Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy

Pol. J. Environ. Stud. Vol. 30, No. 6 (2021), 1-10
DOI: *10.15244/pjoes/134961* ONLINE PUBLICATION DATE:

Original Research

Performance of Multivariate Time Series on Forecasting the Tropospheric Ozone (O_3)

Ahmad Fauzi Raffee[1], Hazrul Abdul Hamid[2]*, Siti Nazahiyah Rahmat[1], Muhammad Ismail Jaffar[1]

[1]Cluster of Water and Environmental Engineering, Faculty of Civil Engineering and Built Environment, Universiti Tun Hussein Onn Malaysia, 86400 Batu Pahat, Johor
[2]Mathematics Division, School of Distance Education, Universiti Sains Malaysia, 11800 Penang

Received: 14 January 2021
Accepted: 25 March 2021

Abstract

The tropospheric ozone (O_3) is known as a hazardous ambient air pollutant that adversely affects human health. Recently, many studies have been devoted to assess and monitor the O_3 concentration due to its impact on society health. This study was carried out to develop a model to forecast O_3 concentration. A comparison between univariate and multivariate time series to examine the most appropriate model was made. The air quality data used in this research was collected from three (3) stations namely Perai, Penang (industrial), Alor Setar, Kedah (urban) and Jerantut, Pahang (background). The selection of background station allows for comparisons to be made with stations closer to anthropogenic emissions. Based on Akaike Information Criterion (AIC), the appropriate multivariate time series to develop a forecasting model for Perai, Alor Setar and Jerantut monitoring stations were vector autoregressive - VAR(3), vector autoregressive - VAR(2) and vector moving average - VMA(2), respectively. The lowest root mean square error (RMSE) is 0.0053 which is for the multivariate time series model in Perai while for normalized absolute error (NAE) and mean absolute error (MAE), the lowest is in Jerantut with 0.0850 and 0.0013 respectively. Validation of the models using three error measures shows that the multivariate time series model performed better compared to the univariate time series model.

Keywords: time series, tropospheric ozone, air pollution

Introduction

The adverse effect of outdoor air pollutant on human health and well-being are consistently reported by many forms of research over the world [1-2]. Particulate matter (PM), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), carbon monoxide (CO) and tropospheric ozone (O_3) are among dominant outdoor air pollutants. Of all pollutants, the O_3 is classified as a secondary pollutant. The O_3 is noxious gaseous form by a series of complex reactions, non-linear, feedback-regulated processes between its precursors, such as nitrogen oxides (NO_X), the volatile organic compound (VOC) at the presence of sunlight [3]. The O_3 pollutant usually recorded higher

*e-mail: hazrul@usm.my

Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy

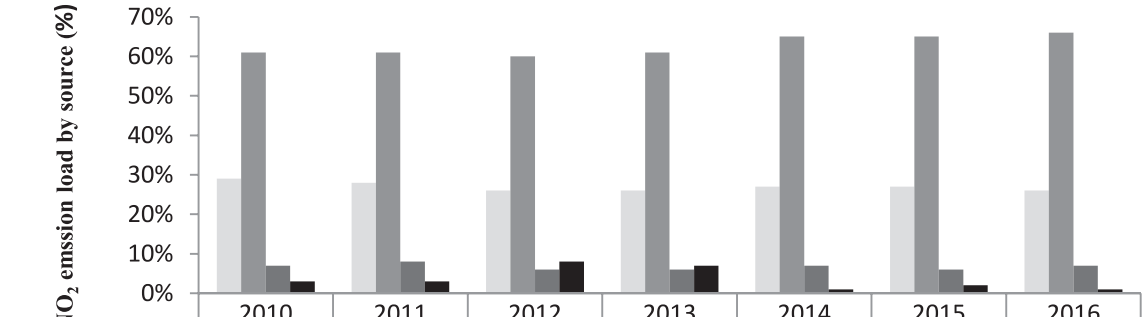
concentrations in the afternoon due to the intensity of ultraviolet (UV) radiation that comes from the sun. The anthropogenic activities are the most common sources of NO_X and VOC. Industrial activities, fossil fuel combustion and biomass are the major sources of NO_2 [4-5]. While, VOC produced from vegetation, motor vehicles exhaust, agricultural and forest fires [6]. The adverse effects of O_3 are well established in recent years [7-8].

The negative effects of O_3 are associated with an effect on human health, climate change, crop yield and ecosystem. The lower atmosphere condition initiated O_3 played critical role in tropospheric chemistry, whereas it being principal precursors to hydroxyl radical (OH) which controlling oxidizing power [5]. More importantly, O_3 is the main factor of photochemical smog and Global warming [9]. The formation, variation and behavior of O_3 concentration strongly influence by several factors such as wind speed, relative humidity, ambient temperature and solar radiation [10]. These factors also played a significant role in dilution and dispersion of O_3 concentration in ambient air. The numerous studies examined the high temperature, low wind speed, minimal rainfall and intensity of solar radiation could increase the O_3 concentrations [11]. In the same way, O_3 can be slightly reduced when degreasing heat of solar radiation, decreased water vapor has reduced the sources of radicals, increasing on cloud cover which is expected to result in a sensitive VOC chemistry [12].

At present, Malaysia initiates to be an industrial country with rapid industrialization, urbanization and economic growth. Due to accommodate human and industrial needs, the precursor of NO_2 released by the power plant for energy supply, processing factory activities and motor vehicle producing O_3 in the country. Currently, the major contributors of NO_2 pollutants are power plants and motor vehicle. These contributors showed slightly increased and degraded each year for power plant and motor vehicle, respectively. Power plants contributed 61% in 2010 and increased to 66% in 2016. However, the NO_2 contribution emission by motor vehicle decreased from 29% (2010) to 26% (2016) as shown in Fig. 1. The overall trend of O_3 in Malaysia reported by the Department of Environment (DoE) had exceeded the Malaysian Ambient Air Quality Guidelines for O_3 at 0.10 ppm in urban areas [13]. This was due to the traffic density and conductive atmosphere which results in the O_3 formation [14].

Nowadays, controlling the source of air pollutants is one of the major challenges in the world. The predicting models for air pollutant concentrations become imperative tools and produce an efficient management and control system in air quality. If any lack of conformity is examined, the related authority can use the data to advise or caution people about the effects [21]. The predicting models for forecasting air pollutants are divided into mathematical and physical models. In the last decade, several researchers widely applied the univariate time series models of an autoregressive moving average (AR) [22], moving average (ARMA) [23] and vector autoregressive moving average (ARIMA) [24] in predicting the future concentration of air pollutants dispersion. However, the application of the multivariate time series method is still very limited.

Besides, the research on air quality forecasting on time series especially relating on O_3 concentration have been conducted previously but focused solely on univariate such as the study by Jamil et al. [25] where applied the univariate time series method of the autoregressive integrated moving average (ARIMA). Additionally, there has been a limited study on multivariate time series such as done in Taiwan and Spain [26-27]. However, all this study was solely on PM_{10} rather than O_3 concentration. Thus, this study was carried out as a comparative study to prove that the



	2010	2011	2012	2013	2014	2015	2016
Motor Vehicle	29%	28%	26%	26%	27%	27%	26%
Power Station	61%	61%	60%	61%	65%	65%	66%
Industrial	7%	8%	6%	6%	7%	6%	7%
Others	3%	3%	8%	7%	1%	2%	1%

Fig. 1. Nitrogen dioxide (NO_2) load emission sources (%) from 2010-2016 (Sources: [14-20]).

Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy

multivariate time series outperformed the univariate time series for forecasting O_3 concentration purposed. The main contribution of this study is to improve the accuracy of predictions as well as to understand the causal relationship between O_3 and other pollutants or meteorological parameters. More accurate forecasting is needed to enable relevant parties to plan strategies in air quality control and also as an early warning for people who may be affected if the air quality level is poor.

This paper is organized as follows. Section 2 describes in detail the location of monitoring stations, air pollutants data including meteorological data and methods of study. Section 3 presents the application of univariate and multivariate time series methods. This section also describes the development of univariate and multivariate time series methods. Finally, Section 4 summarizes the main conclusions drawn.

Materials, Data and Methods

Site Description

Air quality in the country is monitored continuously and manually to detect any changes in the ambient air quality status that may cause harm to human health and the environment. Perai in Penang, Alor Setar in Kedah and Jerantut in Pahang which have been established as industrial, urban and background air monitoring stations, respectively, by the Department of the Environment (DoE), Malaysia were selected in this study. The area surrounding Perai station was developed from a mangrove swamp into the industrial area in northern Peninsular Malaysia. Alor Setar is the state capital of Kedah and the monitoring station is surrounded by large residential areas and business centres. While Jerantut station is surrounded by agricultural areas and traditional Malaysian villages with few local small industries. People who live in urban industrial and urban areas are most affected by air pollution especially children [28]. The O_3 pollutant gave serious attention in the industrial and highly-populated continental region due to the potential human health impact [29]. Thus, people living in these areas may have the possibility be exposed more to air pollutants by anthropogenic sources. The details and description of the stations are illustrated in Table 1 and Fig. 2.

Data

The Department of Environment, Malaysia (DOE) is the responsible agency that monitors the country's air quality. As a part of Malaysian Continuous Air Quality Monitoring (CAQM) program, the O_3 concentration was recorded using Teledyne O_3 Analyser Model 400A UV Absorption. There were two sets of O_3 concentration data utilised in this study. Firstly, the hourly average data set from January 2006 to December 2017 were used for descriptive statistics to examine the behavior pattern of O_3 concentration. Secondly, the hourly data

Fig. 2. Location of the selected monitoring stations.

Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy

Table 1. Descriptions of monitoring stations.

Monitoring station	Coordinates	Category
Sek. Keb. Cederawasih, Taman Inderawasih, Perai	N05º 23.890'-E100º 24.194'	Industrial
Sek. Men. Agama Mergong, Alor Setar	N06º 08.218'-E100º 20.880'	Urban
Pejabat Kaji Cuaca, Batu Embun, Jerantut	N03º 58.238'-E102º 20.863'	Background

was transformed to monthly average data to develop a statistical model for forecasting O_3 using the time series method. A total of 144 monthly data were used with 138 data utilised for forecasting and the remaining for validation purposes. In the multivariate time series method, the other data set that examined to develop the model is particulate matter (PM_{10}), gaseous pollutants (i.e. sulphur dioxide (SO_2), nitrogen dioxide (NO_2) and carbon monoxide (CO)) and meteorological parameters (i.e. relative humidity, temperature and wind speed) also obtained from the DOE. The reliability and quality of all recorded data are guaranteed since it has undergone the quality control established by the standard provided by the DOE.

Method

The procedures of the time series are quite similar. The differences between univariate and multivariate time series procedures are at the estimation part and Granger causality test. The estimation for univariate considered only the O_3 concentration as a single variable, while the multivariate time series consists of multiple single series referred to as component and involving stochastic models to describe and analyze the relationships among data sets. There are three phases in time series modelling which are identification, estimation and testing [30].

The Augmented Dicky-Fuller (ADF) test was used to determine the stationarity of the data series taken into consideration in this study. The ADF expression as in Equation (1) with the hypothesis is H_0: the time series data is non-stationary and H_1: time series data is stationary. If the critical value less than ADF value, the null hypothesis will be rejected [31].

$$\text{ADF} = \alpha_0 + p_1 y_{t-1} + \sum_{j=2}^{p-1} \beta_j \nabla y_{t-j} + e_t \quad (1)$$

...where:
α_0 - Drift Component
e_t - independent and a homogeneous error term

Both time series models of univariate and multivariate have been identified for each type of model namely autoregressive (AR), moving average (MA) and autoregressive moving average (ARMA) for univariate models and vector autoregressive (VAR), vector moving average (VMA) and vector autoregressive moving average (VARMA) for multivariate models. The Akaike Information Criterion (AIC) as for lag length selection was used to identify the appropriate model part by part. The most suitable model was the model that consisted of the smallest AIC value [32]. Then, the most appropriate model which represented the univariate and multivariate models for each monitoring station was provided for comparison purposes. The AIC equation as in Equation (2) [31]

$$AIC = 2k - 2\ln(L) \quad (2)$$

...where:
k - number of estimated parameters in the model.
L - Maximum values of the likelihood function for the model.

The purpose of the estimation procedure was to create the forecasting values of O_3 concentration. Each model involved in this process and produced its equation. The first model of univariate was AR and the expression is shown in Equation (3) where the process of order p is denoted by AR(p) [33].

$$y_t = \sum_{r=1}^{p} \phi_r y_{t-1} + \in_t \quad (3)$$

...where ϕ_1, ..., ϕ_r defined as constant and $\in_t$ as a sequence of independent or uncorrelated random variables with mean 0 and variances σ^2.

The second model of univariate was MA process of order q which denoted as MA(q) and the expression of the model is shown in Equation (4) [33]:

$$y_t = \sum_{s=0}^{q} \theta s + \in_{t-s} \quad (4)$$

...where θ_1, ..., θ_q defined as constant, $\theta_0 = 1$ and $\in_t$ a sequence of independent (or uncorrelated random variables with mean 0 and variances σ^2.

The third univariate model was the combination of AR(p) and MA(q) model which known as autoregressive moving average (ARMA) were denoted as ARMA(p,q) and the equation is shown in Equation (5) [33]:

$$y_t = \sum_{r=1}^{p} \phi_r y_{t-1} = \sum_{s=0}^{q} \theta s + \in_{t-s} \quad (5)$$

Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy

...where the $\in_t$ is white noise. The appropriate θ, ϕ is the process of stationary of the data series.

Besides the three univariate models explained above, another three multivariate time series models were also applied to compare the results. The first model of multivariate namely as VAR. The VAR model describes the situation in which the present value of a series depends on its previous values. This model is an extension of the univariate autoregressive model (AR). The VAR model of order *p*, abbreviated VAR(p) given as in Equation (6) [34]:

$$(I - \Phi_1 B - \cdots - \Phi_p B^p) y_t = a_t \tag{6}$$

...where z_t is an (*m* x 1) vector observed variables, z_t denoted multivariate white noise and Φ is a matrix polynomial of order *p* in the backward shift operator *B*.

The VMA was the second model of the multivariate time series applied in the study. The model was an extension from the MA for univariate forecasting. The phenomena in which events produce an immediate effect that only lasts for short periods is referred to in this model. The abbreviated VMA(*q*) is shown in Equation (7) [34]:

$$y_t = (I - \Theta_1 B - \cdots - \Theta_p B^p) a_t \tag{7}$$

...where z_t is an (*m* x 1) vector observed variables, z_t denoted multivariate white noise and Θ is a matrix polynomial of order *q* in the backward shift operator *B*. The third model was defined as VARMA. Similar to univariate, this model was the combination of VAR and VMA models. In general, the VARMA (*p,q*) process given by Equation (8) [34]:

$$\Phi_p(B) y_t = \Theta_q(B) a_t \tag{8}$$

the autoregressive and moving average matrix polynomial of orders *p* and *q* respectively, where Φ_p and Φ_q is nonsingular *m* x *m* matrices. The process is stationary if the zeros of the determinantal polynomial $|\Phi_p(B)|$ are outside the unit circle.

Besides, for the multivariate model, the Granger causality test was used to determine the influence of other variables of pollutants or meteorological parameters and the equation is shown in Equations (9) and (10) [35]:

$$y_t = g_0 + a_1 y_{t-1} + \ldots + a_p y_{t-1} + b_1 x_{t-1} + \ldots + b_p x_{t-p} + u_t \tag{9}$$

$$x_t = H_o + c_1 x_{t-1} + \ldots + c_p x_{t-p} + d_1 y_{t-1} + \ldots + d_p y_{t-p} + v_t \tag{10}$$

Then, testing H_0: $b_1 = b_2 = \ldots = b_p = 0$, against H_A: *x* Granger causes *y*. Similarly, testing H_0: $d_1 = d_2 = \ldots = d_p = 0$, against H_A: y_t Granger causes x_t. The H_0: b_1 represents the dependent series while the H_o:d_1 represents the independent series. While *a* is the coefficient values of the series. In each case, a rejection of the null implies there is Granger causality. Note that x_t and y_t series are in 'level' form which simply means that the data is not in the 'difference' form where u_t and v_t are white noise error terms.

The final stage in this works was the testing part. To measure the discrepancies between the forecast and actual values, performance indicator (also known as a goodness of fit criteria) regression models namely root mean absolute error (RMSE), normalized absolute error (NAE) and mean absolute error (MAE) were used. RMSE summarizes the difference between the observed and imputed concentrations and is used to provide the average error [36]. NAE is more sensitive in measuring residual error [37] and MAE is the absolute difference between prediction and actual observation on average over the test sample where all individual differences have equal weight [38]. The equation for RMSE, NAE and MAE are shown in Equations (11), (12) and (13) respectively.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - \mathrm{O}_i)^2} \tag{11}$$

$$NAE = \frac{\sum_{i=1}^{N}|P_i - O_i|}{O_i} \tag{12}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|P_i - O_i| \tag{13}$$

...where, *P* is forecast values, *O* is observed values. While, *N* referred to the number of samples. If the values of performance error near zero, it means the forecast models are a better approach compared to another.

Results and Discussion

The pattern and behavior of O_3 concentrations were analyzed using descriptive statistics. The results of the descriptive statistics are presented in Table 2. The unit of measurement is part per million (ppm). The maximum of O_3 concentration exceeded the Malaysia Ambient Air Quality Guideline (MAAQS) at 0.1 ppm for Perai and Alor Setar monitoring stations. The highest concentration recorded was 0.1763 ppm at Perai station followed by 0.1032 ppm at Alor Setar station. The Perai station recorded the highest mean with 0.0190 ppm and Jerantut recorded the lowest mean value with 0.0152 ppm. As expected due to high values of mean, the standard deviation also showed high value in industrial monitoring station (Perai) compared to others.

Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy

Table 2. Descriptive statistics for O_3 concentration by monitoring stations.

Station	Perai	Alor Setar	Jerantut
Mean	0.0190	0.0180	0.0152
Median	0.0136	0.0152	0.0130
Std. Deviation	0.0177	0.0117	0.0104
Maximum	0.1763	0.1032	0.0546

The multivariate times series is distinguished by allowing more than one variable for developing a time series model. This study used only O_3 concentration for developing a univariate time series model. Meanwhile, the multivariate model was analyzed using the O_3 concentration as the dependent variable and particulate matter (PM_{10}), gaseous pollutants (SO_2, NO_2, CO) and meteorological variables (relative humidity, wind speed and temperature) categorized as independent variables. The developed multivariate model, considered the significant variable while the insignificant variable was removed.

The stationarity test was conducted on all variables and the results are summarised in Table 3. This result showed that Perai monitoring station had six stationary variables (i.e. O_3, PM_{10}, SO_2, CO, wind speed, temperature and relative humidity) with significant values less than 0.05. In contrast, Nilai and Jerantut monitoring stations had two and three insignificant variables with significant values of more than 0.05 (i.e. PM_{10} and relative humidity for Nilai monitoring station, while NO_2, temperature and relative humidity for Jerantut monitoring station).

The multivariate times series is distinguished by allowing more than one variable for developing a time series model. This study used only O_3 concentration for developing a univariate time series model. Meanwhile the multivariate model was analyzed using the O_3 concentration as dependent variable and particulate matter (PM_{10}), gaseous pollutants (SO_2, NO_2, CO) and meteorological variables (relative humidity, wind speed and temperature) categorized as independent variables. The developed multivariate model, considered the significant variable while the insignificant variable was removed.

The stationarity test was conducted on all variables and the results are summarised in Table 3. This result showed that Perai monitoring station had six stationary variables (i.e. O_3, PM_{10}, SO_2, CO, wind speed, temperature and relative humidity) with significant values less than 0.05. In contrast, Nilai and Jerantut monitoring stations had two and three insignificant variables with significant values of more than 0.05 (i.e. PM_{10} and relative humidity for Nilai monitoring station, while NO_2, temperature and relative humidity for Jerantut monitoring station).

The results of this stationarity test observed that O_3 concentration was stationary at all monitoring stations and the estimation process for univariate time series can proceed to determine the best model. The multivariate procedure was continued with Granger causality test to examine the significant variables that influenced the concentration of O_3.

In order to examine the independent variables that possible to influence the O_3 concentration, Granger causality was applied to all monitoring stations (Table 4). From the results obtained, it can be concluded that the gaseous pollutants of SO_2 and NO_2 were the influenced variables to O_3 concentration at Perai and Jerantut monitoring stations, respectively. Meanwhile, the meteorological variables namely wind speed (WS) was found to be an influenced factor to O_3 concentration at Alor Setar monitoring station. It also can be concluded that only one variable for each monitoring station was taken into consideration for developing a multivariate model. The SO_2 and NO_2 were expected as variables that influenced the O_3 concentration. This finding was similar to a previous study by Wang et al. [39]. However, the result of Granger causality test for WS at Nilai monitoring station was unexpected. There were limited finding on the significant effect of WS to O_3 concentration. The slow wind speed might be the factor influencing O_3 concentration. Additionally, the slow wind speed allowed more solar radiation hence gave this relation.

Perai monitoring station is surrounded by industrial activities where Perai is known as the main industrial

Table 3. The ADF test statistics for all variables.

Station	ADF value	O_3	PM_{10}	SO_2	NO_2	CO	Wind speed	Tempe-rature	Relative humidity
Perai	ADF	-1.3288	-1.7461	-0.9017	-0.5131	-0.4012	-1.3411	-0.945	-0.007
	Sig	<0.0215	<0.0255	<0.0001	0.7871	0.03364	<0.0014	<0.0030	<0.0001
Alor Setar	ADF	-1.4782	-1.8567	-2.5856	-1.0113	-1.628	-1.7357	-0.7734	-0.543
	Sig	<0.0001	0.1582	0.0013	0.0197	<0.0001	0.0004	<0.0140	<0.8272
Jerantut	ADF	-1.0548	-1.5848	-1.8744	-0.7912	-1.3534	-0.7404	-0.3998	-0.5062
	Sig	<0.0417	0.0239	<0.0001	0.2596	0.0079	0.0228	0.2816	0.8492

*Results in boldface indicate significant values more than 0.05 monitoring station

Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy

Table 4. Granger causality results.

Station	Parameter	t-statistics	Significant
Perai	NO_2	2.665	0.0087
Alor Setar	WS	2.263	0.0254
Jerantut	SO_2	2.515	0.0132

Table 5. Lag selection criteria of univariate and multivariate.

Method	Station	Model	AIC
Univariate	Perai	AR(1)	-7.7975
	Alor Setar	ARMA(1,1)	-6.5656
	Jerantut	MA(2)	-8.7955
Multivariate	Perai	VAR(3)	-23.6990
	Alor Setar	VAR(2)	-8.3943
	Jerantut	VMA(2)	-22.8724

area in Penang. This might be the main cause of gaseous of NO_2 influence the O_3 concentration where it's releasing from industrial processing activities in this area. Meanwhile, at the Jerantut monitoring station, the gaseous of SO_2 influencing the O_3 concentration due to the release from small industrial facilities and combustion of fuel from mobile sources since this area is situated in a rural area. Unexpected relation found in Alor Setar monitoring station reveals that WS as the significant variable influence the O_3 concentration in this area. The monitoring location is located approximately less than 20 km from the Straits of Malacca. This location might be caused by favorable wind speed in this area which influences the O_3 concentration whereas the heat and the sea salt are transported to ground that caused of this relationship.

The procedure was then continued to the identification part. This procedure determines the most appropriate model to develop for both time series methods. The appropriate univariate and multivariate time series based on Akaike Information Criterion (AIC) as well the performance errors are presented in Table 5. The previous study has been suggested that the value of $p + q$ must be equal to or less than 3 [40]. The values of p and q represent the previous value that is used for forecasting purposes. The high lag AIC numbers lead to obtaining the high forecast error value. Based on the results in Table 5, the univariate time series gave the AR(1), ARMA(1,1) and MA(2) as the most appropriate models for Perai, Alor Setar and Jerantut monitoring stations, respectively. Meanwhile, for multivariate time series, a similar model of VAR but different lag numbers were found in Perai and Alor Setar, and VMA for Jerantut monitoring station. The multivariate appropriate models were VAR(3) for Perai and VAR(2) for Alor Setar, while VMA(2) for Jerantut monitoring stations.

The root mean square error (RMSE), normalized absolute error (NAE) and mean absolute error (MAE) were used for verification purposes and the results are also shown in Table 6. The validation used the data from July 2017 to December 2017 by using the monthly simulation data set from January 2006 to June 2017. The results of the performance error also showed that the method of multivariate time series was better compared to the univariate time series for all three monitoring stations. At Perai and Alor Setar stations, all three performance errors for multivariate time series gave good results. The values of RMSE, NAE and MAE were 0.0053, 0.1003 and 0.0022, respectively, for Perai station, and 0.0076, 0.1758 and 0.0031, respectively, for Alor Setar station. Although Jerantut station showed only two multivariate performance errors better than univariate, it was sufficient to conclude that the multivariate time series method was more appropriate to be applied.

These results also indicated that the multivariate time series were applied at all monitoring stations in this study with a model of VAR(3) for Perai, VAR(2) for Alor Setar and VMA(2) for Jerantut monitoring stations. For Alor Setar and Jerantut stations, lag number two represented the two months previous data while lag number three for Perai station represented the three months previous data were taken into consideration to obtain the forecast values. Additionally, even though they gave a similar model of VAR, it should be noted that the influence of variables based on the Granger causality test was different from each monitoring station. This was due to the location and surrounding areas of air monitoring stations.

Finally, the equation for the multivariate time series model was obtained for each monitoring station based on the appropriate model selection. The multivariate equations were developed to forecast O_3 concentration one month ahead of. and given as follows:

Table 6. Performance error results for all monitoring stations.

Station/Error	RMSE		NAE		MAE	
	Univariate	Multivariate	Univariate	Multivariate	Univariate	Multivariate
Perai	0.0055	0.0053	0.1048	0.1003	0.0023	0.0022
Alor Setar	0.0138	0.0076	0.3169	0.1758	0.0056	0.0031
Jerantut	0.0049	0.0366	0.1271	0.0850	0.0020	0.0013

Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy

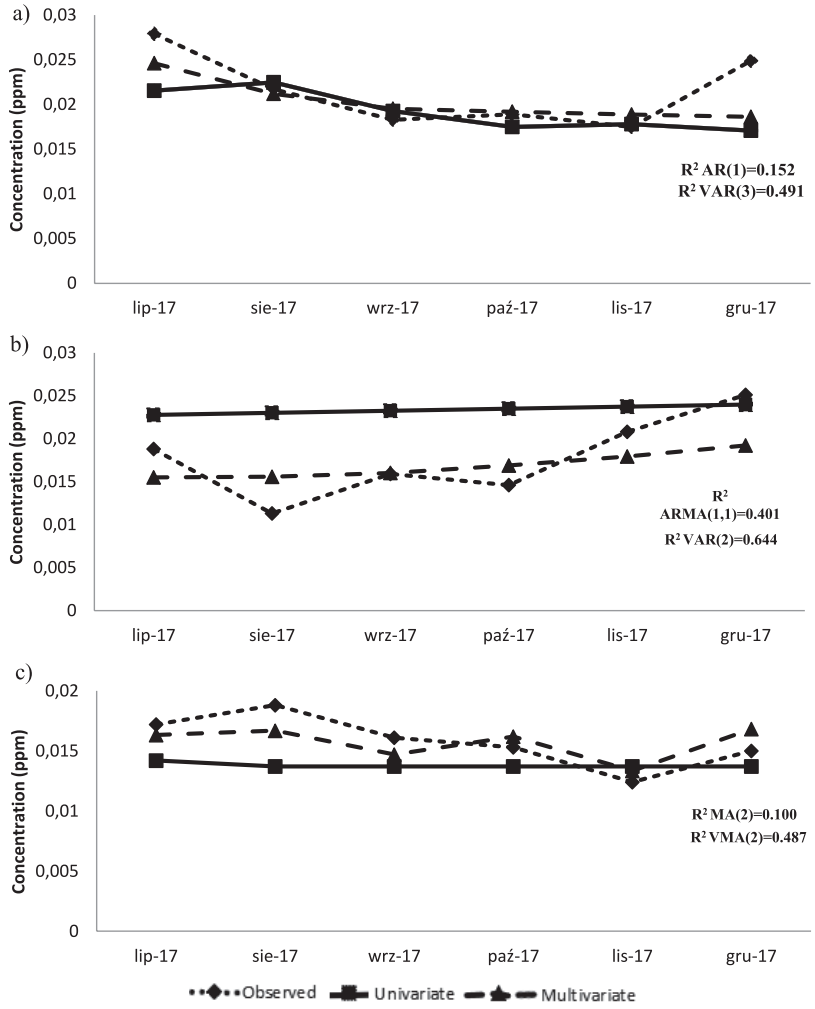


Fig. 3. The actual versus forecast value for a) Perai, b) Alor Setar and c) Jerantut monitoring stations.

For Perai monitoring station:

$$O_3 = 0.007628452 + (0.4691*O_{3,t-1}) + (0.798*NO_{2,t-1}) + (-0.065465* O_{3,t-2}) + (-0.108* NO_{2,t-2}) + (0.0171* O_{3,t-3}) + (-0.46* NO_{2,t-3}) \quad (14)$$

For Alor Setar monitoring station:

$$O_3 = -0.00195626 + (0.533*O_{3,t-1}) + (0.00118*WS_{,t-1}) + (0.25* O_{3,t-2}) + (-0.000295* WS_{,t-2}) \quad (15)$$

For Jerantut monitoring station:

$$O_3 = 0.01470535 - (-0.146* O_{3,\,t-1}) - (0.1* SO_{2,\,t-1}) - (-0.0788* O_{3,\,t-2}) - (0.7* SO_{2,\,t-2}) \quad (16)$$

The observed and predicted values for observation, best univariate and multivariate time series method are shown in Fig. 3. The prediction values for one month in advance can be obtained using the equation provided earlier. Based on the presented graph, the highest difference occurred in December 2017 for Perai and Alor Setar monitoring stations. The huge difference between observed and actual was due to the high recorded value of O_3 concentration at both locations compared to the Jerantut monitoring station. Additionally, the results of the descriptive statistics above proved that the mean and standard deviation at these two stations were high compared to Jerantut station.

The developed multivariate time series models are possible and sufficient to apply for forecast the future concentration of O_3. The forecasting model can be very useful to any related government agency, non-government organization or individual for an early measure to reduce the possible exceed limit of O_3 concentration in the future as well as to plan outdoor activities.

Conclusions

This study selected three different types of air quality monitoring stations to examine the appropriate time series model that can be applied to forecast the O_3 concentration. The Akaike Information Criterion (AIC) was used to examine the appropriate model for univariate is AR(1) with values of AIC is -7.7975,

Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy

ARMA(1,1) with values of AIC is -6.5656 and MA(2) with the value of AIC is -8.7955. Meanwhile, for the multivariate method, the AIC found VAR(3) for Perai, VAR(2) for Alor Setar and VMA(2) for Jerantut as the appropriate models with values of -23.6990, -8.3943 and -22.8724, respectively.

The results of validation using performance errors namely RMSE, MAE and NAE showed that the multivariate overcome the univariate time series. This indicated that all three monitoring stations, the multivariate time series model were better compared to the univariate model. Since the multivariate time series outperformed univariate time series, the different variables that influenced O_3 concentration to forecast future O_3 values were found at each monitoring station. At Perai station, the significant variable was NO_2 while at Jerantut station the significant variable was SO_2. Both variables were expected since these variables known as a precursor of O_3 concentration. Contrary to Alor Setar, where the unexpected significant variables were observed, namely WS. Three equations were successfully developed to forecast O_3 concentration at three monitoring stations. These equations can be used to forecast O_3 concentration based on other parameters that contribute to the reading of O_3 concentration. These models are also useful for better short-term forecasting and can be used by local authorities as a guide in predicting and planning the air quality control strategies.

Acknowledgements

This research was supported by Universiti Sains Malaysia Short Term Grant (PJJAUH/6315089) and Graduate Incentive Scheme (Vote U792) funded by the Research Management Centre (RMC), Universiti Tun Hussein Onn Malaysia. The authors would like to thank the Department of Environment (DOE), Malaysia, for generously providing the air quality data used in this study.

Conflict of Interest

The authors declare no conflict of interest.

References

1. ZHANG D., AUNAN K., MARTIN SEIP H., LARSSEN S, LIU J., Zhang D. The assessment of health damage caused by air pollution and its implication for policy making in Taiyuan, Shanxi, China. Energy Policy, **38** (1), 491, **2010**.
2. D'AMATO G., BAENA-CAGNANI C.E., CECCHI L., ANNESI-MAESANO I., NUNES C., ANSOTEGUI I., D'AMATO M., LICCARDI G., SOFIA M., CANONICA W.G. Climate change, air pollution and extreme events leading to increasing prevalence of allergic respiratory diseases. Multidisciplinary Respiratory Medicine, **8** (1), 12, **2013**.
3. RAFFEE A.F., ABDUL HAMID H., RAHMAT S.N., JAFFAR M.I. Vector autoregressive model: A multivariate time series to forecast the ground level ozone (O_3) concentration in Malaysia. Chiang Mai Journal of Science, **47** (6), 1297, **2020**.
4. LATIF M.T., DOMINICK D., AHAMAD F., AHAMAD N.S., KHAN M.F., JUNENG L., XIANG C.J., NADZIR M.S.M., ROBINSON A.D., ISMAIL M., MEAD M.I., HARRIS N.R.P. Seasonal and Long Term Variations of Surface Ozone Concentrations in Malaysia Borneo. Science of Total Environment, **573**, 494, **2016**.
5. ZAINORDIN N.S., RAMLI N.A., ELBAYOUMI M. Distribution and Temporal Behaviour of O_3 and NO_2 Near Selected Schools in Seberang Perai, Pulau Pinang and Parit Buntar, Perak, Malaysia. Sains Malaysiana, **46** (2), 197, **2017**.
6. EL-FADEL M., ZEIN M., NUWAYHID I., JAMALI D., SADEK S. Environmental management of ozone in Beirut urban areas. Environmental Management and Health, **13** (5), 47, **2002**.
7. MADANIYAZI L., NAGASHIMA T., GUO Y., PAN X., TONG S. Projecting ozone-related mortality in East China. Environment International, **92-93**, 165, **2016**.
8. LEFOHN A.S., MALLEY C.S., SMITH L., WELLS B., HAZUCHA M., SIMON H., NAIK V., MILLS G., SCHULTZ M.G., PAOLETTI E., DE MARCO A., XU X., ZHANG L., WANG T., NEUFELD H.S., MUSSELMAN R.C., TARASICK D., BRAUER M., FENG Z., TANG H., KOBAYASHI K., SICARD P., SOLBERG S., GEROSA G. Tropospheric ozone assessment report: Global ozone metrics for climate change, human health, and crop/ecosystem research. Elementa: Science of the Anthropocene, **6** (1), 2, **2018**.
9. WANG T., XUE L., BRIMBLECOMBE P., FAT Y., LI L., ZHANG L. Ozone pollution in China: A review of concentrations, meteorological influences, chemical precursors, and effects. Science of Total Environment, **575**, 1582, **2017**.
10. ZHANG H., WANG Y., HU J., YING Q., HU X. Relationships between meteorological parameters and criteria air pollutants in three megacities in China. Environmental Research, **140**, 242, **2015**.
11. TOH Y.Y., LIM S.F., VON GLASOW R. The influence of meteorological factors and biomass burning on surface ozone concentrations at Tanah Rata, Malaysia. Atmospheric Environment, **70**, 435, **2013**.
12. LATIF M.T., HUEY L.S., JUNENG L. Variations of surface ozone concentration across the Klang Valley, Malaysia. Atmospheric Environment, **61**, 434, **2012**.
13. DEPARTMENT OF ENVIRONMENT, M. Malaysia Environmental Quality Report 2016, Kuala Lumpur, **2016**.
14. DEPARTMENT OF ENVIRONMENT, M. Malaysia Environmental Quality Report 2015, Kuala Lumpur, **2015**.
15. DEPARTMENT OF ENVIRONMENT, M. Malaysia Environmental Quality Report 2010, Kuala Lumpur, **2010**.
16. DEPARTMENT OF ENVIRONMENT, M. Malaysia Environmental Quality Report 2011, Kuala Lumpur, **2011**.
17. DEPARTMENT OF ENVIRONMENT, M. Malaysia Environmental Quality Report 2012, Kuala Lumpur, **2012**.
18. DEPARTMENT OF ENVIRONMENT, M. Malaysia Environmental Quality Report 2013, Kuala Lumpur, **2013**.
19. DEPARTMENT OF ENVIRONMENT, M. Malaysia Environmental Quality Report 2014, Kuala Lumpur, **2014**.

Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy • Author Copy

20. DEPARTMENT OF ENVIRONMENT, M. "Annual Report 2016" Kuala Limpur, **2016**.
21. AZID A., JUAHIR H., TORIMAN M.E., ENDUT A., KAMARUDIN M.K.A., RAHMAN M.N.A., HASNAM C.N.C., SAUDI A.S.M., Yunus K. Source apportionment of air pollution: A case study in Malaysia. Jurnal Teknologi **72** (1), 83, **2015**.
22. SANSUDDIN N., RAMLI N.A., YAHAYA A.S., YUSOF N.F.F.M., GHAZALI N.A., AL-MADHOUN W.A. Statistical analysis of PM_{10} concentrations at different locations in Malaysia. Environmental Monitoring and Assessment, **180** (1-4), 573, **2011**.
23. ABDUL HAMID H., UL-SAUFIE A.Z., AHMAT H. Characteristic and Prediction of Carbon Monoxide Concentration using Time Series Analysis in Selected Urban Area in Malaysia. MATEC Web Conference, **103**, **2017**.
24. ISMAIL M., IBRAHIM M.Z., IBRAHIM T.A. Time Series Analysis of Surface Ozone Monitoring Records in Kemaman, Malaysia. Sains Malaysiana, **40** (5), 411, **2011**.
25. JAMIL N.I., ROZITA W., MAHIYUDDIN W., IZZAH A.N., LATIF M.T. Forecasting Ozone Concentrations Using Box-Jenkins ARIMA Modeling in Malaysia. American Journal of Environmental Science **14** (3), 118, **2018**.
26. JUNG HSU K. Time Series Analysis of the Interdependence Among Air Pollutants. Atmospheric Environment, **26** (4), 91, **1992**.
27. NIETO P.J.G., LASHERAS F.S., GARCÍA-GONZOLA E., JUEZ F.J.D.C. PM_{10} concentration forecasting in the metropolitan area of Oviedo (Northern Spain) using models based on SVM, MLP, VARMA and ARIMA: A case study. Science of Total Environment, **621**, 753, **2018**.
28. OH I., LEE J., AHN K., KIM J., KIM Y.M., SUN SIM C., KIM Y. Association between particulate matter concentration and symptoms of atopic dermatitis in children living in an industrial urban area of South Korea. Environmental Reseasrch **160**, 462, **2018**.
29. DUEÑAS C., FERNÁNDEZ M.C., CAÑETE S., CARRETERO J., LIGER E. Analyses of ozone in urban and rural sites in Málaga (Spain). Chemosphere, **56** (6), 631, **2004**.
30. COORAY T.M.J.A. Applied Time Series Analysis and Forecasting. Alpha Science, Oxford, USA, **2008**.
31. ABDEL-AZIZ A., FREY H.C. Development of hourly probabilistic utility NOx emission inventories using time series techniques: Part I – univariate approach," Atmospheric Environment, **37**, 5379, **2003**.
32. AGUNG I. G. N. Panel Data Analysis Using EViews. 1st ed.; John Wiley & Sons (Asia) Pte Ltd, Singapore, **2014**.
33. CHATFIELD C. Time series forecasting. 1st ed.; Chapman & Hall/CRC., United States, America, **2000**.
34. BOX G.E.P., JENKINS G.M., REINSEL G.C., LJUNG G.G. Time Series Analysis: Forecasting and Control. 5th Ed.; John Wiley & Sons, Inc., Hobokens, New Jersey, USA, **2016**.
35. RAHMAH M.R., KASHEM M.A. Carbon emissions, energy consumption and industrial growth in Bangladesh : Empirical evidence from ARDL cointegration and Granger causality analysis. Energy Policy, **110**, 600, **2017**.
36. JUNNINEN H., NISKA H., TUPPURAINEN K., RUUSKANEN J., KOLEHMAINEN M. Methods for imputation of missing values in air quality data sets. Atmospheric Environment, **38** (18), 2895, **2004**.
37. SHCHERBAKOV M.V., BREBELS A., SHCHERBAKOVA N.L., TYUKOV A.P., JANOVSKY T.A., KAMAEV V.A. A survey of forecast error measures. World Applied Sciences Journal, **24** (24), 171, **2013**.
38. CHAI T., DRAXLER R.R. Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. Geoscientific model development, **7**, 1247, **2014**.
39. WANG J., YANG Y., ZHANG Y., NIU T., JIANG X., WANG Y., CHE H. Influence of meteorological conditions on explosive increase in O_3 concentration in troposphere. Science of Total Environment, **652** (46), 1228, **2019**.
40. WEI W.W.S. Time Series Analysis: Univariate and Multivariate Methods, 2nd ed.; Pearson Addison Wesle: New York, USA, **2006**.