructed into an m-dimensional phase space.

$$Y_j^m = (x_j, x_{j+\tau}, x_{j+2\tau}, \dots, x_{j+(m-1)\tau}) \tag{4}$$

Delay time τ was set equal to 1 and the minimum embedding dimension m is determined by using Cao method. Cao method can gives two results respectively. It is not only can determine the value of m, it is also can be used to detect the presence of chaotic behavior [4]. Cao method take place in the calculation of two parameters, called E1(m) and E2(m) where m is the variation of the embedding dimension. When E1(m) stops changing when the m value is greater than the value of m0, then m0+1 is the minimum embedding dimension, m. For a random data series, the value of E1(m) will not achieve saturated phase with increasing m. Hence, the graph of m against E1(m) can be used to differentiate whether the nature of time series is chaotic or random. [4] also introduced a calculation of E2(m) for making sure either the data series used is chaotic or random. For random time series, E2(m) will equal to 1 for any m. However for chaotic time series, there will always some E2(m)≠1. Hence, it there exist E2(m) ≠1, then, the observed data series is chaotic. The result in Fig. 4 shows that the value of m0 = 4, E1(m) started saturate within the value 0.95 and 1.00. Thus, the value of m is 5. Since, E2(m) ≠ 1, it is supported enough that the time series used is chaotic. Thus, the prediction model based on chaotic approach is expected to perform well since the nature of the time series is chaotic.

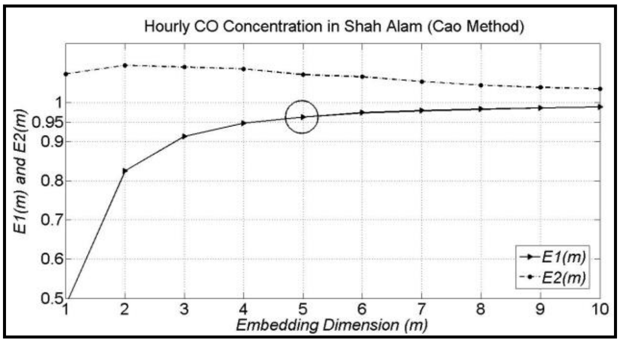

Fig. 4. E1(m) and E2(m) from Cao Method

IV. PREDICTION MODELS

In this paper, three prediction models are used for showing how improved method in determining the number of nearest neighbor used was implemented in making these three prediction models predict better. After the parameter of $\tau = 1$ and $m = 5$ were determined, the last parameter which is k need to be chosen as well. As explained before, *k* was chosen mostly by try and error. In this paper, the previous method in determining the value of k used in past research will be used in the prediction model and will be compared with the optimal number of *k*. Performance of the model is reflected in the calculation of correlation coefficient (*cc*).

A. Local Mean Approximation Method (LMAM)

Prediction process through chaotic approach by using LMAM is interpreted through equation:

$$Y_{j+1}^m = f(Y_j^m) \tag{5}$$

Prediction of Y_{j+1}^m is done based in number of neighbor of Y_j^m. The neighbors of Y_j^m are labelled as Y_j^m where 1< j' <j. For the LMAM, not all neighbors are used to predict Y_{j+1}^m. With parameter k is the number of nearest neighbor, only k nearest number is used. k nearest neighbors are neighbors with the minimum value of Euclidean distance $\| Y_{j'}^m - Y_j^m \|$. After $Y_{j'}^m$ is determined, the one hour ahead $Y_{j'+1}^m$ will be listed. The prediction of Y_{j+1}^m is taken as the average of $Y_{j'+1}^m$ values:

$$\mathbf{Y}_{j+1}^m = \frac{\sum_{q=1}^{k} \left(\mathbf{Y}_{j'_q+1}^m \right)}{k} \tag{6}$$

By testing the number of k use between 1 to 200 as in Fig. 5, the best value of k can be seen through plotting the graph. The highest peak of the plotted graph is the best value of k since it gives the highest value of cc. The best value of k for LMAM is k = 7 and it gives cc = 0.982.

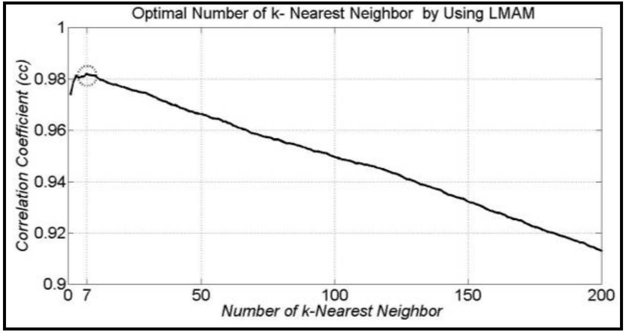

Fig. 5. Prediction by using LMAM with optimal k.

B. Local Linear Approximation Method (LLAM)

For local linear approximation method, a linear equation

$$x_{n+1} = Ax_n + B \tag{7}$$

need to be fitted to the training set (2). The prediction of X_{n+1} is obtained by inserting the value of X_n. The value of A and B is calculated through the least square method. Phase space of (4) is built with τ = 1 and m = 5. Thus the reconstructed phase space is

$$Y_j^5 = (x_j, x_{j+1}, x_{j+2}, x_{j+3}, x_{j+4})$$

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

with $j = 1, 2, 3, ..., L-4$. Since L = 552, the final phase space is

$$Y^5_{548} = (x_{548,}x_{549,}x_{550}, x_{551}, x_{552}) \tag{9}$$

Nearest neighbor to the final phase is sought by calculating the minimum Euclidean distance $\| \mathbf{Y}^5_{552} - \mathbf{Y}^5_{w} \|$ where w < 552. k nearest neighbor are searched and labelled as

$$Y_{n_i} = (Y_{n_1}, Y_{n_2}, Y_{n_3}, ..., Y_{n_k}) \tag{10}$$

A one step forward if Y_n was labelled as

$$Y_{n_i} = (Y_{n_1+1}, Y_{n_2+1}, Y_{n_3+1}, ..., Y_{n_k+1}) \tag{11}$$

m-column values of each Y_{n_i} is searched. The corresponding m-column values of (10) is labelled as

$$x_{n_i} = (x_{n_1}, x_{n_2}, x_{n_3}, ..., x_{n_k}) \tag{12}$$

while the corresponding m-column values of (11) is wrote as

$$x_{n_i} = (x_{n_1+1}, x_{n_2+1}, x_{n_3+1}, ..., x_{n_k+1}) \tag{13}$$

A linear equation $x_{n+1} = Ax_n + B$ of is fitted to both (12) and (13) that formed an linear equation. But, the constant A and B might change depends on the k used. Take k = 50 for illustration of the equation that forms

$$x_{n+1} = 1.1050x_n + 3.0609 \tag{14}$$

In Fig. 6, it shows that the best value of k used that gives the best prediction when k = 3. The lowest value of k need to be chosen as said by [16], the smaller the value of k, the better the prediction value.

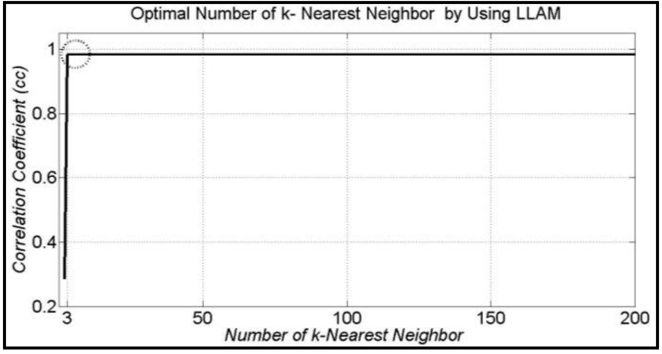

Fig. 6. Prediction by using LLAM with optimal k

C. Improved Local Linear Approximation Method (ILLAM)

In ILLAM prediction model, only one linear equation (7) will be needed. This model has been introduced by [5]. In LLAM prediction model, constant A and B will be evaluated by using X_{train} time series. In ILLAM, X_{train} will be renewed for every new prediction made. This is because the number of data series will be adjusted for new prediction value. For new prediction value, the new equation will be generated as follows:

$$Y^m_{jk^k} = C_n Y^m_{jk} + D_n \tag{15}$$

In other words, LLAM prediction model only generates one equation only. But, in ILLAM, multiple equations were generated depends on the number of X_{train} . Fig. 7 shows the best value of k used for predicting the chosen chaotic data series. It shows that the highest cc = 0.9913 is generated from k = 13.

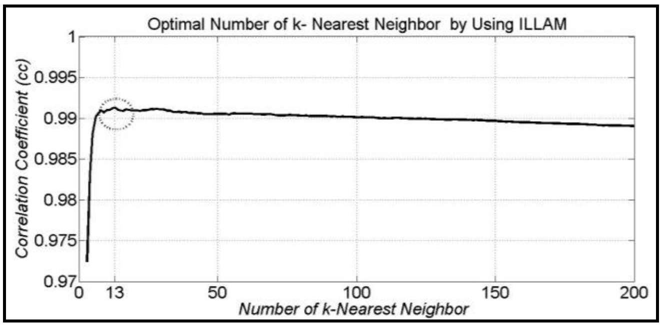

Fig. 7. Prediction by using ILLAM with optimal k

V. DISCUSSION

The chosen range in determining the optimal number of k is between 1 to 200. This is because the previous study randomly try and error up until k = 200. So, this study wanted to see either this improved method in determining the value of k is better than previous method of trying and error. Table II shows the performance of three prediction models with several values of k based on previous study. While in Table III, the performance for those three models is shown by implemented the optimal number of k. By taking cc as performance measure, we can conclude that there is the best value of k for each prediction model that gives the highest cc. The previous method by trying and error is not an exemplary method nowadays. Since, an optimal number of k can be done by this improved method, so this method should be implemented for any prediction model in future.

Specifically for LMAM, the value of cc decreases as k increases (Fig. 5). So, a high value of k is not perfectly give a good cc. By comparing on the best value of with k = 2m, it gives cc = 0.9814. It is clearly that the best value of k = 7 gives a better cc which is 0.9821. So, for LMAM, this approach can be applied for the next study.As for LLAM (Fig. 6), it can been seen that cc does not change at all as the k increase. However, the smallest value of k with the highest value of cc starts from 3. As mentioned before, [16] suggested that a small value gives an excellent prediction. That is why the best value of k is 3.By referring to Fig. 7 and Table III, ILLAM gives the best prediction performance with cc = 99.13. It shows that, ILLAM can make prediction with 99.13% chances of getting close to the real data compared to the model without using optimal value of k gives cc = 0.9910. By this improved method of determining an optimal value of k, ILLAM can have a better prediction model after determining the optimal value of k.

Table- II: Performance of prediction model

Correlation Coefficient (cc)				
Method	k = 2m	k = 50	k = 100	k = 200
LMAM	0.9814	0.9663	0.9496	0.9131
LLAM	0.9873	0.9873	0.9873	0.9873
ILLAM	0.9910	0.9905	0.9905	0.9890

Table- III: Prediction with optimal number of k.

Method	kopt	cc
LMAM	7	0.9821
LLAM	3	0.9873
ILLAM	13	0.9913

VI. CONCLUSION AND FUTURE RESEARCH

In this study, the chaotic behavior of hourly CO concentration time series is detected through phase plot and the Cao method. Three prediction models were used in determining the optimal number of k is ranged between $3 \leq k \leq 13$. Comparison between try and error with the improved method is done through graph plotting. In this study, the value of is set to $\tau = 1$. In the future, the method such as average mutual information and autocorrelation function is suggested to calculate the τ value. In addition, a better prediction model can be created by combining the method of determining an optimal number of time delay τ, embedding dimension m and number of nearest neighbor k.

ACKNOWLEDGMENT

Thanks to the Department of Environment, Malaysia for providing the necessary data. This research is under sponsorship of the research grant with code 2019-0005-102-02 (FRGS/1/STG06/ UPSI/02/1)

REFERENCES

1. O. S. Azeez, B. Pradhan, and H. Z. M. Shafni, “Vehicular CO Emission Prediction using Support Vector Regression Model and GIS,” *Sustainability*, vol. 10, no. 10, pp. 1-18, 2018.
2. H. P .I Abarbanel, *Analysis of Observed Chaotic Data*, Springer Verlag, New York, 1996.
3. C, Sprot, *Chaos and Time Series Analysis*, Oxford University Press, 2003.
4. L. Cao, “Practical Method for Determining the Minimum Embedding Dimension of a Scalar Time Series,” *Physica D: Nonlinear Phenomena*, vol. 110, no. 1-2, 1997, pp. 43-50.
5. N. Z. A. Hamid, and M. S. M. Noorani, “An Improved Prediction Model of Ozone Concentration Time Series Based on Chaotic Approach,” *International Journal of Mathematical and Computational Sciences*, vol. 7, no. 11, 2013, p. 6.
6. N. H. Adenan, and M. S. M. Noorani, “Peramalan Data Siri Masa Aliran Sungai di Dataran Banjir dengan Menggunakan Pendekatan Kalut,” *Sains Malaysiana*, vol. 44, no. 3, 2015, pp. 463-471.
7. N. Z. A. Hamid, and M. S. M. Noorani, “A Pilot Study Using Chaotic Approach to Determine Characteric and Forecasting of PM10 Concentration Time Series,” *Sains Malaysiana*, vol. 43, no. 3, 2014.
8. S. Velickov, *Nonlinear Dynamics and Chaos, Taylor and Francis Group plc*, London, 2004.
9. S. K. Regonda, B. Rajagopalan , U. Lali, M. Clark, and Y. I. Moon, “Local polynomial method for ensemble predict of time series., ” *Nonlinear Process in Geophysics*, vol. 12, 2005, pp 397-406.
10. B. Sivakumar, “A phase-space reconstruction approach to prediction of suspended sediment concentration in rivers,” *Journal of Hydrology*, vol. 258, pp. 149-162, 2002.
11. B. Sivakumar, “Forecasting monthly streamflow dynamics in the western United States: a nonlinear dynamical approach,” *Environmental Modelling & Software*, vol. 18, no. 8-9, pp. 721-728, 2003.
12. N. Z. A. Hamid, M. S. M. Noorani, L. Juneng, and M. T. Latiff, “Prediction of Ozone Concentrations Using Nonlinear Prediction Method,” *Proceedings of the 20th National Symposium on Mathematical Sciences*, 2013, pp. 125-131.
13. N. Z. A. Hamid, M. S. M. Noorani, and N. H. Adenan, “Chaotic Analysis and short-term prediction of ozone pollution in Malaysian urban area,” *Journal of Physics*, vol. 890, 2017.
14. J. Theiler, S. Eubank, L. Alamos, O. P. Trail, and S. Fe, “Don’t bleach the chaotic data,” *Chaos*, vol. 4, no. 1, 1993, pp.1-12.
15. W. N. A. W. M. Zaim, *Forecasting Seasonal Ozone Time Series in Malaysian Higher Education Areas Through Chaotic Approach*, Sultan Idris Education Univeristy, 2018.
16. M. Casdagli, “Chaos and Deterministics versus Stochastic Non-Linear Modelling,” *Santa Fe Institute*, vol. 54, no. 2, 1991, pp. 303-328.
17. J. D. Farmer, and J. J. Sidorovich, *Physical Review Letters* vol. 59, no. 8, pp 123-133, 2002.
18. D. Ruelle, and F. Takens, “On the nature of turbulence,” *Communications in Mathematical Physics*, vol. 20, no. 3, 1971, pp.167-192.
19. B. Sivakumar, “A Phase Space reconstruction approach to prediction of suspended sediment concentrations in river,” *Physica D*, vol. 110, 1997, pp. 43-50.

AUTHORS PROFILE

Ahmad Basri Ruslan I was born in Kuala Terengganu, Terengganu, Malaysia on December 19, 1995. I earned my Bachelor of Science (Mathematics) with Education from Sultan Idris Education University, Perak, Malaysia in 2018. Currently, I’m pursuing my studies in Masters of Science in Applied Mathematics. My field of expertise is dynamical system and chaos theory. Along with my studies in chaos theory and dynamical systems, I’m also working with my lecturer in Applied Mathematics in the field of numerical mathematics.

Nor Zila Abd Hamid was born in Kepala Batas, Penang, Malaysia on November 14, 1983. She earned a Bachelor of Education (Honours) in Mathematics from Universiti Putra Malaysia, Selangor, Malaysia in 2006. She continued her studies at the Masters level and earned a Master Degree in Applied Mathematics from Universiti Kebangsaan Malaysia, Selangor, Malaysia in 2008. Her expertise areas are dynamical systems and chaos theory. Currently, she is a doctor of philosophy student at the Universiti Kebangsaan Malaysia. She is now serving as a lecturer at the Universiti Pendidikan Sultan Idris, Perak, Malaysia.