Journal of Quality Measurement and Analysis *JQMA* 15(2) 2019, 85-93
e-ISSN: 2600-8602 *http://www.ukm.my/jqma*

MODELING MULTIVARIABLE AIR POLLUTION DATA IN MALAYSIA USING VECTOR AUTOREGRESSIVE MODEL

(Memodelkan Data Pencemaran Udara Multipemboleh Ubah di Malaysia Menggunakan Model Autoregresi Vektor)

`ULYA BT ABDUL RAHIM & NURULKAMAL MASSERAN*

ABSTRACT

In this study, the vector autoregressive (VAR) model was used to model and forecast the multivariable air pollution data in Klang area. Stationary test, Hannan–Quinn evaluation criteria, Granger causality test, R^2 coefficient and Root Square Mean Error (RMSE) measurements have been conducted to get the best model and will be used in forecasting. The VAR (7) model is found to be the best model with the highest R^2 and lowest RMSE value recorded for each dependent pollutant variable. Based on the fitted VAR (7) model, the VAR model is able to describe the dynamic behavior of multivariable air pollution data of Klang. Forecasts of up to 12 days ahead were constructed with confidence intervals. The VAR model found to provides good forecast accuracy on the data.

Keywords: air-pollution modeling; VAR model; forecasting

ABSTRAK

Dalam kajian ini, model autoregresi vektor (VAR) digunakan untuk memodel dan meramal data pencemaran udara multipemboleh ubah di kawasan Klang. Ujian kepegunan, kriterium penilaian Hannan-Quinn, ujian kebersebaban Granger, pekali R^2 dan ukuran ralat min kuasa dua (RMSE) telah dijalankan untuk mendapatkan model terbaik dan akan digunakan bagi tujuan peramalan. Model VAR (7) dikenal pasti sebagai model terbaik dengan nilai pekali R^2 tertinggi dan nilai RMSE yang terendah untuk setiap pemboleh ubah pencemar bersandar. Berdasarkan model VAR (7) yang disuaikan, model VAR didapati mampu untuk memerihalkan tingkah laku dinamik data pencemaran udara multipemboleh ubah di Klang. Ramalan sehingga 12 hari ke depan telah dijalankan beserta maklumat selang keyakinan bagi model VAR(7). Model VAR didapati boleh memberikan ketepatan ramalan yang baik terhadap data.

Kata kunci: pemodelan pencemaran udara; model VAR; peramalan

1. Introduction

Air pollution issues are always a matter of concern in Malaysia due to the rapid growth of industry, manufacturing, economy, and transportation. Air pollution involves any chemicals, particulate matter, or biological substances that can cause damage to the environment and discomfort to humans and living organisms when released to the atmosphere. Air pollution will also cause the depletion of the ozone layer, haze, acid rain, warming of the earth and affect the health and safety of people or properties. As stated by Omasa (2002), air pollution is hard to treat and control due to the nature of airborne particles.

According to World Health Organization (2006), four main types of air pollution exist which are nitrogen dioxide (NO_2), sulfur dioxide (SO_2), ozone (O_3), and particulate matter (PM_{10}). Additionally, five main types of air pollution that are commonly used by researchers to measure the air pollution index (API) include PM_{10}, O_3, carbon monoxide (CO_2), SO_2, and NO_2 (Gass *et al.* 2015; Masseran *et al.* 2016). To get the benchmark of air quality status and health, the API scales are used. There are five API scales, which are good (0–50), moderate (51–100), unhealthy (101–200), very unhealthy (201–300), or hazardous (301 and above) status (World Health Organization 2006).

Internationally, considerable research has been conducted to monitor and forecast the air quality status and health by using various methods (Bai *et al.* 2018). Several popularly discussed techniques include auto regressive moving average, compositional time series, auto regressive integrated moving average, artificial neural network, fuzzy time series, generalized autoregressive conditional heteroscedastic model and vector autoregressive model (VAR) (Al-Dhurafi *et al.* 2018; Masseran 2017). However, each model has its own advantages and disadvantages. In this study, we propose the application of the VAR model to API data as it can capture linear interdependencies among multiple time series. Particularly in our cases, the API data involved five different variables, namely, CO, NO_2, O_3, PM_{10}, and SO_2. This research primarily aimed to develop VAR models to predict the daily API data using these five pollutant variables to monitor the air quality status.

2. Study Area

Klang, one of the regions in the state of Selangor Darul Ehsan, is considered as one of the most developed and urbanized areas in Selangor. In 2010, the total population of Klang City has reached 240,016, whereas in Klang District, the population approximates 842,146. Hence, this area has become busy as the center of economy, shipping, residential, and leisure activities.

The rapid growth and development in Klang City have contributed to the positive economic growth and profit to the nation. However, this condition also caused negative impacts to the air quality and health due to industrial and manufacturing activities (Masseran *et al.* 2016). Thus, a reliable and accurate forecasting model must be developed to predict the air quality status and health in the long run. The data from 2002 to 2016, which include CO, NO_2, O_3, PM_{10}, and SO_2, were obtained from the Department of Environment Malaysia (DOE). However, several sets of data from DOE showed a small percentage of missing values. To estimate these missing values, we used a single-imputation method based on the average of the last and next known observations. This method is easy to implement and is reported to provide good results for random missing data (Masseran *et al.* 2013).

3. Methodology

3.1. *Stationary test*

In the VAR model, all response variables should be stationary and contain no unit root (Brooks 2019). The stationary data can be evaluated through several types of unit root test. In general, three types of test, namely, augmented Dickey–Fuller (ADF), Phillips–Perron, and Kwiatkowski–Phillips–Schmidt–Shin tests, are commonly used to assess stationarity of the data (Masseran *et al.* 2012). In this study, we applied ADF test on five dependent variables of API data.

3.2. *Lag order selection*

The lag order selection is also an important step in modeling the data using the VAR model. According to Ary Pani (2016), the optimal lag length in a VAR model can be estimated through several tests of information criteria, such as Akaike information criteria (AIC), Schwarz information (SIC) and Hannan–Quinn (HQ) information criteria. The smaller values of information criteria indicate a better model fit to the data. The equation for each criteria is stated below:

$$AIC = -2\left(\frac{1}{T}\right) + 2(k+1) \tag{1}$$

$$SIC = -2\left(\frac{1}{T}\right) + k\ \frac{\log(T)}{T} \tag{2}$$

$$HQ = -2\left(\frac{1}{T}\right) + 2klog(\frac{\log(T)}{T}) \tag{3}$$

where T is the number of observations and k is the number of parameters involved in the VAR model.

3.3. *Granger causality test*

Granger causality test is used to assess the cause and effect in a given condition. For example, if an event A exists before event B, then A may causes B. However B does not necessary caused by event A. This situation can be explained with the concept of Granger causality test (Ary Pani 2016). The equation of causality test is shown below:

$$Y_t = \sum_{i=1}^{k} a_i\, Y_{t-i} + \sum_{i=1}^{k} \beta_i X_{t-i} + \varepsilon_{1t} \tag{4}$$

where:

- Y_t : Value of variable *Y* at time *t*
- k : Lag length
- a_i : Measure of the influence of Y_{t-i} on X_t
- β_i : Measure of the influence of X_{t-i} on Y_t
- X_{t-i} : Vector of length *k*
- ε_{1t} : Error at time *t*

3.4. *Estimation of the VAR model*

In this study, the restrict function, which was suggested by Pfaff (2008), was used to estimate the VAR model. This function only keeps the significant coefficients. In general, the VAR model at k^{th} order is denoted as VAR (*k*). The VAR model includes *k* lags and *n* variables, can be formulated as follows:

$$Y_t = A_o + A_1 Y_{t-1} + A_2 Y_{t-2} + \cdots + A_k Y_{t-k} \tag{5}$$

where

- $Y_t,\ Y_{t-1}$: Vector $n \times 1$ at time t and $t - i, i$= 1, 2, …, k
- A_o : Intercept
- t : Period of observation
- ε : Residual error

3.5. *Model accuracy*

To determine the performance of the VAR model, we used the root square mean error (RMSE) and R^2 coefficient as measure criteria. According to Barnston (1992), RMSE is the standard deviation of the residuals. Residuals measure how far data points are from the regression line, whereas RMSE measures the spread of residuals. A low RMSE indicates a high accuracy, and the best model with the highest accuracy will be used in forecasting. The formula for calculating RMSE is stated below:

$$RMSE = \sqrt{\frac{1}{L}\sum_{l=1}^{L}(y_{T+l} - \hat{y}_{T+l})^2} \tag{6}$$

where T is the last observation period and l is the lag.

R^2 has the advantage of being scale-free and closely related to RMSE. The value of R^2 coefficient, which ranges between 0 and 1, measures how close of the data are to the fitted line (Cameron & Windmeijer 1997):

$$R^2 = \frac{\sum_{t-1}^{T}(\hat{y}_t - \bar{y})^2}{\sum_{t-1}^{T}(\hat{y}_t - \bar{x})^2 + \sum_{t-1}^{T}(\hat{y}_t - \hat{y}_t)^2} \tag{7}$$

where y_t and $\bar{y}$ are the observed and mean values of the time series data, respectively, and $\hat{y}_t$ is the simulated/predicted value that is obtained from the model (Masseran & Razali 2016). In general, a higher R^2 value means the better model fit of our data.

3.6. *Forecasting using the VAR model*

Forecasting refers to the estimation or prediction of the future. This process predicts the future values of a series using a current information set. The variable $\boldsymbol{y_t}$ is assumed to follow the VAR (p). Then, the forecast $\boldsymbol{\hat{y}_{T+1}}$ is given by the following:

$$\hat{y}_{T+1} = v + A_1 y_T + \cdots + A_P y_{T-P+1} \tag{8}$$

The above equation also defines a forecast for each component of y_{t+1}.

4. Results and Discussions

4.1. *Stationary test*

The results show in Table 1 reveal that that all the series are integrated at order 0. Given that the series are integrated at I(0), the series are not co-integrated. The test statistics indicates that all the variables are stationary because all *p*-values are significant at any significance levels. Given that ADF test fulfills the assumption of stationarity, our data fit the VAR model.

Table 1: Results of ADF test on each series of pollutant variables

Variables	Order	*t*-statistics	Prob.
CO	I(0)	-11.785	0.01***
NO2	I(0)	-10.75	0.01***
O3	I(0)	-10.002	0.01***
PM10	I(0)	-12.403	0.01***
SO2	I(0)	-8.9225	0.01***

Notes:*, **, *** shows the significant variables at 10%, 5%, and 1% of significance levels.

4.2. *Lag order selection*

The number of lag order selection was based on the three valuation criteria, namely, AIC, SIC, and HQ (Table 2). The AIC suggests that lag of 9 is the appropriate lag length, but SC suggests a lag length of 3. By contrast, HQ suggests a lag order of 7 to be used in our model. Usually, AIC is selected over other criterion due to its favorable small-sample forecasting features. However, for a large sample size, HQ works better compared with AIC and SC (Liew 2004). Given our large sample size, we believe that the result from HQ evaluation criteria is appropriate. Thus, a lag order of 7 was selected for our VAR model based on the lowest HQ value. The results are reported in Table 2.

Table 2: Lag order selection for the VAR model

Lag	AIC	SC	HQ
1	29.689	29.726	29.702
2	29.471	29.538	29.494
3	29.387	**29.485***	29.421
4	29.360	29.489	29.405
5	29.342	29.501	29.398
6	29.305	29.495	29.372
7	29.288	29.509	**29.365***
8	29.289	29.540	29.377
9	**29.286***	29.568	29.384
10	29.289	29.602	29.398

Note: the **bold*** value shows the minimum value for each evaluation criterion.

4.3. *Granger causality test*

After obtaining the optimal lag length, a Granger causality test was performed to investigate whether a reciprocal relationship exists between the variables.

Table 3: Granger causality test on each pollutant variable

No	Hypothesis	F-statistics	p-value
1	NO2 does not granger cause CO	14.617	2.2e-16 ***
	CO does not granger cause NO2	4.9547	1.337e-05 ***
2	O3 does not granger cause CO	6.0023	5.504e-07 ***
	CO does not granger cause O3	2.6386	0.01011 **
3	PM10 does not granger cause CO	5.3094	4.579e-06 ***
	CO does not granger cause PM10	11.267	3.017e-14 ***
4	SO2 does not granger cause CO	4.9334	1.425e-05 ***
	CO does not granger cause SO2	3.9706	0.0002455 ***
5	O3 does not granger cause NO2	12.678	3.092e-16 ***
	NO2 does not granger cause O3	3.8967	0.0003042 ***
6	PM10 does not granger cause NO2	3.4724	0.001024 ***
	NO2 does not granger cause PM10	24.711	2.2e-16 ***
7	SO2 does not granger cause NO2	2.4512	0.01652 **
	NO2 does not granger cause SO2	27.08	2.2e-16 ***
8	PM10 does not granger cause O3	2.6367	0.01016 **
	O3 does not granger cause PM10	12.814	2.2e-16 ***
9	SO2 does not granger cause O3	1.8683	0.07046*
	O3 does not granger cause SO2	1.2134	0.2913
10	SO2 does not granger cause PM10	1.2543	0.269
	PM10 does not granger cause SO2	6.2109	2.892e-07 ***

Notes:*, **, and *** indicates the significant variables at 10%, 5%, and 1% significance level.

From the results in Table 3, most API variables have a Granger cause on the other variables because the *p*-value is significant at 1%, 5%, and 10% significance level. For result No. 1, the *p*-value is significant for both variables. NO_2 has a Granger cause on CO, and CO also has a Granger cause on NO_2. This finding indicates that past NO_2 contains information to forecast the current CO, and previous CO also has information to predict the current NO_2.

By contrast, for result No. 9, the *p*-value is non-significant for the second hypothesis. The finding indicates that SO_2 has a Granger cause on O_3, but O_3 lacks a Granger cause on SO_2. Thus, past SO_2 has information to predict the current O_3, but the previous O_3 contains no information to forecast the current SO_2. Thus, this insignificant result was excluded in the estimation of the VAR model. Only variables with significant *p*-value were used for estimation.

4.4. *Estimation using the VAR model*

VAR is estimated separately by using each equation on each variable. After the estimation, the values of R^2 and RMSE were calculated. The summarization of our findings only discussed the estimated and significant coefficients of variables after removing the insignificant coefficients in our model. The overall estimated model is stated in the equation below:

The estimated model for CO is given as follows:

$$\begin{aligned}\widehat{CO}_t = {} & 0.430\widehat{CO}_{t-1} + 0.182\widehat{NO2}_{t-1} + 0.290\widehat{PM10}_{t-1} - 0.060\widehat{SO2}_{t-1} + 0.124\widehat{CO}_{t-2} \\ & - 0.170\widehat{NO2}_{t-2} + 0.037\widehat{O3}_{t-2} - 0.022\widehat{PM10}_{t-2} + 0.055\widehat{SO2}_{t-2} + 0.095\widehat{CO}_{t-3} \\ & - 0.048\widehat{NO2}_{t-4} + 0.060\widehat{CO}_{t-5} - 0.026\widehat{SO2}_{t-5} + 0.080\widehat{CO}_{t-7} - 0.012\widehat{PM10}_{t-7} \\ & + 0.038\widehat{SO2}_{t-7} + 2.461\end{aligned}$$

The estimated model for NO2 is given by the following:

$$\begin{aligned}\widehat{NO2}_t = {} & -0.057\widehat{CO}_{t-1} + 0.548\widehat{NO2}_{t-1} + 0.064\widehat{O3}_{t-1} + 0.080\widehat{NO2}_{t-3} + 0.043\widehat{CO}_{t-5} \\ & - 0.170\widehat{PM10}_{t-5} - 0.025\widehat{SO2}_{t-5} + 0.070\widehat{NO2}_{t-6} + 0.012\widehat{PM10}_{t-6} + 0.031\widehat{SO2}_{t-6} \\ & + 0.062\widehat{NO2}_{t-7} - 0.020\widehat{O3}_{t-7} + 2.642\end{aligned}$$

The estimated model for O3 is given as follows:

$$\begin{aligned}\widehat{O3}_t = {} & 0.077\widehat{NO2}_{t-1} + 0.467\widehat{O3}_{t-1} + 0.011\widehat{PM10}_{t-1} + 0.069\widehat{O3}_{t-2} + 0.066\widehat{O3}_{t-3} + 0.054\widehat{O3}_{t-4} \\ & + 0.076\widehat{O3}_{t-6} - 0.008\widehat{PM10}_{t-6} - 0.027\widehat{SO2}_{t-6} + 0.045\widehat{O3}_{t-7} + 3.163\end{aligned}$$

The estimated model for PM10 is given below:

$$\begin{aligned}\widehat{PM10}_t = {} & 0.321\widehat{CO}_{t-1} + 0.486\widehat{NO2}_{t-1} + 0.310\widehat{O3}_{t-1} + 0.794\widehat{PM10}_{t-1} - 0.127\widehat{SO2}_{t-1} \\ & - 0.745\widehat{NO2}_{t-2} - 0.137\widehat{O3}_{t-2} - 0.174\widehat{PM10}_{t-2} + 0.087\widehat{SO2}_{t-2} + 0.193\widehat{NO2}_{t-3} \\ & + 0.095\widehat{PM10}_{t-3} - 0.114\widehat{CO}_{t-4} - 0.128\widehat{NO2}_{t-4} + 0.039\widehat{PM10}_{t-4} - 0.125\widehat{CO}_{t-7} \\ & + 0.044\widehat{PM10}_{t-7} + 0.071\widehat{SO2}_{t-7} + 9.646\end{aligned}$$

The estimated model for SO2 is given below:

$$\begin{aligned}\widehat{SO2}_t = {} & -0.089\widehat{CO}_{t-1} + 0.288\widehat{NO2}_{t-1} - 0.010\widehat{PM10}_{t-1} + 0.700\widehat{SO2}_{t-1} + 0.073\widehat{CO}_{t-2} \\ & - 0.237\widehat{NO2}_{t-2} - 0.065\widehat{SO2}_{t-2} + 0.106\widehat{SO2}_{t-3} - 0.056\widehat{NO2}_{t-4} + 0.055\widehat{SO2}_{t-4} \\ & + 0.058\widehat{CO}_{t-5} - 0.021\widehat{PM10}_{t-5} + 0.024\widehat{PM10}_{t-6} + 0.036\widehat{SO2}_{t-6} \\ & - 0.051\widehat{NO2}_{t-7} + 0.069\widehat{SO2}_{t-7} + 1.906\end{aligned}$$

4.5. *Model accuracy*

Table 4 summarizes the results of our VAR model considering CO, NO_2, O_3, PM_{10}, and SO_2. Based on the results, the RMSE values for all models are slightly small. This finding means that residuals are spread close to each other, and most of the data are dispersed around the best fit line. The value of RMSE indicates the accuracy of the model in data fitting.

Table 4: Values of RMSE and R^2 coefficient for each VAR model

	CO	**NO_2**	**O_3**	**PM_{10}**	**SO_2**
RMSE	0.010	0.012	0.019	0.012	0.010
R^2 coefficient	0.907	0.950	0.921	0.963	0.928

Based on Table 4, the VAR (7) model with R^2 value greater than 0.9 includes all the pollution variables. Thus, the VAR (7) model can describe more than 90% of the total variation for all the pollution variable. The values of RMSE are also small for all the fitted VAR (7) models. These results indicate that the fitted VAR (7) model is a good model for all the pollution variables involved in this study.

4.6. *Forecasting using the VAR model*

The VAR (7) model was used to provide a forecast 12 days ahead of the air pollutant data. We can choose any future value to be forecast as long as it provide a valid forecasting values from the fitted model. In this study, 12-day ahead is chosen as a forecasting value in order to provide some practical example. The point forecast refers to the mean of the distribution, and the confidence limits describe the spread of the distribution above and below the point forecast.The graphs in Figure 1 show the time series forecasting 12 days ahead and the confidence limit for each pollutant variable (CO, NO_2, O_3, PM_{10}, and SO_2) in the fitted VAR (7) model.

The forecast values of CO for the next 12 days will be in the average range of 8–10, with the lowest and highest confidence intervals of 7 and 10, respectively. The lower and upper confidence limits are between 0 and 20.9, respectively. This range means that the highest future values of CO could be being at or below the upper confidence limit which is about 20.9, and its lowest future values could be being at or below the lower confidence limit is 0. The forecast values of NO_2 for the upcoming 12 days will be in the average range of 9–11, with the lowest and highest confidence interval of 5 and 7, respectively. The upper and lower confidence limits are lies between 18.9 and 3, respectively. These mean that highest future values of NO_2 could be being at or below the upper confidence limit which is about 18.9, and its lowest future values could be being at or below the lower confidence limit is 3.

For O_3, the forecast values for the next 12 days will be in the average range of 15–16, with the lowest and highest confidence interval of 10 and 13, respectively. The upper and lower confidence limits are set to 29.8 and 2.5, respectively. These mean that highest future values of O_3 could be being at or below the upper confidence limit which is about 29.8, and its lowest future values could be being at or below the lower confidence limit is 2.5. For PM_{10}, the forecast values for the upcoming 12 days will be in the average range of 50–57, with the lowest and highest confidence interval of 23 and 39, respectively. The upper and lower confidence limits are between the range of 96.6 and 18.5, respectively. These finding means that highest future values of PM_{10} could be being at or below the upper confidence limit which is about 96.6, and its lowest future values could be being at or below the lower confidence limit is 18.5.

Lastly, the forecast values of SO_2 for the next 12 days will be in the mean range of 9–12, with the lowest and highest confidence interval of 7 and 12, respectively. The upper and lower confidence limits are lies between 25 and 0, respectively. These finding implies that highest future values of SO_2 could be being at or below the upper confidence limit which is about 25, and its lowest future values could be being at or below the lower confidence limit is 0.

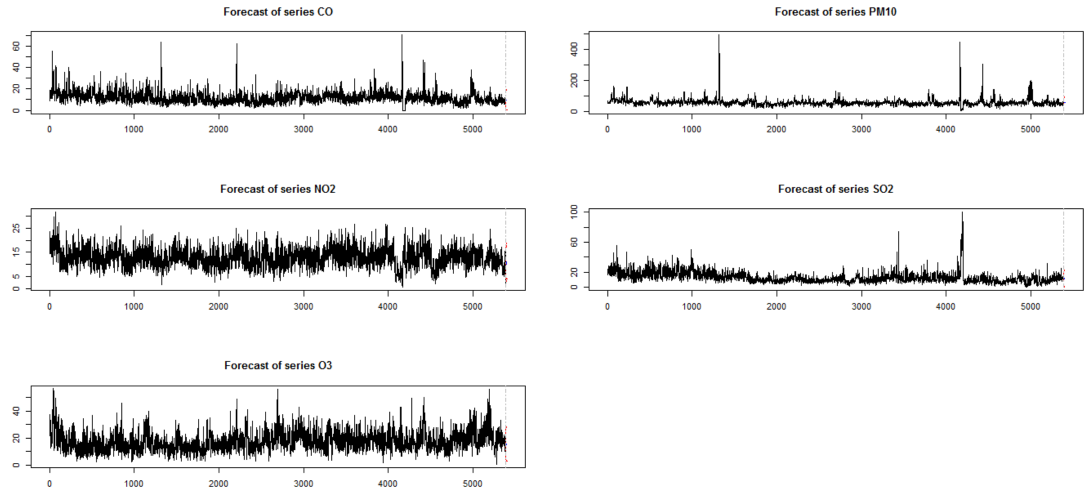

Figure 1: Forecast of CO, NO_2, O_3, PM_{10} and SO_2

These forecast values can serve as basis to the public authorities to monitor the risk of recurrence of extreme air pollution. However, for a wide future time horizon, the confidence interval of forecasting will be larger, which implies that the accuracy of forecasting will decrease. The VAR model constantly needs to be re-estimated to obtain the latest forecasting evaluation of air pollution data values over time to provide a better assessment. Thus, to use the VAR model for air pollution forecasting and decision making, we suggest running this model every day to forecast the future value of air pollution data.

5. Conclusion

This study examined a time series of the VAR model to forecast the 15-year air pollutant data, including CO, NO_2, O_3, PM_{10}, and SO_2, in Klang. The ADF test results showed that all the dependent variables are stationary at order I (0), which implies that the VAR model is suitable to model the air pollutant data. A lag of 7 on the VAR model was selected based on the HQ evaluation criteria. The findings reveal that the VAR (7) model is appropriate and fits the data with the highest R^2 value and lowest RMSE recorded for each dependent model variable.

In consideration of the actual data, an accurate forecasting performance was obtained when the VAR (7) model was used to forecast the five API variables. These forecast values can be a benchmark for the stakeholders to continually monitor the air status and health. The model must be re-estimated regularly to obtain the latest forecasting results of CO, NO_2, O_3, PM_{10}, and SO_2 over time to obtain better predictions. Another important contribution of this paper is showing the possibility of using a large sample size of daily data with a large lag length of up to 7. Overall, we can conclude that the VAR (7) model, which designates the existence of a fluctuation and "shock point" effect, is an appropriate model to use when handling air pollution data.

Acknowledgments

The authors are indebted to the Department of Environment Malaysia for providing the air pollution data that made this paper possible. This research would not be possible without the sponsorship from Universiti Kebangsaan Malaysia (grant number GUP-2017-116).

References

Al-Dhurafi N.A., Masseran N. & Zamzuri Z.H. 2018. Compositional time series analysis for Air Pollution Index data. *Stochastic Environmental Research and Risk Assessment* **32**(10): 2903-2911.

Ary Pani D. 2016. Pemodelan Pencemaran Udara Menggunakan Metode Vector Autoregressive (VAR) di Provinsi Riau. *Jurnal Sains dan Teknologi Industri* **13**(2): 160-167.

Bai L., Wang J., Ma X. & Lu H. 2018. Air pollution forecasts: An overview. *International Journal of Environmental Research and Public Health* **15**(4): 1-44.

Barnston A.G. 1992. Correspondence among the Correlation, RMSE, and Heidke Forecast Verification Measures; Refinement of the Heidke Score. *Weather and Forecasting* **7**: 699-709.

Brooks C. 2019. *Introductory econometrics for finance.* Cambridge: Cambridge University Press

Cameron A.C. & Windmeijer F.A. 1997. An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics* **77**(2): 329-342.

Gass K., Klein M., Sarnat S., Winquist A., Darrow L., Flanders W., Chang H., Mulholland J., Tolbert P. & Strickland M. 2015. Associations between ambient air pollutant mixtures and pediatric asthma emergency department visits in three cities: a classification and regression tree approach. *Environmental Health* **58**:1–14

Liew V.K.S. 2004. Which lag length selection criteria should we employ? *Economics Bulletin* **3**(33): 1-9.

Masseran N. 2017. Modeling fluctuation of PM_{10} data with existence of volatility effect. *Environmental Engineering Science* **34**(11): 816-827

Masseran N. & Razali A.M. 2016. Modeling the wind direction behaviors during the monsoon seasons in Peninsular Malaysia. *Renewable and Sustainable Energy Reviews* **56**:1419–1430

Masseran N., Razali A.M., Ibrahim K. & Latif M. T. 2016. Modeling air quality in main cities of Peninsular Malaysia by using a generalized Pareto model. *Environmental Monitoring and Assessment* **188**(1), Article number 65: 1-12.

Masseran N., Razali A.M., Ibrahim K. & Wan Zin W. Z. 2012. Evaluating the wind speed persistence for several wind stations in Peninsular Malaysia. *Energy* **37**(1): 649-656.

Masseran N., Razali A.M., Ibrahim K., Zaharim A., & Sopian K. 2013. Application of the single imputation method to estimate missing wind speed data in Malaysia. *Research Journal of Applied Sciences, Engineering and Technology* **6**(10): 1780-1784.

Omasa K. 2002. Diagnosis of stomatal response and gas exchange of trees by thermal remote sensing. *Tokyo: Springer.*

Pfaff B. 2008. VAR, SVAR and SVEC models: Implementation within R package vars. *Journal of Statistical Software*, *27*(4), 1-32.

World Health Organization. 2006. WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide: global update 2005: summary of risk assessment. Geneva: World Health Organization.

Statistics Programme
Faculty of Science and Technology
Universiti Kebangsaan Malaysia
43600 UKM Bangi
Selangor DE, MALAYSIA
*E-mail: p89235@siswa.ukm.edu.my, kamalmsn@ukm.my**

*Corresponding author