Prediction Modelling Through Chaotic Approach on Ozone (O₃) Pollutant Time Series in Shah Alam City

Ahmad Basri Ruslan¹, Nor Zila Abd Hamid², Khairunnisa Che Jusoh³, Shahidatul Idayu Shahizam² and Franklin Chai Jun¹

¹Sekolah Menengah Sains Pendang @ Akademi Sains Pendang (ASP), Jalan Jenun, 06700, Pendang, Kedah, Malaysia
²Universiti Pendidikan Sultan Idris, Jabatan Matematik, 35900, Tanjung Malim, Perak, Malaysia

Studies on Ozone (O₃) particles and their effect have become a significant concern. Exposure to high concentrations of O₃ can cause an adverse reaction in the human respiratory system. Therefore, this study aims to predict the monthly O₃ time series in the highly populated area of Shah Alam through a chaotic approach. The phase space plot and Cao methods successfully detected the chaotic behaviour O₃ time series data. The three parameters are determined before the prediction process: time delay t, embedded dimension m, and nearest neighbour k. This study used t = 1 and the value of m is calculated through the Cao method. The last parameter that needs to be determined is by graph plotting k versus correlation coefficient (cc). The combination of parameters t and m will be used for prediction, and the performance measure will be recorded. The prediction process is done using three non-linear methods, namely LMAM, LLAM and ILLAM. LMAM gives the best prediction value by using the combination of parameter t and the value of the performance measure is 0.8486. Trial-and-error method, m = 6, LMAM gives cc = 0.8977 with k = 17, LLAM gives cc = 0.8653 with k = 1 and ILLAM gives cc = 0.8265 with k = 24. Mean-top-bottom method, m = 9, LMAM gives cc = 0.8863 with k = 10, LLAM gives cc = 0.8841 with k = 6 and ILLAM gives cc = 0.9370 with k = 22. This finding indicates that the O₃ time series can be predicted using a chaotic approach and the improved method in determining the value can compute better prediction performance.

Keywords: ozone; chaotic approach; non-linear; phase space plot

1. INTRODUCTION

Ozone (O₃) is naturally formed in between stratosphere and troposphere levels. National Aeronautics Space Administration (NASA) defines O₃ as a natural gas formed from three oxygen atoms (NASA, 2018). O₃ layer is used to protect the earth from ultraviolet rays. The troposphere level is the lowest. The gas composition in the troposphere is adequate for humans to breathe. However, breathing with O₃ is toxic to human health at the troposphere level. It may cause an adverse reaction in the respiratory system (Cisneros et al., 2010). United States Environmental Protection Agency (EPA) founds that high concentrations of O₃ at the troposphere level are dangerous for health and categorised as toxic (EPA, 2017). At the troposphere level, O₃ is formed through a chemical reaction of volatile organic compounds (VOC, i.e. NO₂, CO, etc.) and a free oxygen atom in the air with sunlight as the catalyst. VOC is released by the combustion of vehicle engines and factories from industries. When the emission of waste from vehicles and industries increases, then the formation of O₃ will also increase. The nature of O₃ is not soluble in water. It is freely moving at the troposphere level (Nuvolone et al., 2017). It makes O₃ is highly dangerous at ground level.

There are five types of pollutants observed by the Department of Environment Malaysia (DOE) which are particulate matter below 10 millimetres (PM₁₀), carbon monoxide (CO), nitrogen monoxide (NO), ozone (O₃) and sulfur dioxide (SO₂). At the end of 2019, 8.3 million deaths are reported due to air pollution (McCarthy, 2019). Factors that affected pollution are the economy and the number of

*Corresponding author’s e-mail: ahmadbasirruslan@gmail.com
populations (Myllyvirta, 2020). India and China are the top two countries that are highly populated in the world. These two countries also give the highest number of death due to pollution which is 1.243 million deaths (China) and 1.241 million deaths (India) at the end of 2019. It is proven that a highly populated country may lead to high pollution. Therefore, this study aims to predict the monthly O₃ time series in the highly populated area of Shah Alam through a chaotic approach.

Prediction modelling through a chaotic approach in Malaysia is new. Besides, there are several methods in prediction such as artificial neural network (ANN), multilinear regression, support vector machine, fuzzy logic and many more. These methods require many variables such as meteorology, gas and many more before predictions can be made (Hamid & Noorani, 2014). While through a chaotic approach, this method only requires previous time series data to make the prediction. For example, to predict the CO time series, only previous time series data of CO is required to make the prediction. Several predictions had been successfully done by using chaotic approach onto river flow time series (Adenan & Noorani, 2013), O₃ time series (Hamid & Noorani, 2013), PM₁₀ time series (Hamid & Noorani, 2014), temperature time series (Bahari & Hamid, 2019), sea level time series (Ali & Hamid, 2019), CO time series (Ruslan & Hamid, 2019) and rainfall time series (Mashuri et al., 2019). The Cao method was chosen because: i) does not involve another parameter except τ and ii) is not dependable on the number of time series data used (Cao, 1997). Therefore, the phase space plot and the Cao method are chosen to detect chaotic behaviour in this study.

Before starts with the prediction process, three parameters need to be determined, which are delay time (τ), embedded dimension (m) and the number of nearest neighbours (k). The choice of parameter τ is important to fully capture the structure of the attractor (Velickov, 2004). If τ is too big, different coordinates will not correlate and may cause the loss of information from the original system (Regonda et al., 2005). The previous study uses τ = 1, and the final result shows an excellent correlation coefficient (cc) between the original data and prediction onto hydrology and air pollution data (Sivakumar, 2002; Hamid, 2018). For m , it will be calculated through Cao method for getting the optimum value of m (Velickov, 2004). In the previous study, k is chosen with k = 2m, where m is the embedded dimension (Adenan & Noorani, 2015). Besides, several another few studies used the trial-and-error method in determining the value of parameter k, such as k = 50 (Hamid & Noorani, 2014; Jayawerdana, 1997; Zaim & Hamid, 2017; Hamid et al., 2017), k = 100 (Hamid et al., 2017) and k = 200 (Hamid & Noorani, 2013).

In this study, an optimal of k will be analysed through graph plotting cc against k testing all numbers k from 1 until 200. The value k that gives the maximum value cc will be chosen as the best model and will be recorded. Three methods for predictions will be used, which are Local Mean Approximation Method (LMAM), Local Linear Approximation Method (LLAM) and Improved Local Linear Approximation Method (ILLAM). These three methods are known as chaotic models because they involve reconstructing phase space at m - dimension.
As mentioned above, a different time series will give a different prediction result. This study is focusing on $O_3$ time series in an urban area in Shah Alam. While for the previous study, the focus of the study was in a different geographical area which is in Shah Alam (Hamid et al., 2013). Hence, it will give different prediction results. In addition, the previous study does not find an optimal parameter $k$ before continuing with the prediction. In this study, the optimal number will be used to make the prediction model predict a better value by using $cc$ as the performance measure.

A. Time Series

The monthly reading of $O_3$ real-time series data is prepared by the Department of Environment Malaysia between 1st of June until 30th of June 2014 at Shah Alam Station. The $O_3$ time series are recorded in the unit of part per billion (ppb). Shah Alam is a city and state capital of Selangor, Malaysia, situated within Petaling District and a small part of the neighbouring Klang District. Shah Alam city is filled with residential, commercial centres, an education hub and recreational attractions. Summary from the Department of Statistics Malaysia (2010) shows that Shah Alam is ranked tenth among 20 cities with a large population in Malaysia (DOSM, 2010). Population density affects the number of pollutants released (Idris & Mahmud, 2017). The time-series data are recorded hourly for 30 days. The total number of data used is 720 hours. The $O_3$ time series is recorded in the one-dimensional form vector $X$:

$$X = \{x_1, x_2, x_3, \ldots, x_N\}$$  \hspace{1cm} (1)

$N$ is the total number of data used in hours. In this study, $N = 720$. The overall time series for all 720 consecutive hours is illustrated in Figure 1. Statistical details for $O_3$ concentration time series is recorded in Table 1. Next, the data is divided into two different parts. The first part is the training time series for analysis purposes, and the other is the testing time series for measuring the prediction model’s performance. The training time series is for three weeks long, 552 hours, while the test is one week long, which are 168 hours. These two parts are written as follows:

$$X_{\text{train}} = \{x_1, x_2, x_3, \ldots, x_{552}\}$$  \hspace{1cm} (2)

$$X_{\text{test}} = \{x_{552}, x_{553}, x_{554}, \ldots, x_{720}\}$$  \hspace{1cm} (3)

II. CHAOTIC APPROACH

A. Phase Space Plot

The training time series is recorded accordingly, like Equation (2). With $x_i$ is the concentration of $O_3$ time series, the two-dimensional graph is plotted on an axis $\{x_i, x_{i+1}\}$. Before that, the value for $\tau$ the need to be determined. A few methods previously are used in determining the value of $\tau$ such as average mutual information (Fraser & Swiny, 1986), autocorrelation method (Schuster, 1988) and many more. However, few past studies used $\tau = 1$ real-time series such as flood time series (Lakshmi & Tiwari, 2009), temperature time series (Bahari & Hamid, 2019) and suspended sediment concentration (Sivakumar, 2002). By setting $\tau = 1$, the prediction performance shows an excellent result. Hence, $\tau = 1$ is set to be used in this study.

After set parameter $\tau = 1$, phase space plot $\{x_i, x_{i+1}\}$ is built. The existence of an accumulator or strange attractor on the graph shows the presence of chaotic behaviour on the observed time series (Sivakumar, 2002; Lakshmi & Tiwari, 2009). Figure 2 shows the strange attractor on the phase space plot for $O_3$ time series in Jun 2014 in Shah Alam. Therefore, $O_3$ time series is chaos.
Figure 2. Phase space plot

B. Cao Method

Training time series data are reconstructed into \( m \) - embedded dimension:

\[
Y^m_j = (x_j, x_{j+\tau}, x_{j+2\tau}, \ldots, x_{j+(m-1)\tau})
\]  

The value for \( \tau \) has already set to 1, while parameter \( m \) is set through the Cao method. According to (Regonda et al., 2005), \( m \) is the minimum embedded value needed for explaining the nature of the data. The nature of the data can be characterised as low dimensionally chaotic if the value \( m \) is less than 10 (Hamid & Noorani, 2017). Chaos at a low dimensional means that few variables can only explain the observed time series. Cao method is not limited to determining the value of \( m \). This method also can be used to detect the presence of chaotic behaviour (Cao, 1997). There are two parameters included in the Cao method, which are \( E1(m) \) and \( E2(m) \).

\( E1(m) \) is used to compute the value of \( m \). If the value of \( m \) starts to saturate as \( m \) an increase from \( m_0 \), then \( m_0+1 \) is the minimum embedded dimension \( m \). For a chaotic time series, the graph \( m \) against \( E1(m) \) will shows a saturation state. While for random time series, the graph \( m \) against \( E1(m) \) will not show any saturation state.

As explained earlier, \( E2(m) \) also introduced by (Cao, 1997). For a random time series, the value of \( E2(m) = 1 \) for all \( m \) value. Besides, there are few possibilities where \( E2(m) \neq 1 \). If there is at least one value \( E2(m) \neq 1 \), then the observed time series is chaos.

The overall analysis from the Cao method onto the observed time series are displayed in Figure 3. The saturation value is not specifically stated by (Cao, 1997) since saturation may vary with different time series. For \( O_3 \) time series, the value of \( m_0 \) observed is 2 where \( m \) starts to saturate from 0.95 to 1.00. Since there is saturation state in the observed graph, the observed time series is chaos. Hence, the value of \( m_{0+1} = m = 3 \). For \( E2(m) \) there are few values where \( E2(m) \neq 1 \) which are at \( m = 1, 2, 3, 4, 5, 8 \) and 10. According to Abarbanel (1996) [16], the observed time series is chaos.

It is confirmed that the observed time series is chaos through the phase space plot and Cao method. The presence of chaotic behaviour is recorded in Table 1.

Table 1. Chaos determination

<table>
<thead>
<tr>
<th>Method</th>
<th>Chaos</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase space plot</td>
<td>Yes</td>
</tr>
<tr>
<td>( E1(m) )</td>
<td>Yes</td>
</tr>
<tr>
<td>( E2(m) )</td>
<td>Yes</td>
</tr>
</tbody>
</table>

III. PREDICTION

Three prediction models will be used in this study onto \( O_3 \) time series in Shah Alam. The prediction can be made after the chaotic behaviour is detected in the observed time series. Two parameters \( \tau = 1 \) and \( m = 3 \) has been chosen respectively to be used in prediction. The last parameter that needs to be determined is \( k \). Previous studies on \( O_3 \) time series and another time series use trial-and-error methods to determine the value \( k \). The improvement in this study is by varying the number of \( k \) use \( 1 \leq k \leq 200 \). The combination
of parameters $\tau$, $m$ and $k$, will be used for prediction and the performance measure will be recorded.

### A. Local Mean Approximation Method (LMAM)

Prediction through chaotic approach by using LMAM is based on equation:

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

Prediction for value $Y_{j+1}^m$ is done based on the value of $k$ for $Y_j^m$. The value of $k$ nearest neighbor for $Y_j^m$ are chosen based on the minimum value of Euclidean distance $\|Y_j^m - Y_j^m\|$ with $j' < j$. Let us say $k = 1$ is being used, then the approximation for $Y_{j+1}^m$ is $Y_{j+1}^m$. Prediction is taken as the mean value of $Y_j^m$:

$$Y_{j+1}^m = \frac{1}{k} \sum_{q=1}^{k} Y_{j_q+1}^m$$  \hspace{1cm} (6)

Figure 4 shows the prediction by using $k$ that is ranged between $1 \leq k \leq 200$. The value of $k$ that gives the maximum $cc$ is recorded as well. The best $k$ value is 17 that gives the maximum value of $cc = 0.8486$. The combination of the parameter used are $\tau = 1$, $m = 3$ and $k = 17$ give the maximum prediction performance. Table 3 shows the prediction result by using $k$ the value through $k = 2m$ and trial-and-error method for LMAM.

<table>
<thead>
<tr>
<th>$k$</th>
<th>$cc$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.8379</td>
</tr>
<tr>
<td>100</td>
<td>0.8181</td>
</tr>
<tr>
<td>200</td>
<td>0.7604</td>
</tr>
<tr>
<td>$k = 2m$</td>
<td>0.8261</td>
</tr>
</tbody>
</table>

### B. Local Linear Approximation Method (LLAM)

For prediction by using LLAM, the linear equation that being use is:

$$x_{n+1} = Ax_n + B$$  \hspace{1cm} (7)

Equation (7) is used to be standardised with the training set at (7). Prediction for $x_{n+1}$ is calculated by referring to the value of $x_n$. Constant value $A$ and $B$ are calculated through the smallest square root method. Next, the phase space is constructed by using $\tau = 1$ and $m = 3$. The construction of phase space is:

$$Y_j^3 = (x_j, x_{j+1}, x_{j+2}, x_{j+3}, x_{j+4}x_{j+5})$$  \hspace{1cm} (8)

with $j = 1, 2, 3, 4, ..., N-5$. Since $N = 552$, the last phase space is:

$$Y_{547}^3 = (x_{547}, x_{548}, x_{549}, x_{550}, x_{551}, x_{552})$$  \hspace{1cm} (9)

Linear equation $x_{n+1} = Ax_n + B$ is calculated. The value for $A$ and $B$ are changed accordingly depends on the value of $k$ used. Let say $k$ is 20. It produces equation

$$x_{n+1} = 0.3421x_n + 3.3923$$  \hspace{1cm} (10)

Figure 5 shows that at $k = 1$ gives the maximum $cc = 0.8453$. Even though there are several values of $k$ that gives the same maximum value, the smallest value of $k$ is chosen. A small value of $k$ is enough to make an excellent prediction (Casdagli, 1992). Table 4 shows the prediction result by using $k$ the value through $k = 2m$ and trial-and-error method for LLAM.
C. Improved Local Linear Approximation Method (ILLAM)

This method is first used in predicting O₃ concentration time series, and the results are excellent compared to LLAM (Hamid & Noorani, 2013). Previously, this method used trial-and-error methods to determine the value for \( k \), \( k = 200 \) (Hamid & Noorani, 2013). In this study, the number of \( k \) will be tested from 1 up to 200. ILLAM will use Equation (7). For LLAM, constant values \( A \) and \( B \) are calculated by using a training set. The improvement is made on the training set. The training set is updated for every prediction made, and a new equation is formed as follows:

\[
Y_{jk}^m = C_n Y_{jk}^m + D_n
\]  

Prediction through LLAM only produced one equation for the whole 168 predictions. While for ILLAM, for every 168 predictions, 168 equations will be made. Figure 6 shows the best value of \( k \) where \( k = 146 \) that gives maximum \( pk = 0.8606 \). The value of \( k \) used from 1 to 75 cannot be calculated. The calculation process cannot be solved by using a simultaneous equation and finally produces NaN (Not a Number). Let’s say \( k = 2 \), \( x_{jk} = (5,9) \). By referring to Equation (7), equations as below are formed:

\[
\begin{bmatrix}
5 \\
9
\end{bmatrix} = A + B \begin{bmatrix}
5 \\
5
\end{bmatrix}
\]

Simultaneously, the equation cannot be solved. That is why the calculation cannot be solved and produces NaN for the first 75 \( k \) used.

IV. RESULTS

The time series observed is chaotic. Prediction is made after determining three parameters. The prediction performance gives an excellent result. This study is focused onto the selection of \( k \) value. The improved method in determining the value \( k \) is successfully done. The method of trial-and-error in determining the value of \( k \) are less correlated. Figure 4 until Figure 6 shows the maximum \( cc \) generated by using the best value of \( k \). Table 3 until Table 4 shows the prediction using the trial-and-error method by LMAM, LLAM and ILLAM, respectively. These tables prove that the improved method in determining the value \( k \) gives a better prediction result. Figure 7 until Figure 9 shows the
comparison graphs of the real and predicted data by using LMAM, LLAM and ILLAM.

V. CONCLUSIONS

Phase space plot and Cao method had been used to detect the presence of chaotic behaviour. This study shows that the used time series is chaotic. All the past studies determine the value $k$ by using the trial-and-error method. For this study, the value $k$ varies from 1 to 200. Prediction by using the value of $k$ by using trial-and-error method also had been recorded for comparison purposes. It is proven that the improved method in determining the value $k$ is better than the trial-and-error method. Table 5 shows that for all three methods, a value of $k$ that gives maximum $cc$ for each method. Therefore, it is suggestable to use an improved method by varying the number of $k$ to compute a better prediction performance. This study may help the authority such as Department of Environment Malaysia improve and control Malaysia’s air quality.

<table>
<thead>
<tr>
<th>Method</th>
<th>$k$</th>
<th>$cc$</th>
</tr>
</thead>
<tbody>
<tr>
<td>LMAM</td>
<td>17</td>
<td>0.8486</td>
</tr>
<tr>
<td>LLAM</td>
<td>1</td>
<td>0.8453</td>
</tr>
<tr>
<td>ILLAM</td>
<td>146</td>
<td>0.8606</td>
</tr>
</tbody>
</table>

VI. ACKNOWLEDGMENT

I want to express my utmost gratitude to the Department of Environment Malaysia for contributing the necessary data. The Ministry of Education Malaysia supports this research under the Fundamental Research Grant Scheme (FRGS) with the research code 2019-0005-102-02(FRGS/1/STGo6/UPSI /02/1). Not to forget, I would like to thank the Ministry of Education for Fee Waiver Programme for encouraging participation among teachers for ICSTEM 2021.

VII. REFERENCES


Environmental Protection Agency 2017, Ground-level Ozone Pollution, United States of America.


Hamid, NZA & Noorani, MSM 2014, Suatu kajian perintis menggunakan pendekatan kalut bagi pengesanan sifat dan peramalan siri masa kepekanatan PM10, Sains Malaysiana.


